# Statistical Learning Theory

Joachim Giesen

# Contents

# Framework

### Explained and explanatory variables

We are dealing with vectors of variables

$$(X_1, \dots, X_n, Y_1, \dots, Y_m)$$

that take values in the sets $\mathcal{X}_i$ and $\mathcal{Y}_j$, respectively. Here, the $Y_j$ are the *explained* variables that are also called *response* or *labels* and the $X_i$ are *explanatory* variables that are also called *features*. In the following we deal mostly with the case $m = 1$, i.e., there is only one explained variable $Y$. We also often summarize the vector $(X_1, \dots, X_n)$ in the variable $X$. The space $\mathcal{X}$, where $\mathcal{X} = \prod_{i=1}^{n} \mathcal{X}_i$, is sometimes called the *feature space*. The space $\mathcal{Y}$ is called the *label space*.

### Data generation model

We distinguish two types of data generation models:

1. In *generative models* we assume that there exists a probability density function $p$ on the space $\mathcal{X} \times \mathcal{Y}$. Our access to the model is such that we can query the probability density, i.e., we can perform random experiments that provide us with points in $\mathcal{X} \times \mathcal{Y}$.

2. In *discriminative models* we assume that there exists a family of probability density functions $p_x, x \in \mathcal{X}$ on the space $\mathcal{Y}$. Our access to the model now is that we can query the probability density $p_x$ at any point $x \in \mathcal{X}$. That is, the analyst can choose the explanatory variables $x$ and measure the explained variable $y \in \mathcal{Y}$ at $x$ in a random experiment. The densities $p_x$ are often of the form $g(y, f(x))$, where $f$ is a function on $\mathcal{X}$ and $g$ maps into the unit interval $[0, 1]$. In this case the variables $X$ are also called *covariates*, while the variables $Y$ can always be called *variates* since they are always random.

## Data

Typically we are given data points

$$(x^{(1)}, y^{(1)}), \ldots, (x^{(m)}, y^{(m)})$$

that have been obtained from measurements, i.e., by querying the underlying model. The data is all that we know about the model, though it might be possible to query more data points, maybe even at explanatory variables that we can choose ourselves.

It is common to assume that the measurements are *independent*, which means that the outcome of a measurement does not depend on the outcome of the other measurements. This is also known as the *i.i.d.* (independent identically distributed) assumption.

## Predictors

The ambitious goal would be learning the model (probability densities) from the data. If the learned model $\hat{p}$, from which the data have been generated, approximates the real model $p$ well, then we can use the learned model for predicting the outcome of future measurements, where the typical scenario is that we are given the explanatory variables $x$ and want to predict the explained variables $y$. For instance, in the discriminative case, if $\hat{p}_x$ is the learned model, then we can use the function

$$h : \mathcal{X} \to \mathcal{Y},\, x \mapsto \mathsf{argmax}_{y \in \mathcal{Y}}\, \hat{p}_x(y)$$

as a *predictor*.

In generative models the distinction between explained and explanatory variables is merely syntactical. We can also use the model, or better the approximation $\hat{p}$ of the model, that we have learned from the data for predicting the explanatory variables from the explained variables. Predicting the explained variables from the explanatory variables can be implemented by the following predictor

$$h : \mathcal{X} \to \mathcal{Y},\, x \mapsto \mathsf{argmax}_{y \in \mathcal{Y}}\, \hat{p}(y|X = x).$$

Analogously, we can predict the explanatory variables from the explained variables using the predictor

$$\bar{h} : \mathcal{Y} \to \mathcal{X},\, y \mapsto \mathsf{argmax}_{x \in \mathcal{X}}\, \hat{p}(x|Y = y).$$

Of course it is also possible to predict a mix of explanatory and explained variables from the remaining ones.

Learning a discriminative model that by definition is geared towards serving predetermined prediction queries is an instance of *supervised learning*, while learning a generative model that allows to support more general queries is an instance of *unsupervised learning*.
Often we are satisfied with learning a good predictor and not the whole model from which a predictor can be derived. Any function

$$h : \mathcal{X} \to \mathcal{Y}$$

can serve as a predictor and our task becomes to pick a good one given the observed data. In the following we refer to learning a good predictor when we speak of a learning problem.

## Measures of success

Of course we want to have a good predictor, which leaves us with the problem to define what we mean by good. A common approach is using a loss function

$$L : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_{\geq 0},$$

where the first argument is an observation and the second argument is a prediction. A good predictor is then a predictor $h$ with small expected loss

$$\mathsf{E}_p\big[L(Y|X = x, h(x))\big]$$

with respect to the model density $p$, in the generative case, and with small expected loss

$$\mathsf{E}_{p_x}\big[L(Y, h(x))\big]$$

in the discriminative case.
Alternatively, a good predictor incurs a large loss with only a small probability, i.e., depending on the constant $c > 0$, the probability

$$\mathsf{P}\big[L(Y, h(X)) \geq c\big] \;=\; \int_{\mathcal{X}\times\mathcal{Y}} \mathbf{1}\big[L(y, h(x)) \geq c\big]\, dp(x, y)$$

is small in the generative case, and the probabilities

$$\mathsf{P}\big[L(Y|X = x, h(x)) \geq c\big] \;=\; \int_{\mathcal{Y}} \mathbf{1}\big[L(y, h(x)) \geq c\big]\, dp_x(y)$$

are small in the discriminative case. Here, $\mathbf{1}[\text{condition}]$ denotes the characteristic function

$$\mathbf{1}[\text{condition}] \;=\; \begin{cases} 1 & : \;\; \text{condition is satisfied} \\ 0 & : \;\; \text{condition is not satisfied} \end{cases}$$

Note already here that both measures of goodness are not accessible to us since we have only indirect access to the data generation model through the query oracle.

## Scales and different types of learning problems

We can distinguish learning problems by the properties of the sets $\mathcal{X}$ or $\mathcal{Y}$, respectively. We say that $\mathcal{X}$ induces a

1. *nominal scale*, if $\mathcal{X}$ is a finite set.
2. *ordinal scale*, if $\mathcal{X}$ is a finite, ordered set.
3. *interval scale*, if $\mathcal{X} = \mathbb{R}$ and only the ratio of differences $\frac{|a-b|}{|c-d|}$, for $a, b, c, d \in \mathcal{X}$, is meaningful.
4. *ratio scale*, if $\mathcal{X} = \mathbb{R}_{\geq 0}$ and the ratio of elements in $\mathcal{X}$ is meaningful.

The latter two scales are also called *metric* scales, because distances $|a - b|$, for $a, b \in \mathcal{X}_i$, carry a meaning.
Among the possible learning problems the following received a lot of attention:

1. *Regression problems*, where the explanatory variables are all measured on metric scales and the explained variable is also measured on a metric scale.
2. *Classification problems*, where the explanatory variables are all measured on metric scales and the explained variable is measured on a nominal scale.
3. *Contingency analysis*, where the explanatory variables are all measured on nominal scales and the explained variable is also measured on a nominal scale.
4. *Variance analysis*, where the explanatory variables are all measured on nominal scales and the explained variable is measured on a metric scale.

Note for all these problems the explanatory variables are never measured on mixed scales. Of course such problems are also important in practice.
In the following we will discuss methods for special regression, classification and contingency analysis problems. These methods provide us already with a pretty good tool box that allows to address many of the real world machine learning problems.

# Part I

# Basic Learning Methods

# Chapter 1

# Ordinary least squares regression

Ordinary least squares is a classical regression technique that dates back to Carl Friedrich Gauß who might have used it in 1801 to rediscover the dwarf planet Ceres that had previously been discovered by the Italian monk Guiseppe Piazzi who was able to track it for 41 days before it got lost in the halo of the sun.

### Data generation model

We assume that

$$Y = \theta^\top X + \varepsilon,$$

where

1. $\mathcal{Y} = \mathbb{R}$ and $\mathcal{X} = \{1\} \times \mathbb{R}^n$,
2. $Y$ is a linear function of $X$ with coefficients $\theta \in \mathbb{R}^{n+1}$, and
3. the random noise term $\epsilon$ is normally distributed with mean $0$ and variance $\sigma^2$, i.e., $\varepsilon \sim \mathcal{N}(0, \sigma^2)$. Note that here the noise is not allowed to vary with the covariates $X$ and also not with different instantiations of the random experiment. This assumption is also known as *homoskedasticity*.

That is, we assume that $Y$ is a linear function of $X$ with added stochastic Gaussian noise. This implies that also $Y|X = x$ is a random variable with distribution

$$Y|X = x \sim \mathcal{N}(\theta^\top x, \sigma^2),$$

which means

$$p_x(y;\theta) \;=\; p(y|X=x,\theta) \;=\; \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(y-\theta^\top x)^2}{2\sigma^2}\right).$$

### Maximum likelihood estimate

The whole data generation model, i.e., the family of densities $p_x$, is completely specified by $\sigma > 0$ and $\theta \in \mathbb{R}^{n+1}$. Furthermore, the predictor derived from the model (that is only indirectly accessible for us)

$$h : \mathcal{X} \to \mathcal{Y},\, x \mapsto \mathsf{argmax}_{y\in\mathcal{Y}}\; p_x(y) \;=\; \theta^\top x$$

does not depend on $\sigma$. Hence, the problem of learning this predictor reduces to estimating $\theta$ from the i.i.d. data

$$(x^{(1)}, y^{(1)}), \ldots, (x^{(m)}, y^{(m)}).$$

A natural estimator for $\theta$ is the parameter vector $\hat{\theta}$ that maximizes the probability of the data. This estimator is called the *maximum likelihood estimator* and can be derived as a solution to the following optimization problem

$$\hat{\theta} \;=\; \mathsf{argmax}_{\theta\in\mathbb{R}^{n+1}}\; L(\theta),$$

where

$$L(\theta) \;=\; \prod_{i=1}^{m} p(y^{(i)}|X=x^{(i)},\theta) \;=\; \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(y^{(i)}-\theta^\top x^{(i)})^2}{2\sigma^2}\right)$$

is the likelihood function for the parameter vector $\theta$. The product form of the likelihood function is due to the i.i.d. assumption for the data. An optimum of the likelihood function remains optimal if we apply a monotonically increasing transformation to the likelihood function. Since the likelihood function is the product of exponentials a natural choice for such a transformation is the logarithm. Applying the logarithm to the likelihood function $L(\theta)$ gives us the log-likelihood function $\ell(\theta)$ that reads as

$$\begin{aligned}\ell(\theta) \;=\; \log L(\theta) \;&=\; -\sum_{i=1}^{m}\left(\log(\sqrt{2\pi}\sigma) + \frac{(y^{(i)}-\theta^\top x^{(i)})^2}{2\sigma^2}\right) \\ &=\; -m\log(\sqrt{2\pi}\sigma) - \frac{1}{2\sigma^2}\sum_{i=1}^{m}\left(y^{(i)}-\theta^\top x^{(i)}\right)^2.\end{aligned}$$

Using

$$y = (y^{(1)}, \dots, y^{(m)})^\top \in \mathbb{R}^m, \quad \theta = (\theta_0, \dots, \theta_n)^\top \in \mathbb{R}^{n+1}$$

and

$$X = (x^{(1)}, \dots, x^{(m)}) = \begin{pmatrix} 1 & \dots & 1 \\ x_1^{(1)} & \dots & x_1^{(m)} \\ \vdots & \ddots & \vdots \\ x_n^{(1)} & \dots & x_n^{(m)} \end{pmatrix} \in \mathbb{R}^{(n+1)\times m}$$

we can write $\ell(\theta)$ more compactly as

$$\ell(\theta) = -m \log(\sqrt{2\pi}\sigma) - \frac{1}{2\sigma^2}\|y - X^\top\theta\|^2.$$

Hence, we have

$$\begin{aligned} \hat{\theta} &= \mathsf{argmax}_{\theta\in\mathbb{R}^{n+1}} \; L(\theta) \\ &= \mathsf{argmax}_{\theta\in\mathbb{R}^{n+1}} \; \ell(\theta) \\ &= \mathsf{argmax}_{\theta\in\mathbb{R}^{n+1}} \; -m\log(\sqrt{2\pi}\sigma) - \frac{1}{2\sigma^2}\|y - X^\top\theta\|^2 \\ &= \mathsf{argmax}_{\theta\in\mathbb{R}^{n+1}} \; -\frac{1}{2}\|y - X^\top\theta\|^2 \\ &= \mathsf{argmin}_{\theta\in\mathbb{R}^{n+1}} \; \frac{1}{2}\|y - X^\top\theta\|^2. \end{aligned}$$

A necessary condition for a minimum is the vanishing gradient. We compute

$$\begin{aligned} \nabla_\theta \frac{1}{2}\|y - X^\top\theta\|^2 &= \nabla_\theta \frac{1}{2}\Big(y - X^\top\theta\Big)^\top \Big(y - X^\top\theta\Big) \\ &= \nabla_\theta \frac{1}{2}\Big(y^\top y - y^\top X^\top\theta - \theta^\top X y + \theta^\top X X^\top \theta\Big) \\ &= \nabla_\theta \frac{1}{2}\Big(y^\top y - 2\theta^\top X y + \theta^\top X X^\top\theta\Big) \\ &= -Xy + XX^\top\theta, \end{aligned}$$

and get from the vanishing gradient condition that

$$XX^\top\hat{\theta} = Xy.$$

That is, the maximum likelihood estimate of $\theta$ is the solution of a linear system, where $XX^\top$ is the *second moment matrix*. If the system is underdetermined, i.e., if the rank of the second moment matrix is less than $n+1$, then its solution space is an affine subspace of $\mathbb{R}^{n+1}$, but if the second moment matrix is invertible, i.e., if it has full rank, then the unique solution to the maximum likelihood problem is given as

$$\hat{\theta} = (XX^\top)^{-1}Xy.$$

*Remark:* argmax ... and argmin ... do not need to be unique in general. When we write something like $\hat{\theta} = \text{argmax} \ldots$, then we assign an arbitrary value in argmax ... to $\hat{\theta}$ unless stated otherwise.

## Data preparation

For allowing linear function that do not need to pass through the origin, we have artificially padded the data points $x^{(i)}, \in [m]$ by an additional component with value $1$. Instead we can also introduce the offset term $\theta_0$ explicitly. Let $X \in \mathbb{R}^{n\times m}$ be the data matrix without padding. Then we can write the OLS problem as

$$\hat{\theta} = \text{argmin}_{\theta\in\mathbb{R}^{n+1}} \frac{1}{2}\|y - \theta_0\mathbf{1}_m - X^\top\theta\|^2,$$

where $\mathbf{1}_m \in \mathbb{R}^m$ is the vector whose entries are all $1$. The necessary condition for the optimal value of $\theta_0$ is the vanishing derivative (of a convex, quadratic function)

$$\begin{aligned}\frac{d}{d\theta_0}\frac{1}{2}\|y - \theta_0\mathbf{1}_m - X^\top\theta\|^2 &= \frac{d}{d\theta_0}\frac{1}{2}\left(\theta_0^2\mathbf{1}_m^\top\mathbf{1}_m - 2\theta_0\mathbf{1}_m^\top y + 2\theta_0\mathbf{1}_m^\top X^\top\theta\right)\\ &= m\theta_0 - \mathbf{1}_m^\top y + \mathbf{1}_m^\top X^\top\theta,\end{aligned}$$

which gives

$$\theta_0 = \frac{1}{m}\left(\mathbf{1}_m^\top y - \mathbf{1}_m^\top X^\top\theta\right).$$

The entries of the vector $\mathbf{1}_m^\top X^\top \in \mathbb{R}^m$ are given as

$$\mathbf{1}_m^\top X^\top = \left(\sum_{i=1}^m x_1^{(i)}, \ldots, \sum_{i=1}^m x_n^{(i)}\right).$$

This vector becomes the zero vector, if we center the data points, i.e., if we replace $x^{(i)}$ by

$$x^{(i)} - \frac{1}{m}\sum_{j=1}^m x^{(j)}, \quad i \in [m].$$

Hence, after centering, we get

$$\theta_0 \,=\, \frac{\mathbf{1}_m^\top y}{m} \,=\, \frac{1}{m}\sum_{i=1}^m y_i.$$

Now, if we also center the label vector $y$, then we have $\theta_0 = 0$ and we do no longer need the offset. In the following we assume that the data and label vectors are centered.
It is important to note though that for predictions at $x \in \mathbb{R}^n$ one has to subtract the sample mean also from $x$, and later one has to add the offset, i.e., the average of the observed label vector, back to the prediction.

An additional data transformation, that becomes important in the next paragraph, is making the scales on which the explanatory variables are measured comparable. This is typically done by standardizing, i.e., scaling the features such that all entries on the diagonal of the second moment matrix become 1, that is, the centered data points $x^{(i)}$ are replaced by

$$\left(\frac{x_1^{(i)}}{\sqrt{\sum_{j=1}^m {x_1^{(j)}}^2}}, \,\ldots,\, \frac{x_n^{(i)}}{\sqrt{\sum_{j=1}^m {x_n^{(j)}}^2}}\right)^\top.$$

Alternatively, one can just scale the features simply by replacing $x_j^{(i)}$, $i \in [m], j \in [n]$ with

$$\frac{x_j^{(i)}}{\max_{k=1,\ldots,m}\left\{x_j^{(k)}\right\} - \min_{k=1,\ldots,m}\left\{x_j^{(k)}\right\}}.$$

When computing a prediction at $x \in \mathbb{R}^n$, then $x$ should be transformed in the same way, for instance, when using the first transformation, the sample mean is subtracted from $x$ before the components are each scaled by the inverse of the square root of the corresponding sample variance.

## Ridge regression

The second moment matrix $XX^\top$ can be written as

$$XX^\top \,=\, \sum_{i=1}^m x^{(i)}{x^{(i)}}^\top,$$

where $x^{(i)}{x^{(i)}}^\top$ is a projection matrix that projects any point $x \in \mathbb{R}^n$ onto the one-dimensional subspace spanned by $x^{(i)} \in \mathbb{R}^n$, i.e.,

$$\mathbb{R}^n \to \mathbb{R}^n,\, x \mapsto \big(x^{(i)}{x^{(i)}}^\top\big)x \,=\, \big({x^{(i)}}^\top x\big)x^{(i)}.$$

Thus, the matrix $x^{(i)}{x^{(i)}}^\top$ has rank one, and $XX^\top$ has rank at most $m$. That is, if $m$ is small compared to $n$, then the second moment matrix does not have full rank and is thus not invertible.

By construction, the second moment matrix is symmetric and positive semi-definite, i.e., it holds for all $x \in \mathbb{R}^n$ that $\left(XX^\top\right)^\top = XX^\top$ and $x^\top XX^\top x \geq 0$. Note that the second moment matrix is the Hessian, i.e., the matrix of second order derivatives of the function $\frac{1}{2}\|y - X^\top\theta\|^2$. Hence, the Hessian of this function is positive semi-definite, which means that the function itself is convex and thus, $\hat{\theta} = (XX^\top)^{-1}Xy$ is indeed a minimum.

Another consequence of positive semi-definiteness is that for any $c > 0$ the matrix $XX^\top + c\mathbb{1}_n$ is positive definite and thus invertible. By replacing the second moment matrix $XX^\top$ by $XX^\top + c\mathbb{1}_n$, we get the following estimate for the parameter vector $\theta$,

$$\hat{\theta} = (XX^\top + c\mathbb{1}_n)^{-1}Xy.$$

This estimate is the solution of the regularized maximum likelihood problem

$$\text{min}_{\theta\in\mathbb{R}^n} \frac{1}{2}\|y - X^\top\theta\|^2 + \frac{c}{2}\|\theta\|^2$$

that is known as *ridge regression*. The ridge regression problem is a strictly convex optimization problem that has a unique solution even if the second moment matrix does not have full rank.

*Remarks:* The regularization term $\frac{c}{2}\|\theta\|^2$ treats all the explanatory variables the same. Thus it is important that these variables are measured on comparable scales. Here, we took care of that in the data preparation phase where we rescaled the explanatory variables by their sample variances.

Turning to the regularized problem makes sense even when the matrix $XX^\top$ is invertible. The *condition number* of the matrix $XX^\top$, defined as the quotient of the largest eigenvalue $\lambda_{\max}$ divided by the smallest eigenvalue $\lambda_{\min}$ of the matrix, is a measure of the range of scales that is covered by the data. Since floating point number approximations can adapt well only to small ranges of scales, we can run into numerical problems when attempting to compute $\left(XX^\top\right)^{-1}$, if the range of scales is large. After regularization the condition number of the matrix changes to

$$\frac{\lambda_{\max} + c}{\lambda_{\min} + c}$$

which converges to $1$ for large values of $c$. Of course, for large values of $c$ the ridge regression optimization problem hardly depends on the data anymore since all the weight is on the data independent regularization term $\frac{c}{2}\|\theta\|^2$. Hence, the impact of regularization on the learning problem is making it less dependent on the data and thus also less dependent on small changes in the data. In other words regularization makes the learning problem more *robust/stable*, and thus regularization makes sense also from the learning point of view. We discuss this in more depth in the next chapter.

The practically challenging problem is to figure out the right amount of regularization, i.e., a good value for the regularization parameter $c$. The idea here is to split the data set into two parts, a training set and a so called validation set. Then a solution $\hat{\theta}$ of the ridge regression problem is computed for several values of $c \geq 0$. The performance of the resulting predictors is then compared on the validation set and the predictor with the best performance is chosen.

Choosing a good value for $c$ is an instance of the more general hyper-parameter selection problem. Models for different values of the hyper-parameters are computed on the training data. Afterwards, the validation data are used to chose the best performing among these models. For validating the chosen model one should actually split the data set into three parts *training*, *validation* and *test*. The model that is chosen on the validation set is then validated on the test set. The reason for using a different part of the data for validation is that selecting good hyper-parameter values is also part of learning the model. Also for this part we run the risk of overfitting which can reflect itself in a much better performance on the validation data than on the test data. We come back to this issue with more explanations in later chapters.

# Chapter 2

# Bias-variance trade-off

Any predictor that is computed from probabilistic data will make prediction errors, also called *generalization errors*. There are two basic sources of errors. First, our data generation model can be not adequate. For instance, in a linear regression problem our assumption that the label is a linear function of the features might not be true. The second type of error results from observed data points that are not representative for the data generation model. Since we assume the data generation model to be random, there is some probability that a finite sample is not representative.
The first type of error is often called *bias* or *inductive bias* since it is induced by our assumptions on the data generating mechanism. The second type of error is called *variance* or *estimation* error. The generalization error is composed of the bias and the variance. Interestingly, it is possible to trade-off one type of error against the other. This trade-off is at the heart of machine learning.

In this chapter we analyze the nature of these two types of errors for a slightly more general data generation model than in the OLS case. We assume that

$$Y \,=\, f(X) + \varepsilon,$$

where $\mathcal{Y} = \mathbb{R}$, $f : \mathcal{X} \to \mathbb{R}$, and $\epsilon$ is a random noise term with expectation $\mathsf{E}[\varepsilon] = 0$ and variance $\mathsf{Var}[\varepsilon] = \sigma^2$. Hence, $Y|X = x$ is a random variable with expectation $\mathsf{E}[Y] = f(X)$.
Let $\hat{f}$ be an estimate of $f$ that has been computed from i.i.d. data

$$(x^{(1)}, y^{(1)}), \ldots, (x^{(m)}, y^{(m)}).$$

Since the data are random, also $\hat{f}$ is a random variable. In the spirit of the OLS regression problem, we want to compute the expected quadratic error of

$\hat{f}$ with respect to the assumed model at some fixed value $x \in \mathcal{X}$, where the expectation is over $Y := (Y|X = x) = f(x) + \varepsilon$ and the data that have been used for computing $\hat{f}$. We get

$$\begin{aligned}
\mathsf{E}\big[(Y - \hat{f}(x))^2\big] &= \mathsf{E}\big[(Y - \mathsf{E}[\hat{f}(x)] + \mathsf{E}[\hat{f}(x)] - \hat{f}(x))^2\big] \\
&= \mathsf{E}\big[\big(Y - \mathsf{E}[\hat{f}(x)]\big)^2\big] + \mathsf{E}\big[\big(\mathsf{E}[\hat{f}(x)] - \hat{f}(x)\big)^2\big] \\
&\qquad + 2 \cdot \mathsf{E}\big[\big(Y - \mathsf{E}[\hat{f}(x)]\big)\big(\mathsf{E}[\hat{f}(x)] - \hat{f}(x)\big)\big] \\
&= \mathsf{E}\big[\big(Y - \mathsf{E}[\hat{f}(x)]\big)^2\big] + \mathsf{E}\big[\big(\mathsf{E}[\hat{f}(x)] - \hat{f}(x)\big)^2\big] \\
&= \mathsf{E}\big[\big(Y - f(x) + f(x) - \mathsf{E}[\hat{f}(x)]\big)^2\big] + \mathsf{E}\big[\big(\mathsf{E}[\hat{f}(x)] - \hat{f}(x)\big)^2\big] \\
&= \mathsf{E}\big[\big(Y - f(x)\big)^2\big] + 2 \cdot \mathsf{E}\big[\big(Y - f(x)\big)\big(f(x) - \mathsf{E}[\hat{f}(x)]\big)\big] \\
&\qquad + \mathsf{E}\big[\big(f(x) - \mathsf{E}[\hat{f}(x)]\big)^2\big] + \mathsf{E}\big[\big(\mathsf{E}[\hat{f}(x)] - \hat{f}(x)\big)^2\big] \\
&= \sigma^2 + \big(f(x) - \mathsf{E}[\hat{f}(x)]\big)^2 + \mathsf{E}\big[\big(\mathsf{E}[\hat{f}(x)] - \hat{f}(x)\big)^2\big],
\end{aligned}$$

where we have used the independence of $Y$ and $\hat{f}(x)$ that implies

$$\begin{aligned}
\mathsf{E}\big[\big(Y - \mathsf{E}[\hat{f}(x)]\big)\big(\mathsf{E}[\hat{f}(x)] - \hat{f}(x)\big)\big] &= \mathsf{E}\big[Y - \mathsf{E}[\hat{f}(x)]\big] \cdot \mathsf{E}\big[\mathsf{E}[\hat{f}(x)] - \hat{f}(x)\big] \\
&= \mathsf{E}\big[Y - \mathsf{E}[\hat{f}(x)]\big] \cdot 0 = 0
\end{aligned}$$

for the third equality, and

$$\begin{aligned}
\mathsf{E}\big[\big(Y - f(x)\big)\big(f(x) - \mathsf{E}[\hat{f}(x)]\big)\big] &= \mathsf{E}\big[Y - f(x)\big] \cdot \big(f(x) - \mathsf{E}[\hat{f}(x)]\big) \\
&= 0 \cdot \big(f(x) - \mathsf{E}[\hat{f}(x)]\big) = 0
\end{aligned}$$

and

$$\mathsf{E}\big[\big(Y - f(x)\big)^2\big] = \mathsf{E}\big[\varepsilon^2\big] = \mathsf{Var}[\varepsilon] = \sigma^2$$

for the last equality.

Let us have closer look at the three remaining terms for the expected quadratic error. The term $\sigma^2$ is just the variance of the random variable $Y = f(x) + \varepsilon$ and does not depend of the estimator. The term

$$\big(f(x) - \mathsf{E}[\hat{f}(x)]\big)^2$$

is the *bias* of the estimator, and the term

$$\mathsf{E}\big[\big(\mathsf{E}[\hat{f}(x)] - \hat{f}(x)\big)^2\big]$$

is the *variance* of the estimator. The bias measures the modeling error of the procedure for estimating $f$ from the data. An unbiased estimator $\hat{f}$ should

coincide in expectation with the unknown function $f$. The variance term measures, as its name suggests, the variance of the estimator which is random variable since the data are obtained from random trials.
Hence, up to the variance $\sigma^2$ that is induced by the noise term $\varepsilon$, the quadratic error features two terms that both depend on our procedure for estimating $f$. The procedure can be biased and it exhibits variance since it depends on random data.

## Bias and variance for OLS regression

Let us compute the expected quadratic error of the OLS regressor given by the maximum likelihood estimate $\hat{\theta} \in \mathbb{R}^n$ at some point $x \in \mathbb{R}^n$ assuming the data generation model including data preparation from Chapter 1. We get from the abstract result above that

$$\mathsf{E}\big[(Y - \hat{\theta}^\top x)^2\big] \,=\, \sigma^2 \,+\, \big(\theta^\top x - \mathsf{E}[\hat{\theta}^\top x]\big)^2 \,+\, \mathsf{E}\big[\big(\mathsf{E}[\hat{\theta}^\top x] - \hat{\theta}^\top x\big)^2\big].$$

A closer look at $\mathsf{E}[\hat{\theta}^\top x]$ shows that

$$\begin{aligned}\mathsf{E}[\hat{\theta}^\top x] \,&=\, \mathsf{E}\big[y^\top X^\top (XX^\top)^{-1} x\big] \,=\, \mathsf{E}\big[y^\top\big] X^\top (XX^\top)^{-1} x \\ &=\, \theta^\top X X^\top (XX^\top)^{-1} x \,=\, \theta^\top x.\end{aligned}$$

Since this equality must hold independent of $x$, we even have $\mathsf{E}[\hat{\theta}] = \theta$. It follows that the OLS estimate is unbiased, because the bias term

$$\big(\theta^\top x - \mathsf{E}[\hat{\theta}^\top x]\big)^2 \,=\, \big(\theta^\top x - \theta^\top x\big)^2 \,=\, 0$$

vanishes. Thus the only estimator dependent term in the expected quadratic error is the variance term

$$\begin{aligned}\mathsf{E}\big[\big(\mathsf{E}[\hat{\theta}^\top x] - \hat{\theta}^\top x\big)^2\big] \,&=\, \mathsf{E}\big[\big(\theta^\top x - \hat{\theta}^\top x\big)^2\big] \,=\, \mathsf{E}\big[(\hat{\theta}^\top x)^2\big] \,-\, \big(\theta^\top x\big)^2 \\ &=\, x^\top (XX^\top)^{-1} X \mathsf{E}[yy^\top] X^\top (XX^\top)^{-1} x \,-\, \big(\theta^\top x\big)^2 \\ &=\, x^\top (XX^\top)^{-1} X (X^\top \theta\theta^\top X + \sigma^2 \mathbb{1}_n) X^\top (XX^\top)^{-1} x \,-\, \big(\theta^\top x\big)^2 \\ &=\, x^\top \theta\theta^\top x \,+\, \sigma^2 x^\top (XX^\top)^{-1} x \,-\, \big(\theta^\top x\big)^2 \,=\, \sigma^2 x^\top (XX^\top)^{-1} x.\end{aligned}$$

The Gauß-Markov Theorem shows that we cannot reduce the variance by moving to some other unbiased estimator of the form $My$, where $y \in \mathbb{R}^m$ is the observed label vector and $M \in \mathbb{R}^{n\times m}$ is some matrix that may depend on the data. We assume the following data generation model: $Y = \theta^\top X + \varepsilon$ with $\mathsf{E}[\varepsilon] = 0$ and $\mathsf{Var}[\varepsilon] = \sigma^2$. As before, let $y \in \mathbb{R}^m$ be the observed label vector and $X \in \mathbb{R}^{n\times m}$ be the data matrix whose columns are $x^{(i)} \in \mathbb{R}^n$, $i \in [n]$.

**Theorem 1. [Gauß-Markov]** *Let $\bar{\theta} = M(X)y$ be some linear, unbiased estimator of the true parameter vector $\theta$, where $M(X) \in \mathbb{R}^{n\times m}$ is computed from the data matrix $X$. Then the variance term in the expected quadratic error of this estimator is at least as large as the corresponding term for the maximum likelihood estimator.*

*Proof.* We can always write $M(X) = (XX^\top)^{-1}X + C$ for yet another matrix $C := C(X) \in \mathbb{R}^{n\times m}$. Since we assume that $\bar{\theta}$ is an unbiased estimator, we have

$$\theta \;=\; \mathsf{E}[\bar{\theta}] \;=\; \mathsf{E}[\hat{\theta} + Cy] \;=\; \mathsf{E}[\hat{\theta}] \;+\; C\mathsf{E}[y] \;=\; \theta \;+\; C\mathsf{E}[y] \;=\; \theta \;+\; CX^\top\theta$$

and thus $CX^\top\theta = 0$. The latter equality must hold true independent of $\theta$, because the estimator should be unbiased for any choice of $\theta$ (note that $X$ is not random). This is only possible, if $CX^\top = 0$.
The variance term in the expected quadratic error is given as

$$\begin{aligned}\mathsf{E}\big[(\bar{\theta}^\top x)^2\big] \;-\; (\theta^\top x)^2 \;&=\; \mathsf{E}\big[((\hat{\theta}^\top + y^\top C^\top)x)^2\big] \;-\; (\theta^\top x)^2\\ &=\; \mathsf{E}\big[(\hat{\theta}^\top x)^2\big] \;-\; (\theta^\top x)^2 \;+\; 2\mathsf{E}\big[(\hat{\theta}^\top x)(y^\top C^\top x)\big] \;+\; \mathsf{E}\big[(y^\top C^\top x)^2\big],\end{aligned}$$

where the first two terms together are the variance for the OLS estimator. Since $\mathsf{E}\big[(y^\top C^\top x)^2\big]$ is non-negative, the claim of the theorem would follow, if we can show that

$$\mathsf{E}\big[(\hat{\theta}^\top x)(y^\top C^\top x)\big] \;\geq\; 0.$$

Using $XC^\top = \big(CX^\top\big)^\top = 0^\top = 0$, we compute

$$\begin{aligned}\mathsf{E}\big[(\hat{\theta}^\top x)(y^\top C^\top x)\big] \;&=\; \mathsf{E}\big[x^\top\hat{\theta}y^\top C^\top x\big] \;=\; x^\top\mathsf{E}\big[\hat{\theta}y^\top C^\top\big]x\\ &=\; x^\top\mathsf{E}\big[(XX^\top)^{-1}Xyy^\top C^\top\big]x\\ &=\; x^\top(XX^\top)^{-1}X\mathsf{E}[yy^\top]C^\top x\\ &=\; x^\top(XX^\top)^{-1}X\big(X^\top\theta\theta^\top X + \sigma^2\mathbb{1}_n\big)C^\top x\\ &=\; x^\top\big(\theta\theta^\top + \sigma^2(XX^\top)^{-1}\big)XC^\top x \;=\; 0.\end{aligned}$$

□

Although the maximum likelihood OLS estimate looks optimal since it is unbiased and has minimal variance among the linear, unbiased estimators, we can still improve on it by trading variance for bias. In the next section we show that expected quadratic error of the ridge regression estimate can be lower than the error of the maximum likelihood OLS estimate.

