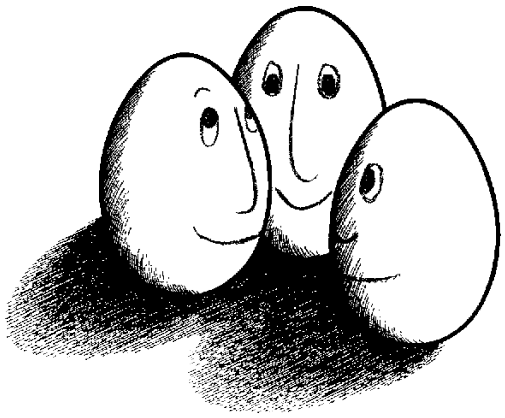
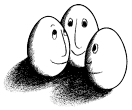
Logistic Regression with Group ℓ_1 vs. Elastic Net Regularization

Bachelorarbeit

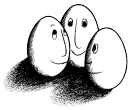
im Fachgebiet Maschinelles Lernen



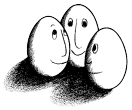
vorgelegt von:	Alexey Egorov
Studienbereich:	Informatik
Matrikelnummer:	128457
Erstgutachter:	Prof. Dr. Katharina Morik
Betreuer:	Sangkyun Lee



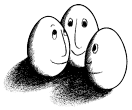
Contents



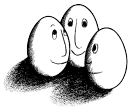
Contents



List of Figures



List of Tables



1 Introduction

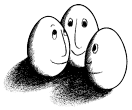
1.1 Motivation

Applications of machine learning in biomedical research has been an important issue in computer science since the 1960s. The complexity of data to be analyzed has been increasing dramatically since more accurate medical approaches and advanced techniques are used [Morik 2010]. Since nowadays there are tools (more details in Chapter 2) that allow us to analyze genetic expression data of patients suffering under various diseases, finding genetic biomarkers has become an useful approach to predict patients' prognosis. Current researches for neuroblastoma focus on finding biomarkers for neuroblastoma cancer to understand its nature [Schramm et al. 2009; Oberthuer et al. 2010; Takita et al. 2011; North et al. 1997]. Since the considered data is very high dimensional, different methods of feature selection can be applied to reduce the number of dimensions [Ng 2004; Zou and Hastie 2003; Schowe 2011]. Feature selection filters those variables which contain no or less important information concerning the learning task. Considering this data we are aiming in comparing two different approaches of feature selection in order to distinguish a better approach to decrease the number of attributes and increase the prediction accuracy on the neuroblastoma data.

1.2 Aim of this thesis

In this thesis we focus on linear models with regularization terms to keep solutions sparse. In course of this thesis we implement two operators for RapidMiner, an open source project by Rapid-I[1], which provides data mining and machine learning procedures. To achieve sparsity it is possible to find important features and filter out unimportant ones. Since our data contains exon expression values which can be sorted by genes they belong to, this can be done with the *group* ℓ_1 regularizer. But as we do not know for sure whether grouping exons by genes is an optimal form to detect linear model with high prediction rate, we compare *group* ℓ_1 to another

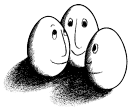
[1] http://rapid-i.com/



type of penalization, *elastic net*. We intend to compare performance of two models trained with these regularization terms and consider some of the features selected in these processes.

1.3 Structure

First in Chapter 2 we mention some prior information about neuroblastoma cancer and describe the data set in more details. In Section 3.1 we focus on the theory behind linear models. In Section 3.2 we explain regularization terms to be used in this thesis. Two optimization methods are introduced in Section 3.3 to solve the problem as described before. In Section 4.1 we talk in more details about both group ℓ_1 and elastic net operators which are implemented for RapidMiner. The evaluation procedure of obtained models and analysis of selected features can be found in Section 4.2.

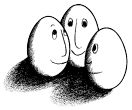


2 Background

As we already described our motivation in Section 1.1 the neuroblastoma study gives us an interesting opportunity to exploring cancer and its genetic origin. Because most of the patients are infants or children they don't have any preexisting/prior diseases or conditions (e.g. smoking or alcohol). Hence this situation gives us a possibility of finding and analyzing biomarkers of neuroblastoma and cancer at all without any further influences on its genetic expression.

2.1 Neuroblastoma cancer

Neuroblastoma is the most common tumor in childhood. It accounts for 8% of all childhood cancers and 15% of pediatric oncology deaths [Oberthuer et al. 2010]. For most infants the disease regresses completely even without any therapy and it only spreads to regional lymph nodes, bone and bone marrow. But on the other hand older patients (older than 1 year old) frequently have metastatic disease despite all intensive therapies. Thus infants have even better prognosis with the presence of metastatic disease [Brodeur 2003].
With the help of divergent genetic and biochemical analysis one was able to distinguish several clinical risk factors as age at diagnosis or stage and molecular risk factors as histology or MYCN amplification. [Schramm et al. 2009]. MYCN amplification turned out to be a powerful predictor of a poor prognosis and "accordingly patients with MYCN-amplified tumors receive a more intense treatment" [Schramm et al. 2005]. Furthermore different other alterations or abnormal patterns of gene expression are giving some hints about outcome and prognosis. Hence there are several reports about genetic predisposition for neuroblastoma [Schramm et al. 2005; Oberthuer et al. 2006; Chen et al. 2008; Asgharzadeh et al. 2006]. Unfortunately they seemed to not be consistent about where the locus lies and which gene deletions or rearrangements are involved in this.



2.2 Dataset

2.2.1 Technology

DNA chip technology or microarray was introduced in the 80's and gave scientists the ability of more detailed insight into genetic information of human cells. Kurian et al. [1999] states that those high-density DNA arrays are capable of analyzing thousands of genes simultaneously. Furthermore by using this method global gene expression can be compared in two populations, for example in a "normal" versus a melanoma cell line.
Affymetrix provides novel whole-transcript expression analysis, called GeneChip, described in more details by Affymetrix[2]. It enables researchers not only to detect the level of expression, but also to determine precisely what is being expressed, including alternative isoforms or genomic deletions. Furthermore it offers exon- and gene-level expression analysis in a single experiment and thus whole-transcript expression analysis can be used to detect and analyze alternative splicing and differential expression of each exon within a gene.

2.2.2 Data structure

Here we consider dataset derived from GeneChip that has been described in previous section. Data used here is originally from two different sources: one part is taken from study center of the German neuroblastoma study group [Schramm et al. 2009]; another part was submitted by R2, a microarray visualization platform developed at the department of human genetics in the Academic Medical Center in Amsterdam.[3]
After preprocessing the total exon array expression data consists of 215525 exon probesets. The dataset contains patient related information that is being used for classification. Data for each patient has a unique ID, label with stages $\{1, 2, 3, 4, 4s\}$ according to the INSS (International Neuroblastoma Staging System) and measured exon data. Each exon has therefore measurements for all 257 patients in this study and can be identified by its 7-digit ID. It is rather difficult to specify good models within this very high dimensional space. Thus our aim is to reduce the number of dimensions to achieve good prognosis prediction.

[2] www.affymetrix.com/promotions/wtexpression/wtexpression.affx
[3] http://r2.amc.nl/

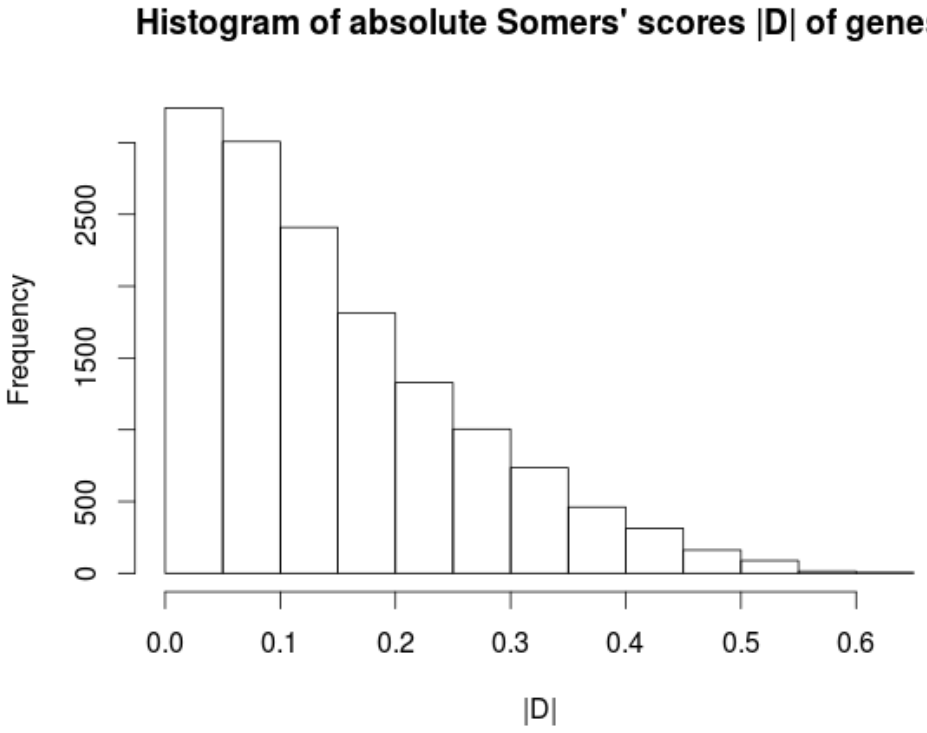


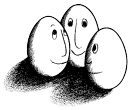
Figure 2.1: Computed rank correlation between each gene vector and the response vector by the extended Somers' D statistics. After filtering out genes with $|D| < 0.3$ approximately 30% of genes are left.[Lee 2012]

2.2.3 Preparation

As a preparation step we filter some of genes based on results of extended version of Somers' D statistics introduced in Newson [2006]. Somers' D means that if you have predictor variable X and outcome variable Y, you may estimate D_{XY} as a performance indicator of X as a predictor of Y. Since this statistics fits our situation, we perform following steps:

1. We compute rank correlation between each gene vector (using gene-level summary data) and the response vector (survival time, with possible right-censoring) by an extended version of Somers' D statistic
2. The scores D varies in [-1,1]. A threshold of $|D| < 0.3$ is then used and filteres around 70% of the range of the genes (Figure 2.1).
3. Exons that belong to the chosen genes (70%) are removed.

This process reduced the dataset from 215525 to 31054 exons and filtered exons that suppose to have very low correlation with a label.
The most often used clinical features are stages of disease ($\{1, 2, 3, 4, 4s\}$), the age of



the patient at diagnosis, survival time and the site of the primary tumor. We used survival time as mentioned above for filtering some features. According to INSS patients with stages $\{1, 2, 4s\}$ are low risk while $\{3, 4\}$ are high risk. As you can see on Figure 2.2 not all patients in stage 4 die and around 40% of them survive longer than 6000 days of study. Oberthuer et al. [2010] questioned INSS stages and were able to improve classification. From this point of view one is interested in classifications between these stages and especially stage 4 may have to be splitted into several subgroups. But due to lack of patients we will not leave out any patients in this research (like for classification between stage 3 and 4 while leaving out patients with stages $\{1, 2, 4s\}$). Consequently we will try to predict low risk group (stages $\{1, 2, 4s\}$) against high risk group (stages $\{3, 4\}$) at an expense of slightly lower accuracy as our main goal is still to compare two approaches which we will introduce later.