## Ridge regression revisited

Of course, we can also compute the bias and variance terms in the expected quadratic error of the logistic regression predictor. Let

$$\hat{\theta} \,=\, (XX^\top + c\mathbb{1}_n)^{-1}Xy \,=:\, M_c y$$

be the ridge regression estimator of the parameter vector $\theta$, then we get again from the abstract result about the bias-variance decomposition of the expected quadratic error that

$$\mathsf{E}\big[(Y - \hat{\theta}^\top x)^2\big] \,=\, \sigma^2 \,+\, \big(\theta^\top x - \mathsf{E}[\hat{\theta}^\top x]\big)^2 \,+\, \mathsf{E}\big[\big(\mathsf{E}[\hat{\theta}^\top x] - \hat{\theta}^\top x\big)^2\big].$$

Expanding the bias term gives

$$\begin{aligned}\big(\theta^\top x - \mathsf{E}[\hat{\theta}^\top x]\big)^2 &= \big(\theta^\top x - \mathsf{E}[y^\top M_c^\top x]\big)^2 \,=\, \big(\theta^\top x - \mathsf{E}[y^\top]M_c^\top x\big)^2 \\ &= \big(\theta^\top x - \theta^\top X M_c^\top x\big)^2 \,=\, \big(\theta^\top(\mathbb{1}_n - XM_c^\top)x\big)^2 \\ &= x^\top\big(\mathbb{1}_n - XM_c^\top\big)^\top\theta\theta^\top\big(\mathbb{1}_n - XM_c^\top\big)x\end{aligned}$$

and expanding the variance term gives

$$\begin{aligned}&\mathsf{E}\big[\big(\mathsf{E}[\hat{\theta}^\top x] - \hat{\theta}^\top x\big)^2\big] \\ &\quad= \mathsf{E}\big[\big(\mathsf{E}[y^\top]M_c^\top x - y^\top M_c^\top x\big)^2\big] \,=\, \mathsf{E}\big[\big(\theta^\top X M_c^\top x - y^\top M_c^\top x\big)^2\big] \\ &\quad= \mathsf{E}\big[\big((\theta^\top X - y^\top)M_c^\top x\big)^2\big] \,=\, \mathsf{E}\big[x^\top M_c(X^\top\theta - y)(X^\top\theta - y)^\top M_c^\top x\big] \\ &\quad= x^\top M_c \mathsf{E}\big[(X^\top\theta - y)(X^\top\theta - y)^\top\big]M_c^\top x \\ &\quad= x^\top M_c \mathsf{E}\big[(X^\top\theta\theta^\top X - y\theta^\top X - X^\top\theta y^\top + yy^\top\big]M_c^\top x \\ &\quad= x^\top M_c\big(X^\top\theta\theta^\top X - X^\top\theta\theta X - X^\top\theta\theta^\top X + \mathsf{E}[yy^\top]\big)M_c^\top x \\ &\quad= x^\top M_c\big(-X^\top\theta\theta^\top X + X^\top\theta\theta^\top X + \sigma^2\mathbb{1}_n\big)M_c^\top x \\ &\quad= \sigma^2 x^\top M_c M_c^\top x.\end{aligned}$$

Since $\lim_{c\to 0} M_c = (XX^\top)^{-1}X$, we have, as expected, that the bias term vanishes in this limit and also the variance becomes the variance of the maximum likelihood OLS estimate. On the other hand, since $\lim_{c\to\infty} M_c = 0$, the bias term becomes $x^\top\theta\theta^\top x$ which is independent of the data. This limit also implies that the variance of the ridge regression estimator goes to zero as $c$ goes to infinity. Hence, there might be values for $c$ where the total expected quadratic error of the ridge regression estimator is smaller than the corresponding error of the maximum likelihood OLS estimate.

Remember that the maximum likelihood OLS estimate needs a regular second moment matrix $XX^\top$. Hence, we can compare the OLS estimate with the ridge regression estimate only in the case that the second moment matrix is regular, although our first motivation for ridge regression was the potential non-regularity of $XX^\top$. Here, we show that even when $XX^\top$ is regular, ridge regression is preferable over maximum likelihood OLS, because it can reduce the expected quadratic error by trading variance for bias. We check this by subtracting the expected quadratic error of the ridge regression estimate from the OLS estimate. Let $L_c = (XX^\top + c\mathbb{1})^{-1}$, then $M_c X^\top = L_c X X^\top$ and the expected error of the ridge regression estimate is

$$\begin{aligned} & x^\top \big( (\mathbb{1}_n - X M_c^\top)^\top \theta\theta^\top (\mathbb{1}_n - X M_c^\top) + \sigma^2 M_c M_c^\top \big) x \\ & \qquad = x^\top L_c \big( c^2 \theta\theta^\top + \sigma^2 X X^\top \big) L_c^\top x . \end{aligned}$$

The expected error of the maximum likelihood OLS estimate is

$$\begin{aligned} x^\top \big( \sigma^2 (XX^\top)^{-1} \big) x &= x^\top L_c \big( \sigma^2 (XX^\top + c\mathbb{1}_n)(XX^\top)^{-1}(XX^\top + c\mathbb{1}_n)^\top \big) L_c^\top x \\ &= x^\top L_c \big( \sigma^2 XX^\top + 2c\sigma^2 \mathbb{1}_n + c^2\sigma^2 (XX^\top)^{-1} \big) L_c^\top x . \end{aligned}$$

Hence, subtracting the first error from the second one gives

$$c \cdot x^\top L_c \big( 2\sigma^2 \mathbb{1}_n + c\sigma^2 (XX^\top)^{-1} - c\theta\theta^\top \big) L_c^\top x .$$

This expression is non-negative, if the matrix

$$2\sigma^2 \mathbb{1}_n \, + \, c\sigma^2 (XX^\top)^{-1} \, - \, c\theta\theta^\top$$

is positive semi-definite. This is the case if $2\sigma^2 \mathbb{1}_n - c\theta\theta^\top$ is positive semi-definite, because $c\sigma^2 (XX^\top)^{-1}$ is positive definite. The matrix $2\sigma^2 \mathbb{1}_n - c\theta\theta^\top$ is positive semi-definite, if $2\sigma^2 - c\|\theta\|^2 \geq 0$, which is equivalent to

$$c \leq \frac{2\sigma^2}{\|\theta\|^2} .$$

Hence, the expected quadratic error of the ridge regression estimate is smaller than the expected quadratic error of the maximum likelihood OLS estimate, if $c \leq 2\sigma^2 / \|\theta\|^2$.

# Chapter 3

# Logistic regression

Despite its name logistic regression is a classification technique and not a regression technique. More precisely it is a technique for binary classification, where the explained variable can take only two values. These values are often chosen as $0$ and $1$, or as $-1$ and $1$, but could be anything like `true` and `false`, or *black* and *white*. Practically, it is convenient to work with numerical values. The mapping to numerical values can always be achieved by using characteristic functions

$$\mathbf{1}[\text{condition}] = \begin{cases} 1 & : \text{ condition is satisfied} \\ 0 & : \text{ condition is not satisfied} \end{cases}$$

For example *black* and *white* can be mapped to $0$ and $1$ by $\mathbf{1}[y = white]$, or equivalently by $\mathbf{1}[\neg(y = black)]$. Here, we want to use the numerical representation by $-1$ and $1$ for the two options, which can always be achieved by the mapping $2 \cdot \mathbf{1}[\text{condition}] - 1$.

### Data generation model

We assume that $\mathcal{Y} = \{\pm 1\}$, $\mathcal{X} = \{1\} \times \mathbb{R}^n$, and

$$p_x(y) = p(y|X = x, \theta) = \frac{1}{1 + \exp(-y\, x^\top \theta)}.$$

*Remark:* Calling this a regression model can be justified as follows: We get for the odds-ratio

$$\frac{p_x(Y = 1)}{p_x(Y = -1)} = \frac{1 - p_x(Y = -1)}{p_x(Y = -1)} = \frac{1}{p_x(Y = -1)} - 1 = \exp(x^\top \theta).$$

Hence, the log-odds-ratio is just $x^\top\theta$. If, instead of $y \in \{\pm 1\}$, we measure at $x$ the logarithm of the ratio of the frequencies of observing $Y = 1$ and $Y = -1$, respectively, which is a number in $\mathbb{R} \cup \{-\infty, \infty\}$, then we actually have a regression problem. Note though, if we randomly sample the features then it is likely that for any given $x \in \mathcal{X}$ we have at most one observation, that is, the observed log-odds-ratios are either $-\infty$ or $\infty$. But since we a free to choose the features at which we sample the labels we can sample the label at any given $x$ several times.

## Maximum likelihood estimate

The whole data generation model, i.e., the family of densities $p_x$, is completely specified by $\theta \in \mathbb{R}^{n+1}$. Hence, the learning problem reduces, as in the ordinary least squares regression case, to estimating $\theta$ from the i.i.d. data

$$(x^{(1)}, y^{(1)}), \dots, (x^{(m)}, y^{(m)}).$$

The likelihood function for the parameter vector $\theta$ is given as

$$L(\theta) \;=\; \prod_{i=1}^m p(y^{(i)} | X = x^{(i)}, \theta) \;=\; \prod_{i=1}^m \left( \frac{1}{1 + \exp\big(-y^{(i)}\, {x^{(i)}}^\top \theta\big)} \right)$$

and the maximum likelihood estimate of $\theta$ is given as

$$\hat\theta \;=\; \mathrm{argmax}_{\theta \in \mathbb{R}^{n+1}}\; L(\theta).$$

Again, it is easier to work with the log-likelihood function

$$\ell(\theta) \;=\; \log L(\theta) \;=\; -\sum_{i=1}^m \log \Big( 1 + \exp\big(-y^{(i)}\, {x^{(i)}}^\top \theta\big) \Big)$$

instead of the likelihood function.
The vanishing gradient condition for a maximum of the log-likelihood function gives

$$\nabla_\theta \ell(\theta) \;=\; \sum_{i=1}^m \frac{y^{(i)} \exp\big(-y^{(i)}\, {x^{(i)}}^\top \theta\big)}{1 + \exp\big(-y^{(i)}\, {x^{(i)}}^\top \theta\big)}\, x^{(i)} \;=\; \sum_{i=1}^m \frac{y^{(i)} \cdot x^{(i)}}{1 + \exp\big(y^{(i)}\, {x^{(i)}}^\top \theta\big)},$$

which is a non-linear system of $n+1$ equations for the $n+1$ unknowns $\theta_0, \dots, \theta_n$. Although, it is difficult to solve the system of equations directly, we can use the gradient in a gradient ascent scheme to solve the original optimization problem

$$\hat\theta \;=\; \mathrm{argmax}_{\theta \in \mathbb{R}^{n+1}}\; \ell(\theta).$$

By computing the second order derivatives of the log-likelihood function, i.e., its Hessian, we see that the optimization problem is concave, because the Hessian

$$\begin{aligned} H\big(\ell(\theta)\big) \;&=\; -\sum_{i=1}^{m} \frac{{y^{(i)}}^2 \exp\big(y^{(i)}\, {x^{(i)}}^\top \theta\big)}{\big(1+\exp\big(y^{(i)}\, {x^{(i)}}^\top \theta\big)\big)^2}\, x^{(i)} {x^{(i)}}^\top \\ &=\; -\sum_{i=1}^{m} \frac{\exp\big(y^{(i)}\, {x^{(i)}}^\top \theta\big)}{\big(1+\exp\big(y^{(i)}\, {x^{(i)}}^\top \theta\big)\big)^2}\, x^{(i)} {x^{(i)}}^\top \end{aligned}$$

is negative semi-definite. Note that the projection matrices $x^{(i)}{x^{(i)}}^\top$ are positive semi-definite and any negative linear combination of convex functions is concave. The Hessian can be negative definite, and thus the log-likelihood function can be strictly concave, only if the number of data points is at least $n+1$.

*Remarks:* Again, we can enforce strict concavity in the maximum likelihood problem, i.e., a negative definite Hessian, by regularization, i.e., replacing the maximization problem for the log likelihood function by

$$\hat{\theta} \;=\; \mathsf{argmax}_{\theta\in\mathbb{R}^{n+1}}\;\ell(\theta) - \frac{c}{2}\|\theta\|^2$$

with regularization parameter $c > 0$. As in the ordinary least squares case, a solution to the regularized logistic regression problem, which is also known as a support vector machine with logistic loss function, is more robust against small changes in the data.

The Hessian can also be used in Newton's method, which is a second order method for solving the maximum likelihood problem. In contrast to gradient methods that only make use of first order derivatives Newton's method also uses the second order derivatives in the Hessian. Newton's method needs fewer iterations to converge, but each iteration is more costly to compute than the iterations in gradient methods. It turns out that Newton's method is faster for small problems while first order methods are faster for large problems.

## Prediction

Once we have estimated the model parameters $\theta$ by $\hat{\theta}$ we can use them for prediction using the predictor

$$h : \mathcal{X} \to \mathcal{Y} = \{\pm 1\},\, x \mapsto \left\{ \begin{array}{rcl} 1 & : & \exp\big(-x^\top\hat{\theta}\big) \;\le\; 1 \\ -1 & : & \exp\big(-x^\top\hat{\theta}\big) \;>\; 1 \end{array}\right.$$

The decision boundary of this classifier is the set

$$\big\{x \in \{1\}\times\mathbb{R}^n \,:\, \exp\big(-x^\top\hat{\theta}\big) = 1\big\},$$

because for points in this set we are indifferent between predicting $-1$ and $1$. In our classifier $h$ we have arbitrarily chosen to predict $1$ for points on the decision boundary, but note that the choice to predict $-1$ is as meaningful. Rewriting the condition for points on the decision boundary by taking the logarithm gives the equivalent characterization

$$\begin{aligned}\big\{x \in \{1\} \times \mathbb{R}^n \,:\, \exp\big(-x^\top \hat{\theta}\big) = 1\big\} &= \big\{x \in \{1\} \times \mathbb{R}^n \,:\, x^\top \hat{\theta} = 0\big\} \\ &= \big\{x \in \mathbb{R}^n \,:\, \hat{\theta}_n x_n + \ldots + \hat{\theta}_1 x_1 + \hat{\theta}_0 = 0\big\},\end{aligned}$$

which is a hyperplane in $\mathbb{R}^n$. That is, all points $x \in \mathbb{R}^n$ that are on the same side of this hyperplane get assigned the same label in $\mathcal{Y}$, and thus logistic regression defines a *linear classifier*.

*Remark:* In Chapter 1 we have discussed data preparation for ordinary least squares regression. We used the two operations, centering and scaling. Centering allowed us to skip the augmentation of the feature vectors by a leading $1$. Centering and scaling together made the scales of the different features comparable which was important for ridge regression, i.e., using Euclidean regularization. The same argument also applies for regularized logistic regression. Hence, centering and scaling should also applied here, in the regularized case. For the standard non-regularized case it is not important to center and scale. Also, centering does not free us from augmenting the feature vectors by a leading $1$, because even for centered feature vectors the optimal decision boundary (hyperplane) still depends on the labels and does not need to pass through the origin.

## Practical performance measures

Our current choice of the binary label space $\mathcal{Y} = \{\pm 1\}$ is only for convenience, because it simplifies some of the maximum likelihood calculations. When evaluating practical performance measure another choice for the label space, namely $\mathcal{Y} = \{0, 1\}$ is more convenient. Hence, we assume now that $\mathcal{Y} = \{0, 1\}$. For evaluating performance measures we need *test data*

$$(\bar{x}^{(1)}, \bar{y}^{(1)}), \ldots, (\bar{x}^{(k)}, \bar{y}^{(k)})$$

that are independent from the training data. Note that in practice we always just have data that we can split into training and test data. The training data are used to learn the model, i.e., estimate the model parameters, and the test data are used to validate the estimated model, i.e., to compute some performance measure for a predictor. Assume that we have the following predictions

$$(\bar{x}^{(1)}, \hat{y}^{(1)}), \ldots, (\bar{x}^{(k)}, \hat{y}^{(k)}),$$

i.e., at each point $\bar{x}^{(i)} \in \mathcal{X}$ we have an observed value $\bar{y}^{(i)}$ and a predicted value $\hat{y}^{(i)}$, for $i \in [k]$

The first performance measure that comes to mind is the frequency of the correctly predicted observations, also called *accuracy*, i.e.,

$$P = \frac{1}{k}\sum_{i=1}^{k}\left(\bar{y}^{(i)}\hat{y}^{(i)} + (1-\bar{y}^{(i)})(1-\hat{y}^{(i)})\right) \in [0,1].$$

Often it makes sense to measure the performance for every label individually. If we arbitrarily consider data points with label $1$ as *positive* and data points with label $0$ as *negative*, then

$$F_{+} = \frac{1}{k}\sum_{i=1}^{k}\hat{y}^{(i)}(1-\bar{y}^{(i)})$$

is the frequency of *false positives*, and

$$F_{-} = \frac{1}{k}\sum_{i=1}^{k}(1-\hat{y}^{(i)})\bar{y}^{(i)}$$

is the frequency of *false negatives*. The frequencies of the *true positives* $T_{+}$ and the *true negatives* $T_{-}$,

$$T_{+} = \frac{1}{k}\sum_{i=1}^{k}\hat{y}^{(i)}\bar{y}^{(i)} \quad \text{and} \quad T_{-} = \frac{1}{k}\sum_{i=1}^{k}(1-\hat{y}^{(i)})(1-\bar{y}^{(i)}),$$

are also called *sensitivity* and *specificity*, respectively.

Other popular, label dependent performance measures are *precision* and *recall*. Precision measures the frequency of the correctly predicted observations for a given class label in $\mathcal{Y}$ among all predictions of the label. Here, we only have the two labels $0$ and $1$, and thus only two precision measures

$$P_1 = \frac{\sum_{i=1}^{k}\bar{y}^{(i)}\hat{y}^{(i)}}{\sum_{i=1}^{k}\hat{y}^{(i)}} \quad \text{and} \quad P_0 = \frac{\sum_{i=1}^{k}(1-\bar{y}^{(i)})(1-\hat{y}^{(i)})}{\sum_{i=1}^{k}(1-\hat{y}^{(i)})}.$$

Recall measures the frequency of the correctly predicted observations for a given class label in $\mathcal{Y}$ among all observations of the label. That is,

$$R_1 = \frac{\sum_{i=1}^{k}\bar{y}^{(i)}\hat{y}^{(i)}}{\sum_{i=1}^{k}\bar{y}^{(i)}} \quad \text{and} \quad R_0 = \frac{\sum_{i=1}^{k}(1-\bar{y}^{(i)})(1-\hat{y}^{(i)})}{\sum_{i=1}^{k}(1-\bar{y}^{(i)})}.$$

Precision and recall can be combined in the so called *F-measure* that is simply defined as the harmonic mean of precision and recall,

$$F_1 = \frac{2P_1R_1}{P_1+R_1} \quad \text{and} \quad F_0 = \frac{2P_0R_0}{P_0+R_0}.$$

## Softmax regression

Softmax regression is the natural extension of logistic regression to explained variables that can take more than two values. For deriving the extension one observes that the data generation model for the binary case can be rewritten as follows: We assume now that $\mathcal{Y} = \{1, 2\}$ and

$$p(Y = 1|X = x, \theta) \;=\; \frac{1}{1 + \exp(-x^\top \theta)} \;=\; \frac{\exp\left(\frac{1}{2}x^\top\theta\right)}{\exp\left(\frac{1}{2}x^\top\theta\right) + \exp\left(-\frac{1}{2}x^\top\theta\right)}$$

and

$$\begin{aligned} p(Y = 2|X = x, \theta) \;&=\; 1 - \frac{1}{1 + \exp(-x^\top \theta)} \;=\; \frac{1}{1 + \exp(x^\top\theta)} \\ &=\; \frac{\exp\left(-\frac{1}{2}x^\top\theta\right)}{\exp\left(\frac{1}{2}x^\top\theta\right) + \exp\left(-\frac{1}{2}x^\top\theta\right)}. \end{aligned}$$

Let $\theta^{(1)} = \frac{1}{2}\theta$ and $\theta^{(2)} = -\frac{1}{2}\theta$, then

$$p(Y = 1|X = x, \Theta) \;=\; \frac{\exp\left(x^\top\theta^{(1)}\right)}{\exp\left(x^\top\theta^{(1)}\right) + \exp\left(x^\top\theta^{(2)}\right)}$$

and

$$p(Y = 2|X = x, \Theta) \;=\; \frac{\exp\left(x^\top\theta^{(2)}\right)}{\exp\left(x^\top\theta^{(1)}\right) + \exp\left(x^\top\theta^{(2)}\right)}.$$

The natural extension to $\mathcal{Y} = [k] = \{1, \ldots, k\}$ is thus

$$p(Y = j|X = x, \Theta) \;=\; \frac{\exp\left(x^\top\theta^{(j)}\right)}{\sum_{l=1}^{k}\exp(x^\top\theta^{(l)})} \qquad \text{for all } j \in [k],$$

with model parameters

$$\theta^{(1)}, \ldots, \theta^{(k)} \;\in\; \mathbb{R}^{n+1},$$

i.e., there are $k \cdot (n + 1)$ parameters in total that can be combined into the parameter matrix $\Theta \in \mathbb{R}^{(n+1)\times k}$. The maximum likelihood function for estimating these parameters from data points

$$(x^{(1)}, y^{(1)}), \ldots, (x^{(m)}, y^{(m)})$$

is given under the i.i.d. assumption as

$$L(\Theta) \;=\; \prod_{i=1}^{m} \frac{\exp\left({x^{(i)}}^\top\theta^{(y^{(i)})}\right)}{\sum_{l=1}^{k}\exp\left({x^{(i)}}^\top\theta^{(l)}\right)} \;=\; \prod_{i=1}^{m}\prod_{j=1}^{k}\left(\frac{\exp\left({x^{(i)}}^\top\theta^{(j)}\right)}{\sum_{l=1}^{k}\exp\left({x^{(i)}}^\top\theta^{(l)}\right)}\right)^{\mathbf{1}[y^{(i)}=j]}.$$

From here we can continue similarly as in the binary case. We leave this as an exercise.

# Chapter 4

# Naive Bayes

Naive Bayes is a contingency analysis technique that can be easily extended for classification. In contrast to ordinary least squares and logistic regression, naive Bayes is a generative model. That is, the naive Bayes model describes a joint probability distribution on the space of explanatory and explained variables. The naive Bayes method is based on a theorem that was discovered in a special case by Reverend Thomas Bayes in the second half of the 18-th century. The general version of the theorem in its modern form was independently found a few years later by Pierre-Simon Laplace.

## Data generation model

We assume that $\mathcal{X} = \{0,1\}^n$, i.e., the space of Bit-strings of length $n$, and $\mathcal{Y} = \{0,1\}$. The naive Bayes model is then given by the following assumption on the joint probability density on $\mathcal{X} \times \mathcal{Y}$:

1. $p(Y=1) = \theta$, and thus $p(Y=0) = 1-\theta$.

2. Naive Bayes assumption:

$$p(x_1,\ldots,x_n|Y=y) \;=\; \prod_{j=1}^{n} p(x_j|Y=y)$$

3. For all $j \in [n] : p(X_j = 1|Y=y) = \theta_{yj}$, and thus $p(X_j = 0|Y=y) = 1-\theta_{yj}$, where $y$ can take the values $0$ and $1$.

The parameters of this model are $\theta, \theta_{0j}$ and $\theta_{1j}$ for $j \in [n]$. That is, there are $2n+1$ parameter that we combine in the parameter vector $\Theta \in [0,1]^{2n+1}$. The number of parameters should be compared to the model without the naive Bayes assumption. The space $\mathcal{X} \times \mathcal{Y}$ has $2^{n+1}$ elements. In the most general case we have a probability for each element in $\mathcal{X} \times \mathcal{Y}$ and these probabilities need to sum up to one. Hence, in the general model there are $2^{n+1} - 1$ parameters. The naive Bayes assumption reduces this number from exponential to linear in the number of variables. This is remarkable, because it is intuitively clear that the number of data points that are necessary for reliably estimating the parameters should grow with the number of parameters. Hence, the general model is not tractable, even for moderately large numbers of variables.

### Maximum likelihood estimate

Assume that we have observed the i.i.d. data

$$(x^{(1)}, y^{(1)}), \ldots, (x^{(m)}, y^{(m)})$$

from a naive Bayes model. The likelihood function for the model parameters $\theta$ and $\theta_{yj}$ is given as follows

$$\begin{aligned}
&L(\Theta) \\
&= \prod_{i=1}^{m} p(x^{(i)}, y^{(i)}) \\
&= \prod_{i=1}^{m} p(x^{(i)}|Y = y^{(i)})p(y^{(i)}) \\
&= \prod_{i=1}^{m} p(x_1^{(i)}, \ldots, x_n^{(i)}|Y = y^{(i)})p(y^{(i)}) \\
&= \prod_{i=1}^{m} p(y^{(i)}) \prod_{j=1}^{n} p(x_j^{(i)}|Y = y^{(i)}) \\
&= \prod_{i=1}^{m} \theta^{y^{(i)}}(1-\theta)^{1-y^{(i)}} \prod_{j=1}^{n} \left(\theta_{1j}^{x_j^{(i)}}(1-\theta_{1j})^{1-x_j^{(i)}}\right)^{y^{(i)}} \left(\theta_{0j}^{x_j^{(i)}}(1-\theta_{0j})^{1-x_j^{(i)}}\right)^{1-y^{(i)}},
\end{aligned}$$

where we have used the definition of conditional probabilities (which leads to Bayes' theorem, if we apply the definition twice)

$$p(x|y) \;=\; \frac{p(x,y)}{p(y)} \left(= \frac{p(y|x)p(x)}{p(y)}\right)$$

or equivalently

$$p(x,y) \;=\; p(x|y)p(y)$$

for the second equality. Consequently, the log-likelihood function is given as

$$\begin{aligned}\ell(\Theta) &= \sum_{i=1}^{m}\sum_{j=1}^{n} y^{(i)}\left(x_j^{(i)}\log\theta_{1j} + (1-x_j^{(i)})\log(1-\theta_{1j})\right)\\ &+ \sum_{i=1}^{m}\sum_{j=1}^{n}(1-y^{(i)})\left(x_j^{(i)}\log\theta_{0j} + (1-x_j^{(i)})\log(1-\theta_{0j})\right)\\ &+ \sum_{i=1}^{m} y^{(i)}\log\theta + (1-y^{(i)})\log(1-\theta).\end{aligned}$$

The maximum likelihood estimate

$$\hat{\Theta} = \text{argmax}_{\Theta\in\mathbb{R}^{2n+1}}\ L(\Theta) = \text{argmax}_{\Theta\in\mathbb{R}^{2n+1}}\ \ell(\theta)$$

can be computed from the vanishing gradient condition for the log-likelihood function as follows: We get $\hat{\theta}$ from

$$\frac{d\ell(\Theta)}{d\theta} = \sum_{i=1}^{m}\frac{y^{(i)}}{\theta} - \frac{1-y^{(i)}}{1-\theta} = 0,$$

which solves to

$$\hat{\theta} = \frac{1}{m}\sum_{i=1}^{m} y^{(i)}.$$

We get $\hat{\theta}_{1j}$ for $j\in[n]$ from

$$\frac{d\ell(\Theta)}{d\theta_{1j}} = \sum_{i=1}^{m}\frac{y^{(i)}x_j^{(i)}}{\theta_{1j}} - \frac{y^{(i)}(1-x_j^{(i)})}{1-\theta_{1j}} = 0,$$

which solves to

$$\hat{\theta}_{1j} = \frac{\sum_{i=1}^{m} y^{(i)}x_j^{(i)}}{\sum_{i=1}^{m} y^{(i)}}.$$

An finally, we get $\hat{\theta}_{0j}$ for $j\in[n]$ from

$$\frac{d\ell(\Theta)}{d\theta_{0j}} = \sum_{i=1}^{m}\frac{(1-y^{(i)})x_j^{(i)}}{\theta_{0j}} - \frac{(1-y^{(i)})(1-x_j^{(i)})}{1-\theta_{0j}} = 0,$$

which solves to

$$\hat{\theta}_{0j} = \frac{\sum_{i=1}^{m}(1-y^{(i)})x_j^{(i)}}{\sum_{i=1}^{m}(1-y^{(i)})}.$$

### Prediction

As in the logistic regression case we can use the estimates $\hat{\theta}$ of the model parameters for prediction using the predictor

$$\begin{aligned} h : \mathcal{X} = \{0,1\}^n \to \mathcal{Y} = \{0,1\},\ x \mapsto & \ \text{argmax}_{y \in \{0,1\}}\ \hat{p}(y|X=x) \\ & = \begin{cases} 0 & : \ \hat{p}(Y=1|X=x) \ < \ \frac{1}{2} \\ 1 & : \ \hat{p}(Y=1|X=x) \ \geq \ \frac{1}{2}, \end{cases} \end{aligned}$$

where

$$\begin{aligned} \hat{p}(Y=1|X=x) &= \frac{\hat{p}(x|Y=1)\hat{p}(Y=1)}{\hat{p}(x)} = \frac{\hat{p}(x|Y=1)\hat{p}(Y=1)}{\hat{p}(x,Y=0)+\hat{p}(x,Y=1)} \\ &= \frac{\hat{p}(x|Y=1)\hat{p}(Y=1)}{\hat{p}(x|Y=0)\hat{p}(Y=0)+\hat{p}(x|Y=1)\hat{p}(Y=1)} \\ &= \frac{\hat{p}(Y=1)\prod_{j=1}^n \hat{p}(x_i|Y=1)}{p(Y=0)\prod_{j=1}^n \hat{p}(x_j|Y=0)+\hat{p}(Y=1)\prod_{j=1}^n \hat{p}(x_j|Y=1)} \\ &= \frac{\hat{\theta}\prod_{j=1}^n \hat{\theta}_{1j}^{x_j}(1-\hat{\theta}_{1j})^{1-x_j}}{(1-\hat{\theta})\prod_{j=1}^n \hat{\theta}_{0j}^{x_j}(1-\hat{\theta}_{0j})^{1-x_j}+\hat{\theta}\prod_{j=1}^n \hat{\theta}_{1j}^{x_j}(1-\hat{\theta}_{1j})^{1-x_j}}. \end{aligned}$$

Note that the last expression for $\hat{p}(Y=1|X=x)$ can be naturally extended from $\mathcal{X} = \{0,1\}^n$ to the unit cube $[0,1]^n$. This extension is useful for defining the decision boundary

$$\left\{ x \in [0,1]^n \ : \ \hat{p}(Y=1|X=x) = \frac{1}{2} \right\},$$

of the classifier $h$. The decision boundary is an $(n-1)$-dimensional surface that divides the unit cube $[0,1]^n$ into two parts. The classifier $h$ predicts the same value $y$ for all points in $\{0,1\}^n$ that fall into the same part. For deriving the form of the $(n-1)$-dimensional decision boundary, we need to solve the equation

$$\begin{aligned} & p(Y=1|X=x) \\ & \qquad = \frac{\hat{\theta}\prod_{j=1}^n \hat{\theta}_{1j}^{x_j}(1-\hat{\theta}_{1j})^{1-x_j}}{(1-\hat{\theta})\prod_{j=1}^n \hat{\theta}_{0j}^{x_j}(1-\hat{\theta}_{0j})^{1-x_j}+\hat{\theta}\prod_{j=1}^n \hat{\theta}_{1j}^{x_j}(1-\hat{\theta}_{1j})^{1-x_j}} = \frac{1}{2} \end{aligned}$$

for $x \in [0,1]^n$. We can use

$$\hat{\theta}\prod_{j=1}^n \hat{\theta}_{1j}^{x_j}(1-\hat{\theta}_{1j})^{1-x_j} = \exp\Big(x^\top\big(\log\hat{\theta}_1 - \log(1-\hat{\theta}_1)\big) + \log 1^\top(1-\hat{\theta}_1) + \log\hat{\theta}\Big)$$

and

$$(1-\hat\theta)\prod_{j=1}^{n}\hat\theta_{0j}^{x_j}(1-\hat\theta_{0j})^{1-x_j}$$
$$= \exp\Big(x^\top\big(\log\hat\theta_0 - \log(1-\hat\theta_0)\big) + \log 1^\top(1-\hat\theta_0) + \log(1-\hat\theta)\Big),$$

where $\hat\theta_0 = (\hat\theta_{0j})_{j\in[n]}$ and $\hat\theta_1 = (\hat\theta_{1j})_{j\in[n]}$, to rewrite the equation as

$$\exp\left(x^\top \log\left(\frac{\hat\theta_0}{\hat\theta_1}\frac{1-\hat\theta_1}{1-\hat\theta_0}\right) + \log\left(\frac{1-\hat\theta}{\hat\theta}\frac{1^\top(1-\hat\theta_0)}{1^\top(1-\hat\theta_1)}\right)\right) \;=\; 1,$$

or by taking the logarithm on both sides

$$x^\top \log\left(\frac{\hat\theta_0}{\hat\theta_1}\frac{1-\hat\theta_1}{1-\hat\theta_0}\right) \;=\; \log\left(\frac{\hat\theta}{1-\hat\theta}\frac{1^\top\hat\theta_1}{1^\top\hat\theta_0}\right),$$

which is linear in $x$. Thus, the decision boundary is again a hyperplane.

## Laplace smoothing

Let us briefly go back to parameter estimation by the maximum likelihood approach. If the feature $x_j$, $j \in [n]$ has not been observed, i.e., if $x_j^{(i)} = 0$ for all the data points $x^{(1)}, \ldots, x^{(m)}$, then we have

$$\hat\theta_{0j} \;=\; \hat\theta_{1j} \;=\; 0,$$

which is after all a reasonable estimate. But the implication of this estimate on our inference is as follows: Assume we want to predict the value of $y$ at $x \in \mathcal{X}$ with $x_j = 1$, then we get

$$p(Y=1|X=x) \;=\; \frac{0}{0}$$

and it is not obvious what to do with this. Of course, a similar problem arises if the feature $x_j$ has been observed in every data point.

One way to deal with the problem is *Laplace smoothing*. Laplace smoothing is a regularization technique that adds $4k$ artificial data points to the set of actual data points, namely $k$ of each of the following four data points

$$(X_1=0,\ldots,X_n=0,Y=0), \quad (X_1=0,\ldots,X_n=0,Y=1),$$
$$(X_1=1,\ldots,X_n=1,Y=0), \quad (X_1=1,\ldots,X_n=1,Y=1).$$

Laplace smoothing is a *Bayesian technique* that allows to incorporate some *prior belief* into the parameter estimation problem. Here, the prior belief is that the true parameters have the values

$$\theta = \frac{1}{2}, \quad \theta_{0j} = \frac{1}{2}, \quad \theta_{1j} = \frac{1}{2} \quad \text{for } j \in [n].$$

By the choice of $k \geq 1$ one can control how strongly the prior belief is reflected in the estimated parameters. For large values of $k$ the prior belief dominates while the influence of the actual data diminishes. On the other hand, for any fixed value of $k$ and a growing number of data points the influence of the prior belief on the estimated parameters diminishes. Hence, $k$ here plays the same role as the regularization parameter $c$ in ridge regression or in regularized logistic regression.

## Multinomial naive Bayes

Naive Bayes as we have described it above can be easily extended to nominal variables with more than two possible outcomes: Let $\mathcal{Y} = [n_0]$ and $\mathcal{X}_i = [n_i]$, $i \in [n]$ for natural numbers $n_0, n_1, \ldots, n_n \in \mathbb{N}$. The parameters of the naive Bayes model then become

$$\theta_i = p(Y = i), \quad \text{for } i \in [n_0]$$

and

$$\theta_{ijk} = p(X_j = k | Y = i), \quad \text{for } i \in [n_0],\, j \in [n],\, k \in [n_j],$$

where

$$\theta_i \geq 0 \text{ and } \sum_{i=1}^{n_0} \theta_i = 1 \quad \text{and} \quad \theta_{ijk} \geq 0 \text{ and } \sum_{k=1}^{n_j} \theta_{ijk} = 1.$$

Given data points

$$(x^{(1)}, y^{(1)}), \ldots, (x^{(m)}, y^{(m)})$$

the maximum likelihood estimates for the parameters become

$$\hat{\theta}_i = \frac{1}{m} \sum_{l=1}^{m} \mathbf{1}[y^{(l)} = i]$$

and

$$\hat{\theta}_{ijk} = \frac{\sum_{l=1}^{m} \mathbf{1}[y^{(l)} = i \wedge x_j^{(l)} = k]}{\sum_{l=1}^{m} \mathbf{1}[y^{(l)} = i]}.$$

Inference queries, i.e., the prediction of the value $y \in \mathcal{Y}$ at $x \in \mathcal{X} = \prod_{j=1}^{n} \mathcal{X}_j$, can be answered by using the predictor

$$h : \mathcal{X} \to \mathcal{Y},\ x \mapsto \mathsf{argmax}_{y \in [n_0]}\ \hat{p}(y|X = x),$$

where

$$\hat{p}\big(Y = i|X = (k_1, \ldots, k_n)\big) \;=\; \frac{\hat{\theta}_i \prod_{j=1}^{n} \hat{\theta}_{ijk_j}}{\sum_{l=1}^{n_0} \hat{\theta}_l \prod_{j=1}^{n} \hat{\theta}_{ljk_j}}.$$

## Gaussian naive Bayes

The naive Bayes idea can also be used for classification. Let $\mathcal{Y} = [k]$ and $\mathcal{X}_j = \mathbb{R}$, $j \in [n]$, i.e., $\mathcal{X} = \mathbb{R}^n$. The Gaussian naive Bayes data generation model assumes:

1. $p(Y = y) = \theta_y \geq 0,\ y \in [k]$, where $\sum_{i=1}^{k} \theta_i = 1$.
2. Naive Bayes assumption:
$$p(x_1, \ldots, x_n|Y = y) \;=\; \prod_{j=1}^{n} p(x_j|Y = y)$$
3. For all $j \in [n]\ :\ X_j|Y = y\ \sim\ \mathcal{N}(\mu_{yj}, \sigma_{yj}^2)$, i.e.,
$$p(X_j = x|Y = y) \;=\; \frac{1}{\sqrt{2\pi}\sigma_{yj}} \exp\left(-\frac{(x - \mu_{yj})^2}{2\sigma_{yj}^2}\right)$$

The parameters of this model are

$$\theta_y,\ \mu_{yj},\ \sigma_{yj} \qquad \text{for } y \in [k],\ j \in [n].$$

Given data points

$$(x^{(1)}, y^{(1)}), \ldots, (x^{(m)}, y^{(m)})$$

the maximum likelihood estimates for these parameters are given as

$$\hat{\theta}_y \;=\; \frac{1}{m} \sum_{i=1}^{m} \mathbf{1}[y^{(i)} = y]$$

and

$$\hat{\mu}_{yj} \;=\; \frac{\sum_{i=1}^{m} \mathbf{1}[y^{(i)} = y] x_j^{(i)}}{\sum_{i=1}^{m} \mathbf{1}[y^{(i)} = y]}, \quad \text{and} \quad \hat{\sigma}_{yj}^2 \;=\; \frac{\sum_{i=1}^{m} \mathbf{1}[y^{(i)} = y]\left(x^{(i)} - \hat{\mu}_{yj}\right)^2}{\sum_{i=1}^{m} \mathbf{1}[y^{(i)} = y]}$$

Inference queries, i.e., the prediction of the value $y \in \mathcal{Y}$ at $x \in \mathcal{X} = \mathbb{R}^n$, can be answered by using the predictor

$$h : \mathcal{X} \to \mathcal{Y},\ x \mapsto \mathsf{argmax}_{y \in [k]}\ \hat{p}(y|X = x),$$

where

$$\begin{aligned} &\hat{p}\big(Y = y|X = (x_1, \ldots, x_n)\big) \\ &\qquad = \frac{\hat{\theta}_y \prod_{j=1}^n \exp\big(-(x_j - \hat{\mu}_{yj})^2/2\hat{\sigma}_{yj}^2\big)/\hat{\sigma}_{yj}}{\sum_{y'=1}^k \hat{\theta}_{y'} \prod_{j=1}^n \exp\big(-(x_j - \hat{\mu}_{y'j})^2/2\hat{\sigma}_{y'j}^2\big)/\hat{\sigma}_{y'j}} \\ &\qquad = \frac{\frac{\hat{\theta}_y}{\prod_{j=1}^n \hat{\sigma}_{yj}} \exp\left(-\sum_{j=1}^n \frac{(x_j - \hat{\mu}_{yj})^2}{2\hat{\sigma}_{yj}^2}\right)}{\sum_{y'=1}^k \frac{\hat{\theta}_{y'}}{\prod_{j=1}^n \hat{\sigma}_{y'j}} \exp\left(-\sum_{j=1}^n \frac{(x_j - \hat{\mu}_{y'j})^2}{2\hat{\sigma}_{y'j}^2}\right)} \end{aligned}$$

In the case $k = 2$, i.e., there are only two possible outcomes for the explained variable $Y$. Let us assume $\mathcal{Y} = \{0, 1\}$ and

$$p(Y = 1) = \theta, \quad \text{and thus } p(Y = 0) = 1 - \theta.$$

Then we have

$$\hat{\theta} = \frac{1}{m} \sum_{i=1}^m y^{(i)}$$

and

$$\begin{aligned} &\hat{p}(Y = 1|X = (x_1, \ldots, x_n)\big) \\ &= \frac{\frac{\hat{\theta}}{\prod_{j=1}^n \hat{\sigma}_{1j}} \exp\left(-\sum_{j=1}^n \frac{(x_j - \hat{\mu}_{1j})^2}{2\hat{\sigma}_{1j}^2}\right)}{\frac{1-\hat{\theta}}{\prod_{j=1}^n \hat{\sigma}_{0j}} \exp\left(-\sum_{j=1}^n \frac{(x_j - \hat{\mu}_{0j})^2}{2\hat{\sigma}_{0j}^2}\right) + \frac{\hat{\theta}}{\prod_{j=1}^n \hat{\sigma}_{1j}} \exp\left(-\sum_{j=1}^n \frac{(x_j - \hat{\mu}_{1j})^2}{2\hat{\sigma}_{1j}^2}\right)}. \end{aligned}$$

The decision boundary

$$\left\{x \in \mathbb{R}^n\ :\ \hat{p}(Y = 1|X = x) = \frac{1}{2}\right\}$$

is a hyperplane only if $\hat{\sigma}_{0j} = \hat{\sigma}_{1j}$ for all $j \in [n]$.

## Gaussian discriminant analysis

The last two assumptions of the binary Gaussian naive Bayes data generation model can be summarized as

$$p(x_1, \dots, x_n | Y = y) \sim \mathcal{N}\big(\mu_y, \mathsf{diag}(\sigma_{1y}, \dots, \sigma_{ny})\big),$$

where $\mu_y = (\mu_{1y}, \dots, \mu_{ny})^\top$ and $\mathsf{diag}(\sigma_{1y}, \dots, \sigma_{ny})$ is a $(n \times n)$-matrix whose non-zero entries are the diagonal elements $\sigma_{1y}, \dots, \sigma_{ny}$. In *Gaussian discriminant analysis* it is no longer assumed that the variables $X_j | Y = y$ are independent, but the distribution of $X | Y = y$ can be any multivariate Gaussian, i.e.,

$$p(x_1, \dots, x_n | Y = y) \sim \mathcal{N}(\mu_y, \Sigma_y),$$

where $\Sigma_y$, $y \in \{0, 1\}$ can be any symmetric, positive definite $(n \times n)$-matrix. That is, in Gaussian discriminant analysis we are giving up on the naive Bayes assumption, namely, that the features are conditionally independent given the label.

*Linear discriminant analysis* is the special case of Gaussian discriminant analysis where $\Sigma_0 = \Sigma_1$. In this case the decision boundary

$$\Big\{ x \in \mathbb{R}^n \,:\, \hat{p}(Y = 1 | X = x) = \frac{1}{2} \Big\}$$

is again a hyperplane, where $\hat{p}(Y = 1 | X = x)$, or more specifically the parameters $\hat{\theta}, \hat{\mu}_0, \hat{\mu}_1$ and $\hat{\Sigma}$, respectively, have been estimated from data.

# Chapter 5

# Exercises

### Chapter 1

**Exercise 1.** For the OLS model, compute the maximum likelihood estimate of the variance $\sigma^2$.

**Exercise 2.** Let $X$ be a data matrix whose columns contain the data points $x^{(1)}, \ldots, x^{(m)}$. From $X$ we can derive the second moment matrix $XX^\top$ and the so called Gram matrix $X^\top X$. Show that both the second moment matrix and the Gram matrix are symmetric and positive semi-definite, and

$$X^\top X = \left(x^{(i)\top} x^{(j)}\right)_{i,j=1,\ldots,m} \quad \text{and} \quad XX^\top = \sum_{i=1}^{m} x^{(i)} x^{(i)\top}.$$

**Exercise 3.** Compute the solution of the ridge regression problem

$$\text{argmin}_{\theta \in \mathbb{R}^{n+1}} \ \frac{1}{2}\|y - X^\top \theta\|^2 + \frac{c}{2}\|\theta\|^2.$$

**Exercise 4.** An infinite floating system is the following set of numbers

$$F(\beta, t) = \left\{x \in \mathbb{Q} \,:\, x = sm\beta^{e-t}\right\} \cup \{0\},$$

where $e \in \mathbb{Z}$ is the *exponent*, $m \in \mathbb{N}$ with $\beta^{t-1} \leq m \leq \beta^t - 1$ is the *mantissa*, $s \in \{\pm 1\}$ is the *sign*, $\beta \in \mathbb{N} \setminus \{0\}$ is the *basis*, and $t \in \mathbb{N} \setminus \{0\}$ is the *precision*. Compute and visualize the elements of $F(2,3)$ with $s = 1$ and $e \in \{-2,-1,0,1,2,3\}$.

## Chapter 3

**Exercise 1.** Write down the objective function of the logistic regression maximum likelihood problem in vectorized form.

Hint: Given data points $(x^{(1)}, y^{(1)}), \dots, (x^{(m)}, y^{(m)}) \in (\{1\} \times \mathbb{R}^n) \times \{\pm 1\}$. Combine the feature vectors into the data matrix $X \in \mathbb{R}^{(n+1)\times m)}$ and the labels into the label vector $y \in \{\pm 1\}^m$. You also need the $m$-dimensional all-ones-vector vec$(1)$.

**Exercise 2.** Derive the maximum likelihood estimate for parameter matrix $\Theta$ of the softmax regression problem.

## Chapter 4

**Exercise 1.** Given data points

$$(x^{(1)}, y^{(1)}), \dots, (x^{(m)}, y^{(m)}) \in \mathcal{X} \times \mathcal{Y} = \mathbb{R}^n \times \{0, 1\}$$

derive the maximum likelihood estimates of the parameters $\hat{\theta}, \mu_0$ and $\mu_1$ for the Gaussian discriminant analysis problem.

**Exercise 2.** Show for the linear discriminant analysis problem that

$$p(Y = 1 | X = x) = \frac{1}{1 + \exp\left(-\theta^\top x\right)},$$

where $\theta$ is some function of the parameters $\hat{\theta}, \mu_0, \mu_1$ and $\Sigma$.
Remark: Use the convention $x_0 = 1$, i.e., the zeroth entry of $x$ is set to $1$.

**Exercise 3.** Use your result from Exercise 2 to show that in linear discriminant analysis the resulting classifier is indeed linear, i.e., the decision boundary is a hyperplane.

**Exercise 4.** Extend the formulas for *precision* and *recall* for the case that the explained variable $y$ can take more than two values.

# Part II

# Feature Maps

# Chapter 6

# Feature maps

So far we have considered mostly $\mathbb{R}^n$ (or $\{1\}\times\mathbb{R}^n$) as the space of our explanatory variables (ordinary least squares regression, logistic regression, and Gaussian discriminant analysis). In all these cases we have made use of the Euclidean structure on $\mathbb{R}^n$, i.e., of the inner product

$$\langle\cdot,\cdot\rangle : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R},\ (x^{(1)}, x^{(2)}) \mapsto \langle x^{(1)}, x^{(2)}\rangle := {x^{(1)}}^\top x^{(2)}.$$

The inner product induces a geometry on $\mathbb{R}^n$, i.e., it allows us to measure distances and angles. We used angles, or more specifically right angles, to define hyperplanes that arose for instance as decision boundaries in logistic regression, naive Bayes and linear discriminant analysis. This approach has two shortcomings:

1. In many cases linear decision boundaries or linear regressors are not flexible enough, and
2. the feature space $\mathcal{X}$ is often not a Euclidean space. For instance, $\mathcal{X}$ could be a space of strings, images, protein structures, ... .

Thus, in the following, we want to move beyond linear decision boundaries or linear regressors and to more general feature spaces $\mathcal{X}$.

### Examples

A straightforward idea towards these goals is to use (non-linear) *feature maps*

$$\phi : \mathcal{X} \to \mathbb{R}^n$$

for dealing with abstract feature spaces $\mathcal{X}$. A feature map maps the points in $\mathcal{X}$ to *feature vectors* in $\mathbb{R}^n$. The learning methods that we have discussed so far can then be used for the feature vectors.

**Binary features.** Typically, the entries of $\phi(x)$, where $x \in \mathcal{X}$, are numerical quantities measured on $x$, but note that our definition also includes feature maps $\phi : \mathcal{X} \to \{0,1\}^n$, where every entry in $\phi(x)$ corresponds to some predicate $c_i$ that is either satisfied by $x \in \mathcal{X}$ or not, i.e.,

$$\phi(x)_i \;=\; \mathbf{1}[c_i(w) = \texttt{true}].$$

**Non-linear feature maps.** Note that also feature maps $\phi : \mathbb{R}^n \to \mathbb{R}^k$ from Euclidean space into Euclidean space make sense, namely, any non-linear map $\phi$ that turns our basic learning methods non-linear. A popular non-linear feature map is

$$\phi : \mathbb{R}^n \to \mathbb{R}^{\binom{n+k}{n}}$$

that maps $x \in \mathbb{R}^n$ to the vector of all monomials of degree at most $k$ evaluated at $x$, i.e., the entries of $\phi(x)$ are of the form

$$x_1^{d_1} \cdot \ldots \cdot x_n^{d_n}$$

with $d \in \mathbb{N}^n$ and $\sum_{i=1}^n d_i \leq k$. In this case, $\theta^\top \phi(x)$, where $\theta \in \mathbb{R}^{\binom{n+k}{n}}$, is a polynomial of degree at most $k$. Note that now the number of parameters, i.e., the entries in $\theta$, grows with $n^k$ and we need $\Theta(n^k)$ arithmetic operations for computing the the inner product $\theta^\top \phi(x)$ which becomes prohibitive already for moderately large values of $k$. We would also assume that we quickly run out of data points for reliably estimating the parameters, but this is not always the case as we will show in the context of support vector machines in Part IV.

**Deep learning feature maps.** In deep learning feature vectors $x \in \mathbb{R}^n$ are traditionally transformed by logistic functions

$$\phi : x \;\mapsto\; \frac{1}{1 + \exp\big(-\theta^\top x + \theta_0\big)} \in \mathbb{R}.$$

To get a feature map into $\mathbb{R}^k$ one simply combines $k$ logistic functions $\phi_i$, $i \in [k]$ into

$$\phi : \mathbb{R}^n \to \mathbb{R}^k,\, x \mapsto \big(\phi_1(x), \ldots, \phi_k(x)\big)^\top,$$

where $\phi_i$ has the parameters $\theta^{(i)} \in \mathbb{R}^n$ and $\theta_0^{(i)} \in \mathbb{R}$. The parameter vectors $\theta^{(i)} \in \mathbb{R}^n$, $i \in [k]$ can be combined into a $(n \times k)$-matrix $\Theta$, whose columns are the vectors $\theta^{(i)} \in \mathbb{R}^n$, $i \in [k]$, and the intercepts $\theta_0^{(i)} \in \mathbb{R}$, $i \in [k]$ can be combined into a vector $\theta^{(0)} \in \mathbb{R}^k$. Using the combined parameters the logistic function feature map can be written in vectorized form as follows

$$x \;\mapsto\; \mathsf{vec}(1) \oslash \Big(\mathsf{vec}(1) + \exp\big(-\Theta^\top x + \theta^{(0)}\big)\Big) \;=\; \bar{x} \in \mathbb{R}^k.$$

The transformation of a data matrix $X \in \mathbb{R}^{n\times m}$ of $m$ data points $x^{(1)}, \ldots, x^{(m)}$ can be written in vectorized form as

$$X \;\mapsto\; \mathsf{mat}(1) \oslash \Big(\mathsf{mat}(1) + \exp\big(-\Theta^\top X + \theta^{(0)}\big)\Big) \;=\; X' \in \mathbb{R}^{k\times m},$$

where $\mathsf{mat}(1) \in \mathbb{R}^{k\times m}$ is the matrix whose entries are all $1$, and $\oslash$ is the componentwise division operator. Whenever the learning method, for instance ordinary least squares or logistic regression, is given in vectorized form, then one can just plug the transformed data matrix instead of the original into the problem .

The transformation process from above can be iterated by feeding the transformed data matrix $X'$ into another layer of logistic units leading to the transformed data matrix $X''$ that can be into the objective function of the learning problem. Iterating the whole process several times results in a model that is called a *deep network*. A transformation layer of the network is the combination of an affine transformation, also called *fully connected layer*, and an *activation function*, here a logistic function.

The data can also be transformed by using other activation functions. Popular choices are $\tanh$-activation functions, i.e.,

$$\mathbb{R}^n \ni x \;\mapsto\; \tanh\big(\Theta^\top x + \theta^{(0)}\big) \in \mathbb{R}^k,$$

or rectified linear units (ReLUs)

$$x \;\mapsto\; \mathsf{max}\big\{\mathsf{vec}(0),\, \Theta^\top x + \theta_0\big\}.$$

Here, the $\tanh$ and the $\max$-function, respectively, are used componentwise. The parameters are $\hat{\Theta} \in \mathbb{R}^{n\times k}$ and the intercept vector $\theta^{(0)} \in \mathbb{R}^k$.

*Remark:* Traditional feature maps are computed from data independent of the learning method. These feature maps do not introduce additional parameters that have to be learned by the learning method. This is also true for most non-linear classical feature maps like the monomial mapping from above. A characteristic feature of feature maps for deep learning is the (often a large number of) additional parameters that have to be learned by the learning method. Introducing the additional parameters typically renders the learning method non-convex, or non-concave in the case of maximization problems, respectively. Typically, we benefit from the added flexibility induced by additional parameters when we have many data points.

### Prediction

It is important to note that once we have trained our model from data that have been transformed using some feature map, we also need to transform any point $x$ in the feature space $\mathcal{X}$ at which we want to predict a label.

# Chapter 7

# Adjoint problems and kernels

Let us come back to the ridge regression and the regularized logistic regression problems. The parameter estimation problems for these problems from i.i.d. data points

$$(x^{(1)}, y^{(1)}), \ldots, (x^{(m)}, y^{(m)})$$

share some commonalities that we want to use here to gain some insights into the structure of their solutions. These insights allow us to significantly generalize the classical regression and classification techniques in the next chapter. Here, the $x^{(i)} \in \mathcal{X} = \mathbb{R}^n$ (or $\mathcal{X} = \{1\} \times \mathbb{R}^n$) are explanatory variables and the $y^{(i)} \in \mathcal{Y}$ are explained variables, i.e., $\mathcal{Y} = \mathbb{R}$ in the regression case and $\mathcal{Y} = \{\pm 1\}$ in the binary classification case. As before we summarize the observed explained variables in a label vector $y \in \mathcal{Y}^m$ and the explanatory variables in a data matrix $X$ whose columns are the data points $x^{(i)}$, $i \in [m]$.

Remark: Note that instead of the data matrix $X$ we can always work with a feature matrix $\Phi \in \mathbb{R}^{n \times m}$, whose columns are $\phi(x^{(i)})$, $i \in [m]$, where $\phi : \mathcal{X} \to \mathbb{R}^n$ is a feature map and the $x^{(i)} \in \mathcal{X}$ are data points (explanatory variables).

The ridge regression problem for the feature vectors can be written as

$$\mathsf{argmin}_{\theta \in \mathbb{R}^n} \;\; \frac{1}{2}\|y - X^\top \theta\|^2 + c\|\theta\|^2,$$

and the regularized parameter estimation problem in logistic regression is given in vectorized form as

$$\mathsf{argmin}_{\theta \in \mathbb{R}^n} \;\; \mathsf{vec}(1)^\top \log\left(\mathsf{vec}(1) + \exp(-y \odot X^\top \theta)\right) + c\|\theta\|^2,$$

where vec$(1)$ is the vector in $\mathbb{R}^n$ whose components are all $1$, $\exp$ and $\log$ are applied componentwise, $\odot$ is the componentwise product, and $X^\top\theta$ is the matrix-vector product of the data matrix $X^\top$ and the parameter vector $\theta$.

Hence, the ridge problem as well as the regularized logistic regression problem are of the form

$$\text{argmin}_{\theta\in\mathbb{R}^n}\ L(X^\top\theta, y) + c\|\theta\|^2,$$

where $L(\cdot,\cdot)$ is a *loss function* that depends on the vector $\theta^\top X$ and the label vector $y$. We call problems of this form Euclidean regularized learning problems, since the regularization term $\|\theta\|^2$ is the squared Euclidean norm of the parameter vector $\theta$. The key property of Euclidean regularized learning problems for parameter vectors $\theta\in\mathbb{R}^n$ is the following theorem.

**Theorem 2. [Representer Theorem]** *For any loss function $L(\cdot,\cdot)$ if*

$$\hat{\theta}\ =\ \text{argmin}_{\theta\in\mathbb{R}^n}\ L(\theta^\top X, y) + c\|\theta\|^2$$

*exists, then $\hat{\theta} = X\hat{a}$ for some $\hat{a}\in\mathbb{R}^m$.*

*Proof.* Since $\hat{\theta}$ can be decomposed as $\hat{\theta}_{\parallel} + \hat{\theta}_{\perp}$, where $\hat{\theta}_{\parallel}$ is contained in the space spanned by the columns of the data matrix $X$, and $\hat{\theta}_{\perp}$ is contained in the orthogonal complement of this space in $\mathbb{R}^n$, we have that

1. $\hat{\theta}_{\parallel} = X\hat{a}$ for some $\hat{a}\in\mathbb{R}^m$,
2. $X^\top\hat{\theta} = X^\top\hat{\theta}_{\parallel} + X^\top\hat{\theta}_{\perp} = X^\top\hat{\theta}_{\parallel}$, and
3. $\|\hat{\theta}\|^2 = \|\hat{\theta}_{\parallel} + \hat{\theta}_{\perp}\|^2 = \|\hat{\theta}_{\parallel}\|^2 + \|\hat{\theta}_{\perp}\|^2$.

Assuming $\hat{\theta}_{\perp} \neq 0$ leads to

$$\begin{aligned} L(X^\top\hat{\theta}, y) + c\|\hat{\theta}\|^2 =&\ L(X^\top\hat{\theta}_{\parallel}, y) + c\|\hat{\theta}_{\perp}\|^2 + c\|\hat{\theta}_{\parallel}\|^2 \\ >&\ L(X^\top\hat{\theta}_{\parallel}, y) + c\|\hat{\theta}_{\parallel}\|^2, \end{aligned}$$

a contradiction to the optimality of $\hat{\theta}$. □

The representer theorem holds in the case $c = 0$, i.e., without the Euclidean regularization term. While Euclidean regularized learning problems become strongly convex for convex loss functions if $c > 0$, this is no longer the case for $c = 0$. In the latter case the minimum does not need to be unique. Still, it holds that the problem has at least one solution of the form $X\hat{a}$, i.e., it is a linear combination of the data points $x^{(i)}$, $i\in[m]$.

An immediate consequence of the representer theorem is that we can optimize over $a \in \mathbb{R}^m$ instead of $\theta \in \mathbb{R}^n$, namely by substituting $\theta = Xa$ in the original optimization problem, which results in an equivalent *adjoint* formulation

$$\hat{a} \;=\; \text{argmin}_{a\in\mathbb{R}^m}\; L(X^\top Xa, y) + c \cdot a^\top X^\top Xa$$

of the Euclidean regularized learning problem. The adjoint problem formulation allows the so called *kernel trick* that comes with significant (computational) benefits

## Positive definite kernels

A statistically and algorithmically efficient, but not as flexible alternative to feature maps are *positive definite kernels*.

A *kernel* on a set $\mathcal{X}$ is just a bivariate function

$$k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}.$$

It is called *positive definite*, if the *kernel matrix*

$$\big(k(x^{(i)}, x^{(j)})\big)_{i,j=1,\ldots,m}$$

is symmetric and positive semi-definite for all $n \in \mathbb{N}$ and $x^{(1)}, \ldots, x^{(m)} \in \mathcal{X}$. It is called *strictly positive definite*, if the kernel matrix is positive definite for all distinct $x^{(1)}, \ldots, x^{(m)} \in \mathcal{X}$.

The key observation now is that the adjoint formulation of Euclidean regularized learning problems like ridge regression or regularized logistic regression depends on the data matrix $X$, whose columns are the data points $x^{(i)}$, $i \in [m]$, only through the inner products of the data points. These inner products are stored in the symmetric, positive semi-definite *Gram matrix* $X^\top X$. The so called *kernel trick* just means replacing the Gram matrix $X^\top X$ by the kernel matrix

$$K \;=\; \Big(k(x^{(i)}, x^{(j)})\Big)_{i,j\in[m]}$$

for some positive definite kernel function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. After this replacement, the adjoint problem reads as

$$\hat{a} \;=\; \text{argmin}_{a\in\mathbb{R}^m}\; L(Ka, y) + c \cdot a^\top Ka.$$

*Remarks:* The adjoint optimization problem remains strictly convex as long as the kernel matrix $K$ is positive definite. In this case the number of parameters that need to be estimated in the adjoint problem is $m$, i.e., the same as the number of data points. Learning problems where the number of parameters can grow with the number of data points are called *nonparametric learning problems*, in contrast to *parametric learning problems*, where the number of parameters is fixed. Note though that here the number of (free) parameters can only grow to the maximum rank of $K$. In the next chapter we will see that there are kernels, where the rank of $K$ can grow with indefinitely with the number of data points, while for others there is an upper bound. The difference is in the dimension of the feature space that is associated with the kernel. The rank can grow indefinitely only if the associated feature space is infinite dimensional.

## Prediction

Remember that we get the optimal parameter vector $\hat{\theta} \in \mathbb{R}^n$ of an Euclidean regularized learning problem from the optimal solution $\hat{a} \in \mathbb{R}^m$ of its adjoint formulation as $\hat{\theta} = X\hat{a}$.

The predictor (regressor) in ridge regression can be expressed using $\hat{a}$ as

$$h : \mathbb{R}^n \to \mathbb{R},\; x \mapsto x^\top \hat{\theta} \;=\; x^\top X \hat{a} \;=\; \sum_{i=1}^{m} \hat{a}_i \, x^\top x^{(i)}.$$

When we are using a kernel matrix $K$ instead of a Gram matrix in the adjoint problem formulation, then we do not have an original problem anymore. Hence, there is also no optimal solution $\hat{\theta}$, though we still have an optimal solution $\hat{a}$ of the adjoint problem. But as we have seen above, $\hat{a}$ can still be used to define the regressor

$$h : \mathcal{X} \to \mathbb{R},\; x \mapsto \sum_{i=1}^{m} \hat{a}_i \, k(x, x^{(i)}),$$

when we are using the kernel function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. Again, we have just replaced inner products by evaluations of the kernel function.

Similarly, the predictor (classifier) in regularized logistic regression can be expressed using $\hat{a}$ as

$$h : \mathcal{X} \to \{\pm 1\},\; x \mapsto \begin{cases} 1 & : \quad X^\top X \hat{a} \;=\; \sum_{i=1}^{m} \hat{a}_i \, x^\top x^{(i)} \;\geq\; 0 \\ -1 & : \quad \text{otherwise} \end{cases}$$

and thus as

$$h : \mathcal{X} \to \{\pm 1\},\ x \mapsto \begin{cases} 1 & : \ \sum_{i=1}^{m} \hat{a}_i \, k(x, x^{(i)}) \ \geq \ 0 \\ -1 & : \ \text{otherwise} \end{cases}$$

when using the kernel function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$.

# Chapter 8

# Kernels and feature maps

Kernels and feature maps are closely related in the sense that every feature map defines a positive definite kernel and every strictly positive kernel has an associated feature map. The latter feature map does not necessarily take values in some Euclidean space, but can take values in a more general *Hilbert space*. A Hilbert space is essentially a vector space with an inner product that allows to define distances and angles just like in Euclidean geometry. See the supplemental material in Section 10 for a brief review of the basics of Euclidean geometry.

We start by showing that one always gets a positive definite kernel from a feature map.