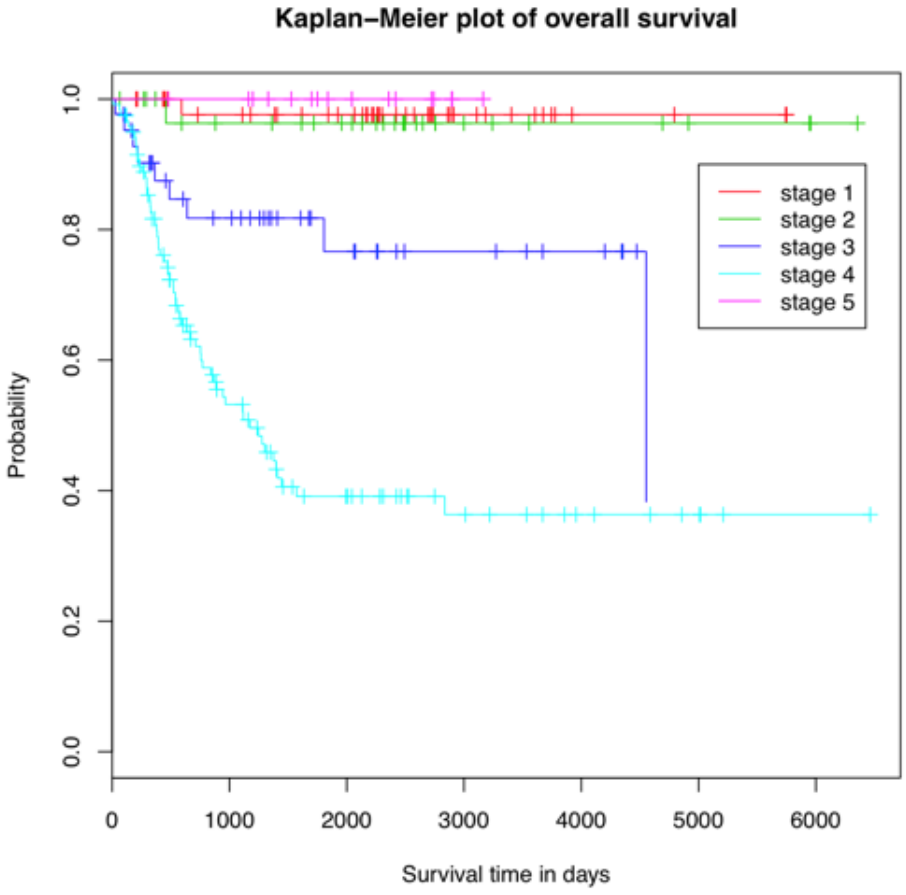
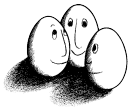


Figure 2.2: Kaplan Meyer plot of the overall survival. In this plot stage 4s is called stage 5. Vertical marks indicate a loss. [Lee 2012]



3 Method

To achieve our aim of learning a good prediction model for neuroblastoma studies we are going to use logistic regression as a cost function that we will discuss in more details in Section 3.1. In order to reduce number of features we are going to add two different regularization terms to the cost function, elastic net and group ℓ_1. We will describe them and solution of created optimization problems in Sections 3.2 and 3.4 respectively.

3.1 Logistic regression

Linear regression solves the following problem set. We consider for all N patients $\{(x_i, y_i)\}_{i=1}^N$ where $y_i \in \mathbb{R}$ is the label, $x_i = (\hat{x}_i, 1) \in \mathbb{R}^p$ is a set of $p-1$ features $\hat{x}_i \in \mathbb{R}^{p-1}$ and 1 is the intercept for the optimization problem. With a linear function one can describe the dependency of y_i and x_i:

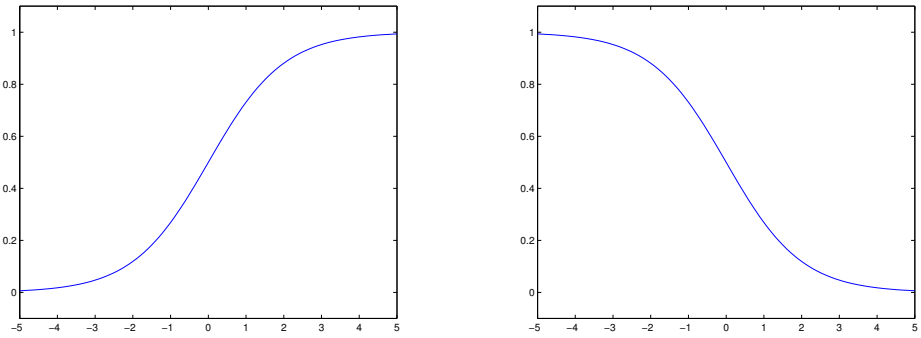
$$\begin{aligned} f(x_i) &= w[1]x_i[1] + w[2]x_i[2] + \cdots + w[p]x_i[p] \\ &= \sum_{j=1}^{p} w[j]x_i[j] \ . \end{aligned}$$

where $z_i[j]$ means j-th element in the i-th vector z. Using this function we can easily estimate best fitting solution for w by minimizing following equation:

$$\min_{w \in \mathbb{R}^p} \sum_{i=1}^{N} (y_i - f(x_i))^2 \ .$$

This estimation provides us w. As we want to compare probabilities of label y_i for some certain value, the function mentioned above is not appropriate. Instead we need a model that gives us posterior probabilities of the classes via linear functions that map values to an intervall $[0, 1]$ [Hastie et al. 2003, Chapter 4.4]. The sigmoid function has such properties as one can see in Figure 3.1. Considering function $f(x)$, sigmoid function h is described as followed:

$$h(f(x)) = \frac{1}{1 + e^{-f(x)}} \ . \tag{3.1}$$



(a) $\frac{1}{1+e^{-f(x)}}$ (b) $\frac{1}{1+e^{f(x)}}$

Figure 3.1: Sigmoid function with both negative (a) and positive (b) exponent.

Using the above sigmoid function it is possible to determine the probability of the i-th label obtaining the value y_i: $\mathbb{P}[Y_i = y_i]$. To simplify the equation in our case we will use binary labels such as already suggested in Section 2.2.3. We summarize low-risk stages $\{1, 2, 4s\}$ to label -1 and high-risk stages $\{3, 4\}$ to label $+1$. This leads us to following equations:

$$\begin{aligned}\mathbb{P}[Y_i = +1] &= \frac{1}{1+e^{-f(x)}},\\ \mathbb{P}[Y_i = -1] &= 1 - \frac{1}{1+e^{-f(x)}} = \frac{e^{-f(x)}}{1+e^{-f(x)}} = \frac{1}{1+e^{f(x)}} \quad .\end{aligned}$$

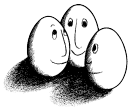
Thus we can summarize it to

$$\mathbb{P}[Y_i = y_i] = \frac{1}{1+e^{-y_i f(x_i)}} \quad . \tag{3.2}$$

The log-likehood for N observations $L(w) = \sum_{i=1}^{N} \log \mathbb{P}[Y_i = y_i]$ can be used to estimate w more precisely. In order to achieve a good estimation $L(w)$ is maximized over w:

$$\begin{aligned}\max_{w\in\mathbb{R}^p} L(w) &= \max_{w\in\mathbb{R}^p} \sum_{i=1}^{N} \log\left(\frac{1}{1+e^{-y_i f(x_i)}}\right)\\ &= \max_{w\in\mathbb{R}^p} -\sum_{i=1}^{N} \log(1+e^{-y_i f(x_i)})\\ &= -\min_{w\in\mathbb{R}^p} \sum_{i=1}^{N} \log(1+e^{-y_i f(x_i)}) \quad .\end{aligned} \tag{3.3}$$

This convex problem is ill-posed if p is much larger than N. Since in case of this research there are 31054 dimensions (exons), we introduce regularization terms to reduce number of dimensions in the next section.



3.2 Regularization terms

As described in Chapter 2 number of dimensions in biological data we are analyzing is much larger than number of examples resulting in an effect so called the "curse of dimensionality" that was coined by Bellman in 1961 [Hastie et al. 2003, Chapter 2.5]. Thus a subset of features should be found that is able to produce a model with low prediction error and good performance. For this purpose there is LASSO approach, a shrinkage method based on ℓ_1, that produces sparse solutions [Hastie et al. 2003, Chapter 3.4.2]. Another penalization term ℓ_2 is used in ridge regression and it provides coefficients for every variable [Hastie et al. 2003, Chapter 3.4.1]. $\lambda > 0$ is used as a tuning parameter.

$$\ell_1: \quad \lambda \sum_{j=1}^{p} |w[j]| = \lambda \|w\|_1 \qquad \ell_2: \quad \lambda \sum_{j=1}^{p} (w[j])^2 = \lambda \|w\|_2^2 \ .$$

Choosing higher values for λ sets some coefficients exactly to zero in case of ℓ_1. Hence for this work ℓ_1 is a more appropriate choice than ℓ_2 to reduce the number of dimensions.
Nevertheless given our biological data of thousands of exons we want to consider them in groups as we know the mapping of exons to genes before the splicing process. Therefore ℓ_1 can set some of exons of one gene to zero while setting some other exons of the same gene to non-zero values. As we know which exons belong to certain genes one can consider "grouped lasso" as described in Hastie et al. [2003, Chapter 3.8.4]. First every exon is represented as elements $x_i[j]$ of the vector x_i with $1 \leq i \leq N$ and $1 \leq j \leq p$. Furthermore we define S as a collection of sets of indices $\{1, 2, \ldots, p\}$ where each set $s \in S$ corresponds to a group defined by a gene. Every exon belongs consequently to a certain set s where $|s|$ is the size of s. For all $h = 1 \ldots |s|$ for some set s we define

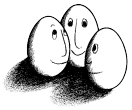
$$w^{[s]}[h] = w[j[h]] \text{ with } j[h] \in s \text{ and } 1 \leq j[h] \leq p \tag{3.4}$$

Using Definition (3.4) we can define *group* ℓ_1 regularization

$$\textit{Group } \ell_1: \quad \lambda \sum_{s \in S} \|w^{[s]}\|_2 \ .$$

The idea of setting groups of related exons to zero is be compared to another regularization method, *elastic net*. Zou and Hastie [2005] introduces it as a linear combination of ℓ_1 and ℓ_2:

$$\textit{Elastic net}: \quad \lambda_1 \|w\|_1 + \lambda_2 \|w\|_2^2$$



where $\lambda_1 > 0$ and $\lambda_2 > 0$. Elastic net selects variables with ℓ_1 and encourages highly correlated features to be selected together with ℓ_2. For that reason one may not need any prior knowledge of grouping. This concept is compared on the real data to group ℓ_1 and the results are discussed in Section 5.

3.3 Optimization

In Equation 3.3 we maximized log-likelihood to distinguish optimal solution for logistic regression. After introducing two regularizations terms in Section 3.2 we append them to the consisting optimization problem. This way we have

$$F_1(\lambda) = -\min_{w\in\mathbb{R}^p} \sum_{i=1}^{N} \log(1 + e^{-y_i f(x_i)}) + \lambda \sum_{s\in S} \|w^{[s]}\|_2 \quad , \tag{3.5}$$

$$F_2(\lambda_1, \lambda_2) = -\min_{w\in\mathbb{R}^p} \sum_{i=1}^{N} \log(1 + e^{-y_i f(x_i)}) + \lambda_1\|w\|_1 + \lambda_2\|w\|_2^2 \quad . \tag{3.6}$$

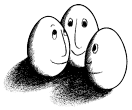
In the following two optimization approaches are presented and compared with respect to the Equations (3.5) and (3.6).