**Lemma 1.** *Let $\mathbb{H}$ be a Hilbert space and $\phi : \mathcal{X} \to \mathbb{H}$ a feature map. Then*

$$k : \mathcal{X} \times \mathcal{X} \to \mathbb{R},\ (x^{(1)}, x^{(2)}) \mapsto \langle \phi(x^{(1)}), \phi(x^{(2)}) \rangle$$

*is a positive definite kernel.*

*Proof.* Let $x^{(1)}, \ldots, x^{(m)} \in \mathcal{X}$ and

$$K \;=\; \big(k(x^{(i)}, x^{(j)})\big)_{i,j=1,\ldots,m} \;=\; \big(\langle \phi(x^{(i)}), \phi(x^{(j)}) \rangle\big)_{i,j=1,\ldots,m}$$

be the associated kernel matrix. We need to show that $K$ is symmetric and positive semi-definite. The symmetry follows immediately from the symmetry of the inner product on the Hilbert space $\mathbb{H}$. Positive semi-definiteness holds, if we have

$$a^\top K a \geq 0$$

for all $a \in \mathbb{R}^m$. We compute that

$$\begin{aligned} a^\top K a &= \sum_{i,j=1}^{m} a_i a_j k(x^{(i)}, x^{(j)}) \\ &= \sum_{i,j=1}^{n} a_i a_j \big\langle \phi(x^{(i)}), \phi(x^{(j)}) \big\rangle \\ &= \Big\langle \sum_{i=1}^{m} a_i \phi(x^{(i)}), \sum_{j=1}^{m} a_j \phi(x^{(j)}) \Big\rangle \\ &= \Big\langle \sum_{i=1}^{m} a_i \phi(x^{(i)}), \sum_{i=1}^{m} a_i \phi(x^{(i)}) \Big\rangle \\ &= \Big\| \sum_{i=1}^{m} a_i \phi(x^{(i)}) \Big\|^2 \;\geq\; 0. \end{aligned}$$

□

Given a kernel $k$, we call a feature map $\phi$ *associated* to $k$, if

$$k(x^{(1)}, x^{(2)}) \;=\; \big\langle \phi(x^{(1)}), \phi(x^{(2)}) \big\rangle.$$

for all $x^{(1)}, x^{(2)} \in \mathcal{X}$. The remarkable fact, that we formalize in the following theorem, is that every strictly positive definite kernel has an associated feature map.

**Theorem 3. [Moore-Aronszajn]** *Let $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a strictly positive definite kernel, then $k$ has an associated feature map $\phi$.*

*Proof.* For a proof we construct a Hilbert space $\mathbb{H}$ and a feature map $\phi : \mathcal{X} \to \mathbb{H}$ associated to $k$. Denote the space of all functions from $\mathcal{X}$ to $\mathbb{R}$ by $\mathbb{R}^{\mathcal{X}}$. We can turn $\mathbb{R}^{\mathcal{X}}$ into a vector space by defining the addition and scalar multiplication pointwise, i.e.,

$$\begin{aligned} (f+g)(x) &= f(x) + g(x) \\ (\alpha f)(x) &= \alpha f(x) \end{aligned}$$

for $f, g \in \mathbb{R}^{\mathcal{X}}$ and $\alpha \in \mathbb{R}$. Obviously, $\mathbb{R}^{\mathcal{X}}$ contains all functions of the form

$$\phi_x : \mathcal{X} \to \mathbb{R},\; x' \mapsto k(x, x').$$

Our inner product space $\mathbb{H} \subseteq \mathbb{R}^{\mathcal{X}}$ is then defined as the space that contains all finite linear combinations of functions $\phi_x$, $x \in \mathcal{X}$. The inner product itself is defined by

$$\langle f, g \rangle = \sum_{i=1}^{n} \sum_{j=1}^{m} \alpha_i \beta_j k(x^{(i)}, x^{(j)}),$$

for

$$f = \sum_{i=1}^{n} \alpha_i \phi_{x^{(i)}},\; g = \sum_{j=1}^{m} \beta_j \phi_{x^{(j)}} \in \mathbb{H}.$$

Note that this definition is independent of the representation of $f$ and $g$ since we have

$$\langle f, g \rangle = \sum_{i=1}^{n} \alpha_i g(x^{(i)}) = \sum_{j=1}^{m} \beta_j f(x^{(j)}),$$

where the second expression is independent of the representation of $g$ and the third expression is independent of the representation of $f$.
We need to check that the definition satisfies all three properties of inner products.

1. Symmetry: We have $k(x^{(i)}, x^{(j)}) = k(x^{(j)}, x^{(i)})$, because every kernel matrix for a positive definite kernel is symmetric. From this it follows immediately that $\langle f, g \rangle = \langle g, f \rangle$.

2. Positivity: We have

   $$\langle f, f \rangle = \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j k(x^{(i)}, x^{(j)}) = \alpha^\top K \alpha \geq 0,$$

   where $K = \big(k(x^{(i)}, x^{(j)})\big)_{i,j=1,\ldots,n}$ is positive semi-definite since $k$ is positive. We also have

   $$\langle f, f \rangle = 0 \quad \text{if and only if} \quad f = 0,$$

   because $k$ is strictly positive.

3. Bilinearity: Follows immediately from the definition.

Technically, $\mathbb{H}$ is not necessarily a Hilbert space since not every Cauchy sequence needs to converge in $\mathbb{H}$, but we can turn $\mathbb{H}$ into a Hilbert space by completing it, i.e., by adding the limit points of non-convergent Cauchy sequences. This procedure is similar to the completion of the set of rational numbers $\mathbb{Q}$ that gives us the set of real numbers $\mathbb{R}$.

Finally, the feature map can be defined as

$$\phi : \mathcal{X} \to \mathbb{H},\ x \mapsto \phi_x.$$

This feature map is associated to $k$, because

$$\langle \phi(x^{(1)}), \phi(x^{(2)}) \rangle \ =\ \langle \phi_{x^{(1)}}, \phi_{x^{(2)}} \rangle \ =\ k(x^{(1)}, x^{(2)}),$$

where the first equality follows directly from the definition of the feature map $\phi$ and the second equality follows from our definition of the inner product on $\mathbb{H}$. □

The Hilbert space that we have constructed in the proof of Theorem 3 is called a *reproducing kernel Hilbert space*.

### Examples of kernels and associated feature maps

(1) Let $\mathcal{X} = \mathbb{R}^n$ and

$$k : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R},\ (x^{(1)}, x^{(2)}) \mapsto \left( {x^{(1)}}^\top x^{(2)} \right)^2.$$

We have

$$\begin{aligned} k(x^{(1)}, x^{(2)}) &= \left( {x^{(1)}}^\top x^{(2)} \right)^2 \\ &= \left( \sum_{i=1}^n x_i^{(1)} x_i^{(2)} \right)^2 \\ &= \left( \sum_{i=1}^n x_i^{(1)} x_i^{(2)} \right) \left( \sum_{j=1}^n x_j^{(1)} x_j^{(2)} \right) \\ &= \sum_{i=1}^n \sum_{j=1}^n x_i^{(1)} x_i^{(2)} x_j^{(1)} x_j^{(2)} \\ &= \sum_{i=1}^n \sum_{j=1}^n (x_i^{(1)} x_j^{(1)})(x_i^{(2)} x_j^{(2)}). \end{aligned}$$

Thus,

$$\phi : \mathbb{R}^n \to \mathbb{R}^{n^2},\ (x_1, \ldots, x_n) \mapsto (x_1^2, x_1x_2, \ldots, x_1x_n, x_2x_1, \ldots, x_nx_1, \ldots, x_n^2)$$

is an associated feature map to $k$.

(2) Let $\mathcal{X} = \mathbb{R}^n$ and

$$k : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R},\ (x^{(1)}, x^{(2)}) \mapsto \left(x^{(1)^\top} x^{(2)} + 1\right)^2.$$

We have

$$k(x^{(1)}, x^{(2)}) \;=\; \sum_{i=1}^{n}\sum_{j=1}^{n}(x_i^{(1)}x_j^{(1)})(x_i^{(2)}x_j^{(2)}) \;+\; \sum_{i=1}^{n}(\sqrt{2}x_i^{(1)})(\sqrt{2}x_i^{(2)}) \;+\; 1.$$

Thus,

$$\begin{aligned}&\phi : \mathbb{R}^n \to \mathbb{R}^{n^2+n+1},\\ &(x_1, \dots, x_n) \mapsto \big(x_1^2, x_1x_2, \dots, x_1x_n, x_2x_1, \dots, x_n^2, \sqrt{2}x_1, \dots, \sqrt{2}x_n, 1\big)\end{aligned}$$

is a feature map for $k$.

(3) Let $\mathcal{X} = \mathbb{R}^n$, then

$$k : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R},\ (x^{(1)}, x^{(2)}) \mapsto \left(x^{(1)^\top} x^{(2)} + 1\right)^k$$

has an associated feature map

$$\phi : \mathbb{R}^n \to \mathbb{R}^{n^k + n^{k-1} + \ldots + n + 1}$$

that essentially maps $x$ to the corresponding vector of monomials of degree at most $k$, where permutations of the indices make up additional dimensions. Hence, the dimension of the reproducing kernel Hilbert space is larger than $\binom{n+k}{n}$.
Note that evaluating $\langle \phi(x^{(1)}), \phi(x^{(2)}) \rangle$ requires $\Theta(n^k)$ arithmetic operations, whereas evaluating $k(x^{(1)}, x^{(2)})$, i.e., evaluating $\left(x^{(1)^\top} x^{(2)} + 1\right)^k$, only needs $O(n + \log k)$ arithmetic operations.

(4) Let $\mathcal{X} = \mathbb{R}$, then we have for the one-dimensional *Gaussian kernel*

$$k\big(x^{(1)}, x^{(2)}\big) \;=\; \exp\left(-\frac{1}{2\sigma^2}\big(x^{(1)} - x^{(2)}\big)^2\right)$$

that

$$\begin{aligned}
\exp\left(-\frac{1}{2\sigma^2}\big(x^{(1)}-x^{(2)}\big)^2\right) &= \sum_{i=0}^{\infty}(-1)^i\,\frac{\big(x^{(1)}-x^{(2)}\big)^{2i}}{(2\sigma^2)^i i!}\\
&= \sum_{i=0}^{\infty}(-1)^i\,\frac{\Big({x^{(1)}}^2-2x^{(1)}x^{(2)}+{x^{(2)}}^2\Big)^i}{(2\sigma^2)^i i!}\\
&= \sum_{i=0}^{\infty}(-1)^i\,\frac{\Big({x^{(1)}}^2+{x^{(2)}}^2-2x^{(1)}x^{(2)}\Big)^i}{(2\sigma^2)^i i!}\\
&= \sum_{i=0}^{\infty}\frac{(-1)^i}{(2\sigma^2)^i i!}\sum_{j=0}^{i}\binom{i}{j}\Big({x^{(1)}}^2+{x^{(2)}}^2\Big)^j\Big(-2x^{(1)}x^{(2)}\Big)^{i-j}\\
&= \sum_{i=0}^{\infty}\sum_{j=0}^{i}(-1)^j\,\frac{\Big({x^{(1)}}^2+{x^{(2)}}^2\Big)^j}{(2\sigma^2)^j j!}(-1)^{i-j}\,\frac{\Big(-2x^{(1)}x^{(2)}\Big)^{i-j}}{(2\sigma^2)^{i-j}(i-j)!}\\
&= \sum_{i=0}^{\infty}\sum_{j=0}^{i}(-1)^j\,\frac{\Big({x^{(1)}}^2+{x^{(2)}}^2\Big)^j}{(2\sigma^2)^j j!}\,\frac{\Big(2x^{(1)}x^{(2)}\Big)^{i-j}}{(2\sigma^2)^{i-j}(i-j)!}\\
&= \exp\left(-\frac{{x^{(1)}}^2+{x^{(2)}}^2}{2\sigma^2}\right)\exp\left(\frac{2x^{(1)}x^{(2)}}{2\sigma^2}\right)\\
&= \exp\left(-\frac{{x^{(1)}}^2+{x^{(2)}}^2}{2\sigma^2}\right)\sum_{i=0}^{\infty}\frac{\big(2x^{(1)}x^{(2)}\big)^i}{(2\sigma^2)^i i!}\\
&= \exp\left(-\frac{{x^{(1)}}^2}{2\sigma^2}\right)\exp\left(-\frac{{x^{(2)}}^2}{2\sigma^2}\right)\sum_{i=0}^{\infty}\frac{\big(x^{(1)}x^{(2)}\big)^i}{\sigma^{2i} i!}\\
&= \sum_{i=0}^{\infty}\exp\left(-\frac{{x^{(1)}}^2}{2\sigma^2}\right)\exp\left(-\frac{{x^{(2)}}^2}{2\sigma^2}\right)\frac{\big(x^{(1)}x^{(2)}\big)^i}{\sigma^{2i} i!}\\
&= \sum_{i=0}^{\infty}\exp\left(-\frac{{x^{(1)}}^2}{2\sigma^2}\right)\frac{{x^{(1)}}^i}{\sigma^i\sqrt{i!}}\exp\left(-\frac{{x^{(2)}}^2}{2\sigma^2}\right)\frac{{x^{(2)}}^i}{\sigma^i\sqrt{i!}}.
\end{aligned}$$

Thus,

$$\phi:\mathbb{R}\to\ell_2,\ x\mapsto\left(\exp\left(-\frac{x^2}{2\sigma^2}\right)\frac{x^i}{\sigma^i\sqrt{i!}}\right)_{i\in\mathbb{N}}$$

is a feature map for $k$. Here, $\ell_2$ is the infinite dimensional Hilbert space of square-summable sequences.

*Remark:* The Gaussian kernel has the following straightforward generalization to $n$ dimensions

$$k\big(x^{(1)}, x^{(2)}\big) \;=\; \exp\left(-\frac{1}{2\sigma^2}\big\|x^{(1)} - x^{(2)}\big\|^2\right),$$

where $x^{(1)}, x^{(2)} \in \mathcal{X} = \mathbb{R}^n$. We leave it as an exercise to show that

$$\phi : \mathbb{R} \to \ell_2,\; x \mapsto \left(\exp\left(-\frac{\|x\|^2}{2\sigma^2}\right)\frac{x^{i_1}\cdot \ldots \cdot x^{i_n}}{\sigma^i\sqrt{i_1! \cdot \ldots \cdot i_n!}}\right)_{(i_1,\ldots,i_n)\in N_i,\, i\in\mathbb{N}}$$

is an associated feature map to $k$. Here,

$$N_i \;=\; \left\{(i_1, \ldots, i_n) \in \mathbb{N}^n \;:\; \sum_{i=1}^{n} i_i = i\right\}.$$

# Chapter 9

# Exercises

## Chapter 7

**Exercise 1.** What is the difference between the sigmoid function and the hyperbolic tangent.

## Chapter 10

**Exercise 1.** Let $k_1$ and $k_2$ be positive definite kernels on $\mathcal{X}$, and let $\alpha > 0$. Show that $k_1 + k_2$ and $\alpha k_1$ are also positive definite kernels.

**Exercise 2.** Let $S$ be a finite set and $2^S$ the power set (set of all subsets) of $S$. Show that the so called set kernel

$$k : P(S) \times P(S) \to \mathbb{R}, (A, B) \mapsto |A \cap B|$$

is indeed a positive definite kernel on $P(S)$. Also, discuss how to evaluate the set kernel efficiently.

# Chapter 10

# Supplemental material

### Brief review of Euclidean geometry

The Euclidean inner product has the three properties of inner products:

1. Symmetry: $\langle x^{(1)}, x^{(2)} \rangle = \langle x^{(2)}, x^{(1)} \rangle$
2. Positivity: $\langle x, x \rangle \geq 0$ and $\langle x, x \rangle = 0$, if and only if $x = 0$
3. Bilinearity: $\langle a_1 x^{(1)} + a_2 x^{(2)}, x^{(3)} \rangle \; = \; a_1 \langle x^{(1)}, x^{(3)} \rangle + a_2 \langle x^{(2)}, x^{(3)} \rangle$

The inner product can be used for defining a norm on $\mathbb{R}^n$ which then in turn can be used for defining a metric on $\mathbb{R}^n$. The norm of $x \in \mathbb{R}^n$ is simply defined as

$$\|x\| \; = \; \sqrt{\langle x, x \rangle},$$

which we have used before. The norm satisfies

1. Positivity: $\|x\| \geq 0$ and $\|x\| = 0$ if and only if $x = 0$
2. Homogeneity: $\|ax\| = |a| \cdot \|x\|$
3. Triangle inequality: $\|x^{(1)} + x^{(2)}\| \leq \|x^{(1)}\| + \|x^{(1)}\|$

The triangle inequality follows from the Cauchy-Schwarz inequality.

**Lemma 2. [Cauchy-Schwarz]** *For any $x^{(1)}, x^{(2)} \in \mathbb{R}^n$ it holds that*

$$\left|\langle x^{(1)}, x^{(2)} \rangle\right| \; \leq \; \|x^{(1)}\| \|x^{(2)}\|.$$

*Proof.* Note that the inequality applies, if $x^{(2)} = 0$. Hence, we assume now that $x^{(2)} \neq 0$. We have for any $a \in \mathbb{R}$ that

$$0 \le \langle x^{(1)} - ax^{(2)}, x^{(1)} - ax^{(2)} \rangle = \|x^{(1)}\|^2 - 2a\langle x^{(1)}, x^{(2)} \rangle + a^2\|x^{(2)}\|^2.$$

Setting $a = \frac{\langle x^{(1)}, x^{(2)} \rangle}{\|x^{(2)}\|^2}$ gives

$$0 \le \|x^{(1)}\|^2 - \frac{\langle x^{(1)}, x^{(2)} \rangle^2}{\|x^{(2)}\|^2},$$

or equivalently

$$\langle x^{(1)}, x^{(2)} \rangle^2 \le \|x^{(1)}\|^2 \|x^{(2)}\|^2,$$

from which the Cauchy-Schwarz inequality follows by taking the square root of both sides. □

The triangle inequality now follows from

$$\begin{aligned} \|x^{(1)} + x^{(2)}\|^2 &= \langle x^{(1)} + x^{(2)}, x^{(1)} + x^{(2)} \rangle \\ &= \|x^{(1)}\|^2 + 2\langle x^{(1)}, x^{(2)} \rangle + \|x^{(2)}\|^2 \\ &\le \|x^{(1)}\|^2 + 2\|x^{(1)}\|\|x^{(2)}\| + \|x^{(2)}\|^2 \\ &= \big(\|x^{(1)}\| + \|x^{(2)}\|\big)^2. \end{aligned}$$

The *angle* between $x^{(1)}$ and $x^{(2)}$, or more precisely the cosine of the angle between the two vectors, is defined as

$$\cos\alpha = \frac{\langle x^{(1)}, x^{(2)} \rangle}{\|x^{(1)}\|\|x^{(2)}\|}$$

That is, we now know how to measure angles. It remains to define how to measure distances. The norm on $\mathbb{R}^n$ induces a metric on $\mathbb{R}^n$ as follows,

$$d(x^{(1)}, x^{(2)}) = \|x^{(1)} - x^{(2)}\|.$$

The metric satisfies

1. Symmetry: $d(x^{(1)}, x^{(2)}) = d(x^{(2)}, x^{(1)})$
2. Positivity: $d(x^{(1)}, x^{(2)}) \ge 0$ and $d(x^{(1)}, x^{(2)}) = 0$ if and only if $x^{(1)} = x^{(2)}$
3. Triangle inequality: $d(x^{(1)}, x^{(2)}) \le d(x^{(1)}, x^{(3)}) + d(x^{(3)}, x^{(2)})$

Angles and distances are related through the cosine theorem.

**Theorem 4. [Cosine Theorem]** *For any $x^{(1)}, x^{(2)}, x^{(3)} \in \mathbb{R}^n$ and the angle $\alpha$ at $x^{(1)}$, i.e., between $x^{(2)} - x^{(1)}$ and $x^{(3)} - x^{(1)}$, it holds that*

$$d(x^{(2)}, x^{(3)})^2 = d(x^{(1)}, x^{(2)})^2 + d(x^{(1)}, x^{(3)})^2 - 2d(x^{(1)}, x^{(2)})d(x^{(1)}, x^{(3)}) \cos\alpha.$$

*Proof.* We have

$$\begin{aligned} d(x^{(2)}, x^{(3)})^2 &= \|x^{(2)} - x^{(3)}\|^2 \\ &= \|x^{(2)} - x^{(1)} + x^{(1)} - x^{(3)}\|^2 \\ &= \langle x^{(2)} - x^{(1)} + x^{(1)} - x^{(3)}, x^{(2)} - x^{(1)} + x^{(1)} - x^{(3)} \rangle \\ &= \|x^{(2)} - x^{(1)}\|^2 + \|x^{(1)} - x^{(3)}\|^2 + 2\langle x^{(2)} - x^{(1)}, x^{(1)} - x^{(3)} \rangle \\ &= d(x^{(1)}, x^{(2)})^2 + d(x^{(1)}, x^{(3)})^2 - 2\langle x^{(2)} - x^{(1)}, x^{(3)} - x^{(1)} \rangle \\ &= d(x^{(1)}, x^{(2)})^2 + d(x^{(1)}, x^{(3)})^2 - 2d(x^{(1)}, x^{(2)})d(x^{(1)}, x^{(3)}) \cos\alpha. \end{aligned}$$

□

# Part III

# Probably Approximately Correct (PAC) Learning

# Chapter 11

# The PAC learning framework

The formal PAC learning model was introduced by Leslie Valiant in 1984 as a framework for the mathematical analysis of machine learning. We already know the ingredients of the framework: a data generation model, a hypothesis class like the class of linear functions, and a loss function on the hypothesis class.

### Data generation models and loss functions

We allow general generative models, i.e., probability density functions $p$ on $\mathcal{X} \times \mathcal{Y}$, and loss functions

$$L : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$$

for evaluating predictors $h : \mathcal{X} \to \mathcal{Y}$.

*Remark:* Examples are the 0-1-loss function

$$L(y^{(1)}, y^{(2)}) \;=\; \mathbf{1}[y^{(1)} \neq y^{(2)}]$$

for classification problems, where $\mathcal{Y}$ is a finite set, and the square loss function

$$L(y^{(1)}, y^{(2)}) \;=\; \big(y^{(1)} - y^{(2)}\big)^2$$

for regression problems with $\mathcal{Y} = \mathbb{R}$ that we have seen already when we discussed ordinary least squares regression.

The *expected loss*, sometimes also called the *expected risk*, of a predictor $h$ is defined as

$$L_p(h) \;=\; \int_{\mathcal{X}\times\mathcal{Y}} L(h(x), y) p(x, y)\, dxdy.$$

Obviously, we would like to choose a predictor with small expected loss, but unfortunately we cannot compute the expected loss of a predictor since we do not know the probability density function $p$. Our access to $p$ is only indirectly through sample points. Given a sequence $S$ of data points $\big((x^{(i)}, y^{(i)})\big)_{i\in[m]}$ that have been sampled from $p$ we can compute

$$L_S(h) \;=\; \frac{1}{m}\sum_{i=1}^{m} L(h(x^{(i)}), y^{(i)}),$$

which is called the *empirical loss* of the predictor $h$. A natural idea is using the empirical loss $L_S(h)$ as a proxy for the expected loss $L_p(h)$ and thus choosing a predictor with small empirical loss. But this is not always a good idea, because of *overfitting*.

### The problem of overfitting

The following, simple example shows that the naive, straightforward implementation of the idea of choosing a predictor with small empirical loss can lead to bad results. Let $\mathcal{X} = [0,1]$, i.e., the unit interval, let $p$ the uniform distribution on $[0,1]$, i.e., the probability to observe $x$ in the interval $[a,b] \subseteq [0,1]$ is $b-a$, and let

$$f : [0,1] \to \{0,1\},\; x \mapsto \mathbf{1}\left[x \leq \frac{1}{2}\right]$$

be a labeling function. That is, here we assume that the explained variables in $\mathcal{Y}$ are determined by the explanatory variables, i.e., the only randomness is in accessing the explanatory variables in $\mathcal{X}$. Given a sequence $S$ data points $\big((x^{(i)}, y^{(i)})\big)_{i\in[m]}$, we can define the predictor

$$h : [0,1] \to \{0,1\},\; x \mapsto \sum_{i=1}^{m} y^{(i)}\mathbf{1}\big[x = x^{(i)}\big],$$

after we have eliminated any redundancies in the sample, i.e., multiples of data points. The empirical loss of this predictor is $0$ while its expected loss is $\frac{1}{2}$. That is, predictors with small empirical loss do not need to generalize well beyond the observed data. Hence, such a predictor is actually not very good at predicting the labels of new data points. This phenomenon is known as *overfitting*, and the big question is how to avoid overfitting.

## Restricted hypothesis classes

The key idea for mitigating the problem of overfitting is restricting the class of allowed predictors before we have observed any data. So far we have allowed all functions

$$\mathcal{Y}^{\mathcal{X}} = \{h : \mathcal{X} \to \mathcal{Y}\}$$

as possible predictors. Restricting the class of allowed predictors to a subset $H$ of $\mathcal{Y}^{\mathcal{X}}$ before we have seen any data points can mitigate the problem of overfitting. For instance, assume at the moment that $H \subseteq \{0,1\}^{[0,1]}$ is a *finite hypothesis class* of possible predictors. Since we need to fix $H$ a priori, i.e., before we have seen any data points, it is unlikely that, given a sequence of data points $S$, $H$ will contain the overfitting predictor from above, because in order to make sure that $H$ contains this predictor, $H$ should contain all indicator functions of finite subsets of $[0,1]$ and thus $H$ cannot be a finite set. In the next chapter we show that restricting $H$ to finite sets actually prevents overfitting. Before we can do so we need to formalize what we mean by overfitting. This is done in the defintion of PAC learnability.

## PAC learnability

We call a hypothesis class

$$H \subseteq \mathcal{Y}^{\mathcal{X}} = \{h : \mathcal{X} \to \mathcal{Y}\},$$

*PAC learnable*, if there exists a function

$$m : \mathbb{R}_+ \times (0,1) \to \mathbb{R}_{\geq 0},\ (\varepsilon, \delta) \mapsto m(\varepsilon, \delta)$$

and a learning algorithm that on input of $m \geq m(\varepsilon, \delta)$ i.i.d. data points returns a predictor $\hat{h} \in H$ that satisfies

$$\mathsf{P}\left[\left\{S \in (\mathcal{X} \times \mathcal{Y})^m \ : \ L_p(\hat{h}) \leq \mathsf{min}_{h \in H} L_p(h) + \varepsilon\right\}\right] \geq 1 - \delta$$

for any probability density function $p$ on $\mathcal{X} \times \mathcal{Y}$, which also provides us with a product measure on $\big(\mathcal{X} \times \mathcal{Y}\big)^m$, and for any choice of $\varepsilon > 0, \delta \in (0,1)$. The function $m$ is called the *sample complexity* of the hypothesis class and algorithm. We call any algorithm for choosing a predictor from sample points *probably approximately correct*, if it returns a predictor $\hat{h}$ that satisfies the inequality in definition of PAC learnability when run on a random sample of size at least $m(\varepsilon, \delta)$. Approximately, because the algorithm gives only a predictor with *accuracy* $\varepsilon$ compared to a best possible predictor from the class $H$, and this accuracy even only holds with probability $1-\delta$. The probability $1-\delta$ is also called our *confidence* in the learning algorithm.

*Remark:* Our definition of PAC learnability is sometimes called *agnostic PAC learnability*. The original definition of PAC learnability assumes that

$$\text{min}_{h\in H} L_p(h) = 0,$$

which is known as the *realizability assumption*. We still get with probability at least $1-\delta$ that $L_p(\hat{h}) \leq \varepsilon$, if the realizability condition holds. Hence the name agnostic PAC learnable.

# Chapter 12

# Uniformity and the ERM rule

In the this chapter we show that finite hypothesis classes are PAC learnable, more specifically we show using the *uniform convergence property* that the so called *empirical risk minimization (ERM) rule* achieves PAC learnability for finite hypothesis classes.

## Empirical risk minimization (ERM) rule

One specific learning approach is the *empirical risk minimization (ERM)* rule. The straightforward idea of the ERM rule is as follows: If $H$ is a hypothesis class, and $S$ is a sequence of data points, then let

$$\hat{h} = \text{argmin}_{h \in H} L_S(h)$$

be a predictor that minimizes the empirical risk. If $L_S$ does not have a unique minimizer in $H$, then just pick any of the minimizers.

*Remark:* Technically, the ERM rule is not an algorithm. We need an algorithm for computing a predictor with minimal empirical risk. It turns out that an efficient algorithm for minimizing the empirical risk does not always exist. In computational learning theory the existence of an efficient algorithm is required for PAC learnability. Here, we neglect computational issues.

## Uniform convergence

Since the outcome $\hat{h}$ of a learning algorithm depends on its input, a random sample $S$, it is a random variable itself. Hence, *non-representative samples* can lead the learning algorithm astray, resulting in a predictor whose expected loss is large. Let us make the notion of *non-representative samples* more precise. Given

$\varepsilon > 0$, we call a sequence $S$ of sample points $\varepsilon$-representative for a hypothesis class $H$ and loss function $L$, if it holds true that

$$\forall h \in H \;:\; |L_S(h) - L_p(h)| \leq \varepsilon.$$

Directly from this definition we get the following lemma.

**Lemma 3.** *Let $S$ be an $\frac{\varepsilon}{2}$-representative sequence of sample points for the hypothesis class $H$ and loss function $L$, and let*

$$\hat{h} = \text{argmin}_{h \in H} \; L_S(h).$$

*Then,*

$$L_p(\hat{h}) \leq \textit{min}_{h \in H} \; L_p(h) + \varepsilon.$$

*Proof.* We have for any $h \in H$ (and thus also for a minimizer $\hat{h}$ of the empirical loss) that

$$L_p(\hat{h}) \leq L_S(\hat{h}) + \frac{\varepsilon}{2} \leq L_S(h) + \frac{\varepsilon}{2} \leq L_p(h) + \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = L_p(h) + \varepsilon,$$

where we have used in the first and third inequality that $S$ is $\frac{\varepsilon}{2}$-representative. The second inequality follows from the definition of $\hat{h}$ as a minimizer of the empirical loss. □

Given a loss function $L$, we say that a hypothesis class $H$ has the *uniform convergence property*, if there exists a function

$$m : \mathbb{R}_{+} \times (0,1) \to \mathbb{R}_{\geq 0}, \; (\varepsilon, \delta) \mapsto m(\varepsilon, \delta)$$

such that any sample of size $m \geq m(\varepsilon, \delta)$ drawn i.i.d. from $p$ is $\varepsilon$-representative for $H$ and $L$ with probability at least $1 - \delta$ for any probability density function $p$ on $\mathcal{X} \times \mathcal{Y}$ and any choice of $\varepsilon > 0$, $\delta \in (0,1)$.

**Theorem 5.** *Let $H$ be a hypothesis class that has the uniform convergence property with function $m'(\varepsilon, \delta)$. Then $H$ is PAC learnable with sample complexity $m(\varepsilon, \delta) = m'(\varepsilon/2, \delta)$.*

*Proof.* It follows immediately from Lemma 3 that the ERM rule is a PAC learner for $H$. □

### Finite hypothesis classes are PAC learnable

We want to use Theorem 5 for proving that finite hypothesis classes are PAC learnable by the ERM rule.

**Theorem 6.** *Given $\mathcal{X}$, $\mathcal{Y}$, a loss function*

$$L : \mathcal{Y} \times \mathcal{Y} \to [0,1]$$

*and a finite hypothesis class $H \subseteq \mathcal{Y}^{\mathcal{X}}$. Then $H$ is PAC learnable through the ERM rule.*

*Proof.* By Theorem 5 it suffices to show that $H$ has the uniform convergence property. That is, for a given accuracy $\varepsilon \in (0,1)$ and $\delta \in (0,1)$ we need to find a function

$$m : (0,1) \times (0,1) \to \mathbb{R}_{\geq 0},\ (\varepsilon, \delta) \mapsto m(\varepsilon, \delta)$$

such that

$$\mathsf{P}\big[\big\{S \in (\mathcal{X} \times \mathcal{Y})^m \ : \ \forall h \in H \ : \ |L_S(h) - L_p(h)| \leq \varepsilon\big\}\big] \ \geq \ 1 - \delta$$

for $m \geq m(\varepsilon, \delta)$ and any probability density $p$ on $\mathcal{X} \times \mathcal{Y}$. Alternatively, we can make sure that

$$\mathsf{P}\big[\big\{S \in (\mathcal{X} \times \mathcal{Y})^m \ : \ \exists h \in H \ : \ |L_S(h) - L_p(h)| > \varepsilon\big\}\big] \ < \ \delta$$

for $m \geq m(\varepsilon, \delta)$. Using a union bound and Hoeffding's inequality (see the supplemental material) we get that

$$\begin{aligned}
&\mathsf{P}\big[\big\{S \in (\mathcal{X} \times \mathcal{Y})^m \ : \ \exists h \in H \ : \ |L_S(h) - L_p(h)| > \varepsilon\big\}\big] \\
&= \mathsf{P}\left[\bigcup_{h \in H} \{S \in (\mathcal{X} \times \mathcal{Y})^m \ : \ |L_S(h) - L_p(h)| > \varepsilon\}\right] \\
&\leq \sum_{h \in H} \mathsf{P}\big[\{S \in (\mathcal{X} \times \mathcal{Y})^m \ : \ |L_S(h) - L_p(h)| > \varepsilon\}\big] \\
&= \sum_{h \in H} \mathsf{P}\left[\left\{S = \Big(\big(x^{(i)}, y^{(i)}\big)\Big)_{i \in [m]} \ : \ \left|\frac{1}{m}\sum_{i=1}^{m} L(h(x^{(i)}), y^{(i)}) - L_p(h)\right| > \varepsilon\right\}\right] \\
&\leq \sum_{h \in H} 2\exp\big(-2m\varepsilon^2\big) \ = \ 2|H|\exp\big(-2m\varepsilon^2\big).
\end{aligned}$$

Hence, we have

$$\mathsf{P}\big[\big\{S \in (\mathcal{X} \times \mathcal{Y})^m \ : \ \forall h \in H \ : \ |L_S(h) - L_p(h)| \leq \varepsilon\big\}\big] \ \geq \ 1 - \delta$$

if we choose

$$m > \frac{\log\left(2|H|/\delta\right)}{2\varepsilon^2}.$$

It follows also from Theorem 5 that the sample complexity of learning finite hypothesis spaces is

$$m > \frac{2\log\left(2|H|/\delta\right)}{\varepsilon^2}.$$

□

## Approximation-estimation trade-off

Let $S$ be an i.i.d. sample from $p$ and let $\hat{h} = \text{argmin}_{h\in H} L_S(h)$ be the ERM-predictor. We can decompose the expected loss of $\hat{h}$ as follows

$$L_p(\hat{h}) = \text{argmin}_{h\in H} L_p(h) + \varepsilon_{est} =: \varepsilon_{app} + \varepsilon_{est},$$

where $\varepsilon_{est}$ is called the *estimation error* and $\varepsilon_{app} = \text{argmin}_{h\in H} L_p(h)$ is called the *approximation error* or *inductive bias* since it depends on the a priori choice of the hypothesis class $H$. In order to get a good predictor, both the estimation and the approximation error must be small. But minimizing the estimation and the approximation error are somewhat conflicting goals. For instance, in the case of finite hypothesis classes, we can reduce the approximation error (or inductive bias) by increasing the hypothesis class $H$, but from Theorem 6 we get that

$$\varepsilon_{est} > \sqrt{\frac{2\log\left(|H|/\delta\right)}{m}},$$

i.e., the estimation error grows with the size of the hypothesis class, or more specifically $\varepsilon_{est} \in \Omega\left(\sqrt{\log\left(|H|\right)}\right)$. That is, we are facing a trade-off decision when choosing a hypothesis class: small classes exhibit a large approximation error (inductive bias), while large classes entail large estimation errors.

In the next chapter we show that there are also infinite hypothesis classes that are PAC learnable, for instance the class of linear functions that we have used extensively before are PAC learnable. We show that it is not the finiteness of the hypothesis class that characterizes PAC learnability, but finiteness of a complexity measure called *VC-dimension*. Hence, the trade-off we are facing when choosing a hypothesis class is between the inductive bias and the complexity of the hypothesis class, because the estimation error grows with increasing complexity of the hypothesis class while the approximation error decreases. This trade-off decision is called the *bias-complexity trade-off*. The trade-off highlights the importance of using prior knowledge, if available, for choosing a hypothesis class.

# Chapter 13

# Vapnik-Chervonenkis (VC) theory

So far our measure of complexity for hypothesis classes was just the size of the classes. The VC-dimension, that we introduce here, is a more subtle concept of complexity that allows to characterize PAC learnable hypothesis classes for binary classification problems, i.e., subsets of $\mathcal{Y}^{\mathcal{X}}$, where $\mathcal{Y} = \{0,1\}$.
The name VC-dimension honors the Russian mathematicians Vladimir Vapnik and Alexey Chervonenkis who discovered the importance of this parameter for the theory of statistical learning.

### Range spaces and VC-dimension

A *range space* is a pair $\big(\mathcal{X}, R\big)$, where $R$ is a subset of the power set $2^{\mathcal{X}}$ of $\mathcal{X}$, i.e., $R$ is a set of subsets of $\mathcal{X}$. The sets in $R$ are called *ranges*. Any hypothesis class $H \subseteq \{0,1\}^{\mathcal{X}}$ has a naturally associated range space

$$R = \big\{\{x \in \mathcal{X} \,:\, h(x) = 1\} \,:\, h \in H\big\},$$

from which we can recover the hypothesis class $H$ as follows

$$\big\{h : \mathcal{X} \to \{0,1\},\, x \mapsto \mathbf{1}[x \in A] \,:\, A \in R\big\} \,=\, H.$$

Given a range space $\big(\mathcal{X}, R\big)$ and $S \subseteq \mathcal{X}$, we define the *projection* of $R$ onto $S$ as

$$R|S \,=\, \{A \cap S \,:\, A \in R\}.$$

We say that $S \subseteq \mathcal{X}$ is *shattered* by $R$, if the projection of $R$ onto $S$ is the power set $2^S$ of $S$. The *VC-dimension* of the range space, denoted as $VC(R)$, is the cardinality of a largest subset of $\mathcal{X}$ that is shattered by $R$. Note that this cardinality can be infinite. In this case we say that the VC-dimension is infinite and write $VC(R) = \infty$.

*Remarks:* If $R = \emptyset$, then no subset of $\mathcal{X}$ is shattered, not even the empty set, and we define $VC(R) = -1$ in this case. Note that if $R = \emptyset$, then $R|S = \emptyset$ for any $S \subseteq \mathcal{X}$. For shattering the empty set we need $R|(S = \emptyset) = \{\emptyset\}$.
The VC-dimension $VC(H)$ of a hypothesis class $H \subseteq \{0,1\}^{\mathcal{X}}$ is just the VC-dimension of the associated range space.

## Sauer's Lemma

The main theorem of PAC learning that we prove in Chapter 14 generalizes Theorem 6 that establishes PAC learnability of finite hypothesis classes to hypothesis classes for binary classification problems of finite VC-dimension. Key to the proof of the generalization is Sauer's Lemma that we state in terms of range spaces instead of hypothesis classes.

**Lemma 4. [Sauer]** *Let $(\mathcal{X}, R)$ be a range space of VC-dimension at most $d < \infty$, then*

$$\big|R|S\big| \;\leq\; \sum_{i=0}^{d} \binom{m}{i} \;:=\; \Phi_d(m)$$

*for all $S \subseteq \mathcal{X}$ with $|S| = m$.*

*Proof.* We first observe that

$$\Phi_d(m) \;=\; \begin{cases} 0 & : \; d = -1 \\ 1 & : \; m = 0 \wedge d \geq 0 \\ \Phi_d(m-1) + \Phi_{d-1}(m-1) & : \; \text{otherwise.} \end{cases}$$

The observation follows immediately from properties of binomial coefficients, namely, for $d, m \geq 1$,

$$\begin{aligned} \sum_{i=0}^{d} \binom{m}{i} &= \binom{m}{0} + \sum_{i=1}^{d} \left[ \binom{m-1}{i} + \binom{m-1}{i-1} \right] \\ &= \binom{m-1}{0} + \sum_{i=1}^{d} \binom{m-1}{i} + \sum_{i=1}^{d} \binom{m-1}{i-1} \\ &= \sum_{i=0}^{d} \binom{m-1}{i} + \sum_{i=0}^{d} \binom{m-1}{i}. \end{aligned}$$

For the border cases we use of the conventions $0! = 1$ and $\binom{m}{i} = 0$ for $i > m$. Second, for $d = -1$, i.e., $R = \emptyset$, the assertion is true, because we have $R|S = \emptyset$ and thus $\big|R|S\big| = 0 = \Phi_{-1}(m)$. Hence, in the following we can assume that $R \neq \emptyset$. We now prove that $\big|R|S\big| \leq \Phi_d(m)$ by induction on $m$. The proof makes use of a third observation, namely, that the VC-dimension of the range space $\big(S, R|S\big)$, where $S \subseteq \mathcal{X}$ and $|S| = m$, is also at most $d$. Otherwise, we have a subset $S' \subseteq S \subseteq \mathcal{X}$ with $|S'| > d$ that is shattered by $R|S$ in $S$ and thus also shattered by $R$ in $\mathcal{X}$. A contradiction to $VC(R) \leq d$.

Base case: If $m = 0$, then $S = \emptyset$ and thus $R|S = \{\emptyset\}$. Hence, we have that

$$\big|R|S\big| \;=\; 1 \;=\; \Phi_d(0) \;=\; \Phi_d(m).$$

Inductive step: Assume that the assertion holds for $m - 1$ and consider the two derived range spaces

$$\big(S \setminus \{x\}, R|(S \setminus \{x\})\big) \quad \text{and} \quad \big(S \setminus \{x\}, (R|S)^x\big)$$

for some fixed $x \in S$, where

$$(R|S)^x \;=\; \big\{A \in R|S \;:\; x \notin A \,\wedge\, A \cup \{x\} \in R|S\big\}.$$

Note that the ranges in $(R|S)^x$ are exactly those ranges in $R|(S \setminus \{x\})$ that have two preimages under the map

$$R|S \;\ni\; A \mapsto A \setminus \{x\} \;\in\; R|(S \setminus \{x\}),$$

all other ranges have a unique preimage. Consequently,

$$\big|R|S\big| \;=\; \big|R|(S \setminus \{x\})\big| + \big|(R|S)^x\big|.$$

We have by the induction hypothesis that

$$\big|R|(S \setminus \{x\})\big| \;\leq\; \Phi_d(m-1).$$

If $S' \subseteq S \setminus \{x\}$ is shattered by $(R|S)^x$, then $S' \cup \{x\}$ is shattered by $R|S$. Hence, $\big(S \setminus \{x\}|(R|S)^x\big)$ has VC-dimension at most $d - 1$, because otherwise the VC-dimension of $\big(S, R|S\big)$ would be larger than $d$, and thus we have

$$\big|(R|S)^x\big| \;\leq\; \Phi_{d-1}(m-1),$$

again by the induction hypothesis. It follows that

$$\big|R|S\big| \;\leq\; \Phi_d(m-1) + \Phi_{d-1}(m-1) \;=\; \Phi_d(m),$$

which yields the claim of Sauer's Lemma. □

The bound in Sauer's Lemma is tight: Let $\mathcal{X}$ some set and consider the range space $\left(\mathcal{X}, \bigcup_{i=1}^{d} \binom{\mathcal{X}}{d}\right)$, where $\binom{\mathcal{X}}{d}$ is the set of all subsets of $\mathcal{X}$ of size at most $d$. Obviously, a set with more than $d$ elements cannot be shattered and thus the VC-dimension of this range space is at most $d$. For any finite subset $S \subseteq \mathcal{X}$ the projection of the ranges onto $S$ is the set $\bigcup_{i=0}^{d} \binom{S}{d}$ whose size is $\Phi_d\big(|S|\big)$.

The function

$$\tau_R : \mathbb{N} \to \mathbb{N},\, m \mapsto \max_{S \subseteq \mathcal{X}\,:\,|S|=m} |R|S|$$

is called the *shatter coefficient* or *growth function* of the range space $\big(\mathcal{X}, R\big)$. By definition, this function grows exponentially, i.e., $\tau_R(m) = 2^m$, if $VC(R) = \infty$. The following upper bound on $\Phi_d(m)$,

$$\Phi_d(m) \;=\; \sum_{i=0}^{d} \binom{m}{i} \;\leq\; \sum_{i=0}^{d} \frac{m^i}{i!} \;\leq\; \sum_{i=0}^{d} \frac{m^i d!}{(d-i)!i!} \;=\; \sum_{i=0}^{d} \binom{d}{i} m^i \;=\; (m+1)^d,$$

shows that it grows polynomially, i.e., $\tau_R(m) \in O\big(m^d\big)$, if the VC-dimension $VC(R)$ is finite. The upper bound on $\Phi_d(m)$ can be improved to $\big(em/d\big)^d$ for $d \leq m$ , where $e$ is Euler's constant.

Sometimes it is easier than directly working with VC-theory to work with Rademacher complexity which is an alternative, but closely related approach to the VC-dimension for defining the complexity/capacity of a hypothesis class. In Chapter 18 we will see that Rademacher complexity is also interesting in its own right.

### Rademacher complexity

Let $H \subseteq \mathcal{Y}^{\mathcal{X}}$, where $\mathcal{Y} \subseteq \mathbb{R}$, be a hypothesis class, let $p$ be a probability distribution on $\mathcal{X}$, and let $S = \big(x^{(i)}\big)_{i\in[m]}$ be sequence of i.i.d. sample points drawn from $p$. The *empirical Rademacher complexity* of $H$ is defined as

$$R_S(H) \;=\; \mathsf{E}_\sigma \left[ \sup_{h\in H} \Big| \frac{1}{m} \sum_{i=1}^{m} \sigma_i h(x^{(i)}) \Big| \right],$$

where $\sigma$ is drawn from the uniform distribution on $\{\pm 1\}^m$, i.e., every $\sigma_i$, $i \in [m]$ is drawn independently from the uniform distribution on $\{\pm 1\}$.
The *expected Rademacher complexity* of $H$ is defined as

$$R_m(H) \;=\; \mathsf{E}_S\big[R_S(H)\big],$$

where $S$ is a sequence of sample points drawn independently from $p$.

In principle, the expected Rademacher complexity of a hypothesis class can be related to its VC-dimension. The following theorem bounds the expected Rademacher complexity for a fairly general class of hypothesis classes in terms of their associated growth functions.

**Theorem 7.** *Let $H \subseteq \{\pm 1\}^{\mathcal{X}}$ be a hypothesis class such that $h \in H$ implies that also $-h \in H$, and let $\tau_H$ be the growth function of its associated range space $(\mathcal{X}, R)$, where*

$$R = \big\{\{x \in \mathcal{X} \,:\, h(x) = 1\} \,:\, h \in H\big\}.$$

*Then,*

$$R_m(H) \le \sqrt{\frac{2\log\big(\tau_H(m)\big)}{m}}.$$

*Proof.* Let $S = \big(x^{(i)}\big)_{i\in[m]}$ be a sequence of i.i.d. sample points in $\mathcal{X}$ that have been drawn from the probability distribution $p$ on $\mathcal{X}$ that is also used to compute expected Rademacher complexity $R_m(H)$. If we define

$$A = \big\{\big(h(x^{(i)}\big)_{i\in[m]} \,:\, h \in H\big\} \subseteq \{\pm 1\}^m,$$

then we have

$$\begin{aligned} R_S(H) &= \mathsf{E}_\sigma\left[\mathsf{sup}_{h\in H}\left|\frac{1}{m}\sum_{i=1}^m \sigma_i h(x^{(i)})\right|\right] = \mathsf{E}_\sigma\left[\mathsf{max}_{a\in A}\left|\frac{1}{m}\sum_{i=1}^m \sigma_i a_i\right|\right] \\ &= \mathsf{E}_\sigma\left[\mathsf{max}_{a\in A}\,\frac{1}{m}\sum_{i=1}^m \sigma_i a_i\right], \end{aligned}$$

where the last equality simply holds because with $h \in H$ we also have $\bar{h} = -h \in H$ (which always predicts the opposite of the prediction of $h$), i.e., with $a \in A$ we also have $-a \in A$. Now we consider the random variables

$$X_a = \frac{1}{m}\sum_{i=1}^m \sigma_i a_i = \sum_{i=1}^m Y_{ai}, \quad a \in A$$

that are themselves sums of random variables $Y_{ai} = \frac{\sigma_i a_i}{m}$. In the following we only consider the randomness that comes from $\sigma_i$ but not the randomness from $a_i$. The $Y_{ai}$, $a \in A, i \in [m]$ take values in the interval $[-1/m, 1/m]$ and we have $\mathsf{E}_\sigma[Y_{ai}] = 0$ and thus by the linearity of expectation also $\mathsf{E}_\sigma[X_a] = 0$. Hence, we can apply Hoeffding's Lemma (see the supplemental material) and get for $s \geq 0$ that

$$\mathsf{E}_\sigma\big[\exp(sY_{ai})\big] \leq \exp\big(s^2/2m^2\big).$$

We compute

$$\begin{aligned}
\exp\Big(\mathsf{E}_\sigma\big[\max_{a\in A} sX_a\big]\Big) &\leq \mathsf{E}_\sigma\big[\exp\big(\max_{a\in A} sX_a\big)\big] \\
&= \mathsf{E}_\sigma\big[\max_{a\in A}\ \exp(sX_a)\big] \\
&\leq \sum_{a\in A} \mathsf{E}_\sigma\big[\exp(sX_a)\big] \\
&= \sum_{a\in A} \mathsf{E}_\sigma\left[\exp\left(s\sum_{i=1}^m Y_{ai}\right)\right] \\
&= \sum_{a\in A} \mathsf{E}_\sigma\left[\prod_{i=1}^m \exp(sY_{ai})\right] \\
&= \sum_{a\in A} \prod_{i=1}^m \mathsf{E}_\sigma\big[\exp(sY_{ai})\big] \\
&\leq \sum_{a\in A} \prod_{i=1}^m \exp\big(s^2/2m^2\big) \\
&= \sum_{a\in A} \exp\big(s^2/2m\big) \\
&= |A| \exp(s^2/2m),
\end{aligned}$$

where the first inequality is Jensen's Inequality for convex functions. Hence, we have

$$\begin{aligned}
\mathsf{E}_\sigma\big[\max_{a\in A} X_a\big] &= \frac{1}{s}\mathsf{E}_\sigma\big[\max_{a\in A} sX_a\big] \\
&\leq \frac{1}{s}\left(\log(|A|) + \frac{s^2}{2m}\right) = \frac{\log(|A|)}{s} + \frac{s}{2m}.
\end{aligned}$$

From which we get by choosing $s = \sqrt{2m\log(|A|)}$ that

$$\mathsf{E}_\sigma\big[\max_{a\in A} X_a\big] \leq \sqrt{\frac{2\log(|A|)}{m}}.$$

Finally, we want to bound the cardinality $|A|$ of $A$. Let $\big(\mathcal{X}, R\big)$ be the range space associated with $H$. Then $|A|$ is exactly the size $|R|S|$ of the projection of $R$ onto $S = \big(x^{(i)}\big)_{i\in[m]}$. The size of the projection is by the definition of the growth function at most $\tau_H(m)$. Hence, we get

$$R_S(H) \;=\; \mathsf{E}_\sigma\big[\mathsf{max}_{a\in A}\, X_a\big] \;\leq\; \sqrt{\frac{2\log\big(\tau_H(m)\big)}{m}},$$

where the first equlity has been established earlier. The assertion of the theorem now follows from the observation that that the right hand side of the inequality above does not depend on the sample $S$ but only on $m$. Thus,

$$R_m(H) \;=\; \mathsf{E}_S\big[R_S(H)\big] \;\leq\; \sqrt{\frac{2\log\big(\tau_H(m)\big)}{m}}.$$

□

If the VC-dimension $d$ of $H$ is finite, then it follows immediately from Theorem 7 that

$$R_m(H) \;\leq\; \sqrt{\frac{2d\log(em/d)}{m}}$$

where we have used Sauer's Lemma in the following form

$$\tau_H(m) \;\leq\; \Phi_d(m) \;\leq\; \big(em/d\big)^d.$$

# Chapter 14

# Main theorem of PAC learning

Before we can prove the main result of PAC learning for binary classification problems, that states that hypothesis classes of finite VC-dimension are PAC learnable, we need one more lemma that is proved by using Rademacher theory.

**Lemma 5.** *Given $\mathcal{X}$, $\mathcal{Y} = \{0,1\}$ and the loss function*

$$L : \mathcal{Y} \times \mathcal{Y} \to \{\pm 1\},\ (y^{(1)}, y^{(2)}) \mapsto 2 \cdot \mathbf{1}\big[y^{(1)} \neq y^{(2)}\big] - 1.$$

*Let $H \subseteq \mathcal{Y}^{\mathcal{X}}$ be a hypothesis class such that with $h \in H$ also $\bar{h} = 1 - h \in H$, and let $\tau_H$ be the growth function associated with $H$, i.e., the growth function of the range space $\big(\mathcal{X}, R\big)$ associated with $H$. Then, for every probability density function $p$ on $\mathcal{X} \times \mathcal{Y}$, every $h \in H$, every $m \geq 1$, and every $\delta \in (0,1)$,*

$$\mathsf{P}\left[\left\{ S \in (\mathcal{X} \times \mathcal{Y})^m \,:\, |L_p(h) - L_S(h)| \leq 2\sqrt{\frac{2\log\big(\tau_H(m)\big)}{m\delta^2}} \right\}\right] \,\geq\, 1 - \delta.$$

*Proof.* It is sufficient to show that

$$\mathsf{E}_S\Big[\mathsf{sup}_{h \in H}\, \big|L_p(h) - L_S(h)\big|\Big] \,\leq\, 2\sqrt{\frac{2\log\big(\tau_H(m)\big)}{m}}.$$

Since the random variable $\mathsf{sup}_{h \in H}|L_p(h) - L_S(h)|$ is non-negative it follows from Markov's Inequality that the probability that this random variable is larger than $1/\delta$ times its expectation is at most $\delta$, which is equivalent to the assertion of the lemma.

Since $L_p(h) = \mathsf{E}_{\bar{S}}[L_{\bar{S}}(h)]$ for a random sample $\bar{S} = \big((\bar{x}^{(i)}, \bar{y}^{(i)})\big)_{i \in [m]}$ we have

$$\begin{aligned} \mathsf{E}_S\Big[\mathsf{sup}_{h\in H}\,\big|L_p(h) - L_S(h)\big|\Big] &= \mathsf{E}_S\Big[\mathsf{sup}_{h\in H}\,\big|\mathsf{E}_{\bar{S}}[L_{\bar{S}}(h)] - L_S(h)\big|\Big] \\ &= \mathsf{E}_S\Big[\mathsf{sup}_{h\in H}\,\big|\mathsf{E}_{\bar{S}}[L_{\bar{S}}(h) - L_S(h)]\big|\Big]. \end{aligned}$$

Since for any random variable $X$ it holds true that $\big|\mathsf{E}[X]\big| \leq \mathsf{E}\big[|X|\big]$ we have

$$\big|\mathsf{E}_{\bar{S}}[L_{\bar{S}}(h) - L_S(h)]\big| \;\leq\; \mathsf{E}_{\bar{S}}\big[|L_{\bar{S}}(h) - L_S(h)|\big],$$

and because

$$\mathsf{E}_{\bar{S}}\big[|L_{\bar{S}}(h) - L_S(h)|\big] \;\leq\; \mathsf{E}_{\bar{S}}\Big[\mathsf{sup}_{h\in H}\,|L_{\bar{S}}(h) - L_S(h)|\Big]$$

for all $h \in H$, we also have

$$\mathsf{sup}_{h\in H}\,\mathsf{E}_{\bar{S}}\big[|L_{\bar{S}}(h) - L_S(h)|\big] \;\leq\; \mathsf{E}_{\bar{S}}\Big[\mathsf{sup}_{h\in H}\,|L_{\bar{S}}(h) - L_S(h)|\Big].$$

Hence, we have

$$\begin{aligned} \mathsf{E}_S\Big[\mathsf{sup}_{h\in H}\,|L_p(h) - L_S(h)|\Big] &\leq \mathsf{E}_S\Big[\mathsf{E}_{\bar{S}}\big[\mathsf{sup}_{h\in H}\,|L_{\bar{S}}(h) - L_S(h)|\big]\Big] \\ &= \mathsf{E}_{S,\bar{S}}\Big[\mathsf{sup}_{h\in H}\,\big|L_{\bar{S}}(h) - L_S(h)\big|\Big], \end{aligned}$$

and it suffices to bound

$$\begin{aligned} &\mathsf{E}_{S,\bar{S}}\big[\mathsf{sup}_{h\in H}\,|L_{\bar{S}}(h) - L_S(h)|\big] \\ &\qquad = \mathsf{E}_{S,\bar{S}}\left[\mathsf{sup}_{h\in H}\,\Big|\frac{1}{m}\sum_{i=1}^{m} L\big(h(\bar{x}^{(i)}), \bar{y}^{(i)}\big) - L\big(h(x^{(i)}), y^{(i)}\big)\Big|\right]. \end{aligned}$$

Since $S$ and $\bar{S}$ are independent i.i.d. samples of size $m$ the latter expectation is the same as

$$\mathsf{E}_{S,\bar{S}}\left[\mathsf{sup}_{h\in H}\,\Big|\frac{1}{m}\sum_{i=1}^{m} \sigma_i \left(L\big(h(\bar{x}^{(i)}), \bar{y}^{(i)}\big) - L\big(h(x^{(i)}), y^{(i)}\big)\right)\Big|\right],$$

for any $\sigma \in \{\pm 1\}^m$. Since this holds for any such $\sigma$ it also holds for the expectation, if we sample $\sigma$ (uniformly) from $\{\pm 1\}^m$. Hence, the latter expectation is the same as

$$\mathsf{E}_\sigma\left[\mathsf{E}_{S,\bar{S}}\left[\mathsf{sup}_{h\in H}\,\Big|\frac{1}{m}\sum_{i=1}^{m} \sigma_i \left(L\big(h(\bar{x}^{(i)}), \bar{y}^{(i)}\big) - L\big(h(x^{(i)}), y^{(i)}\big)\right)\Big|\right]\right].$$

Since computing the expectation over $\sigma$ is a finite sum, we can can compute it before the expectation over $S$ and $\bar{S}$, i.e. move the summation over $\sigma$ into the integral for for computing the expectation over $S$ and $\bar{S}$. Hence, the latter expectation is the same as

$$\mathsf{E}_{S,\bar{S}}\left[\mathsf{E}_\sigma\left[\mathsf{sup}_{h\in H}\left|\frac{1}{m}\sum_{i=1}^m \sigma_i\left(L\big(h(\bar{x}^{(i)}),\bar{y}^{(i)}\big)-L\big(h(x^{(i)}),y^{(i)}\big)\right)\right|\right]\right].$$

Using the inequality $|a-b|\leq|a|+|b|$ we can upper bound the latter expectation by

$$\mathsf{E}_{S,\bar{S}}\left[\mathsf{E}_\sigma\left[\mathsf{sup}_{h\in H}\left|\frac{1}{m}\sum_{i=1}^m \sigma_i L\big(h(\bar{x}^{(i)}),\bar{y}^{(i)}\big)\right|+\left|\frac{1}{m}\sum_{i=1}^m \sigma_i L\big(h(x^{(i)}),y^{(i)}\big)\right|\right]\right],$$

which is the same as

$$2\cdot\mathsf{E}_S\left[\mathsf{E}_\sigma\left[\mathsf{sup}_{h\in H}\left|\frac{1}{m}\sum_{i=1}^m \sigma_i L\big(h(x^{(i)}),y^{(i)}\big)\right|\right]\right],$$

since the samples $S$ and $\bar{S}$ are drawn independently from the same distribution. In Theorem 7 we have shown that

$$\mathsf{E}_S\left[\mathsf{E}_\sigma\left[\mathsf{sup}_{h\in H}\left|\frac{1}{m}\sum_{i=1}^m \sigma_i L\big(h(x^{(i)}),y^{(i)}\big)\right|\right]\right]\;\leq\;\sqrt{\frac{2\log\big(\tau_H(m)\big)}{m}}.$$

Since this bound does only depend on the size $m$ of the sample $S$, we finally get that

$$\begin{aligned}\mathsf{E}_S&\Big[\mathsf{sup}_{h\in H}\big|L_p(h)-L_S(h)\big|\Big]\\ &\leq\;2\cdot\mathsf{E}_S\left[\mathsf{E}_\sigma\left[\mathsf{sup}_{h\in H}\left|\frac{1}{m}\sum_{i=1}^m \sigma_i L\big(h(x^{(i)}),y^{(i)}\big)\right|\right]\right]\\ &\leq\;2\sqrt{\frac{2\log\big(\tau_H(m)\big)}{m}},\end{aligned}$$

which implies the assertion of the lemma. □

## Hypothesis classes of finite VC-dimension are PAC learnable

Now we are finally prepared to prove that hypothesis classes of finite VC-dimension are PAC learnable.

**Theorem 8.** *Given $\mathcal{X}$, $\mathcal{Y} = \{0,1\}$ and the loss function*

$$L : \mathcal{Y} \times \mathcal{Y} \to \{\pm 1\},\ (y^{(1)}, y^{(2)}) \mapsto 2 \cdot \mathbf{1}\big[y^{(1)} \neq y^{(2)}\big] - 1.$$

*If $H \subseteq \mathcal{Y}^{\mathcal{X}}$ is a hypothesis class such that with $h \in H$ also $\bar{h} = 1 - h \in H$ and $VC(H) = d < \infty$, then $H$ is PAC learnable.*

*Proof.* By Theorem 5 it suffices to prove that $H$ has the uniform convergence property. That is, we need to establish that there exists a function

$$m : \mathbb{R}_+ \times (0,1) \to \mathbb{R}_{\geq 0},\ (\varepsilon, \delta) \mapsto m(\varepsilon, \delta)$$

such that any sample of size $m \geq m(\varepsilon, \delta)$ drawn i.i.d. from $p$ is $\varepsilon$-representative for $H$ and $L$ with probability at least $1 - \delta$ for any probability density function $p$ on $\mathcal{X} \times \mathcal{Y}$ and any choice of $\varepsilon > 0,\ \delta \in (0,1)$.
From Lemma 5 we know that

$$|L_p(h) - L_S(h)| \,\leq\, 2\sqrt{\frac{2\log\big(\tau_H(m)\big)}{m\delta^2}}$$

with probability at least $1 - \delta$, and from Sauer's Lemma we know that

$$\tau_H(m) \,\leq\, \Phi_d(m) \,\leq\, \big(em/d\big)^d.$$

Thus we have with probability at least $1 - \delta$ that

$$|L_p(h) - L_S(h)| \,\leq\, 2\sqrt{\frac{2d\log\big(em/d\big)}{m\delta^2}}.$$

To ensure that this bound is at most $\varepsilon > 0$ we need that

$$m \,\geq\, \frac{8d\log\big(em/d\big)}{(\delta\varepsilon)^2} \,=\, \frac{8d\log(m)}{(\delta\varepsilon)^2} + \frac{8d\log\big(e/d\big)}{(\delta\varepsilon)^2}.$$

Using that for $a > 1$ and $b > 0$ it follows from $x \geq 4a\log(2a) + 2b$ that $x \geq a\log(x) + b$ (Exercise!) we get, if we set $a = 8d/(\delta\varepsilon)^2$ and $b = 8d\log(e/d)/(\delta\varepsilon)^2$, that it is sufficient to have

$$m \,\geq\, \frac{32d}{(\delta\varepsilon)^2}\log\left(\frac{16d}{(\delta\varepsilon)^2}\right) + \frac{16d\log\big(e/d\big)}{(\delta\varepsilon)^2},$$

from which we get the function $m$ we were looking for. □

*Remark:* The sample complexity that results from Theorem 8 can be improved. The best possible sample complexity is

$$m(\varepsilon, \delta) \,\in\, \Theta\left(\frac{d + \log(1/\delta)}{\varepsilon^2}\right).$$

### Validation

Assume that we have computed a classifier $\hat{h} \in H$ in a given hypothesis class $H$ from i.i.d. training data points $S = \Big((x^{(i)}, y^{(i)})\Big)_{i\in[m]}$. Of course we would like to know how good $\hat{h}$ is, that is, we want to assess the expected loss $L_p(\hat{h})$ of $\hat{h}$. We still do not have access to $L_p(\hat{h})$, but only to the empirical loss $L_S(\hat{h})$. Assume that $\mathcal{Y} = \{0, 1\}$ and let $L$ be the 0-1 loss-function. Then we get from Hoeffding's Inequality, see the supplemental material, for any $\varepsilon > 0$ that

$$\mathsf{P}\left[\left\{\bar{S} \in (\mathcal{X}\times\mathcal{Y})^m \,:\, |L_p(\hat{h}) - L_{\bar{S}}(\hat{h})| \ge \varepsilon\right\}\right] \;\le\; 2\exp\left(-\frac{m\varepsilon^2}{2}\right).$$

In other words, the empirical loss on the test data deviates from the expected loss by at most $\varepsilon$ with probability at least $1-\delta$ for a prescribed $\delta \in (0, 1)$, if $m \ge 2\log(2/\delta)/\varepsilon^2$. Hence, for $m$ large enough, the empirical loss $L_S(\hat{h})$ is a good proxy for $L_p(\hat{h})$. The loss is made up from two terms, the approximation error $\varepsilon_{app} = \min_{h\in H} L_p(h)$ and the estimation error $\varepsilon_{est}$. If we know the finite VC-dimension of $H$ and use a *probably approximately correct* algorithm, then, by Theorem 8, we can control the estimation error $\varepsilon_{est}$ with high prescribed probability $1-\delta$, by making the training set sufficiently large. In this case, we can fairly safely claim that a large empirical loss $L_S(\hat{h})$ is due to a large approximation error $\varepsilon_{app}$. There are a two options for reducing the approximation error,

1. switching to a larger hypothesis class (increasing the VC dimension), or
2. measuring more or other features, that is changing the feature space $\mathcal{X}$ and consequently also the hypothesis class.