3.3.1 Stochastic gradient descent (SGD)

One of the wide spread approaches is stochastic gradient descent (SGD) suggested in 50's [Robbins and Monro 1951; Kiefer and Wolfowitz 1952]. For ease of explanation let us define $\phi(w) = \min_w L(w) + R(w)$ where $L(w) = \sum_{i=1}^{N} \log(1 + e^{-y_i f(x_i)})$ is loss function and $R(w)$ is one of regularizers for group ℓ_1 or elastic net. Our aim is to find best fitting weight vector for $\min_w \phi(w)$. There are two conditions for using SGD: $w \in W$, $W \subset \mathbb{R}^p$ while W should be convex and compact (Appendix). SGD as an iterative approach recieves first estimated "guess" of the weight vector and works for $M > 0$ iterations. With knowledge of k-th estimation of w one can compute w_{k+1} by solving the following minimization problem:

$$w_{k+1} = \arg\min_{w\in W} \{\langle g_k, w - w_k\rangle + \frac{\eta_k}{2}\|w - w_k\|_2^2\} \quad .$$

where $g_k \in \partial\phi(w_k)$, $\eta_k = \frac{\sqrt{k}}{\theta}$ for some $\theta > 0$ given and $0 \leq k \leq M$. SGD has a disadvantage of treating loss function and regularization term as one loss function. Thus in case of nondifferentiable regularization terms like group ℓ_1 SGD is hardly generating any sparse solution we aimed to get by regularization.



3.3.2 Regularized dual average (RDA)

Another approach which fits for our minimization problem better was introduced in Xiao [2010]. It is called regularized dual averaging (RDA). It can be seen as SGD with some improvements. First we summarize Equations (3.5) and (3.6) as $\min_w L(w) + R(w)$ where $L(w) = \sum_{i=1}^{N} \ell_i(w)$ is a total differentiable loss function with $\ell_i(w) = \log(1 + e^{-y_i f(x_i)})$ and $R(w)$ is nondifferentiable regularizer $R(w) = \lambda \sum_{s \in S} \|w^{[s]}\|_2$ for group ℓ_1 and $R(w) = \lambda_1 \|w\|_1 + \lambda_2 \|w\|_2^2$ for elastic net. Some differentiable regularizer can be included into $\ell_i(w)$ for ease of notice. RDA calculates stepwise next better fitting solution in a form

$$w_{k+1} = \arg\min_{w \in \mathbb{R}^p} \{\langle \bar{g}_k, w - w_1 \rangle + R(w) + \frac{\alpha_k}{2} \|w - w_1\|_2^2\}$$

where $\bar{g}_k = \frac{1}{k} \sum_{j=1}^{k} \nabla \ell_j(w)$, $0 \leq k \leq M$ and $\alpha_k = \frac{\theta'}{\sqrt{k}}$ for some constant $\theta' > 0$. Introducing the term $R(w)$ into the calculation of the next step leads to no loss of sparsity in contrast to SGD and therefore we will apply this method to optimize (3.5) and (3.6).

Algorithm 1 RDA optimization as used for elastic net

Input:

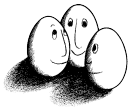
- nonnegative and nondecreasing sequence $\{\alpha_k\}$

Initialize: set $w_1 = \arg\min_w \|w\|_2^2 = 0$ and $\bar{g}_0 = 0$

for k=1,2,3,... **do**

1. Given a sample $b_k \in \{1, 2, \ldots, N\}$ compute a subgradient $g_k \in \partial \ell_{b_k}(w_k)$
2. Update the average sub gradient:
$$\bar{g}_k = \frac{k-1}{k} \bar{g}_{k-1} + \frac{1}{k} g_k$$
3. Compute the next weight vector:
$$w_{k+1} = \arg\min_w \left\{ \langle \bar{g}_k, w \rangle + R(w) + \frac{\alpha_k}{2} \|w\|_2^2 \right\}$$
with $R(w)$ is a regularizer.

end for



3.4 RDA solution

In previous section we introduced RDA solving optimization problems with non-differentiable functions/regularizers in contrast to SGD. Algorithm 1 summarizes RDA. In the following we present more detailed solutions for logistic regression with group ℓ_1 and elastic net respectively relating to RDA process. In both cases we consider some iteration k of RDA process as defined in Section 3.3.2. Although we concentrate on explaining how to compute next estimation of the weight vector for the following RDA iteration, it is only guaranteed that average over all found weight vectors

$$\bar{w}_k = \frac{1}{k} \sum_{d=1}^{k} w_d$$

is converging against the optimal solution [Xiao 2010, Section 4]. Thus we use $\bar{w}_k$ for prediction purposes and to find manifold (more details in Section 3.4.3).

3.4.1 Elastic net solution

Based on a closed-form solution for ℓ_1-RDA method suggested by Xiao [2010, Appendix A] we split a p-dimensional problem into p of 1-dimensional problems:

$$\min_{w_k[j] \in \mathbb{R}} g_k[j] w_k[j] + \lambda_1 |w_k[j]| + \lambda_2 w_k[j]^2 + \frac{\alpha_k}{2} w_k[j]^2$$

with $j \in 1 \dots p$, $\lambda_1 > 0$, $\lambda_2 > 0$ and $\alpha_k = \frac{\theta'}{\sqrt{k}}$ for some constant $\theta' > 0$.In order to find a solution we build a derivative and consider the optimality condition:

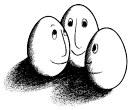
$$g_k[j] + \lambda_1 \psi + \rho \; w_k[j] \ni 0$$

where $\rho = 2\lambda_2 + \alpha_k$ and ψ is the subdifferential for $|w_k[j]|$ that can be described as following for three different cases:

$$\partial |w_k[j]| = \begin{cases} \{\psi \in \mathbb{R} | -1 \leq \psi \leq 1\} & \text{if } w_k[j] = 0, \\ \{1\} & \text{if } w_k[j] > 0, \\ \{-1\} & \text{if } w_k[j] < 0. \end{cases}$$

After considering these three cases in detail (see Xiao [2010, Appendix A]) we achieve the following closed-form solution for each w_j:

$$w_k[j] = \begin{cases} 0 & \text{if } |g_k[j]| \leq \lambda_1, \\ -\frac{1}{\rho}(g_k[j] - \lambda_1 \; sgn(g_k[j])) & \text{otherwise} \end{cases}$$



where gradient $g_k[j]$ with $1 \leq j \leq p$ is defined as followes:

$$\begin{aligned} g_k[j] &= \frac{\partial f_k}{\partial w[j]}(w_k[j]) = \frac{\partial \log(1 + e^{-y_{i_k} \sum_{m=1}^{p} x_{i_k}[m] w_k[m]})}{\partial w[j]}(w_k[j]) \\ &= \frac{e^{-y_{i_k} \sum_{m=1}^{p} x_{i_k}[m] w_k[m]}}{1 + e^{-y_{i_k} \sum_{m=1}^{p} x_{i_k}[m] w_k[m]}}(-y_{i_k})\ x_{i_k}[j] \\ &= \frac{-y_{i_k}}{1 + e^{y_{i_k} \sum_{m=1}^{p} x_{i_k}[m] w_k[m]}}\ x_{i_k}[j]. \end{aligned}$$

3.4.2 Group ℓ_1 solution

For group ℓ_1 we use the same approach suggested by Xiao to solve the optimization problem with ℓ_1 regularizer. This time the situation is slightly different. We use the same notation as defined in Section 3.2. Then we can split the p-dimensional problem into $|s|$-dimensional subproblems where $|s|$ is the size of s:

$$\min_{w_k^{[s]} \in \mathbb{R}^{|s|}} \langle g_k^{[s]}, w_k^{[s]} \rangle + \lambda \|w_k^{[s]}\|_2 + \frac{\alpha_k}{2} \|w_k^{[s]}\|_2^2$$

Thus to calculate each of the elements' values we consider optimality criterion

$$g_k^{[s]}[v] + \lambda \frac{w_k^{[s]}[v]}{\|w_k^{[s]}\|_2} + \alpha_k w_k^{[s]}[v] = 0$$

with $v = 1 \ldots |s|$. After some transformations we first achieve

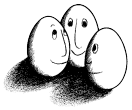
$$\begin{aligned} w_k^{[s]}[v] \left(\frac{\lambda}{\|w_k^{[s]}\|_2} + \alpha_k \right) &= -g_k^{[s]}[v] \\ w_k^{[s]}[v] &= -\left(\frac{\lambda}{\|w_k^{[s]}\|_2} + \alpha_k \right)^{-1} g_k^{[s]}[v] \end{aligned} \tag{3.7}$$

As we know the following definition

$$\|w_k^{[s]}\|_2^2 = \sum_{i=1}^{|s|} (w_k^{[s]}[v])^2 \tag{3.8}$$

we apply the knowledge from (3.7) and have

$$\begin{aligned} \|w_k^{[s]}\|_2^2 &= \sum_{i=1}^{|s|} \left(\frac{\lambda}{\|w_k^{[s]}\|_2} + \alpha_k \right)^{-2} (g_k^{[s]}[v])^2 \\ &= \left(\frac{\lambda}{\|w_k^{[s]}\|_2} + \alpha_k \right)^{-2} \sum_{i=1}^{|s|} (g_k^{[s]}[v])^2 \end{aligned} \tag{3.9}$$



Taking the square root of (3.9) we have

$$\|w_k^{[s]}\|_2 = \left(\frac{\lambda}{\|w_k^{[s]}\|_2} + \alpha_k\right)^{-1} \|g_k^{[s]}\|_2 \tag{3.10}$$

After multiplying $\left(\frac{\lambda}{\|w_k^{[s]}\|_2} + \alpha_k\right)$ on both sides we have

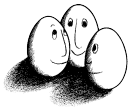
$$\begin{aligned} \|w_k^{[s]}\|_2 \left(\frac{\lambda}{\|w_k^{[s]}\|_2} + \alpha_k\right) &= \|g_k^{[s]}\|_2 \\ \Rightarrow \lambda + \alpha_k \|w_k^{[s]}\|_2 &= \|g_k^{[s]}\|_2 \\ \Rightarrow \|w_k^{[s]}\|_2 &= \frac{\|g_k^{[s]}\|_2 - \lambda}{\alpha_k} \end{aligned} \tag{3.11}$$

If $\|g_k^{[s]}\|_2 > \lambda$, then $\|w_k^{[s]}\|_2 > 0$ and we deduce from (3.7) and (3.11)

$$\begin{aligned} w_k^{[s]}[v] &= -\left(\frac{\lambda}{\frac{\|g_k^{[s]}\|_2 - \lambda}{\alpha_k}} + \alpha_k\right)^{-1} g_k^{[s]}[v] \qquad \text{with } \frac{\|g_k^{[s]}\|_2 - \lambda}{\alpha_k} > 0 \\ &= -\left(\frac{\lambda \alpha_k}{\|g_k^{[s]}\|_2 - \lambda} + \alpha_k\right)^{-1} g_k^{[s]}[v] \\ &= -\left(\frac{\lambda \alpha_k + \alpha_k(\|g_k^{[s]}\|_2 - \lambda)}{\|g_k^{[s]}\|_2 - \lambda}\right)^{-1} g_k^{[s]}[v] \\ &= -\frac{\|g_k^{[s]}\|_2 - \lambda}{\alpha_k \|g_k^{[s]}\|_2} g_k^{[s]}[v] \\ &= -\frac{1 - \frac{\lambda}{\|g_k^{[s]}\|_2}}{\alpha_k} g_k^{[s]}[v] \end{aligned}$$

Otherwise in case of $\|g_k^{[s]}\|_2 \leq \lambda$ we set all the values of a group s to zero. We can summarize this closed-form solution as followes:

$$w_k^{[s]}[v] = \begin{cases} -\frac{1 - \frac{\lambda}{\|g_k^{[s]}\|_2}}{\alpha_k} g_k^{[s]}[v] & \text{if } \|g_k^{[s]}\|_2 > \lambda, \\ 0 & \text{otherwise.} \end{cases}$$