The bound in Theorem 8 can be overly pessimistic in the sense that often much fewer training data points are sufficient for a small estimation error. We have discussed before, in the context of logistic regression, how to use an independent (often smaller) test data set $V$ for validating a classifier. Let $V = \Big((\bar{x}^{(i)}, \bar{y}^{(i)})\Big)_{i\in[k]}$ be the i.i.d. test data that are sampled independently from the training data. Together, the training data set $S$ and validation data set $V$ can be used for disentangling the approximation and estimation error as follows: We can decompose the expected loss of $\hat{h}$ into three more accessible terms

$$L_p(\hat{h}) \;=\; \big(L_p(\hat{h}) - L_V(\hat{h})\big) + \big(L_V(\hat{h}) - L_S(\hat{h})\big) + L_S(\hat{h}).$$

Let us have a closer look at the three terms. The first term can be bounded using Hoeffding's inequality as above and should be small. Note that, in general, Hoeffding's inequality needs much fewer sample points to guarantee a good bound with high probability than Theorem 8 which needs the number of sample points to grow with the VC-dimension. If the second term is large, then it is reasonable to suspect a large estimation error. The third term can also be written as follows

$$L_S(\hat{h}) \;=\; \big(L_S(\hat{h}) - L_S(h^*)\big) + \big(L_S(h^*) - L_p(h^*)\big) + L_p(h^*),$$

where $h^* \in \operatorname{argmin}_{h\in H} L_p(h)$ and thus we have $L_p(h^*) = \varepsilon_{app}$. If $h$ has been computed from the ERM rule, then $\big(L_S(h) - L_S(h^*)\big) \leq 0$. The term $\big(L_S(h^*) - L_p(h^*)\big)$ can be bounded again by Hoeffding's inequality and should be small. Hence, we have wit high probability that $L_S(\hat{h}) \leq \varepsilon_{app} + \varepsilon$, where $\varepsilon$ can be arbitrarily small, if $m$ is sufficiently large. That is, if $L_S(h)$ is large, then the approximation error is also large. Thus, a large third term in the decomposition of $L_p(\hat{h})$ indicates a large approximation error. It is important to note though that the approximation error can be large although $L_S(\hat{h})$ is small. Still, if we suspect that the approximation error is large, then we should try changing the hypothesis class or even the feature space. When the estimation error is large, then we should either *enlarge the training data set*, or *reduce the hypothesis class* (which of course may increase the approximation error). Remember our discussion of the bias-variance trade-off in the context of the ridge regression problem, it can pay off to increase the bias, here the approximation error, because this could mean an even larger reduction of the variance, here estimation error.

*Remark:* By taking a union bound, we can extend Hoeffding's approximation bound on the expected loss by the empirical loss from one to several classifiers. Let $H$ be a finite hypothesis class, then we have

$$\mathsf{P}\left[\left\{S \in (\mathcal{X}\times\mathcal{Y})^k \,:\, \exists h \in H \,:\, |L_p(h) - L_S(h)| \geq \varepsilon\right\}\right] \;\leq\; 2|H|\exp\left(-\frac{k\varepsilon^2}{2}\right).$$

for any probability density $p$ on $\mathcal{X}\times\mathcal{Y}$, and any $\varepsilon > 0$. This is interesting in the context of regularized learning methods. Typically, one samples the regularization parameter, randomly or on some finite grid, and evaluates the predictors for the respective regularization parameter values on the test data set. Often the predictor that performs best on the test set is chosen as the final predictor. Note that this last step also constitutes some form of learning and thus can be prone to overfitting. Hence, the final predictor should be evaluated on a third independent i.i.d. data set. Typically, the third data set is called the test set, while the second data set is called *validation* or *dev* set.

### Hypothesis classes of infinite VC-dimension are not PAC learnable

If $VC(H) = \infty$, then $\tau_H(m) = 2^m$ and $m$ vanishes from the bound in Lemma 5 and thus, we cannot use Lemma 5 to get a sample complexity, if $VC(H) = \infty$. The following theorem, known as a *no-free-lunch theorem*, states that this is not a shortcoming of Lemma 5, because it is not possible to get a sample complexity in the case that $VC(H) = \infty$. Hence, hypothesis classes of infinite VC-dimension are not PAC learnable.

**Theorem 9. [No-free-lunch]** *Let $H \subseteq \mathcal{Y}^{\mathcal{X}}$, where $\mathcal{Y} = \{0,1\}$, be a hypothesis class with $VC(H) = d$, and let $L$ be the 0-1-loss function on $\mathcal{Y} \times \mathcal{Y}$. Then, for any algorithm that computes a predictor $\hat{h}(S) \in H$ on input of a sequence $S$ of sample points in $\mathcal{X} \times \mathcal{Y}$ there exists a probability distribution $p$ on $\mathcal{X} \times \mathcal{Y}$ such that*

$$P\left[\left\{S \in (\mathcal{X} \times \mathcal{Y})^m \;:\; L_p\big(\hat{h}(S)\big) \;>\; \mathit{min}_{h \in H}\, L_p(h) + \frac{1}{8}\right\}\right] \;\geq\; \frac{1}{7}$$

*for $m \leq \lfloor d/2 \rfloor$.*

*Proof.* For any given algorithm we construct a *bad* probability distribution $p$ such that

$$L_p(\hat{h}) - \mathrm{min}_{h \in H}\, L_p(h) \;>\; \frac{1}{8}$$

with probability at least $1/7$ on sequences of $m < d/2$ sample points drawn independently from $p$. Since $VC(H) = d$, there exists a sequence $\big(x^{(i)}\big)_{i \in [d]}$ of distinct points in $\mathcal{X}$ such that the sets $\{x^{(1)}, \ldots, x^{(m)}\}$, where $m < d$, are shattered by the ranges associated with $H$. Let $m < d$, let $S = \{x^{(1)}, \ldots, x^{(m)}\}$, and let $H' \subseteq H$ be a set of $2^m$ predictors whose associated ranges shatter the set $S$. That is, for every possible labeling of the points in $S$ there is exactly one function in $H'$ that achieves that labeling. Let $p_S$ be the density

$$p_S(x) \;=\; \frac{1}{m} \sum_{\bar{x} \in S} \mathbf{1}[x = \bar{x}]$$

on $\mathcal{X}$, and let $p_h,\ h \in H'$ be the following densities on $\mathcal{X} \times \mathcal{Y}$,

$$p_h(x,y) \;=\; p_S(x) \cdot \mathbf{1}[h(x) = y].$$

Obviously, we have $L_h(h) = 0$ for all $h \in H'$, where $L_h(h)$ is the expected loss of $h$ with respect to the density $p_h$.

In the following we are going to show that

$$\max_{h\in H'} \mathsf{E}_{\bar{S}}\big[L_h\big(\hat{h}(\bar{S})\big)\big] \;\geq\; \frac{1}{4},$$

where $\hat{h}(\bar{S})$ is the predictor that has been computed by the algorithm from a sequence $\bar{S}$ of $\lfloor m/2 \rfloor$ of random samples drawn independently from $p_h$. Let $h \in H'$ be the maximizer, then the bound on the expectation readily implies the assertion of the theorem

$$\begin{aligned}\mathsf{P}_h\left[\left\{\bar{S}\in(\mathcal{X}\times\mathcal{Y})^{\lfloor m/2\rfloor} \;:\; L_h\big(\hat{h}(\bar{S})\big) \;>\; \min_{\bar{h}\in H} L_h(\bar{h}) + \frac{1}{8}\right\}\right]&\\ = \mathsf{P}_h\left[\left\{\bar{S}\in(\mathcal{X}\times\mathcal{Y})^{\lfloor m/2\rfloor} \;:\; L_h\big(\hat{h}(\bar{S})\big) > \frac{1}{8}\right\}\right] \;&\geq\; \frac{1}{7},\end{aligned}$$

because by Markov's Inequality and $L\big(\hat{h}(\bar{S})\big) \in [0,1]$ we have

$$\begin{aligned}\mathsf{P}_h\left[\left\{\bar{S} \;:\; L_h\big(\hat{h}(\bar{S})\big) > \frac{1}{8}\right\}\right] \;&=\; \mathsf{P}_h\left[\left\{\bar{S} \;:\; 1 - L_h\big(\hat{h}(\bar{S})\big) < \frac{7}{8}\right\}\right]\\ &=\; 1 - \mathsf{P}_h\left[\left\{\bar{S} \;:\; 1 - L_h\big(\hat{h}(\bar{S})\big) \geq \frac{7}{8}\right\}\right]\\ &\geq\; 1 - \frac{\mathsf{E}_{\bar{S}}\big[1 - L_h\big(\hat{h}(\bar{S})\big)\big]}{7/8}\\ &\geq\; 1 - \frac{1-1/4}{7/8} \;=\; 1 - \frac{6}{7} \;=\; \frac{1}{7}.\end{aligned}$$

Let $S_{h,1}, \ldots S_{h,k}$ be all the sequences of $\lfloor m/2 \rfloor$ sample points $(x, y)$, where $x \in S$ and $y = h(x)$. There are $k = m^{\lfloor m/2 \rfloor}$ different sequences and by our choice of the density $p_S$, every sequence $S_{h,i}$ has the same probability of $1/k$ to be drawn from $p_h$. Hence,

$$\mathsf{E}_{\bar{S}}\big[L_h\big(\hat{h}(\bar{S})\big)\big] \;=\; \frac{1}{k}\sum_{i=1}^{k} L_h\big(\hat{h}(S_{h,i})\big)$$

and thus

$$\begin{aligned}
\mathsf{max}_{h\in H'}\ \mathsf{E}_{\bar{S}}\big[L_h\big(\hat{h}(\bar{S})\big)\big] &\geq \frac{1}{2^m}\sum_{h\in H'}\mathsf{E}_{\bar{S}}\big[L_h\big(\hat{h}(\bar{S})\big)\big]\\
&= \frac{1}{2^m}\sum_{h\in H'}\frac{1}{k}\sum_{i=1}^{k}L_h\big(\hat{h}(S_{h,i})\big)\\
&= \frac{1}{k}\sum_{i=1}^{k}\frac{1}{2^m}\sum_{h\in H'}L_h\big(\hat{h}(S_{h,i})\big)\\
&\geq \mathsf{min}_{i\in[k]}\ \frac{1}{2^m}\sum_{h\in H'}L_h\big(\hat{h}(S_{h,i})\big).
\end{aligned}$$

Next we bound $L_h\big(\hat{h}(S_{h,i})\big)$ for the minimizer $i\in[k]$ in the inequality from above and for any $h\in H'$. Let

$$S_{h,i} = \Big((\bar{x}^{(j)}, y^{(j)})\Big)_{j\in[\lfloor m/2\rfloor]} = \Big((\bar{x}^{(j)}, h(x^{(j)}))\Big)_{j\in[\lfloor m/2\rfloor]}$$

and let $T = S\setminus\big\{\bar{x}^{(1)},\dots,\bar{x}^{(\lfloor m/2\rfloor)}\big\}$, where $T$ is by construction independent of $h$. Then we have $|T|\geq\lceil m/2\rceil$ and by the definition of the definition of the 0-1-loss function that

$$\begin{aligned}
L_h\big(\hat{h}(S_{h,i})\big) &= \frac{1}{m}\sum_{x\in S}\mathbf{1}\big[\hat{h}(S_{h,i})(x)\neq h(x)\big]\\
&\geq \frac{1}{m}\sum_{x\in T}\mathbf{1}\big[\hat{h}(S_{h,i})(x)\neq h(x)\big]\\
&\geq \frac{1}{2|T|}\sum_{x\in T}\mathbf{1}\big[\hat{h}(S_{h,i})(x)\neq h(x)\big]
\end{aligned}$$

Hence,

$$\begin{aligned}
\mathsf{max}_{h\in H'}\, \mathsf{E}_{\bar{S}}\big[L_h\big(\hat{h}(\bar{S})\big)\big] &\geq \frac{1}{2^m}\sum_{h\in H'} L_h\big(\hat{h}(S_{h,i})\big)\\
&\geq \frac{1}{2^m}\sum_{h\in H'}\frac{1}{2|T|}\sum_{x\in T}\mathbf{1}\big[\hat{h}(S_{h,i})(x)\neq h(x)\big]\\
&= \frac{1}{2|T|}\sum_{x\in T}\frac{1}{2^m}\sum_{h\in H'}\mathbf{1}\big[\hat{h}(S_{h,i})(x)\neq h(x)\big]\\
&\geq \frac{1}{2}\,\mathsf{min}_{x\in T}\,\frac{1}{2^m}\sum_{h\in H'}\mathbf{1}\big[\hat{h}(S_{h,i})(x)\neq h(x)\big]
\end{aligned}$$

Finally, let $x\in T$ be the minimizer in inequality from above. We can partition the functions in $H'$ into $2^{m-1}$ pairs $\{h,h'\}$ that disagree only on $x$, i.e., $h(x)\neq h'(x)$ only for $x$. For any such pair we have $S_{h,i}=S_{h',i}$, because $x$ does by the definition of $T$, that is independent of $h$, not appear in these sequences. Thus,

$$\mathbf{1}\big[\hat{h}(S_{h,i})(x)\neq h(x)\big]+\mathbf{1}\big[\hat{h}(S_{h',i})(x)\neq h'(x)\big]\;=\;1,$$

which implies

$$\frac{1}{2^m}\sum_{h\in H'}\mathbf{1}\big[\hat{h}(S_{h,i})(x)\neq h(x)\big]\;=\;\frac{1}{2}.$$

Hence,

$$\frac{1}{2}\,\mathsf{min}_{x\in T}\,\frac{1}{2^m}\sum_{h\in H'}\mathbf{1}\big[\hat{h}(S_{h,i})(x)\neq h(x)\big]\;=\;\frac{1}{2}\cdot\frac{1}{2}\;=\;\frac{1}{4},$$

from which we finally get

$$\mathsf{max}_{h\in H'}\,\mathsf{E}_{\bar{S}}\big[L\big(\hat{h}(\bar{S})\big)\big]\;\geq\;\frac{1}{4}.$$

□

*Remark:* In this chapter we have actually shown that the following are equivalent for a given hypothesis class $H$: (1) $VC(H)<\infty$, (2) $H$ has the uniform convergence property, (3) the ERM rule is a successful PAC learner for $H$, and (4) $H$ is PAC learnable,

# Chapter 15

# VC-dimension of some hypothesis classes

So far we have considered hypothesis classes $H \subseteq \{0,1\}^{\mathcal{X}}$. Such classes have naturally associated range spaces $\big(\mathcal{X}, R\big)$ and we can recover $H$ from $R$. Here, we consider hypothesis classes $H \subseteq \mathbb{R}^{\mathcal{X}}$, i.e., $\mathcal{Y} = \mathbb{R}$, that also have associated range spaces $\big(\mathcal{X}, R\big)$, whose ranges are given as

$$R = \big\{\{x \in \mathcal{X} \,:\, h(x) \geq 0\} \,:\, h \in H\big\}.$$

Although we cannot recover $H$ from $R$, we can still use $R$ for defining the VC-dimension of $H$, namely by $VC(H) = VC(R)$. Note that $\mathbb{R}^{\mathcal{X}}$ carries the structure of a real vector space, as we have seen when constructing feature maps from kernels in the proof of the Moore-Aronszajn Theorem. Here, we consider (finite dimensional) subspaces $H \subseteq \mathbb{R}^{\mathcal{X}}$ and relate their vector space dimension dim$(H)$ to their VC-dimension in the following theorem.

**Theorem 10.** *Let $H \subseteq \mathbb{R}^{\mathcal{X}}$ a vector space with dim$(H) = d < \infty$. Then $VC(H) \leq d$.*

*Proof.* We need to show that no set $S = \{x^{(1)}, \ldots, x^{(d+1)}\} \subseteq \mathcal{X}$ with $d+1$ points can be shattered by the ranges associated with $H$. If we define

$$\phi : H \to \mathbb{R}^{d+1},\, h \mapsto \big(h(x^{(1)}), \ldots, h(x^{(d+1)})\big)^\top,$$

then $\phi(H)$ is a linear subspace of $\mathbb{R}^{d+1}$, whose dimension is at most $d$. Hence, there must exist $a \in \mathbb{R}^{d+1} \setminus \{0\}$ that is orthogonal to $\phi(H)$. That is, for any $h \in H$ we have

$$a^\top \phi(h) \,=\, \sum_{i=1}^{d+1} a_i h(x^{(i)}) \,=\, 0.$$

We can assume that there exists $i \in [d+1]$ such that $a_i < 0$, otherwise just replace $a$ by $-a$. If the ranges associated with $H$ shatter $S$, then there exists $h \in H$ such that

$$\{x^{(i)} \,:\, h(x^{(i)}) \geq 0\} \,=\, \{x^{(i)} \,:\, a_i \geq 0\},$$

but this implies

$$0 \,\leq \sum_{i \in [d+1] \,:\, a_i \geq 0} a_i h(x^{(i)}) \,=\, - \sum_{i \in [d+1] \,:\, a_i < 0} a_i h(x^{(i)}) \,<\, 0,$$

which is a contradiction. Hence, $S$ cannot be shattered by the range space associated with $H$. □

From Theorem 10 we immediately get the following corollary.

**Corollary 1.** *Let* $\big(\mathbb{R}^n, R\big)$*, where*

$$R \,=\, \big\{\{x \in \mathbb{R}^n \,:\, a^\top x + b \geq 0\} \,:\, a \in \mathbb{R}^n \setminus \{0\}, b \in \mathbb{R}\big\},$$

*be the range space of closed half-spaces in* $\mathbb{R}^n$*. Then* $VC(R) \leq n+1$*.*

*Proof.* Let $H'$ be the following set of functions on $\mathbb{R}^n$,

$$H' \,=\, \big\{h : \mathbb{R}^n \to \mathbb{R},\, x \mapsto a^\top x + b \,:\, a \in \mathbb{R}^n, b \in \mathbb{R}\big\}$$

and let $R'$ be the set of associated ranges. Then we have $R \subseteq R'$ and $H'$ is a linear subspace of $\mathbb{R}^{\mathbb{R}^n}$ of dimension $n+1$. Note that the functions

$$x \mapsto e_i^\top x,\; i \in [n], \;\text{ and }\; x \mapsto 1$$

form a basis of $H'$. Here, $e_i$, $i \in [n]$ are the standard basis vectors of $\mathbb{R}^n$. Hence, we have by the definition of VC-dimension and Theorem 10 that

$$VC(R) \,\leq\, VC(R') \,=\, VC(H') \,\leq\, n+1.$$

□

Of course the proof of the corollary also implies that the hypothesis class of linear functions

$$H' \,=\, \big\{h : \mathbb{R}^n \to \mathbb{R},\, x \mapsto a^\top x + b \,:\, a \in \mathbb{R}^n, b \in \mathbb{R}\big\}$$

has VC-dimension at most $n+1$.

**Corollary 2.** *Let $\mathbb{R}[x_1, \ldots, x_n]_k$ be the set of all polynomials with degree at most $k \in \mathbb{N}$ on $\mathbb{R}^n$. We have*

$$VC\big(\mathbb{R}[x_1, \ldots, x_n]_k\big) \;\le\; \binom{n+k}{n}.$$

*Proof.* The range space associated with $\mathbb{R}[x_1, \ldots, x_n]_k$ is $\big(\mathbb{R}^n, R\big)$, where

$$R \;=\; \big\{\{x \in \mathbb{R}^n \;:\; p(x) \ge 0\} \;:\; p \in \mathbb{R}[x_1, \ldots, x_n]_k\big\}.$$

Let $M \subset \mathbb{R}[x_1, \ldots, x_n]_k$ be the set of monomials of degree at most $k$ (including the constant monomial $1$) and let $m = |M|$. The *Veronese map*

$$\phi : \mathbb{R}^n \to \mathbb{R}^m,\; x \mapsto \big(q(x)\big)_{q \in M},$$

maps $x \in \mathbb{R}^n$ to the vector of its images under all monomials. We want to show that if a set $S \subset \mathbb{R}^n$ is shattered by $R$, then $\phi(S)$ is shattered by halfspaces in $\mathbb{R}^m$. If $S$ is shattered by $R$, then for any $S' \subseteq S$ there exists a polynomial $p \in \mathbb{R}[x_1, \ldots, x_n]_k$ such that $p$ is non-negative on $S'$ and negative on $S \setminus S'$. The polynomial $p$ has a unique representation

$$p \;=\; \sum_{q \in M} a_q \, q$$

as a linear combination of monomials. Let

$$H_p \;=\; \Big\{\bar{x} \in \mathbb{R}^m \;:\; a^\top \bar{x} = \sum_{q \in M} a_q \bar{x}_q \ge 0\Big\}.$$

Then $H_p \cap \phi(S) = \phi(S')$. Since $\phi$ is injective we have found a set of size $|S|$ in $\mathbb{R}^m$, namely $\phi(S)$, that is shattered by closed halfspaces. Thus we have $|S| \le m$ by Corollary 1. Note that the bound is $m$ and not $m+1$, because the constant monomial, that serves as the offset of the hyperplanes, is included+ n $M$. The claim now follows from $m = \binom{n+k}{n}$, i.e., there are $\binom{n+k}{n}$ monomials of degree at most $k$. □

Remark: $n$-variate monomials of degree at most $k$ can be counted by induction over $n$. If $n = 1$, then we need to count the univariate monimials of degree at most $k$. These momomials are easily enumerated as $1, x_1, x_1^2, \ldots, x_1^k$. Hence, there are $k + 1 = \binom{1+k}{1}$ such monomials. For the inductive step, assume that the assertion holds for $n-1$. The variable $x_n$ can have the degrees $0$ to $k$. If the variable $x_n$ has degree $k$, then all other variables need to have degree $0$. By the induction hypothesis the number of $(n-1)$-variate monomials of degree at most $0$ is $\binom{n-1+0}{n-1}$ . If the variable $x_n$ has degree $k-1$, then all other variables need to have degree $0$, except for one that has degree $1$. By the induction hypothesis the number of $(n-1)$-variate monomials of degree at most $1$ is $\binom{n-1+1}{n-1}$. In general, by the induction hypothesis the number of $n$-variate monomials of degree at most $k$, where $x_n$ has degree $i$, is $\binom{n-1+i}{n-1}$ . Thus there are

$$\sum_{i=0}^{k}\binom{n-1+i}{n-1} = \sum_{i=0}^{k}\frac{(n-1+i)!}{(n-1)!i!} = \sum_{i=0}^{k}\binom{n-1+i}{i}.$$

$n$-variate monomials of degree at most $k$. Finally, by induction over $k$ we get

$$\sum_{i=0}^{k}\binom{n-1+i}{i} = \binom{n+k}{n}.$$

If $k=0$, then we have

$$\sum_{i=0}^{0}\binom{n-1+i}{i} = \binom{n-1}{0} = 1 = \binom{n+0}{0}.$$

For the inductive step from $k-1$ to $k$ observe that

$$\begin{aligned}\sum_{i=0}^{k}\binom{n-1+i}{i} &= \sum_{i=0}^{k-1}\binom{n-1+i}{i} + \binom{n-1+k}{k}\\ &= \binom{n+k-1}{n} + \binom{n-1+k}{k}\\ &= \binom{n+k-1}{k-1} + \binom{n+k-1}{k}\\ &= \binom{n+k}{n},\end{aligned}$$

where we have used, in the last equality, the recursive formula for binomial coefficients that we have also used in the proof of Sauer's Lemma.

# Chapter 16

# Exercises

### Chapter 13

**Exercise 1** Let $\mathcal{X} = \{0,1\}^n, \mathcal{Y} = \{0,1\}$ and

$$L : \mathcal{Y} \times \mathcal{Y} \to [0,1]$$

a loss function, for instance the 0-1 loss. What is the sample complexity of the hypothesis class

$$H = \mathcal{Y}^{\mathcal{X}} = \{h : \mathcal{X} \to \mathcal{Y}\}?$$

### Chapter 15

**Exercise 1.** Show that

$$\binom{m}{i} = \binom{m-1}{i} + \binom{m-1}{i-1}.$$

**Exercise 2.** Let $\mathcal{X}$ be a sample space and

$$H = \big\{h : \mathcal{X} \to \{0,1\} \ : \ |h^{-1}(1)| = k\big\}$$

be the hypothesis class of functions that take exactly $k$ times the value $1$. What is the VC-dimension of $H$, i.e., the VC-dimension of the associated range space?

**Exercise 3.** Determine the VC-dimension of of axis aligned squares in the plane, i.e., the VC-dimension of the range space $(\mathbb{R}^2, R)$, where

$$R = \big\{\{x \in \mathbb{R}^2 \ : \ a_1 \leq x_1 \leq b_1 \wedge a_2 \leq x_2 \leq b_2\} \ : \ b_1 - a_1 = b_2 - a_2\big\}.$$

**Exercise 4.** Determine the VC-dimension of of axis aligned rectangles in the plane, i.e., the VC-dimension of the range space $(\mathbb{R}^2, R)$, where

$$R = \big\{\{x \in \mathbb{R}^2 \,:\, a_1 \leq x_1 \leq b_1 \wedge a_2 \leq x_2 \leq b_2\} \,:\, a_1 < b_1 \wedge a_2 < b_2\big\}.$$

**Exercise 5.** Let

$$H = \{x \mapsto \theta^\top x \,:\, \theta \in \mathbb{R}^n,\, \|\theta\| = 1\} \subset \mathbb{R}^{\mathbb{R}^n}$$

be the hypothesis class of linear functions without offset, and let $S = \big(x^{(i)}\big)_{i \in [m]}$ sample points drawn independently from some probability distribution on $\mathbb{R}^n$. Determine (an upper bound on) the empirical Rademacher complexity $R_S(H)$ of $H$.

Hint: Use Cauchy-Schwarz' and Jensen's inequalities.

## Chapter 17

**Exercise 1.** Is the upper bound in Corollary 1 on the VC-dimension of closed half-spaces in $\mathbb{R}^n$ tight?

**Exercise 2.** Show that there are exactly $\binom{n+k}{n}$ monomials of degree at most $k$ in $\mathbb{R}^n$.

Hint: Try an induction proof over the dimension $n$.

**Exercise 3.** Use the *lifting map*

$$l : \mathbb{R}^n \to \mathbb{R}^{n+1}, x \mapsto (x, \|x\|^2)$$

for deriving an upper bound for the VC-dimension of closed balls in $\mathbb{R}^n$, i.e., derive an upper bound for the VC-dimension of the range spaces $\big(\mathbb{R}^n, R\big)$, where

$$R = \big\{\{x \in \mathbb{R}^n \,:\, \|x - c\| \leq r\} \,:\, c \in \mathbb{R}^n,\, r \geq 0\big\}.$$

Hint: What is the image of a closed ball in $\mathbb{R}^n$ under the lifting map.

**Exercise 4.** Is the upper bound in Exercise 3 sharp?

# Chapter 17

# Supplemental material

## Hoeffding's Inequality

We start with the elementary *Markov Inequality* for non-negative random variables.

**Lemma 6. [Markov Inequality]** *Let $X$ be a non-negative random variable, i.e., $\mathcal{X} \subseteq [0, \infty)$. Then we have for any $t > 0$ that*

$$\mathsf{P}[X \geq t] \leq \frac{\mathsf{E}[X]}{t}.$$

*Proof.* We have

$$\begin{aligned}
\mathsf{E}[X] &= \int_{\mathcal{X}} xp(x)\,dx \\
&= \int_{\mathcal{X}} x \cdot \big(\mathbf{1}[x < t] + \mathbf{1}[x \geq t]\big)\, p(x)dx \\
&= \int_{\mathcal{X}} x \cdot \mathbf{1}[x < t]\, p(x)dx + \int_{\mathcal{X}} x \cdot \mathbf{1}[x \geq t]\, p(x)dx \\
&\geq \int_{\mathcal{X}} x \cdot \mathbf{1}[x \geq t]\, p(x)dx \\
&\geq \int_{\mathcal{X}} t \cdot \mathbf{1}[x \geq t]\, p(x)dx \\
&= t \cdot \int_{\mathcal{X}} \mathbf{1}[x \geq t]\, p(x)dx \\
&= t \cdot \mathsf{P}[X \geq t]
\end{aligned}$$

□

From Markov's Inequality we also get inequalities for a real valued random variable $X$, i.e., $\mathcal{X} \subseteq \mathbb{R}$. An immediate consequence is Chebyshev's Inequality

$$\mathsf{P}\big[|X - \mathsf{E}[X]| \geq t\big] \;=\; \mathsf{P}\big[(X - \mathsf{E}[X])^2 \geq t^2\big] \;\leq\; \frac{\mathsf{E}\big[\big(X - \mathsf{E}[X]\big)^2\big]}{t^2} \;=\; \frac{\mathsf{Var}[X]}{t^2}.$$

The moments of a real valued random variable $X$ are the powers

$$1 = X^0,\, X^1,\, X^2,\, X^3,\, \ldots.$$

Hence, Markov's Inequality is a bound that uses the first moment of $X$, while Chebyshev's Inequality is a bound that uses the first two moments of $X$. If we apply Markov's Inequality to the *moment generating function* $\exp(sX)$, $s > 0$ of $X$, then we have by the monotonicity of the exponential function that

$$\mathsf{P}[X \geq t] \;=\; \mathsf{P}\big[\exp(sX) \geq \exp(st)\big] \;\leq\; \frac{\mathsf{E}[\exp(sX)]}{\exp(st)},$$

which is known as *Chernoff's Bound*. Hence, for bounding the probability that $X$ is larger than $t$ we would like to have a good upper bound on $\mathsf{E}[\exp(sX)]$. Hoeffding's Lemma provides such a bound.

**Lemma 7. [Hoeffding]** *Let $X$ be a real valued random variable with $\mathcal{X} = [a,b] \subset \mathbb{R}$ and $\mathsf{E}[X] = 0$. Then we have for every $s \geq 0$ that*

$$\mathsf{E}[\exp(sX)] \;\leq\; \exp\big(s^2(b-a)^2/8\big).$$

*Proof.* Any $x \in [a,b]$ can be written as

$$\begin{aligned} x \;&=\; \frac{(b-a)x}{b-a} \;=\; \frac{(b-a)x - ba + ab}{b-a} \\ &=\; \frac{b(x-a) + a(b-x)}{b-a} \;=\; \frac{x-a}{b-a}\,b + \frac{b-x}{b-a}\,a. \end{aligned}$$

Note that

$$\frac{(x-a)}{b-a},\; \frac{(b-x)}{b-a} \;\geq\; 0 \quad \text{and} \quad \frac{(x-a)}{b-a} + \frac{(b-x)}{b-a} \;=\; 1.$$

From the convexity of the exponential function we get that

$$\begin{aligned} \exp(sx) \;&=\; \exp\left(\frac{x-a}{b-a}sb + \frac{b-x}{b-a}sa\right) \\ &\leq\; \frac{x-a}{b-a}\exp(sb) \;+\; \frac{b-x}{b-a}\exp(sa). \end{aligned}$$

Hence, we get using the definition $t = -a/(b-a)$ that

$$\begin{aligned}
\mathsf{E}\big[\exp(sX)\big] \;&\leq\; \frac{\mathsf{E}[X]-a}{b-a}\exp(sb) \;+\; \frac{b-\mathsf{E}[X]}{b-a}\exp(sa)\\
&=\; \frac{-a}{b-a}\exp(sb) \;+\; \frac{b}{b-a}\exp(sa)\\
&=\; t\exp(sb) + (1-t)\exp(sa)\\
&=\; t\exp\big(s(1-t)(b-a)\big) + (1-t)\exp\big(-st(b-a)\big)\\
&=\; t\exp\big(s(b-a)\big)\exp\big(-st(b-a)\big) + (1-t)\exp\big(-st(b-a)\big)\\
&=\; \Big(1-t+t\exp\big(s(b-a)\big)\Big)\exp\big(-st(b-a)\big)\\
&=:\; \exp\big(-tu + \log\big(1-t+t\exp(u)\big) \;=:\; \exp\big(f(u)\big),
\end{aligned}$$

where $u = s(b-a) \geq 0$. We have $f(0) = 0$ and

$$f'(u) \;:=\; \frac{df}{du} \;=\; -t + \frac{t\exp(u)}{1-t+t\exp(u)} \;=\; -t + \frac{t}{(1-t)\exp(-u)+t},$$

and thus also $f'(0) = 0$. Furthermore,

$$f''(u) \;:=\; \frac{df'}{du} \;=\; \frac{t(1-t)\exp(-u)}{\big((1-t)\exp(-u)+t\big)^2} \;\leq\; \frac{1}{4},$$

where the last inequality is equivalent to $\big((1-t)\exp(-u)-t\big)^2 \geq 0$. Hence, we get from a Taylor approximation for some $\bar{u} \in [0,u]$ that

$$f(u) \;=\; f(0) + uf'(0) + \frac{u^2}{2}f''(\bar{u}) \;=\; \frac{u^2}{2}f''(\bar{u}) \;\leq\; \frac{u^2}{8} \;=\; \frac{s^2(b-a)^2}{8}.$$

□

Now we have everything in place to state and prove Hoeffding's Inequality that bounds the probability that a sum of independent, real valued random variables deviates substantially from its expectation.

**Theorem 11.** *Let $X_i$, $i \in [n]$ be independent random variables with $\mathcal{X}_i = [a_i, b_i] \subset \mathbb{R}$, and let $S_n = \sum_{i=1}^n X_i$. Then we have for any $t > 0$ that*

$$\mathsf{P}\big[S_n - \mathsf{E}[S_n] \geq t\big] \;\leq\; \exp\left(-\frac{2t^2}{\sum_{i=1}^n (b_i-a_i)^2}\right)$$

*and*

$$\mathsf{P}\big[S_n - \mathsf{E}[S_n] \leq -t\big] \;\leq\; \exp\left(-\frac{2t^2}{\sum_{i=1}^n (b_i-a_i)^2}\right).$$

*Proof.* Using Chernoff's Bound, the independence assumption, and Hoeffding's Lemma, we get that

$$\begin{aligned}
\mathsf{P}\big[S_n - \mathsf{E}[S_n] \geq t\big] &\leq \frac{\mathsf{E}\Big[\exp\big(s(S_n - \mathsf{E}[S_n])\big)\Big]}{\exp(st)} \\
&= \exp(-st)\,\mathsf{E}\left[\exp\left(s\sum_{i=1}^{n}(X_i - \mathsf{E}[X_i])\right)\right] \\
&= \exp(-st)\,\mathsf{E}\left[\prod_{i=1}^{n}\exp\big(s(X_i - \mathsf{E}[X_i])\big)\right] \\
&= \exp(-st)\prod_{i=1}^{n}\mathsf{E}\Big[\exp\big(s(X_i - \mathsf{E}[X_i])\big)\Big] \\
&\leq \exp(-st)\prod_{i=1}^{n}\exp\big(s^2(b_i - a_i)^2/8\big) \\
&= \exp(-st)\exp\left(\frac{s^2}{8}\sum_{i=1}^{n}(b_i - a_i)^2\right) \\
&= \exp\left(-\frac{2t^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right),
\end{aligned}$$

if we choose

$$s = \frac{4t}{\sum_{i=1}^{n}(b_i - a_i)^2}.$$

The second inequality follows analogously, because the sample space of $-X_i$ is $-\mathcal{X}_i = [-b_i, -a_i]$ and the interval $[-b_i, -a_i]$ has again the length $-a_i - (-b_i) = b_i - a_i$. □

*Remark:* Using a union bound we immediately get from Theorem 11 that

$$\mathsf{P}\Big[\big|S_n - \mathsf{E}[S_n]\big| \geq t\Big] \leq 2\exp\left(-\frac{2t^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right).$$

# Part IV

# Geometry in Learning

# Chapter 18

# Support vector machines

Support vector machines (SVMs) are a linear binary Euclidean classification method that are derived directly from statistical learning theory, more specifically from Rademacher complexity theory. SVMs have a nice geometric interpretation that provides additional insights into their statistical performance. Before we can derive SVMs from a so called uniform margin bound we need some more results from Rademacher theory that we have to develop first.

### Relating empirical and expected Rademacher complexities

Similarly, as with the expected and empirical loss, we only have access to the empirical Rademacher complexity, but would like to relate it to the expected Rademacher complexity. Our working horse for relating the expected and the empirical loss in terms of the VC-dimension of the hypothesis class, or more specifically its associated growth function, in Lemma 5 was Hoeffding's Inequality. For relating the corresponding Rademacher complexities we use McDiarmid's Inequality which is an extension of Hoeffding's Inequality (see the supplemental material).

**Lemma 8.** *Let $H \subseteq \mathcal{Y}^{\mathcal{X}}$, where $\mathcal{Y} = [0, 1]$, be a hypothesis class. Then we have for every probability density function $p$ on $\mathcal{X}$, every $\delta \in (0, 1)$ and every $m \geq 1$ that*

$$\mathsf{P}\left[\left\{S \in \mathcal{X}^m \,:\, R_m(H) - R_S(H) \leq \sqrt{\frac{\log(1/\delta)}{2m}}\right\}\right] \geq 1 - \delta.$$

*Proof.* It is enough to show that

$$\mathsf{P}\left[\left\{S \in \mathcal{X}^m \,:\, R_m(H) - R_S(H) \geq \sqrt{\frac{\log(1/\delta)}{2m}}\right\}\right] \leq \delta.$$

Let

$$f(x^{(1)},\dots,x^{(m)}) \;=\; R_S(H) \;=\; \mathsf{E}_\sigma\left[\mathsf{sup}_{h\in H}\left|\frac{1}{m}\sum_{i=1}^m \sigma_i h(x^{(i)})\right|\right],$$

then $f$ satisfies the bounded difference assumption with $c_i = 1/m$, $i\in[m]$. Hence, by McDiarmid's Inequality

$$\mathsf{P}\Big[\big\{(x^{(i)})_{i\in[m]}\in\mathcal{X}^m \;:\; \mathsf{E}[f]-f(x^{(1)},\dots,x^{(m)})\geq t\big\}\Big] \;\leq\; \exp\left(-\frac{2t^2}{\sum_{i=1}^m c_i^2}\right).$$

Using $c_i = 1/m$ for $i\in[m]$, setting $\delta =\; \exp(-2mt^2)$ and solving for $t$ gives $t=\sqrt{\frac{\log(1/\delta)}{2m}}$, from which the assertion of the lemma follows. □

The lemma will be used in the proof of the following theorem, whose proof is similar to the proof of Lemma 5, where Rademacher averages already appeared implicitly.

**Theorem 12.** *Let $H\subseteq\mathcal{Y}^\mathcal{X}$, where $\mathcal{Y}=[0,1]$, be a hypothesis class. Then we have for every probability distribution $p$ on $\mathcal{X}$, every $h\in H$, every $\delta\in(0,1)$ and every $m\geq 1$ that*

$$P\left[\left\{S\in\mathcal{X}^m \;:\; E[h]\leq\frac{1}{m}\sum_{x\in S}h(x)+2R_S(H)+3\sqrt{\frac{\log(2/\delta)}{2m}}\right\}\right] \;\geq\; 1-\delta.$$

*Here, $E[h]$ is short for $E[h(X)]$, where $X$ is the random variable that takes values in $\mathcal{X}$ whose distribution is $p$.*

*Proof.* Obviously, we have for every $h\in H$ and sequence $S=\big(x^{(i)}\big)_{i\in[m]}$ of sample points in $\mathcal{X}$ that

$$\mathsf{E}[h] \;\leq\; \frac{1}{m}\sum_{x\in S}h(x) \;+\; \mathsf{sup}_{g\in H}\left\{\mathsf{E}[g]-\frac{1}{m}\sum_{i=1}^m g(x^{(i)})\right\}.$$

Hence, it remains to upper bound

$$\mathsf{sup}_{g\in H}\left\{\mathsf{E}[g]-\frac{1}{m}\sum_{i=1}^m g(x^{(i)})\right\}.$$

From McDiarmid's Inequality we get with probability at least $1-\delta/2$ that

$$\begin{aligned}\mathsf{sup}_{g\in H}\left\{\mathsf{E}[g]-\frac{1}{m}\sum_{i=1}^m g(x^{(i)})\right\} \;\leq\; &\mathsf{E}_S\left[\mathsf{sup}_{g\in H}\left\{\mathsf{E}[g]-\frac{1}{m}\sum_{i=1}^m g(x^{(i)})\right\}\right]\\ &+\sqrt{\frac{\log(2/\delta)}{2m}}\end{aligned}$$

by setting $\delta = 2\exp(-2mt^2)$ and thus $t = \sqrt{\frac{\log(2/\delta)}{2m}}$, because

$$\mathcal{X}^m \ni (x^{(1)}, \ldots, x^{(m)}) \;\mapsto\; \mathsf{sup}_{g\in H} \left\{ \mathsf{E}[g] - \frac{1}{m}\sum_{i=1}^{m} g(x^{(i)}) \right\}$$

satisfies the bounded difference assumption with $c_i = 1/m$, $i \in [m]$. We still need to upper bound

$$\mathsf{E}_S \left[ \mathsf{sup}_{g\in H} \left\{ \mathsf{E}[g] - \frac{1}{m}\sum_{i=1}^{m} g(x^{(i)}) \right\} \right].$$

Let $\bar{S} = \big(\bar{x}^{(i)}\big)_{i\in[m]}$ another sequence of i.i.d. sample points, then we have

$$\mathsf{E}[g] \;=\; \mathsf{E}_{S'} \left[ \frac{1}{m}\sum_{i=1}^{m} g(\bar{x}^{(i)}) \right]$$

and obtain

$$\begin{aligned}
&\mathsf{E}_S \left[ \mathsf{sup}_{g\in H} \left\{ \mathsf{E}[g] - \frac{1}{m}\sum_{i=1}^{m} g(x^{(i)}) \right\} \right] \\
&\quad= \mathsf{E}_S \left[ \mathsf{sup}_{g\in H} \left\{ \mathsf{E}_{S'} \left[ \frac{1}{m}\sum_{i=1}^{m} g(\bar{x}^{(i)}) \right] - \frac{1}{m}\sum_{i=1}^{m} g(x^{(i)}) \right\} \right] \\
&\quad= \mathsf{E}_S \left[ \mathsf{sup}_{g\in H} \left\{ \mathsf{E}_{S'} \left[ \frac{1}{m}\sum_{i=1}^{m} g(\bar{x}^{(i)}) - \frac{1}{m}\sum_{i=1}^{m} g(x^{(i)}) \right] \right\} \right] \\
&\quad\le \mathsf{E}_S \left[ \mathsf{E}_{S'} \left[ \mathsf{sup}_{g\in H} \left\{ \frac{1}{m}\sum_{i=1}^{m} g(\bar{x}^{(i)}) - \frac{1}{m}\sum_{i=1}^{m} g(x^{(i)}) \right\} \right] \right] \\
&\quad= \mathsf{E}_\sigma \left[ \mathsf{E}_S \left[ \mathsf{E}_{S'} \left[ \mathsf{sup}_{g\in H} \left\{ \frac{1}{m}\sum_{i=1}^{m} \sigma_i g(\bar{x}^{(i)}) - \frac{1}{m}\sum_{i=1}^{m} \sigma_i g(x^{(i)}) \right\} \right] \right] \right] \\
&\quad\le \mathsf{E}_\sigma \left[ \mathsf{E}_S \left[ \mathsf{E}_{S'} \left[ \mathsf{sup}_{g\in H} \left\{ \Big| \frac{1}{m}\sum_{i=1}^{m} \sigma_i g(\bar{x}^{(i)}) \Big| + \Big| \frac{1}{m}\sum_{i=1}^{m} \sigma_i g(x^{(i)}) \Big| \right\} \right] \right] \right] \\
&\quad\le \mathsf{E}_\sigma \left[ \mathsf{E}_S \left[ \mathsf{sup}_{g\in H} \Big| \frac{2}{m}\sum_{i=1}^{m} \sigma_i g(x^{(i)}) \Big| \right] \right] \\
&\quad= 2\cdot \mathsf{E}_S \left[ \mathsf{E}_\sigma \left[ \mathsf{sup}_{g\in H} \Big| \frac{1}{m}\sum_{i=1}^{m} \sigma_i g(x^{(i)}) \Big| \right] \right] \\
&\quad= 2\cdot \mathsf{E}_S \big[ R_S(H) \big] \;=\; 2R_m(H),
\end{aligned}$$

where the $\sigma_i$, $i \in [m]$ are drawn independently from the uniform distribution on $\{\pm 1\}$. By Lemma 8 we have

$$R_m(H) \;\le\; R_S(H) + \sqrt{\frac{\log(2/\delta)}{2m}}$$

with probability at least $1 - \delta/2$. Hence, the assertion of the theorem follows from a union bound. □

### Margins and margin bounds

We can use real valued functions in $\mathbb{R}^{\mathcal{X}}$ also for binary classification problems with $\mathcal{Y} = \{\pm 1\}$ by considering the sign of the function values. Let $H \subseteq \mathbb{R}^{\mathcal{X}}$ be a symmetric hypothesis class, i.e., $h \in H$ implies $-h \in H$, and let $(x^{(1)}, y^{(1)}), \ldots, (x^{(m)}, y^{(m)})$ be sample points from $\mathcal{X} \times \mathcal{Y}$. Using Theorem 12, we can bound the misclassification probability for any $h \in H$ and any probability distribution $p$ on $\mathcal{X} \times \{\pm 1\}$ with probability at least $1 - \delta$ as follows

$$\begin{aligned}\mathsf{P}\Big[\big\{(x,y) \in \mathcal{X} \times \{\pm 1\} \,:\, y \neq \mathsf{sign}(h(x))\big\}\Big] \;&=\; \mathsf{E}\big[\mathbf{1}[(-Yh(X))]\big] \\ &\le\; \frac{1}{m}\sum_{i=1}^{m} \mathbf{1}\big[-y^{(i)}h(x^{(i)})\big] + 2R_S(H) + 3\sqrt{\frac{\log(2/\delta)}{2m}}\end{aligned}$$

where we have used the symmetry of the hypothesis class $H$ for the empirical Rademacher term $2R_S(H)$.

The upper bound consists of three terms, but only the first term

$$\frac{1}{m}\sum_{i=1}^{m} \mathbf{1}\big[-y^{(i)}h(x^{(i)})\big]$$

depends on a specific hypothesis $h \in H$. This term just counts the number of sample points that are misclassified by $h$. Hence, if we want to choose a hypothesis that minimizes the upper bound on the misclassification probability, then we have to find $h \in H$ that misclassifies the least number of sample points. Unfortunately, finding such a hypothesis is a *hard optimization* problem in general.

In order to obtain tractable optimization problems we can weaken the upper bound on the misclassification problem by approximating the non-continuous indicator functions $\mathbf{1}\big[-y^{(i)}h(x^{(i)})\big]$ by simple piecewise linear functions. The piecewise linear approximations are based on the concept of *margin*. For $h \in H$ and $(x, y) \in \mathcal{X} \times \mathcal{Y}$ we call $yh(x)$ the *margin* of $h$ at $(x, y)$. Note that data points $(x, y)$ that are close to the decision boundary $\{x \in \mathcal{X} \,:\, h(x) = 0\}$ of the classifier $h$ have a small margin. Let $\gamma > 0$ be a *prescribed margin* $\gamma > 0$, then the *slack variable*

$$\xi(h) \,=\, \mathsf{max}\big\{\gamma - yh(x),\, 0\big\}, \quad i \in [m],\, h \in H$$

is a measure of how the classifier $h \in H$ fails to achieve the margin $\gamma$. That is, for $0 < \xi(h) < \gamma$ the point $x$ is correctly classified by $h$, i.e., $h(x)$ has the correct sign, but fails to achieve the margin $\gamma$. For $\xi(h) \geq \gamma$ the data point $x$ is misclassified by $h$, and for $\xi(h) = 0$ the classifier $h$ achieves the margin $\gamma$ at $x$.

We have the following uniform prescribed margin dependent generalization bound.

**Theorem 13.** *Let $H \subseteq \mathbb{R}^{\mathcal{X}}$ be a symmetric hypothesis class and let $\gamma > 0$ a prescribed margin. Then we have for every probability distribution $p$ on $\mathcal{X} \times \{\pm 1\}$, every $h \in H$, every $\delta \in (0, 1)$ and every $m \geq 1$ that*

$$\begin{aligned} \mathsf{P}\Bigg[\Bigg\{ & S = \big((x^{(i)}, y^{(i)})\big)_{i\in[m]} \in (\mathcal{X} \times \{\pm 1\})^m \;: \\ & \mathsf{P}\Big[\big\{(x, y) \in \mathcal{X} \times \{\pm 1\} \,:\, \mathsf{sign}(h(x)) \neq y\big\}\Big] \\ & \qquad \leq \frac{1}{m\gamma}\sum_{i=1}^{m} \xi_i(h) + \frac{2}{\gamma} R_S(H) + 3\sqrt{\frac{\log(2/\delta)}{2m}} \Bigg\}\Bigg] \,\geq\, 1 - \delta, \end{aligned}$$

*where $\xi_i(h)$ is the slack variable for $h$ at $(x^{(i)}, y^{(i)})$.*

*Proof.* The function

$$f : \mathbb{R} \to [0, 1],\, x \mapsto \begin{cases} 1 & : \quad x > 0 \\ 1 + x/\gamma & : \quad -\gamma \leq x \leq 0 \\ 0 & : \quad x < -\gamma. \end{cases}$$

is a piecewise linear upper bound of the indicator function $\mathbf{1}[x \geq 0]$. It follows that with probability at least $1 - \delta$,

$$\begin{aligned} \mathsf{E}\big[\mathbf{1}[-Yh(X)]\big] \,&\leq\, \mathsf{E}\big[f(-Yh(X))\big] \\ &\leq\, \frac{1}{m}\sum_{i=1}^{m} f\big(-y^{(i)}h(x^{(i)})\big) + 2R_S(f \circ H) + 3\sqrt{\frac{\log(2/\delta)}{2m}} \\ &\leq\, \frac{1}{m}\sum_{i=1}^{m} f\big(-y^{(i)}h(x^{(i)})\big) + \frac{2}{\gamma}R_S(H) + 3\sqrt{\frac{\log(2/\delta)}{2m}}, \end{aligned}$$

where we have used Theorem 12 in the second inequality that only holds with probability $1-\delta$, and the so called *Contraction Lemma* for Rademacher complexities in the last inequality that is, like the first inequality, non-probabilistic. The Contraction Lemma states that

$$R_S(\phi \circ H) \leq c R_S(H)$$

for any Lipschitz continuous function $\phi : \mathbb{R} \to \mathbb{R}$ with Lipschitz constant $c > 0$. Remember that $\phi$ is Lipschitz continuous with constant $c$, if we have for every $x, x' \in \mathbb{R}$ that $|\phi(x) - \phi(x')| \leq c|x - x'|$. The function $f$ is Lipschitz continuous with Lipschitz constant $1/\gamma$.
The claim of the theorem now follows from

$$f\big(-y^{(i)}h(x^{(i)})\big) \;\leq\; \frac{\xi_i(h)}{\gamma},$$

which can be checked in a simple case distinction:

1. If $-y^{(i)}h(x^{(i)}) > 0$, then

$$f\big(-y^{(i)}h(x^{(i)})\big) \;=\; 1 \;\leq\; \frac{\gamma - y^{(i)}h(x^{(i)})}{\gamma} \;=\; \frac{\xi_i(h)}{\gamma}.$$

2. If $-\gamma \leq -y^{(i)}h(x^{(i)}) \leq 0$, then

$$f\big(-y^{(i)}h(x^{(i)})\big) \;=\; 1 + \frac{-y^{(i)}h(x^{(i)})}{\gamma} \;=\; \frac{\gamma - y^{(i)}h(x^{(i)})}{\gamma} \;=\; \frac{\xi_i(h)}{\gamma}.$$

3. If $-y^{(i)}h(x^{(i)}) \leq -\gamma$, then

$$f\big(-y^{(i)}h(x^{(i)})\big) \;=\; 0 \;=\; \xi_i(h) \;=\; \frac{\xi_i(h)}{\gamma}. \qquad \square$$

*Remark:* The condition in Theorem 13 that $H$ is a symmetric hypothesis class is not really important. It just allows us to write the Rademacher complexity term that appears in the upper bound in a simple form. We are mostly interested in the fist term that explicitly depends on $h$ and that we want to minimize by choosing an appropriate $h \in H$ in the following.

### Support vector machines

Support vector machines are derived by minimizing the upper bound on the misclassification probability in Theorem 13 for the following hypothesis class of linear functions

$$H \;=\; \big\{x \mapsto \theta^\top x + \theta_0 \;:\; \theta \in \mathbb{R}^n,\, \|\theta\|^2 = 1,\, \theta_0 \in \mathbb{R}\big\} \;\subset\; \mathbb{R}^{\mathbb{R}^n}.$$

*Remark:* $H$ is essentially the class of linear functions, except that we restrict the parameter vector $\theta$ to the unit sphere instead of the set of non-zero vectors. The rationale behind this restriction is that for any $\lambda > 0$ the decision boundary that results from the choice of $\lambda\theta \in \mathbb{R}^n \setminus \{0\}$ and $\lambda\theta_0 \in \mathbb{R}$ is independent of $\lambda$. Hence, it is enough to consider $\theta$ with $\|\theta\| = 1$.

Assume that we are given a sample $\big((x^{(i)}, y^{(i)})\big)_{i\in[m]}$ that has been drawn i.i.d. from an unknown probability distribution on $\mathbb{R}^n \times \{\pm 1\}$. Naturally, in order to minimize the bound in Theorem 13, one should pick the classifier $h \in H$ that minimizes the slack

$$\begin{aligned}\sum_{i=1}^m \xi_i(h) \;&=\; \sum_{i=1}^m \mathsf{max}\big\{\gamma - y^{(i)}h(x^{(i)}),\, 0\big\}\\ &=\; \sum_{i=1}^m \mathsf{max}\big\{\gamma - y^{(i)}(\theta^\top x^{(i)} + \theta_0),\, 0\big\}.\end{aligned}$$

That is, we need to solve the following optimization problem

$$(\hat\theta, \hat\theta_0) \;=\; \mathsf{argmin}_{\theta,\theta_0} \sum_{i=1}^m \mathsf{max}\big\{\gamma - y^{(i)}(\theta^\top x^{(i)} + \theta_0),\, 0\big\} \quad \text{subject to} \quad \|\theta\|^2 \;=\; 1.$$

This optimization problem still looks hard because of of the non-differentiable $\max$-function. In the following we transform the problem into an equivalent differentiable form, actually, into a *convex quadratic program* that can be solved efficiently. We can fix the prescribed margin at $1$ and scale $\theta_0$ and the norm of $\theta$ accordingly by $1/\gamma$. That is,

$$\begin{aligned}(\hat\theta, \hat\theta_0) \;=\; &\mathsf{argmin}_{\theta,\theta_0} \sum_{i=1}^m \mathsf{max}\big\{1 - y^{(i)}(\theta^\top x^{(i)} + \theta_0/\gamma),\, 0\big\}\\ &\text{subject to} \quad \|\theta\|^2 \;=\; 1/\gamma^2.\end{aligned}$$

By introducing a vector of slack variables $\xi \in \mathbb{R}^m$, we can further rewrite the optimization problem as

$$\begin{aligned}
\text{min}_{\theta,\theta_0,\xi} \quad & \sum_{i=1}^m \xi_i \\
\text{subject to} \quad & y^{(i)}(\theta^\top x^{(i)} + \theta_0/\gamma) \geq 1 - \xi_i \text{ and } \xi_i \geq 0,\ i \in [m] \\
& \|\theta\|^2 = 1/\gamma^2.
\end{aligned}$$

In this problem formulation the objective function and the constraint functions are differentiable. Hence, we can use that a local optimal solution of an equality constrained optimization problem satisfies that the gradient of the objective function is a linear combination of the gradients of the equality constraints. Using the Lagrange multiplier theorem, we get the following equivalent optimization problem

$$\begin{aligned}
\text{min}_{\theta,\theta_0,\xi} \quad & \sum_{i=1}^m \xi_i + \frac{\|\theta\|^2}{2c_\gamma} \\
\text{subject to} \quad & y^{(i)}(\theta^\top x^{(i)} + \theta_0/\gamma) \geq 1 - \xi_i \text{ and } \xi_i \geq 0,\ i \in [m],
\end{aligned}$$

where $1/2c_\gamma$ is the so called Lagrange multiplier (written in a complicated form) that depends on $\gamma$. This problem is almost the standard *soft-margin support vector machine* (SVM). In the standard soft-margin SVM the dependency on the prescribed margin $\gamma$ is dropped, i.e., the offset $\theta_0$ is not scaled by $1/\gamma$ and the regularization parameter (Lagrange multiplier) $c$ does no longer depend on $\gamma$. Furthermore, the regularization parameter $c$ is placed in front of the loss term instead of the $\ell_2$-norm regularization term. Hence, the standard soft-margin SVM problem reads as

$$\begin{aligned}
\text{min}_{\theta,\theta_0,\xi} \quad & \frac{1}{2}\|\theta\|^2 + c\sum_{i=1}^m \xi_i \\
\text{subject to} \quad & y^{(i)}(\theta^\top x^{(i)} + \theta_0) \geq 1 - \xi_i \text{ and } \xi_i \geq 0,\ i \in [m].
\end{aligned}$$

In the literature you also find the equivalent non-differentiable problem formulation

$$(\hat{\theta}, \hat{\theta}_0) = \text{argmin}_{\theta,\theta_0} \frac{1}{2}\|\theta\|^2 + c\sum_{i=1}^m \text{max}\{1 - y^{(i)}(\theta^\top x^{(i)} + \theta_0),\, 0\}.$$

Note that here, in contrast to the logistic regression problem (cf. Chapter 3), the regularization term that renders the optimization problem strictly convex enters naturally through the specification of the hypothesis class, i.e., the class of linear functions.

## Geometric interpretation

Support vector machines have a nice geometric interpretation that also provides some intuition about their statistical properties. We start our discussion of the geometry of support vector machines with the linearly separable case. Assume that the sample $\big((x^{(i)}, y^{(i)})\big)_{i\in[m]}$ is linearly separable. Then there is a feasible but not necessarily optimal solution, where all the variables $\xi_i$, $i \in [m]$ are zero. If we restrict our search for an optimal solution for $\theta$ and $\theta_0$ to hyperplanes with zero slack, then we can drop the variables $\xi_i$ and need to solve only the following *hard-margin support vector machine*

$$\begin{aligned} &\text{min}_{\theta,b} \quad \frac{1}{2}\|\theta\|^2 \\ &\text{subject to} \quad y^{(i)}(\theta^\top x^{(i)} + \theta_0) \,\geq\, 1, \quad i \in [m]. \end{aligned}$$

For obtaining insight into the geometry of hard-margin SVMs, let

$$P \,=\, \{i \in [m] \,:\, y^{(i)} = 1\} \quad \text{and} \quad Q \,=\, \{i \in [m] \,:\, y^{(i)} = -1\},$$

and let $\theta \in \mathbb{R}^n \setminus \{0\}$ and $\theta_0 \in \mathbb{R}$ be the parameters of a hyperplane that separates the points $x^{(i)}$, $i \in P$ from the points $x^{(i)}$, $i \in Q$. That is, without loss of generality we can assume that

$$\theta^\top x^{(i)} + \theta_0 \,>\, 0 \text{ for } i \in P, \text{ and } \theta^\top x^{(i)} + \theta_0 \,<\, 0 \text{ for } i \in Q.$$

We have $d = \text{min}\big\{|\theta^\top x^{(i)} + \theta_0| \,:\, i \in [m]\big\} > 0$ and

$$\theta^\top x^{(i)} + \theta_0 \,\geq\, d \text{ for } i \in P, \text{ and } \theta^\top x^{(i)} + \theta_0 \,\leq\, -d \text{ for } i \in Q.$$

If we scale $\theta$ and $\theta_0$ by $1/d$, then we get

$$\theta^\top x^{(i)} + \theta_0 \,\geq\, 1 \text{ for } i \in P, \text{ and } \theta^\top x^{(i)} + \theta_0 \,\leq\, -1 \text{ for } i \in Q.$$

The *geometric margin* of the hyperplane

$$H_{\theta,\theta_0} \,=\, \big\{x \in \mathbb{R}^n \,:\, \theta^\top x + \theta_0 = 0\big\}$$

is the maximum distance that $H$ can be shifted either in the direction of $\theta$ or in the direction of $-\theta$, before hitting one of the sample points $x^{(i)}$, $i \in [m]$. This distance is exactly the distance of the hyperplanes

$$H_P \,=\, \big\{x \in \mathbb{R}^n \,:\, \theta^\top x + \theta_0 = 1\big\} \quad \text{and} \quad H_Q \,=\, \big\{x \in \mathbb{R}^n \,:\, \theta^\top x + \theta_0 = -1\big\}$$

which is $d(H_P, H_Q) = 2/\|\theta\|$. Hence, we get the parameters of the separating hyperplane that maximizes the geometric margin as the solution of the following optimization problem

$$\begin{array}{ll} \text{max}_{\theta,\theta_0} & 2/\|\theta\| \\ \text{subject to} & y^{(i)}(\theta^\top x^{(i)} + \theta_0) \geq 1, \quad i \in [m], \end{array}$$

which is equivalent to the hard-margin SVM, i.e., instead of maximizing $2/\|\theta\|$ we can also minimize $\frac{1}{2}\|\theta\|^2$ over the same set of constraints. Hence, the hard-margin SVM maximizes the geometric margin. For a better understanding of the geometry of the soft-margin SVM it helps to look at the dual SVM problem that we discuss in the next chapter.

# Chapter 19

# Dual support vector machines

We have formulated soft- and hard-margin SVMs as constrained, convex optimization problems. For such problems there exists a well established duality theory with far reaching implications and applications. Here, we provide a short introduction into Lagrangian duality theory for constrained, convex optimization problems. Later we apply this theory to SVMs and use it for obtaining some insight into their geometry.

## Lagrangian duality

We consider optimization problems of the form

$$\min_{x\in\mathbb{R}^n} f(x) \quad \text{subject to} \quad c_i(x) \leq 0,\ i \in [m],$$

where $f, c_i : \mathbb{R}^n \to \mathbb{R}$ are convex and continuously differentiable. The *Lagrangian* of this problem is the function

$$L : \mathbb{R}^n \times \mathbb{R}^m_{\geq 0} \to \mathbb{R},\ (x, a) \mapsto f(x) + \sum_{i=1}^{m} a_i c_i(x).$$

Let $(\hat{x}, \hat{a})$ be a saddle point of the Lagrangian, i.e., we have

$$L(\hat{x}, a) \leq L(\hat{x}, \hat{a}) \leq L(x, \hat{a}) \text{ for all } x \in \mathbb{R}^n,\ a \in \mathbb{R}^m_{\geq 0}.$$

Note that such a saddle point does not need to exist, but if it exists then we have the following lemma.

**Lemma 9.** *Let $(\hat{x}, \hat{a})$ be a saddle point of the Lagrangian $L$. Then we have*

1. $L(\hat{x}, \hat{a}) = f(\hat{x})$.

2. *$f(\hat{x}) \leq f(x)$ for all feasible $x \in \mathbb{R}^n$.*

3. *$\hat{x}$ is a feasible solution of the constrained optimization problem.*

*Proof.* 1. We need to show that $\hat{a}_i c_i(\hat{x}) = 0$ for $i \in [m]$. Assume that $\hat{a}_i c_i(\hat{x}) \neq 0$. Then we have $\hat{a}_i > 0$. If $c_i(\hat{x}) < 0$, then we can decrease $\hat{a}_i$ to make $L(\hat{x}, \hat{a})$ larger, and if $c_i(\hat{x}) > 0$, then we can increase $\hat{a}_i$ to make $L(\hat{x}, \hat{a})$ larger. This is a contradiction to the saddle point condition, which asserts that

$$\mathsf{max}_{a \geq 0} L(\hat{x}, a) \;=\; L(\hat{x}, \hat{a}).$$

2. We have

$$f(\hat{x}) \;=\; L(\hat{x}, \hat{a}) \;\leq\; L(x, \hat{a}) \;=\; f(x) + \sum_{i=1}^{m} a_i c_i(x) \;\leq\; f(x),$$

where the last inequality follows from $a_i \geq 0,\, i \in [m]$ and the feasibility of $x$, i.e., we have $c_i(x) \leq 0$ for all $i \in [m]$.

3. From the first assertion we know that $\hat{a}_i c_i(\hat{x}) = 0$ for all $i \in [m]$. Assume that $c_i(\hat{x}) > 0$, which renders $\hat{x}$ infeasible. Then we can increase $\hat{a}_i$, which by 1. has to be zero, to make $L(\hat{x}, \hat{a})$ larger. This is again a contradiction to the saddle point condition, see the first assertion. □

From the saddle point condition we also get

$$\begin{aligned}
\mathsf{max}_{a \in \mathbb{R}^m_{\geq 0}} \mathsf{min}_{x \in \mathbb{R}^n} L(x, a) \;&\leq\; \mathsf{max}_{a \geq 0} L(\hat{x}, a) \\
&=\; L(\hat{x}, \hat{a}) \\
&=\; \mathsf{min}_{x \in \mathbb{R}^n} L(x, \hat{a}) \\
&\leq\; \mathsf{max}_{a \in \mathbb{R}^m_{\geq 0}} \mathsf{min}_{x \in \mathbb{R}^n} L(x, a),
\end{aligned}$$

which implies

$$\mathsf{max}_{a \geq 0} \mathsf{min}_{x \in \mathbb{R}^n} L(x, a) \;=\; L(\hat{x}, \hat{a}) \;=\; f(\hat{x}).$$

Hence, the *dual optimization problem*

$$\mathsf{max}_{a \in \mathbb{R}^m} \mathsf{min}_{x \in \mathbb{R}^n} L(x, a) \quad \text{subject to} \quad a \;\geq\; 0$$

has the same optimal value as the *primal problem*. The Lagrangian is by our assumptions a convex, differentiable function. Thus we have

$$\hat{x} \;=\; \mathsf{argmin}_{x \in \mathbb{R}^n} \; L(x, a) \quad \text{if and only if} \quad \nabla_x L(x, a)|_{x=\hat{x}} \;=\; 0,$$

and get the following equivalent formulation of the dual problem

$$\mathsf{max}_{a \in \mathbb{R}^m} L(x, a) \quad \text{subject to} \quad a \geq 0 \;\text{ and }\; \nabla_x L(x, a)|_{x=\hat{x}} \;=\; 0.$$

### Dual hard-margin SVM

Let us start our discussion of dual SVMs with the hard-margin SVM. The Lagrangian for the hard-margin SVM reads as

$$L(\theta, \theta_0, a) \;=\; \frac{1}{2}\|\theta\|^2 - \sum_{i=1}^{m} a_i\big(y^{(i)}(\theta^\top x^{(i)} + \theta_0) - 1\big).$$

We get from the saddle point condition that

$$\nabla_\theta L(\theta, \theta_0) \;=\; \theta - \sum_{i=1}^{m} a_i y^{(i)} x^{(i)} \;=\; 0$$

and

$$\frac{dL(\theta, \theta_0)}{d\theta_0} \;=\; -\sum_{i=1}^{m} a_i y^{(i)} \;=\; 0.$$

The first condition can also be written as $\theta = \sum_{i=1}^{m} a_i y^{(i)} x^{(i)}$. Plugging both conditions back into the Lagrangian gives the following form for the dual problem

$$\begin{array}{ll} \text{max}_{a\in\mathbb{R}^m} & -\dfrac{1}{2}\displaystyle\sum_{i,j=1}^{m} a_i a_j y^{(i)} y^{(j)} {x^{(i)}}^\top x^{(j)} \;+\; \displaystyle\sum_{i=1}^{m} a_i \\ \text{subject to} & \displaystyle\sum_{i=1}^{m} a_i y^{(i)} \;=\; 0 \;\text{ and }\; a_i \;\geq\; 0,\, i \in [m]. \end{array}$$

It remains to establish the existence of a saddle point. The existence follows from the Karush-Kuhn-Tucker condition that we state here without a proof.