3.4.3 Convergence criterion and manifold identification

If we assume w^* as an optimal solution, then the difference of the expected function value at $\bar{w}_k$ and the optimal function value after k iterations can be bounded as followes [Xiao 2010, Theorem 3]:

$$\mathbf{E}\phi(\bar{w}_k) - \phi(w^*) \leq \frac{1}{\sqrt{k}}\left(\theta' D^2 + \frac{G^2}{\theta'}\right)$$

where $\|g_k\|_2 \leq G$ with $\forall k \geq 1$ is a uniform upper bound on the norms of the subgradients of g_k and for some $D > 0$ that satisfies $h(w^*) \leq D^2$. Hence after k iterations our estimate is at most $\mathcal{O}\left(\frac{1}{\sqrt{k}}\right)$ away from the optimum.
According to Lee and Wright [2012, Definition 3] we can define a manifold $\mathcal{M}$ for a special case of $R(w) = \|w\|_1$. Manifold definitions for elastic net and group ℓ_1 are similar to ℓ_1. Given set of indices $\{1, 2, \ldots, p\}$ (in our case these are the indices of the features) and some $\bar{z} \in \mathbb{R}^p$ we can define a partition of those indices into 3 disjoint subsets $\mathcal{P} \subseteq \{1, 2, \ldots, p\}$, $\mathcal{N} \subseteq \{1, 2, \ldots, p\}$ and $\mathcal{Z} \subseteq \{1, 2, \ldots, p\}$ that implies $\mathcal{P} \cap \mathcal{N} = \mathcal{N} \cap \mathcal{Z} = \mathcal{P} \cap \mathcal{Z} = \emptyset$. Hence we have

$$\begin{aligned}\{1, 2, \ldots, p\} &= \mathcal{P} \cup \mathcal{N} \cup \mathcal{Z} \\ &\text{with } \bar{z}_r = 0 \; \forall r \in \mathcal{Z}, \quad \bar{z}_r > 0 \; \forall r \in \mathcal{P}, \quad \bar{z}_r < 0 \; \forall r \in \mathcal{N}\end{aligned}$$

To fulfill Lee and Wright [2012, Definition 3] a map $H : \mathbb{R}^p \to \mathbb{R}^q$ can be constructed as $H(z) = z[r]_{r \in \mathcal{Z}}$ with $q = card(\mathcal{Z})$. Thus a manifold $\mathcal{M}$ is then

$$\mathcal{M} = \{z \in \mathbb{R}^p \mid z_r = 0 \; \forall r \in \mathcal{Z}\} = \{z \in \mathbb{R}^p \mid H(z) = 0\}.$$

Using this definition we estimate an upper bound for the probability to detect manifold $w_k \in \mathcal{M}$ after k iterations in a general convex case that matches group ℓ_1 penalization [Lee and Wright 2012, Theorem 16]:

$$\mathbb{P}(w_k \in M) \in \mathcal{O}\left(1 - \frac{1}{\sqrt[4]{k}}\right). \tag{3.12}$$

In a similar way we use Lee and Wright [2012, Theorem 17] to estimate an upper bound for the probability to detect manifold $w_k \in \mathcal{M}$ after k iterations in a strongly convex case that corresponds to our elastic net penalization:

$$\mathbb{P}(w_k \in M) \in \mathcal{O}\left(1 - \sqrt{\frac{6 + \ln k}{k}}\right). \tag{3.13}$$

In both convex and strongly convex cases $\mathcal{P}(w_k \in M)$ is increasing and converging to 1. In Figure 3.2 we can see the two curves supporting this proposition. Hence

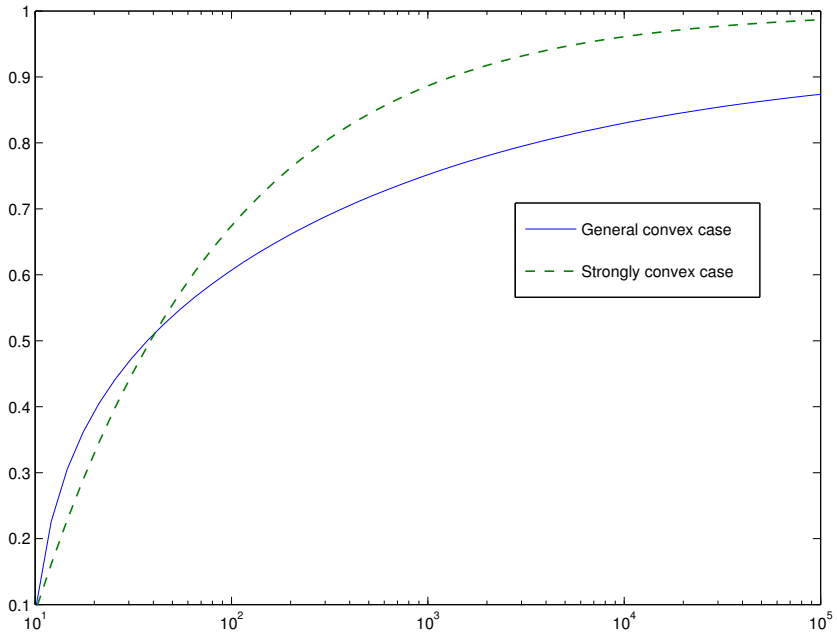
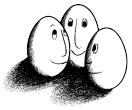


Figure 3.2: Upper bound of the probability for RDA to identify a manifold in k iterations in a convex case is $\mathcal{O}\left(1 - \frac{1}{\sqrt[4]{k}}\right)$. For strongly convex problems RDA converges to an optimal solution after k iterations with a probability bounded by $\mathcal{O}\left(1 - \sqrt{\frac{6+\ln k}{k}}\right)$. The comparison of these upper bounds identifies faster convergence of RDA in strongly convex problems.

elastic net as we can see is able to identify manifold faster than group ℓ_1 and at 10^5 iterations the probability is about 1. Group ℓ_1 as a general convex problem needs more iterations to be performed to achieve the same level of probability as elastic net.



4 Experiments

4.1 Rapidminer operators

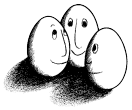
Two operators for Rapidminer have been implemented:

- group ℓ_1 operator[4]:
  - input: example set on which a model should be learned
  - output: learned model, input example set
  - further arguments: λ, θ, *epochs*, "iterations until convergence";
- elastic net operator[5]:
  - input: example set on which a model should be learned
  - output: learned model, input example set
  - further arguments: λ_1, λ_2, θ, *epochs*, "iterations until convergence".

In both operators we use an parameter called *epoch*. One *epoch* defines one iteration over all patients of the training set (e.g. 70% of 257). After each *epoch* the order of the patients is shuffled in the internal data structure.
Another extra parameter is called "iterations until convergence". We introduced it in Section 3.4.3. Hence we stop learning process after a set of selected features is found that has not changed for the last n iterations. To have upper bound of iterations in real situation both discussed features are correlated in the following way:

- in case a set of selected features is found that has not changed within the last n iterations, the training is stopped;
- in case we were not able to find a set of non-changing features we stop the training after $2k$ epochs.

[4]https://bitbucket.org/AEgorov/logistic-regression-with-group-l1/
[5]https://bitbucket.org/AEgorov/logistic-regression-with-elastic-net/



4.2 Experiment design

In order to detect the best fitting solution we have to figure out the parameters given to the operators that optimize the classifier. To achieve these two goals we perform a validation process consisting of two validation processes:

- cross-validation (called "inner validation" later)
- validation with random sampling (called "outer validation" later)

In the following we explain these validation processes and how they work together on group ℓ_1 operator and its tuning parameter λ.

1. First we use random sampling to split the given data into training set TR and test set TE.
   a) In the inner validation we consider TR and apply 10-fold cross-validation for a given λ on it. To do this we optimize the function (3.5) using group ℓ_1 operator.
   b) This step is repeated on different values of λ from a predefined interval. After computing different models and comparing their accuracy we select λ that produces produces model with the highest performance on this TR.
2. After optimal λ for TR is fixed, we compute the weight vector w for TR as above and test its accuracy on TE.

Repeating steps 1 to 2 for 10 times (10-fold outer validation) each time splitting given data set into different TR and TE sets will give us 10 possible models using feature selection.

	group ℓ_1	elastic net
λ / λ_1	0.00001, 0.0533427, 0.1066753, 0.160008, 0.213341, 0.266673, 0.320006, 0.373339, 0.426671, 0.480004, 0.533337, 0.586669, 0.640002, 0.693335, 0.746667, 0.8	0.0000001, 0.00900009, 0.01800008, 0.02700007, 0.03600006, 0.04500005, 0.05400004, 0.06300003, 0.07200002, 0.0810000, 0.09
λ_2	n/a	0.0001
θ	1	1
Epoch	200	100
IUC	150	30 - 60

Table 4.1: Values used in experiments. IUC = "Iterations until convergence". Values in this table are rounded and were more precise in the experiments.

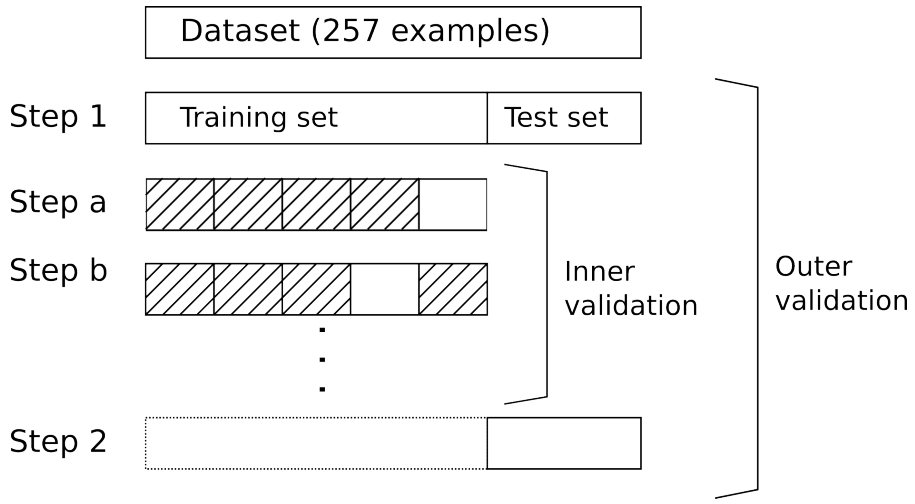


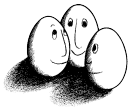
Figure 4.1: Validation with random sampling and inner cross-validation. After randomly spitting the data into training and test sets inner cross-validation of several parameters is completed. Performance of a model with parameter chosen in cross-validation is measured.

Figure 4.1 shows a draft of the described processes where the dataset is split randomly into different training and test sets after every fold in "Step 1" of outer validation. Then inner validation ("Step a" and "Step b") is repeated with different parameters values, before parameters producing the highest performance values are evaluated in "Step 2" of outer validation.

In case of elastic net we optimize function (3.6) using elastic net operator. The process described above is repeated the same way, but using two tuning parameters (λ_1 and λ_2).

In Table 4.1 you can find parameter values used for described operators. *epoch*-values are just guide values as it depends on "Iterations until convergence". We described it in previous section. Though "Iterations until convergence" varies between the two operators, achieved accuracy is comparable. This result corresponds with the different convergence rate that we stated in Section 3.4.3 which is needed to identify the manifold in group ℓ_1 and elastic net. Hence elastic net has a lower epoch and "Iterations until convergence" values.

As an overall result of the experiments described above we achieve sparse weight vectors representing features selected by group ℓ_1 (genes containing exons) and elastic net (exons). In this way in 10 outer validation steps we produce 10 classifiers with each of them having different performance values. In Table 4.2 you can find performance results of the found classifiers and number of selected exons for both approaches and for group ℓ_1 also the number of selected groups (genes). According to mean values group ℓ_1 achieves slightly higher values for accuracy, f-measure and sensitivity (see Appendix for explanation) while elastic net with average twice as

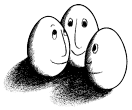


	ID	Accuracy	F-Measure	Sensitivity	Specificity	Groups	Exons
group ℓ_1	g1	0.794872	0.846154	0.956522	0.5625	175	5662
	g7	0.833333	0.876191	0.901961	0.703704	91	3062
	g10	0.807692	0.859813	0.958333	0.566667	79	2738
	g11	0.782051	0.831683	0.807692	0.730769	95	3192
	g13	0.858974	0.904348	0.962963	0.625	15	517
	g18	0.846154	0.87234	0.82	0.892857	89	2876
	g20	0.794872	0.826087	0.883721	0.685714	84	2929
	g23	0.807692	0.862385	0.959184	0.551724	86	2970
	g27	0.807692	0.869565	0.943396	0.52	115	3576
	g29	0.794872	0.84	0.875	0.666667	148	4615
	μ	**0.8128**	**0.8589**	**0.9069**	**0.6506**	**97.7**	**3213.7**
elastic net	d1	0.782051	0.831683	0.823529	0.703704	—	1933
	d2	0.820513	0.862745	0.916667	0.666667	—	2076
	d5	0.807692	0.845361	0.773585	0.88	—	2344
	h1	0.794872	0.849057	0.918367	0.586207	—	1780
	h12	0.833333	0.868687	0.934783	0.6875	—	2826
	h18	0.820513	0.862745	0.862745	0.740741	—	40
	h20	0.846154	0.888889	0.96	0.642857	—	6477
	h23	0.807692	0.864865	0.888889	0.625	—	48
	h24	0.782051	0.828283	0.872340	0.645161	—	31
	h29	0.782051	0.838095	0.862745	0.62963	—	45
	μ	**0.8077**	**0.8540**	**0.8814**	**0.6807**	—	**1760**

Table 4.2: Results of the cross-validation process 4.1. ID is an internal name of each iteration. Groups and exons account number of selected groups and exons respectively. Since elastic net does not need any grouping knowledge, we do not obtain any information of some kind of groups selected. All numbers have been rounded to fit this representation.

little exons selected than in group ℓ_1 can only produce higher specificity.

Though each of trained classifiers can demonstrate high accuracy, it is essentially to find set of genes or exons that can be used to predict stage and consequently survival of new patients. Unfortunately Ein-Dor et al. [2005] stated that different sets of genes/exons can produce similar classifiers depending on the chosen training set. This means that since in every new iteration of outer validation new training and test sets are randomly chosen, we can get models with different features, but similar performance. We observed similar behavior with classifiers in our experiments: all 10 classifiers prepared with group ℓ_1-operator do not have a single gene in common. Hence comparing these results is not useful. Therefore two further approaches are applied to evaluate the selected features:



1. We build an average over all average weight vectors according to one approach, build a union over all selected features in all runs and take them into account;
2. Select n features with the highest weights produced in step 1 (in case of group ℓ_1 we consider average group weight).

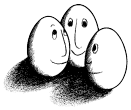
4.3 Comparing results

4.3.1 Prediction performance analysis

We figured out the most important set of features chosen in 10 runs of our in Figure 4.1 represented process. For the group ℓ_1 approach the built union consists of 329 genes while they contain 9449 exons in total. With elastic net penalization we were able to select 8113 exons in 10 outer validation runs altogether. We produced a new dataset containing only the united features of one approach and used these datasets to train our final models and test their accuracy. For this purpose Rapidminer operators mentioned above were used. This time we set λ_1 and λ_2 to zero to only use logistic regression function and not filtering out more features. Furthermore "iterations until convergence" was also set to zero as the set of features would not change without a regularization term. This case of "iterations until convergence"$= 0$ is implemented in the operators so that only the number of epochs is crucial. We set $\theta = 1$ and $epoch = 200$ for both new data sets. To evaluate the prediction power of these models we perform another validation process. The scheme stays similar to the previous process:

- dataset of 257 patients with a smaller set of features (exons) is split 30 times into training and test sets;
- the learned model is applied on the test set and prediction performance is measured.

We calculated accuracy, f-measure, sensitivity and specificity to measure the performance. Mean and standard deviation of those values give us a possibility to compare the produced prediction models.
We used both bootstraping and random sampling in this evaluation process. Steyerberg et al. [2001] suggests bootstrapping as the most efficient validation in terms of logistic regression analysis and small sample sets. Nevertheless random sampling shows comparably good results in case of group ℓ_1 as one can see in Table 4.3. Both validations emphasized that the model learned on features selected with elastic net approach is able to predict the low risk group better than the model built with the group ℓ_1 selected features (specificity is around 10 % higher). Concerning high risk



group prediction (sensitivity), even though bootstraping and random sampling are performing differently using the elastic net approach, they still present us with high performance value which is comparable to that from group ℓ_1. The overall accuracy of elastic net is still higher in both validation methods.

In Figure 4.4 we can find 25 genes with highest weights selected by group ℓ_1. We consider the exons chosen by elastic net approach and take those into account that are contained in top 25 genes mentioned above. In the following we examine data from several genes that are associated with neuroblastoma cancer. In Section 4.3.2 we mention some of those publications and what has been found so far about those genes according to neuroblastoma. In the following we observe weights calculated by the two approaches and evaluate and analyze the expression data of exons from some genes.

Althoug elastic net has no knowledge about grouping of exons to genes, it is working surprisingly well. In Figure 4.3 we observe that elastic net identified 26 out of 33 exons in gene MAP3K5 as important for classification. Mean expression values for labels 1 and -1 of exons 28 to 33 are almost overlapping inducing that their expression values are not differing so much to be used for classification. Thus it appears to make sense that elastic net does not select those features. Exon 7 has a similar behavior and hence was set to zero by the elastic net operator. Weights of exons 1, 4, 12, 14, 20 and 25 are quiet different in group ℓ_1 and elastic net. To reveal why this is happening a more precise analysis of the expression data of those particular exons would be needed.

Similar behavior can be observed in Figure 4.4. The curves of found weights for the gene TMEFF2 are partly similar. Though exon 14 has the smallest difference between mean expression values for label 1 and label -1 and is also not selected by elastic net operator. Interesting in this case is that the weight value by group ℓ_1 for this exon was set high compared to other weights in this gene by this approach. Also weights of exon 11 have a big gap in the figure confirming different penalization

			Accuracy	F-Measure	Sensitivity	Specificity
Group ℓ_1	Random sampling	μ	0.7983	0.8431	0.8660	0.6839
		std	0.1468	0.1527	0.1626	0.1406
	Bootstrapping	μ	0.7997	0.8447	0.8729	0.6780
		std	0.1465	0.1526	0.1590	0.1473
Elastic net	Random sampling	μ	0.8252	0.8588	0.8497	0.7839
		std	0.1515	0.1556	0.1620	0.1719
	Bootstrapping	μ	0.8443	0.8784	0.8855	0.7738
		std	0.1521	0.1570	0.1621	0.1626

Table 4.3: Final mean and standard deviation of 30 iterations of bootstrapping and random sampling with both group ℓ_1 and elastic net.

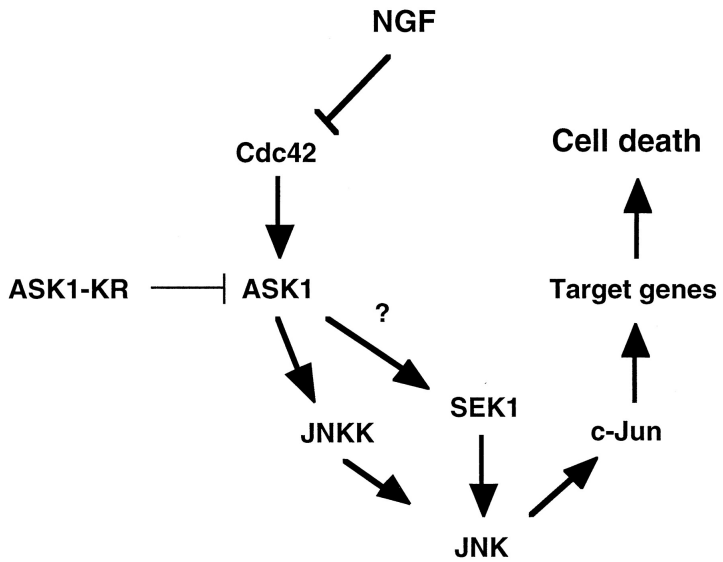


Figure 4.2: The name giving function of ASK1 is "apoptosis" which means a process of programmed cell death. Hence this figure shows a pathway of ASK1 with cell death at the end.[Kanamoto et al. 2000]

strategies of both approaches.

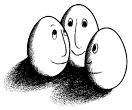
Two histones appear in the top rankings of group ℓ_1 penalized features. In Figure 4.5 represented weights of HIST1H1A for elastic net seem to rely on the difference between mean label expression values unlike group ℓ_1. It is surprinsingly to see that gene HIST1H3B that was selected by group ℓ_1 with the highest weight consisting of one single exon was not selected by elastic net in any of 10 runs of the validation process at all.

In the similar way Figures 4.6 and 4.7 point out the weights of NEGR1 and second highest weighted gene SLC24A2 respectively. In both cases we detect differences in group ℓ_1 and elastic net penalization terms setting several features weights dissimilar. Further results for PTPRZ1 and SRR can be seen in Figure 4.8 and Figure 4.9 respectively.

4.3.2 Biological importance

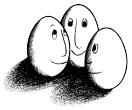
In Section 4.3.1 we analyzed weights and expression values of the genes and exons selected by group ℓ_1 and elastic net. In the following we consider publications about neuroblastoma or cancer research in general mentioning several of the genes as seen in Figure 4.4.

One of the selected genes was MAP3K5, mitogen-activated protein kinase kinase kinase 5. It is encoding apoptosis signal-regulating kinase (ASK1). The name giving function of ASK1 is "apoptosis" and this means a process of programmed cell death. Its crucial importance in cancer research has been cited in several publications. Kanamoto et al. [2000] investigated general role of ASK1 enzym and detected



Gene ID	Gene name	Name expanation	Weight
2946215	HIST1H3B	histone cluster 1, H3b	-8.208
3200762	SLC24A2	solute carrier family 24 (sodium/potassium/calcium exchanger), member 2	-5.0801
2946194	HIST1H1A	histone cluster 1, H1a	-4.8779
3762753	CA10	carbonic anhydrase X	-4.7127
4026075	GABRA3	gamma-aminobutyric acid (GABA) A receptor, alpha 3	4.0264
3456805	GTSF1	gametocyte specific factor 1	4.0049
4026119	MAGEA10	melanoma antigen family A, 10	3.6744
2602770	DNER	delta/notch-like EGF repeat containing	-3.6167
3127156	GFRA2	GDNF family receptor alpha 2	3.4531
3361811	STK33	serine/threonine kinase 33	3.3473
3706219	SRR	serine racemase	-3.2492
3452478	AMIGO2	adhesion molecule with Ig-like domain 2	-3.1524
3985717	PLP1	proteolipid protein 1	-2.9398
2592598	TMEFF2	transmembrane protein with EGF-like and two follistatin-like domains 2	-2.746
2349402	AMY2B	amylase, alpha 2B (pancreatic)	-2.7031
3021377	PTPRZ1	protein tyrosine phosphatase, receptor-type, Z polypeptide 1	-2.6879
2975867	MAP3K5	mitogen-activated protein kinase kinase kinase 5	-2.5206
3216276	SLC35D2	solute carrier family 35, member D2	-2.5061
2418078	NEGR1	neuronal growth regulator 1	-2.4708
3825609	NCAN	neurocan	2.4678
3373070	LOC441601	septin 7 pseudogene	2.3774
2980516	CNKSR3	CNKSR family member 3	-2.3299
2378180	C1orf107	chromosome 1 open reading frame 107	2.2586
2343025	AK5	adenylate kinase 5	-2.2561
3013565	DYNC1I1	dynein, cytoplasmic 1, intermediate chain 1	2.2534

Table 4.4: Top 25 weighted genes. The weights has been computed with the final model using only features selected with group ℓ_1 penalization of logistic regression in previous iterations of the process represented in Figure 4.1.



pathway which is showed in Figure 4.2. Arvidsson et al. [2001] even already analyzed ASK1 resistant neuroblastoma. Some other research of ASK1 in malignant fibrous histiocytomas was done in Chibon et al. [2004]. Further research of growth inhibition of colon cancer cells has been done by Kuwamura et al. [2007].
TMEFF2, transmembrane protein with EGF-like and two follistatin-like domains 2, was analyzed in Ali and Knauper [2007] which is implicated in cell signaling, neuronal cell survival, tumor suppression and Alzheimer disease. The connection between ADAM, a disintegrin and metalloproteinase, and TMEFF2 in prostate cancer is studied in that publication. According to this publication TMEFF2 may intrude some ADAM's pathways in ErbB signaling. Since ErbB family members are considered in Okada and Koizumi [1997] and Richards et al. [2010] as part of neuroblastoma research, it may be of great interest to include TMEFF2 into more detailed studies of its possible impact on neuroblastoma cancer cells.
Glial cell line-derived neurotrophic factor (GDNF) family that contains GFRA2 gene selected in this study is described to have some impact on neuroblastoma [Hansford and Marshall 2005].
Serine racemase (SRR) takes part in the regulation process of glutamate-N-methyl-D-aspartate (NMDA) [Wolosker et al. 1999] while NMDA as a glutamate receptor was discovered in neuroblastoma cell line in 1997 [North et al. 1997].
NEGR1 (neuronal growth regulator 1) is next to MYEOV one of the novel candidate gene targets in neuroblastoma. The results and findings of Takita et al. [2011] revealed significantly lower expression of this gene in neuroblastomas at an advanced stage of the disease and thus suggest a possible prognostic value for NEGR1 in neuroblastoma.
The two histones, HIST1H3B and HIST1H1A, that were high weighted by group ℓ_1 were mentioned in some publications according the neuroblastoma research. In Cotterman and Knoepfler [2009] HIST1H3B is stated in connection with MYCN expression. HIST1H1A can be found in [Pieler et al. 1981; Ajiro et al. 1990] as part of the research of histone H1 in neuroblastoma cells.

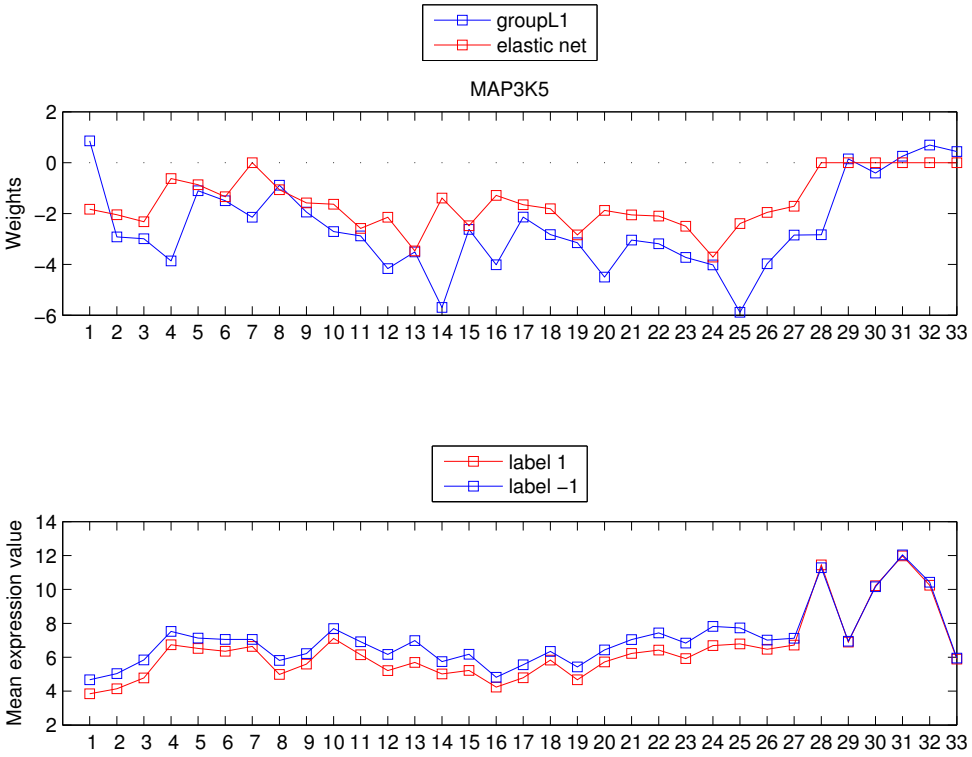


Figure 4.3: Elastic net identified 26 out of 33 exons in MAP3K5 as important while setting 7 exons to zero.

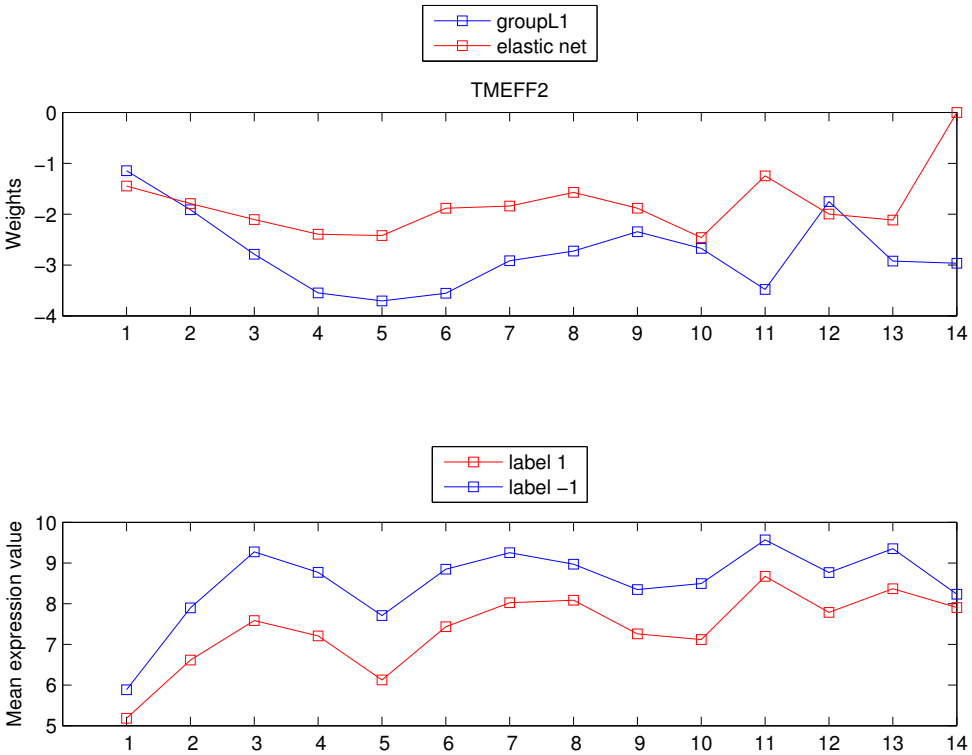


Figure 4.4: The curves of weights for TMEFF2 are similar in both approaches except exon 14 which is set to zero by elastic net and exon 11 that shows a big gap between the two approaches.

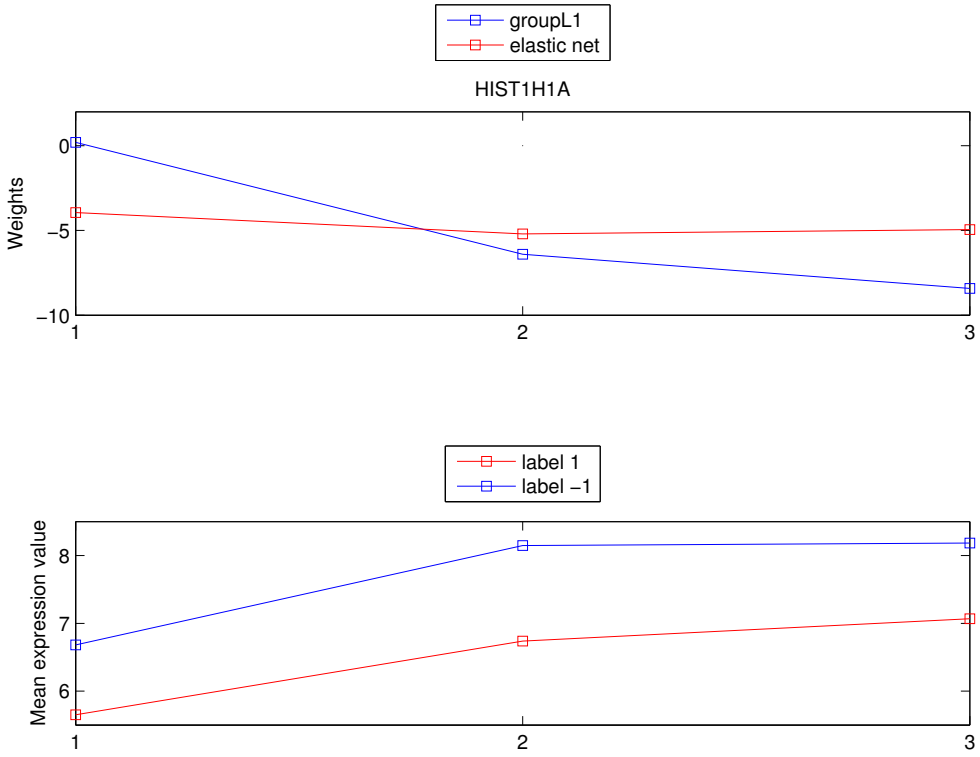


Figure 4.5: No exons in HIST1H1A were as to zero by elastic net, but the weights selected with group ℓ_1 seem not to correlate so much with the difference of mean label expression.

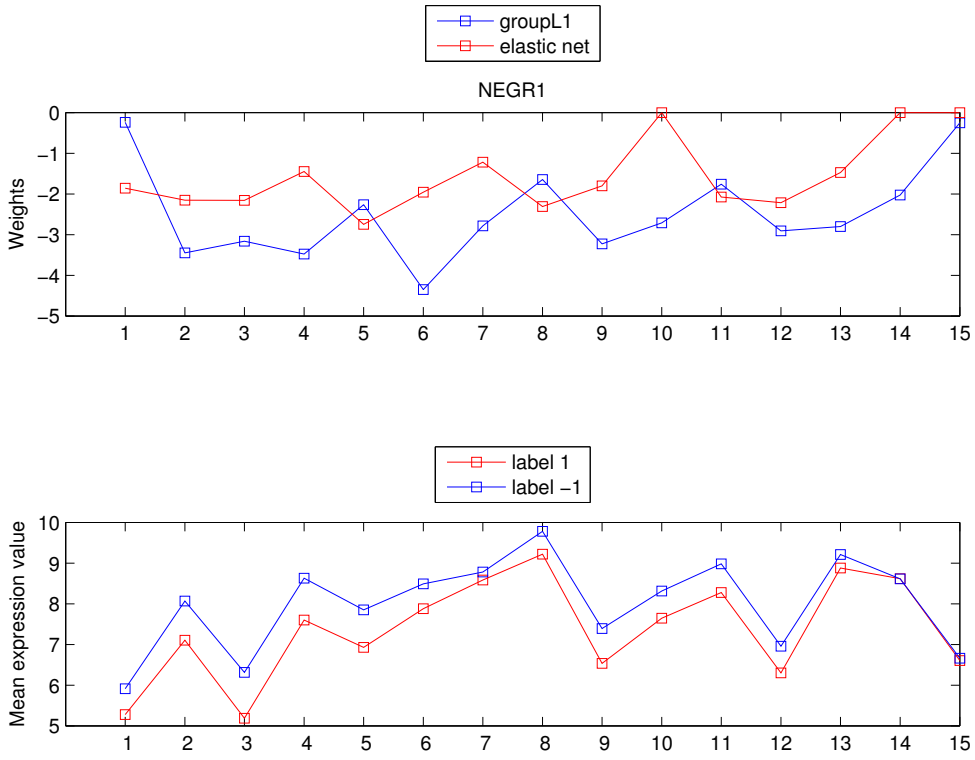


Figure 4.6: NEGR1 is one of the examples where we can detect several differences in group ℓ_1 against elastic net penalization terms.

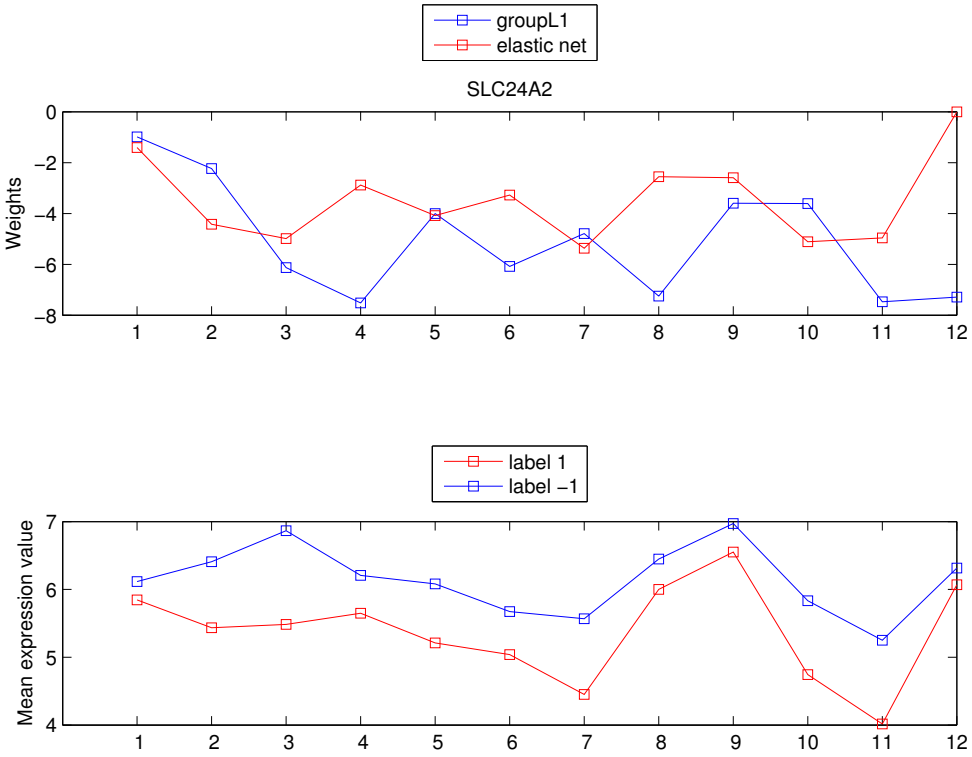


Figure 4.7: SLC24A2 gene is the second highest weighted gene by group ℓ_1 showing similar difference between the two regularizers as NEGR1 gene.

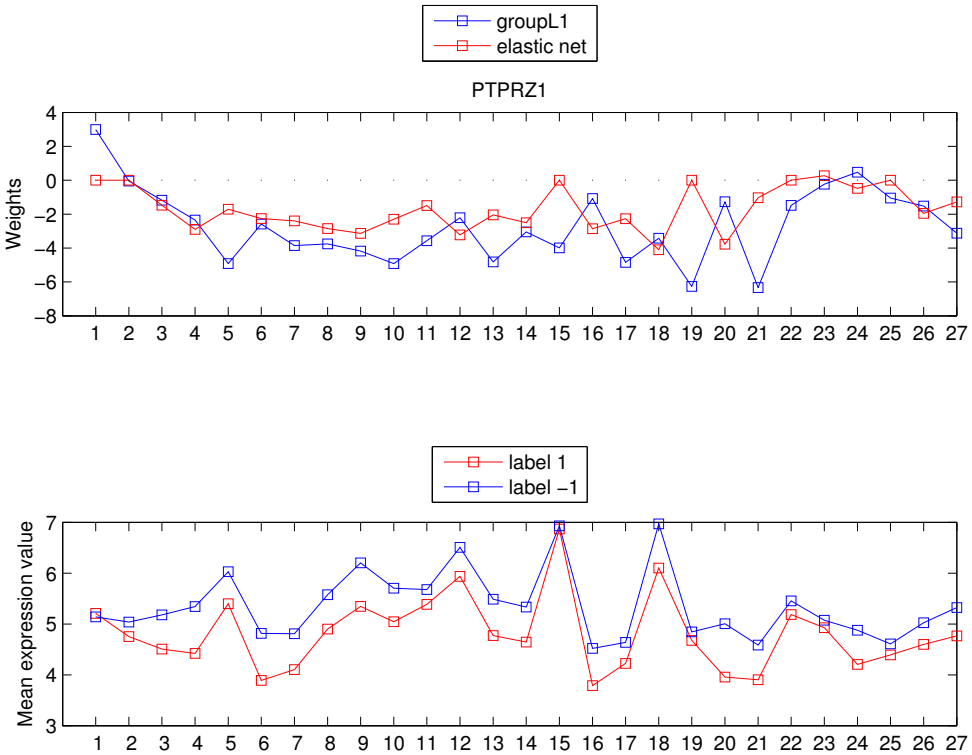


Figure 4.8: 6 out of 27 exons in PTPRZ1 gene has been set to zero by elastic net. Several exons show big gaps between the weights produced with different regularizers.

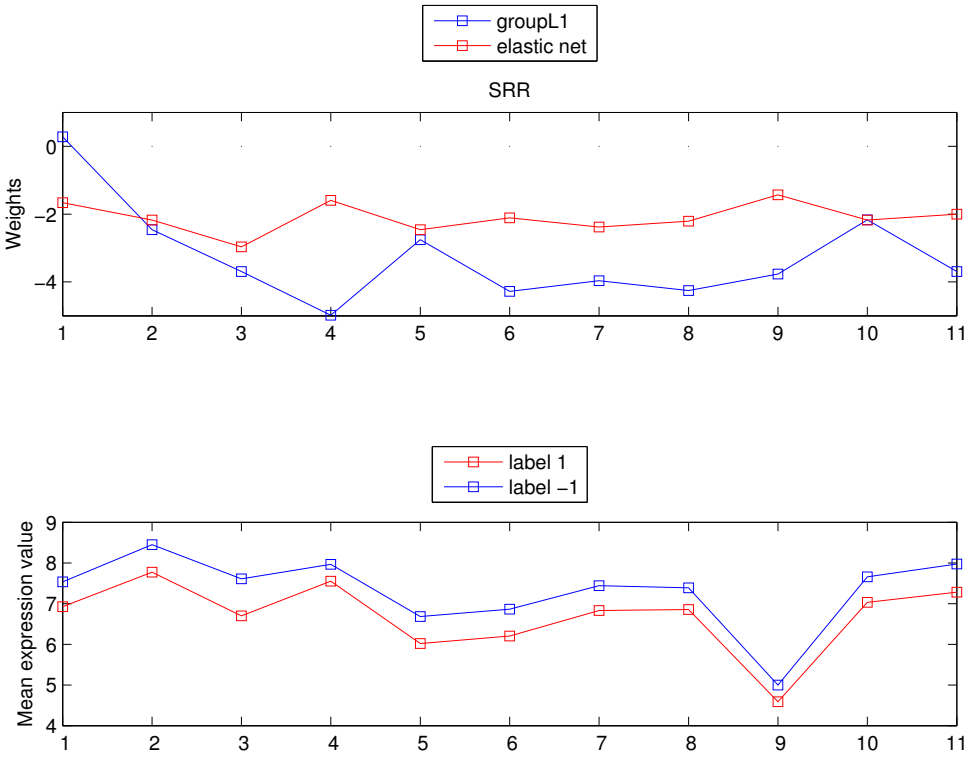
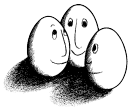


Figure 4.9: In SRR gene no features were set to zero by elastic net which can be explained by considering mean label expression values, although in case of group ℓ_1 exon 1 is set almost to zero.

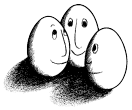


5 Discussion

In the course of this thesis we introduced two penalization terms for logistic regression, group ℓ_1 and elastic net. Our motivation was driven by a real life application of feature selection as a part of the neuroblastoma research. Using RDA algorithm we were able to derive an appropriate solution for the previously defined optimization problem that was containing nondifferentiable regularizers. Two Rapidminer operators were implemented to produce models using logistic regression with one of the regularizers. The computed solutions were sparse with both approaches delivering a comparably good performance.

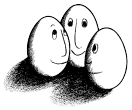
After analyzing results achieved in this work we identified that elastic net is able to make better predictions while still selecting similar features on a gene level, but choosing different weights for single exons. Though elastic net performed well, the running time of that approach was higher compared to group ℓ_1. Thus on the one hand some further analysis of running time for both methods can be included in next research. Furthermore, using enhanced algorithm RDA+ suggested by Lee and Wright [2012] could even lead to faster convergence of manifold identification in the implemented operators.

On the other hand since both approaches produce comparably good predictors, a combination of choosing features first on the gene level (group ℓ_1) and then filtering out further features on the exon level inside the groups (ℓ_1) could be analyzed in comparison to the results stated in this work. This approach, sparse group lasso, was described together with group ℓ_1 in Friedman et al. [2010]. A variation of sparse group lasso using elastic net instead of ℓ_1 as further penalization term could also be of interest and compared to sparse group lasso and results from this thesis.



Acknowledgement

This thesis would not be possible without Prof. Dr. Morik who raised my interest to the application of machine learning in the biomedical research. A special thank goes to Sangkyun Lee who supervised and advised me through all the work on this thesis. My colleagues at the chair of artificial intelligence at the TU Dortmund always helped and supported me. I am also grateful to my friends for their aid in reviewing this thesis and giving useful comments and hints. Last but not the least, the warm support of my family strengthened me though all the time.



Bibliography

K Ajiro, K Shibata, and Y Nishikawa. Subtype-specific cyclic amp-dependent histone h1 phosphorylation at the differentiation of mouse neuroblastoma cells. *J Biol Chem*, 265(11):6494–6500, 1990. 4.3.2

Nazim Ali and Vera Knauper. Phorbol ester-induced shedding of the prostate cancer marker transmembrane protein with epidermal growth factor and two follistatin motifs 2 is mediated by the disintegrin and metalloproteinase-17. *J Biol Chem*, 282(52):37378–37388, 2007. 4.3.2

Y Arvidsson, T S Hamazaki, H Ichijo, and K Funa. Ask1 resistant neuroblastoma is deficient in activation of p38 kinase. *Cell Death Differ*, 8(10):1029–1037, 2001. 4.3.2

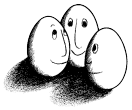
S. Asgharzadeh, R. Pique-Regi, R. Sposto, H. Wang, Y. Yang, H. Shimada, K. Matthay, J. Buckley, A. Ortega, and R.C. Seeger. Prognostic significance of gene expression profiles of metastatic neuroblastomas lacking mycn gene amplification. *Journal of the National Cancer Institute*, 98(17):1193–1203, 2006. 2.1

Garrett M Brodeur. Neuroblastoma: biological insights into a clinical enigma. *Nat Rev Cancer*, 3(3):203–216, 2003. 2.1

Q.R. Chen, Y.K. Song, J.S. Wei, S. Bilke, S. Asgharzadeh, R.C. Seeger, and J. Khan. An integrated cross-platform prognosis study on neuroblastoma patients. *Genomics*, 92(4):195, 2008. 2.1

Frédéric Chibon, Odette Mariani, Josette Derré, Aline Mairal, Jean-Michel Coindre, Louis Guillou, Xavier Sastre, Florence Pédeutour, and Alain Aurias. Ask1 (map3k5) as a potential therapeutic target in malignant fibrous histiocytomas with 12q14–q15 and 6q23 amplifications. *Genes, Chromosomes and Cancer*, 40 (1):32–37, 2004. 4.3.2

Rebecca Cotterman and Paul S. Knoepfler. N-myc regulates expression of pluripotency genes in neuroblastoma including <italic>lif, klf2</italic>, <italic>klf4</italic>, and <italic>lin28b</italic>. *PLoS ONE*, 4(6):e5799, 2009. 4.3.2



Liat Ein-Dor, Itai Kela, Gad Getz, David Givol, and Eytan Domany. Outcome signature genes in breast cancer: is there a unique set? *Bioinformatics*, 21(2): 171–178, 2005. 4.2

J. Friedman, T. Hastie, and R. Tibshirani. A note on the group lasso and a sparse group lasso. *arXiv preprint arXiv:1001.0736*, 2010. 5

Loen M Hansford and Glenn M Marshall. Glial cell line-derived neurotrophic factor (gdnf) family ligands reduce the sensitivity of neuroblastoma cells to pharmacologically induced cell death, growth arrest and differentiation. *Neurosci Lett*, 389(2):77–82, 2005. 4.3.2

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, corrected edition, 2003. 3.1, 3.2

T Kanamoto, M Mota, K Takeda, LL Rubin, K Miyazono, H Ichijo, and CE Bazenet. Role of apoptosis signal-regulating kinase in regulation of the c-jun n-terminal kinase pathway and apoptosis in sympathetic neurons. *Molecular and cellular biology*, 20(1):196–204, 2000. 4.2, 4.3.2

Kiefer and Wolfowitz. Stochastic estimation of the maximum of a regression function. *The Annals of Mathematical Statistics*, 23(3):462–466, 1952. 3.3.1

Kathreena M. Kurian, Christine J. Watson, and Andrew H. Wyllie. Dna chip technology. *Journal of Pathology*, 187:267–271, 1999. 2.2.1

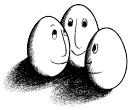
Hikaru Kuwamura, Kazunari Tominaga, Masayuki Shiota, Reiko Ashida, Takafumi Nakao, Eiji Sasaki, Toshio Watanabe, Yasuhiro Fujiwara, Nobuhide Oshitani, Kazuhide Higuchi, Hidenori Ichijo, Tetsuo Arakawa, and Hiroshi Iwao. Growth inhibition of colon cancer cells by transfection of dominant-negative apoptosis signal-regulating kinase-1. *Oncol Rep*, 17(4):781–786, 2007. 4.3.2

Sangkyun Lee. Internal report for SFB876 project C1. 2012. 2.1, 2.2

Sangkyun Lee and Stephen J. Wright. Manifold identification in dual averaging methods for regularized stochastic online learning. *Journal of Machine Learning Research*, 13:1705–1744, 2012. 3.4.3, 3.4.3, 3.4.3, 5

Katharina Morik. Medicine: Applications in machine learning. In Claude Sammut and Geoffrey I. Webb, editors, *Encyclopedia of Machine Learning*, pages 654–661. Springer, 2010. 1.1

R. Newson. Confidence intervals for rank statistics: Somers' d and extensions. *Stata Journal*, 6(3):309–334, 2006. 2.2.3



Andrew Y. Ng. Feature selection, l1 vs. l2 regularization, and rotational invariance. In *Proceedings of the twenty-first international conference on Machine learning*, ICML '04, pages 78–. ACM, 2004. 1.1

Jorge Nocedal and Stephen J. Wright. *Numerical Optimization*. Springer, New York, USA, 2000. A

W G North, M J Fay, J Du, M Cleary, J D Gallagher, and F V McCann. Presence of functional nmda receptors in a human neuroblastoma cell line. *Mol Chem Neuropathol*, 30(1-2):77–94, 1997. 1.1, 4.3.2

A. Oberthuer, B. Hero, F. Berthold, D. Juraeva, A. Faldum, Y. Kahlert, S. Asgharzadeh, R. Seeger, P. Scaruffi, G. P. Tonini, I. Janoueix-Lerosey, O. Delattre, G. Schleiermacher, J. Vandesompele, J. Vermeulen, F. Speleman, R. Noguera, M. Piqueras, J. Benard, A. Valent, S. Avigad, I. Yaniv, A. Weber, H. Christiansen, R. G. Grundy, K. Schardt, M. Schwab, R. Eils, P. Warnat, L. Kaderali, T. Simon, B. DeCarolis, J. Theissen, F. Westermann, B. Brors, and M. Fischer. Prognostic impact of gene expression-based classification for neuroblastoma. *Journal of Clinical Oncology*, 28(21):3506–3515, 2010. 1.1, 2.1, 2.2.3

Andre Oberthuer, Frank Berthold, Patrick Warnat, Barbara Hero, Yvonne Kahlert, Rudiger Spitz, Karen Ernestus, Rainer Konig, Stefan Haas, Roland Eils, Manfred Schwab, Benedikt Brors, Frank Westermann, and Matthias Fischer. Customized oligonucleotide microarray gene expression-based classification of neuroblastoma patients outperforms current clinical risk stratification. *J Clin Oncol*, 24(31): 5070–5078, 2006. 2.1

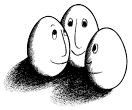
N Okada and S Koizumi. Tyrosine phosphorylation of erbb4 is stimulated by aurintricarboxylic acid in human neuroblastoma sh-sy5y cells. *Biochem Biophys Res Commun*, 230(2):266–269, 1997. 4.3.2

Christian Pieler, Guenther R. Adolf, and Peter Swetly. Accumulation of histone h1° during chemically induced differentiation of murine neuroblastoma cells. *European Journal of Biochemistry*, 115(2):329–333, 1981. 4.3.2

Kristen N Richards, Patrick A Zweidler-McKay, Nadine Van Roy, Frank Speleman, Jesus Trevino, Peter E Zage, and Dennis P M Hughes. Signaling of erbb receptor tyrosine kinases promotes neuroblastoma growth in vitro and in vivo. *Cancer*, 116(13):3233–3243, 2010. 4.3.2

Herbert Robbins and Sutton Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, 22(3):400–407, 1951. 3.3.1

Benjamin Schowe. Feature selection for high-dimensional data in rapidminer. In Simon Fischer and Ingo Mierswa, editors, *Proceedings of the 2nd RapidMiner*



Community Meeting And Conference (RCOMM 2011), Aachen, 2011. Shaker Verlag. 1.1

A. Schramm, J. H. Schulte, L. Klein-Hitpass, W. Havers, H. Sieverts, B. Berwanger, H. Christiansen, P. Warnat, B. Brors, J. Eils, R. Eils, and A. Eggert. Prediction of clinical outcome and biological characterization of neuroblastoma by expression profiling. *Oncogene*, 24(53):7902–12, 2005. 2.1

Alexander Schramm, Ingo Mierswa, Lars Kaderali, Katharina Morik, Angelika Eggert, and Johannes H. Schulte. Reanalysis of neuroblastoma expression profiling data using improved methodology and extended follow-up increases validity of outcome prediction. *Cancer letters*, 282(1):55–62, 09 2009. 1.1, 2.1, 2.2.2

Ewout W. Steyerberg, Frank E. Harrell Jr., Gerard J.J.M. Borsboom, M.J.C. Eijkemans, Yvonne Vergouwe, and J.Dik F. Habbema. Internal validation of predictive models: Efficiency of some procedures for logistic regression analysis. *Journal of Clinical Epidemiology*, 54(8):774 – 781, 2001. 4.3.1

Junko Takita, Yuyan Chen, Jun Okubo, Masashi Sanada, Masatoki Adachi, Kentaro Ohki, Riki Nishimura, Ryoji Hanada, Takashi Igarashi, Yasuhide Hayashi, and Seishi Ogawa. Aberrations of negr1 on 1p31 and myeov on 11q13 in neuroblastoma. *Cancer Sci*, 102(9):1645–1650, 2011. 1.1, 4.3.2

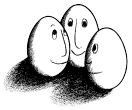
Herman Wolosker, Seth Blackshaw, and Solomon H. Snyder. Serine racemase: A glial enzyme synthesizing d-serine to regulate glutamate-N-methyl-d-aspartate neurotransmission. *Proceedings of the National Academy of Sciences*, 96(23): 13409–13414, 1999. 4.3.2

Lin Xiao. Dual averaging methods for regularized stochastic learning and online optimization. *Journal of Machine Learning Research*, 11:2543–2596, 2010. 3.3.2, 3.4, 3.4.1, 3.4.3

Wen Zhu, Nancy Zeng, and Ning Wang. Nesug 2010 sensitivity, specificity, accuracy, associated confidence interval and roc analysis with practical sas implementations. *Life Sciences*, pages 1–9, 2010. A

H. Zou and T. Hastie. Regression shrinkage and selection via the elastic net, with applications to microarrays. *Journal of the Royal Statistical Society: Series B. v67*, pages 301–320, 2003. 1.1

Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67 (2):301–320, 2005. 3.2



Appendix

Accuracy, specificity and sensitivity There are several terms that are commonly used along with the description of sensitivity, specificity and accuracy. They are true positive (TP), true negative (TN), false negative (FN), and false positive (FP). If a predicted value is "true" and actual value is also "true", the result is considered "true positive" (TP). Both "true positive" and "true negative" suggest a consistent result between the predicted and actual value (also called standard of truth). If we predict "true" value while it is actually "false", the result is "false positive" (FP). Similarly, if the predicted result is "false", but actual value is "true", the result is "false negative" (FN). Sensitivity, specificity and accuracy are described in terms of TP, TN, FN and FP:

- Sensitivity = TP/(TP + FN)
- Specificity = TN/(TN + FP)
- Accuracy = (TN + TP)/(TN+TP+FN+FP)

Source: [Zhu et al. 2010]

Bounded set The set F is *bounded* if there is some real number $M > 0$ such that $||x|| \leq M$ for all $x \in F$.

Closed set The set F is *closed* if for all possible sequences of points x_k in F, all limit points of x_k are elements of F.

Compact set The set F is *compact* if every sequence x_k of points in F has at least one limit point and all such limit points are in F. Thus: $F \in \mathbb{R}^n$ is closed and bounded $\Rightarrow F$ is compact.

Convex set A set $S \in \mathbb{R}^n$ is a *convex set* if the straight line segment connecting any two points in S lies entirely inside S. Formally, for any two points $x \in S$, we define $\alpha x + (1 - \alpha)y \in S$ for all $\alpha \in [0, 1]$.
Source: [Nocedal and Wright 2000, Chapter 1, Convexity]