**Theorem 14. [Karush-Kuhn-Tucker]** *Given an optimization problem*

$$\textit{min}_{x\in\mathbb{R}^n}\, f(x) \quad \textit{subject to} \quad c_i(x) \,\leq\, 0,\, i \in [m],$$

*where $f, c_i : \mathbb{R}^n \to \mathbb{R}$ are convex and continuously differentiable. Then $\hat{x} \in \mathbb{R}^n$ is an optimal solution of the primal optimization problem, if and only if there exists $\hat{a} \in \mathbb{R}^m_{\geq 0}$ such that $(\hat{x}, \hat{a})$ is a saddle point of the Lagrangian associated with the optimization problem.*

The Karush-Kuhn-Tucker (KKT) Theorem guarantees the existence of a saddle point, if the optimization problem has an optimal solution. Hence, for establishing the existence of a saddle point for the hard-margin SVM it is enough to establish the existence of an optimal solution. The existence of an optimal solution follows from the existence of a feasible solution which itself follows from our assumption that the point sets $\{x^{(i)} \,:\, i \in P\}$ and $\{x^{(i)} \,:\, i \in Q\}$ are linearly separable. Let $(\bar{\theta}, \bar{\theta}_0)$ be a feasible solution, then we only need to search for an optimal solution in the set

$$C \;=\; \big\{\theta \in \mathbb{R}^n \;:\; \|\theta\| \le \|\bar{\theta}\|\big\} \times \big\{\theta_0 \;:\; |\theta_0| \le d\|\bar{\theta}\|\big\},$$

where $d = \max\big\{\|x^{(i)}\| \;:\; i \in [m]\big\}$. Note that any hyperplane that separates the point sets $\{x^{(i)} \,:\, i \in P\}$ and $\{x^{(i)} \,:\, i \in Q\}$ has at most distance $d$ from the origin. Hence, we have

$$\frac{|\theta_0|}{\|\theta\|} \;\le\; d, \;\; \text{or equivalently} \;\; |\theta_0| \;\le\; d\|\theta\| \;\le\; d\|\bar{\theta}\|.$$

Since $C$ is compact and the objective function $\frac{1}{2}\|\theta\|^2$ of the hard-margin SVM is continuous, there exists an optimal solution of the hard-margin SVM within $C$.

*Remark:* Now that we have established the existence of a saddle point for hard margin SVM, we can use Lemma 9 to recover the optimal solution of the optimal solution $\hat{a}$ of the dual hard-margin SVM. The saddle point condition implies $\nabla_\theta L(\theta, \theta_0) = 0$ and thus

$$\hat{\theta} \;=\; \sum_{i=1}^{m} \hat{a}_i y^{(i)} x^{(i)}$$

is optimal for the primal hard-margin SVM.
The optimal $\hat{\theta}_0$ can be obtained from the first assertion of Lemma 9 that implies

$$\hat{a}_i\big(y^{(i)}(\hat{\theta}^\top x^{(i)} + \hat{\theta}_0) - 1\big) \;=\; 0 \;\text{ for all }\; i \in [m].$$

If $\hat{a}_i \neq 0$, then we must have

$$y^{(i)}(\hat{\theta}^\top x^{(i)} + \hat{\theta}_0) - 1 \;=\; 0,$$

from which we get by using ${y^{(i)}}^2 = 1$ that $\hat{\theta}_0 = y^{(i)} - \hat{\theta}^\top x^{(i)}$.

## Polytope distance problem

The *polytope distance problem* asks to compute the distance between conv$(P)$ and conv$(Q)$, where

$$\text{conv}(P) \;=\; \left\{ \sum_{i:y^{(i)}=1} a_i x^{(i)} \;:\; \sum_{i:y^{(i)}=1} a_i = 1 \,\wedge\, \forall i : a_i \ge 0 \right\}$$

and

$$\mathsf{conv}(Q) \;=\; \left\{ \sum_{i:y^{(i)}=-1} a_i x^{(i)} \;:\; \sum_{i:y^{(i)}=-1} a_i = 1 \,\wedge\, \forall i:\, a_i \geq 0 \right\}.$$

We get, if we minimize the quadratic distance instead of the distance

$$\begin{aligned} \frac{1}{2}\mathsf{dist}\big(\mathsf{conv}(P), \mathsf{conv}(Q)\big)^2 \;&=\; \mathsf{min}_{p\in P, q\in Q} \quad \frac{1}{2}\|p-q\|^2 \\ &=\; \mathsf{min}_{a\in\mathbb{R}^m_{\geq 0}} \quad \frac{1}{2}\sum_{i,j=1}^{m} a_i a_j y^{(i)} y^{(j)} {x^{(i)}}^{\top} x^{(j)} \\ &\qquad\qquad \text{subject to} \;\; \sum_{i:y^{(i)}=1} a_i \;=\; 1 \;=\; \sum_{i:y^{(i)}=-1} a_i, \end{aligned}$$

where we have used that

$$\begin{aligned} \|p-q\| \;&=\; \Big\| \sum_{i:y^{(i)}=1} a_i x^{(i)} - \sum_{i:y^{(i)}=-1} a_i x^{(i)} \Big\|^2 \\ &=\; \Big\| \sum_{i:y^{(i)}=1} a_i y^{(i)} x^{(i)} + \sum_{i:y^{(i)}=-1} a_i y^{(i)} x^{(i)} \Big\|^2 \\ &=\; \Big\| \sum_{i=1}^{m} a_i y^{(i)} x^{(i)} \Big\|^2 \\ &=\; \sum_{i,j=1}^{m} a_i a_j y^{(i)} y^{(j)} {x^{(i)}}^{\top} x^{(j)}. \end{aligned}$$

It turns out that the polytope distance problem is equivalent to the dual hard-margin SVM.

**Theorem 15.** *The dual hard-margin SVM and the polytope distance problem are equivalent, i.e., their optimal solutions only differ by scaling.*

*Proof.* We need to show that we can compute by scaling an optimal solution for the polytope distance problem from an optimal solution for the dual hard-margin SVM and vice versa.
The dual hard-margin SVM can be written as

$$\begin{aligned} &\mathsf{min}_{a\in\mathbb{R}^m_{\geq 0}} \quad \frac{1}{2}\sum_{i,j=1}^{m} a_i a_j y^{(i)} y^{(j)} {x^{(i)}}^{\top} x^{(j)} \;-\; \sum_{i=1}^{m} a_i \\ &\text{subject to} \quad \sum_{i:y^{(i)}=1} a_i \;=\; \sum_{i:y^{(i)}=-1} a_i. \end{aligned}$$

Hence, a feasible solution to polytope distance problem is also feasible for the dual hard-margin SVM, but not the other way around. That is, the feasible region for the hard-margin SVM contains the feasible region for the polytope distance problem.
Let $\bar{a}$ be the optimal solution for the polytope distance problem and let $\hat{a}$ be the optimal solution for the dual hard-margin SVM. Let

$$s = \sum_{i:y^{(i)}=1} \hat{a}_i = \sum_{i:y^{(i)}=-1} \hat{a}_i.$$

We have $s > 0$, because otherwise all the $\hat{a}_i$, $i \in [m]$ must be zero and thus also the optimal value of the objective function of the dual hard margin SVM is zero. This is not possible, because optimal value of the dual hard margin SVM is the same as the optimal value of the hard margin SVM. If the latter value is zero, then we have for the optimal solution $\hat{\theta} = 0$. It is easy to see that for $\hat{\theta} = 0$ the constraints of the hard margin SVM cannot be satisfied. If we scale $\hat{a}$ by $1/s$, then $\hat{a}$ becomes feasible for the polytope distance problem. We claim that the scaled $\hat{a}$ is optimal for the polytope distance problem. For a contradiction assume that this is not the case, then we have

$$\frac{1}{2}\sum_{i,j=1}^{m} \bar{a}_i\bar{a}_j y^{(i)}y^{(j)}{x^{(i)}}^\top x^{(j)} < \frac{1}{2s^2}\sum_{i,j=1}^{m} \hat{a}_i\hat{a}_j y^{(i)}y^{(j)}{x^{(i)}}^\top x^{(j)},$$

from which we get

$$\begin{aligned} \frac{1}{2}\sum_{i,j=1}^{m} \hat{a}_i\hat{a}_j y^{(i)}y^{(j)}{x^{(i)}}^\top x^{(j)} - \sum_{i=1}^{m}\hat{a}_i \\ &= \frac{1}{2}\sum_{i,j=1}^{m} \hat{a}_i\hat{a}_j y^{(i)}y^{(j)}{x^{(i)}}^\top x^{(j)} - 2s \\ &> \frac{s^2}{2}\sum_{i,j=1}^{m} \hat{a}_i\bar{a}_j y^{(i)}y^{(j)}{x^{(i)}}^\top x^{(j)} - 2s \\ &= \frac{s^2}{2}\sum_{i,j=1}^{m} \bar{a}_i\bar{a}_j y^{(i)}y^{(j)}{x^{(i)}}^\top x^{(j)} - s\sum_{i=1}^{m}\bar{a}_i. \end{aligned}$$

Since $\bar{a}$ scaled by $s$ is also feasible for the dual hard-margin SVM the latter inequality contradicts the minimality of $\hat{a}$ for the dual hard-margin SVM. Hence, $\hat{a}$ scaled by $1/s$ is optimal for the polytope distance problem. Thus, we get get an optimal solution for polytope distance problem from an optimal solution for the hard-margin SVM by scaling the latter by $1/s$.

We also have that scaling $\bar{a}$ by $s > 0$ is feasible and optimal for the dual hard-margin SVM. Note though that we do not know the scaling factor $s$, if we only have the optimal solution for the polytope distance problem. Thus, it remains to compute the scaling factor from the optimal solution for the polytope distance problem. Let $t > 0$ be a scaling factor, then $t\bar{a}$ is feasible for the dual hard-margin SVM and we get for the value of the objective function as a function of $t$ that

$$f(t) \;=\; \frac{t^2}{2}\sum_{i,j=1}^{m} \bar{a}_i\bar{a}_j y^{(i)}y^{(j)}{x^{(i)}}^\top x^{(j)} \;-\; t\sum_{i=1}^{m}\bar{a}_i.$$

Minimizing $f(t)$ with respect to $t$ gives the optimal scaling factor

$$s \;=\; \frac{\sum_{i=1}^{m}\bar{a}_i}{\sum_{i,j=1}^{m}\bar{a}_i\bar{a}_j y^{(i)}y^{(j)}{x^{(i)}}^\top x^{(j)}} \;=\; \frac{2}{\mathsf{dist}\big(\mathsf{conv}(P),\mathsf{conv}(Q)\big)^2},$$

and thus $s\bar{a}$ is optimal for the dual hard-margin SVM. □

## Dual soft-margin SVM

The Langrangian for the soft-margin SVM reads as

$$L(\theta,\theta_0,\xi,a,b) \;=\; \frac{1}{2}\|\theta\|^2+c\cdot\sum_{i=1}^{m}\xi_i-\sum_{i=1}^{m}a_i\big(y^{(i)}(\theta^\top x^{(i)}+\theta_0)-1+\xi_i\big)-\sum_{i=1}^{m}b_i\xi_i,$$

where $a \in \mathbb{R}^m_{\geq 0}$ and $b \in \mathbb{R}^m_{\geq 0}$. We get from the saddle point condition that

$$\begin{aligned}
\nabla_\theta L(\theta,\theta_0,\xi) \;&=\; \theta - \sum_{i=1}^{m} a_i y^{(i)}x^{(i)} \;=\; 0\\
\nabla_\xi L(\theta,\theta_0,\xi) \;&=\; c\cdot\mathbf{1} - a - b \;=\; 0\\
\frac{dL(\theta,\theta_0,\xi)}{d\theta_0} \;&=\; -\sum_{i=1}^{m} a_i y^{(i)} \;=\; 0.
\end{aligned}$$

The second condition and the condition that $b \geq 0$ can be combined into

$$c - a_i \;=\; b_i \;\geq\; 0,\;\; \text{or, equivalently}\;\; a_i \;\leq\; c,\;\; \text{for all}\;\; i\in[m].$$

Plugging all three conditions back into the Lagrangian and eliminating $b$ by the argument from above gives the following form for the dual problem

$$\begin{array}{ll} \max_{a\in\mathbb{R}^m} & -\dfrac{1}{2}\displaystyle\sum_{i,j=1}^{m} a_i a_j y^{(i)} y^{(j)} {x^{(i)}}^\top x^{(j)} + \sum_{i=1}^{m} a_i \\ \text{subject to} & \displaystyle\sum_{i=1}^{m} a_i y^{(i)} = 0 \text{ and } 0 \leq a_i \leq c,\, i \in [m]. \end{array}$$

As in the case of dual hard margin SVMs, if we scale an optimal solution $\bar{a} \in \mathbb{R}^m_{\geq 0}$ of the dual soft-margin SVM by $s = \sum_{i:y^{(i)}=1} \bar{a}_i = \sum_{i:y^{(i)}=-1} \bar{a}_i$, then we get an optimal solution to the following *reduced polytope distance problem*

$$\begin{array}{ll} \min_{a\in\mathbb{R}^m} & \dfrac{1}{2}\displaystyle\sum_{i,j=1}^{m} a_i a_j y^{(i)} y^{(j)} {x^{(i)}}^\top x^{(j)} \\ \text{subject to} & \displaystyle\sum_{i:y^{(i)}=1} a_i = 1 = \sum_{i:y^{(i)}=1} a_i \text{ and } 0 \leq a_i \leq \bar{c},\, i \in [m], \end{array}$$

where $\bar{c} = c/s$. In the reduced polytope distance problem we are computing the distance between the reduced convex hulls

$$\mathrm{conv}_{\bar{c}}(P) = \left\{ \sum_{i:y^{(i)}=1} a_i x^{(i)} \; : \; \sum_{i:y^{(i)}=1} a_i = 1 \wedge \forall i : 0 \leq a_i \leq \bar{c} \right\}$$

and

$$\mathrm{conv}_{\bar{c}}(Q) = \left\{ \sum_{j:y^{(i)}=-1} a_j x^{(j)} \; : \; \sum_{j:y^{(i)}=-1} a_j = 1 \wedge \forall i : 0 \leq a_j \leq \bar{c} \right\}.$$

Obviously, the convex hulls are only reduced, if $\bar{c} < 1$. In this case the convex hulls shrink away from its extreme points, i.e., the vertices on the respective boundaries of the convex hulls. The reduced convex hulls remain convex polytopes as long as $\bar{c} \geq \max\{1/|P|, 1/|Q|\}$. Thus, the reduced polytope distance problem becomes infeasible if $\bar{c} < \max\{1/|P|, 1/|Q|\}$. The proof of Theorem 15 shows that the reduced polytope distance problem for the given data points and choice of $\bar{c} = c/s$ is feasible as long as the dual soft-margin SVM is feasible.

In non-degenerate situations, i.e., in situations that are combinatorially stable under small perturbations of the data points, the optimal solution $\hat{a}$ of the polytope distance problem has at most $n + 1$ non-zero entries. That is, the number of non-zeroes is independent of the number $m$ of data points. The data points that correspond to the non-zero entries are called *support vectors*. The support vectors are extreme points of the polytopes. It is also true for the reduced polytope distance problem that the distance is determined by at most $n+1$ extreme points of the reduced polytopes (reduced convex hulls), however, these extreme points are now non-trivial convex combinations of the data points. This is reflected in the fact that the solution to the reduced polytope distance problems has more than $n+1$ non-zero entries. This renders the solution statistically more robust, because if there are more data points on which the solution depends, then it is less likely that these data points are non-representative for the respective distributions.

# Chapter 20

# Polytope distance problems

The polytope distance problem naturally has a sparse solution in the sense that the number of (data) points that determine the solution is small compared to the total number of data points, if the number data points is large compared to the dimension. Sparse solutions are also of interest in regression problems, especially in the high-dimensional setting where the number $n$ of dimensions (features) is large compared to the number $m$ of data points.

Features in an OLS regression problem, see Chapter 1, can be deselected by setting the corresponding entries in the parameter vector $\theta$ to zero. The direct feature selection approach is to bound the number of non-zero entries in $\theta$ by $k < n$ and search for the best solution among all parameter vectors with at most $k$ non-zero entries. Unfortunately, this is approach is computationally infeasible already for moderately small values of $k$ as there are already $\binom{n}{k}$ possibilities to place $k$ zeroes among the $n$ entries of $\theta$.
The direct feature selection approach can be reformulated as the following regularized optimization problem

$$\text{argmin}_{\theta\in\mathbb{R}^n}\ \frac{1}{2}\|y - X^\top\theta\|^2 + \lambda\|\theta\|_0,$$

where $\|\theta\|_0$ is the number of non-zero entries in $\theta$ and $\lambda > 0$ a regularization parameter. Intuitively, the vector $\hat{\theta}$ will be more sparse, i.e., have more zero-entries, when $\lambda$ is large. One should be careful to note here that $\|\cdot\|_0$ is not a norm since it is not homogeneous. Also, since $\|\cdot\|_0$ is by its nature not amenable to numerical methods, solving this regularized problem basically amounts to the feature selection approach from above. A computationally feasible alternative is replacing the $\ell_0$-regularization term by a $\ell_1$-regularization term. The resulting optimization problem is called the LASSO (least absolute shrinkage and selection operator). The effect of the $\ell_1$-regularization term is twofold: some entries of $\theta$ are set to zero (deselected), which is exactly what we want, but also the remaining entries of $\theta$ are shrunken towards zero similarly as in ridge regression. Statistically, the LASSO behaves similarly as ridge regression: It is a biased estimator, but for small values of the regularization parameter $\lambda$ the expected quadratic error of the LASSO estimate is smaller than the generalization error of the maximum likelihood OLS estimate. The regularization term restricts the total, absolute size of the parameters and by that reduces the variance of the estimator.

In the following we focus feature selection capability of the LASSO and an extension that is called the *elastic net*. We show that both problems can be formulated as a polytope distance problems, where one of the polytopes is just a single point, namely the origin. The feature selection capabilities are then a direct consequence of the sparsity of polytope distance problem solutions.

## The LASSO

The LASSO is given as the following optimization problem

$$\text{min}_{\theta\in\mathbb{R}^n}\ \frac{1}{2}\|y - X^\top\theta\|^2 + \lambda\|\theta\|_1,$$

where $X \in \mathbb{R}^{n\times m}$ is a centered and standardized data matrix, $y \in \mathbb{R}^m$ is a centered label vector, $\lambda > 0$ is a regularization parameter, and $\theta \in \mathbb{R}^n$ is the parameter vector that we want to estimate.

The $\ell_1$-norm is not differentiable, but the LASSO can be transformed into a constrained, differentiable problem by splitting the parameter vector $\theta \in \mathbb{R}^n$ into a positive and a negative part, i.e., $\theta = \theta^+ - \theta^-$, where $\theta^+, \theta^- \in \mathbb{R}^n_{\geq 0}$. Using the parameter vector $\theta^+$ and $\theta^-$ the LASSO becomes

$$\begin{aligned} &\text{min}_{\theta\in\mathbb{R}^n} \quad \frac{1}{2}\|y - X^\top(\theta^+ - \theta^-)\|^2 + \lambda\sum_{i=1}^n \left(\theta_i^+ + \theta_i^-\right) \\ &\text{subject to} \quad \theta^+,\, \theta^- \geq 0. \end{aligned}$$

Now that we have a formulation with differentiable objective function and differentiable constraints we can invoke the optimality condition (saddle point condition for the Lagrangian) and replace the regularization term in the objective function by another constraint. We get

$$\begin{aligned} &\min_{\theta\in\mathbb{R}^n} && \frac{1}{2}\|y - X^\top(\theta^+ - \theta^-)\|^2 \\ &\text{subject to} && \sum_{i=1}^{n}\left(\theta_i^+ + \theta_i^-\right) \le t \;\text{ and }\; \theta^+,\, \theta^- \ge 0, \end{aligned}$$

where $t > 0$ is function of the regularization parameter $\lambda$. If we want we can now undo the splitting of the parameter vector $\theta$ into a positive and a negative part. Doing so results in the equivalent constrained formulation of the LASSO for the original parameter vector $\theta$ that reads as

$$\min_{\theta\in\mathbb{R}^n} \frac{1}{2}\|y - X^\top\theta\|^2 \quad \text{subject to } \|\theta\|_1 \le t.$$

If we scale the data matrix $X$ by $t$, then we obtain the following reformulation of the constrained version of the LASSO

$$\min_{\theta\in\mathbb{R}^n} \frac{1}{2}\left\|y - t\cdot X^\top\theta\right\|^2 \quad \text{subject to } \|\theta\|_1 \le 1.$$

The unit $\ell_1$-ball is polytope, also known as the cross polytope $C_n$. The vertices of the cross polytope are $\{\pm e_i \,:\, i \in [n]\}$. Since any polytope can be represented as the convex hull of its vertices, we can parameterize the cross polytope over the unit simplex $\Delta_{2n-1}$ as

$$C_n = \left\{(\mathbb{1}_n, -\mathbb{1}_n)a \,:\, a \in \Delta_{2n-1}\right\}.$$

That is, any $\theta \in C_n$ has a representation as $(\mathbb{1}_n, -\mathbb{1}_n)a$ with $a \in \Delta_{2n-1}$. Hence, the LASSO can be rewritten as

$$\min_{a\in\mathbb{R}^{2n}} \frac{1}{2}\left\|y + (-t\cdot X^\top, t\cdot X^\top)a\right\|^2 \quad \text{subject to } a \in \Delta_{2n-1}.$$

If we define $Z = y\mathbf{1}^\top + (-t\cdot X^\top, t\cdot X^\top) \in \mathbb{R}^{m\times 2n}$, then the problem becomes

$$\min_{a\in\mathbb{R}^{2n}} \frac{1}{2}\|Za\|^2 \quad \text{subject to } a \in \Delta_{2n-1},$$

since we have $y\mathbf{1}^\top a = y$ for $a \in \Delta_{2n-1}$.

The geometric interpretation of this problem is as follows: We are looking for the distance to the origin of the polytope in $\mathbb{R}^m$ that is given as the convex hull of the rows of the scaled data matrix $t \cdot X$ and the rows of the scaled and negated data matrix $-t \cdot X$ shifted each by the label vector $y$. We assume that $y \neq 0$.

If $m \geq n$, that is we have more data points than features/dimensions, then the dimension of the polytope is only $n$ since the rows of $X$ and $-X$ together span only an $n$-dimensional linear subspace of $\mathbb{R}^m$. In this case, if $y \neq 0$, then the distance of the polytope to the origin is, in general, determined by at most $n$ vertices of the polytope. That is, the optimal $\hat{a}$ has at most $n$ non-zero entries, and thus also the optimal parameter vector $\hat{\theta}$ has at most $n$ non-zeroes, because $\hat{\theta}_i = \hat{a}_i - \hat{a}_{n+i}$, $i \in [n]$.
The more interesting case is $n > m$ when we have more features than data points. If the origin is not contained in the polytope, which is the case if $t$ small enough, then the distance is determined by at most $m$ vertices of the polytope and thus the optimal $\hat{a} \in \Delta_{2n-1}$ has at most $m$ non-zero entries. Hence, in this case, the optimal parameter vector $\hat{\theta}$ can also have at most $m$ non-zero entries, because $\hat{\theta}_i = \hat{a}_i - \hat{a}_{n+i}$, $i \in [n]$. Note that, when $t$ goes to zero, then the distance of the polytope to the origin converges to $\|y\|$ since the polytope shrinks to the point $y$.
If $t$ is so large or, equivalently, $\lambda$ is so small that the polytope contains the origin, then distance of the polytope to the origin becomes zero. Let $\bar{x}^{(i)} \in \mathbb{R}^m, i \in [n]$ be the rows of the data matrix $X$. We have

$$y + t \cdot \sum_{i=1}^{n} (\hat{a}_{n+i}\bar{x}^{(i)} - \hat{a}_i\bar{x}^{(i)}) \;=\; 0,$$

which is, as expected, equivalent to

$$y \;=\; t \cdot \sum_{i=1}^{n} (\hat{a}_i - \hat{a}_{n+i})\bar{x}^{(i)} \;=\; t \cdot \sum_{i=1}^{n} \hat{\theta}_i \bar{x}^{(i)} \;=\; t \cdot X^\top \theta.$$

*Remark:* The Representer Theorem does, unfortunately, not hold for $\ell_1$-regularized problems. Also, note that adding a $\ell_1$-regularization term of the form $\lambda\|a\|_1$ to the adjoint problem, where $\lambda > 0$ is a regularization parameter, does not perform feature selection but data point selection. Nevertheless, feature selection is possible in the adjoint formulation, though not in its kernelized variant. The Gram matrix $X^\top X$ can also be written as

$$X^\top X \;=\; \sum_{j=1}^{n} \Psi_j \Psi_j^\top,$$

where $\Psi_j$ is the $j$-th row of the matrix $X$. That is, $X^\top X$ has been written as a sum of $n$ rank-1 matrices $\Psi_j \Psi_j^T$ that correspond to the dimensions (features) that are index by $j \in [n]$. Now let $w \in [0,1]^n$ be a weight vector that weights the $j$-th dimension by $0 \leq w_j \leq 1$. Let $X^{(w)} \in \mathbb{R}^{n\times m}$ be the matrix of weighted data points, i.e., the $i$-the column of $X^{(w)}$ is the componentwise product $w \odot x^{(i)} \in \mathbb{R}^n$ of the weight vector $w$ and the $i$-th data point $x^{(i)}$. The weighted data matrix $X^{(w)}$ can also be written as the matrix product $X^{(w)} = WX$, where $W = \text{diag}(w) \in \mathbb{R}^{n\times n}$ is the diagonal matrix whose diagonal is the weight vector $w$. Hence, we have

$$X^{(w)^\top} X^{(w)} = \sum_{j=1}^{m} w_j^2 \Psi_j \Psi_j^T.$$

Sparse solutions can now be promoted by adding an $\ell_1$-regularization term for the weight vector $w$ which results in the following feature selective, adjoint problem

$$(\hat{w}, \hat{a}) = \\ \text{argmin}_{w\in[0,1]^n,\, a\in\mathbb{R}^m} \quad L(a^\top X^\top W^2 X, y) + c \cdot a^\top X^\top W^2 X a + \lambda \|w\|_1.$$

## The elastic net

The LASSO has a two main limitations. Firstly, it deselects features (dimensions) to aggressively. For $n > m$ the LASSO always selects at most only $m$ features, i.e., $\theta$ has at least $n - m$ zeroes, and secondly, for a group of highly correlated features the LASSO tends to select only one feature from the group. The elastic net, that in a sense combines the LASSO and ridge regression, has been proposed to overcome these limitations. The elastic net is given as

$$\text{min}_{\theta\in\mathbb{R}^{n+1}} \frac{1}{2}\|y - X^\top \theta\|^2 + \frac{c}{2}\|\theta\|^2 + \lambda\|\theta\|_1,$$

where $c > 0$ is another regularization parameter. Using the same reasoning and notation as for the LASSO from above, we can rewrite the elastic net problem as

$$\text{min}_{a\in\mathbb{R}^{2n}} \frac{1}{2}\|Za\|^2 + \frac{ct^2}{2}\big\|(\mathbb{1}_n, -\mathbb{1}_n)a\big\|^2 \quad \text{subject to } \; a \in \Delta_{2n-1}.$$

The two terms of the objective function can be merged into a single term as follows

$$\begin{aligned}
&\frac{1}{2}\|Za\|^2 + \frac{ct^2}{2}\big\|(\mathbb{1}_n, -\mathbb{1}_n)a\big\|^2 \\
&\quad= \frac{1}{2}a^\top Z^\top Za + \frac{ct^2}{2}a^\top \begin{pmatrix} \mathbb{1}_n \\ -\mathbb{1}_n \end{pmatrix} (\mathbb{1}_n, -\mathbb{1}_n)a \\
&\quad= \frac{1}{2}a^\top \left(Z^\top Z + ct^2 \begin{pmatrix} \mathbb{1}_n \\ -\mathbb{1}_n \end{pmatrix} (\mathbb{1}_n, -\mathbb{1}_n)\right) a \\
&\quad= \frac{1}{2}a^\top \left(\begin{pmatrix} Z \\ (t\sqrt{c}\,\mathbb{1}_n, -t\sqrt{c}\,\mathbb{1}_n) \end{pmatrix}^\top \begin{pmatrix} Z \\ (t\sqrt{c}\,\mathbb{1}_n, -t\sqrt{c}\,\mathbb{1}_n) \end{pmatrix}\right) a \\
&\quad= \frac{1}{2}a^\top W^\top Wa \;=\; \frac{1}{2}\|Wa\|^2,
\end{aligned}$$

where

$$W \;=\; \begin{pmatrix} Z \\ (t\sqrt{c}\,\mathbb{1}_n, -t\sqrt{c}\,\mathbb{1}_n) \end{pmatrix} \in \mathbb{R}^{(m+n)\times 2n}.$$

Using this reformulation of the objective function allows to write the elastic net problem as

$$\min_{a\in\mathbb{R}^{2n}} \frac{1}{2}\|Wa\|^2 \quad \text{subject to } \; a \in \Delta_{2n-1}.$$

In this problem we compute the distance of a polytope in $\mathbb{R}^{m+n}$ that is given as the convex hull of the $2n$ rows of the matrix $W$. If the origin is not contained in the polytope, the distance is determined by at most $\min\{m, n\} + n$ points and thus, in this case, the optimal $\hat{a} \in \Delta_{2n-1}$ has at most $\min\{m, n\} + n$ non-zero entries. That is, the parameter vector $\hat{\theta}$, where $\hat{\theta}_i = \hat{a}_i - \hat{a}_{n+i}$, $i \in [n]$, can have $n$ zeroes, even in the case that $m < n$. Hence, the sparsity of $\hat{\theta}$ is not restricted, although it tends to be sparse because of the symmetries in $W$. Especially, if $\hat{a}_i = \hat{a}_{n+i}$, then the $(m+i)$-th component in the convex combination of the columns vanishes and thus does not contribute the distance of this convex combination to the origin. If $\hat{a}_i = \hat{a}_{n+i}$, then $\hat{\theta}_i = 0$. This mitigates the first shortcoming of the LASSO, namely, its strict bound on the number of non-zeroes in $\hat{\theta}$, in the case that $m < n$.

Let us now discuss the second shortcoming of the LASSO, namely, not selecting all highly-correlated features. Assume that the features are standardized such that their sample variances are $1/m$, that is, $\sum_{j=1}^m x_i^{(j)^2} = 1$ for $i \in [n]$. We call the $i$-th and the $j$-th feature correlated if the sample correlation coefficient

$$\rho_{ij} \;=\; \sum_{l=1}^{m} x_i^{(l)} \cdot x_j^{(l)}$$

is close to $1$. We can prove the following theorem that bounds the difference of the optimal coefficients $\hat{\theta}_i$ and $\hat{\theta}_j$ by the correlation coefficient provided that both features are selected.

**Theorem 16.** *Assume that the features are standardized and that $\hat{\theta}_i \cdot \hat{\theta}_j > 0$ for the optimal solution $\hat{\theta}$ of the elastic net. Then*

$$|\hat{\theta}_i - \hat{\theta}_j| \;\leq\; \frac{\|y\|}{c}\sqrt{2\big(1-\rho_{ij}\big)}.$$

*Proof.* Let

$$L(\theta) \;=\; \frac{1}{2}\|y - X^\top\theta\|^2 + \frac{c}{2}\|\theta\|^2 + \lambda\|\theta\|_1$$

be the objective function of the elastic net. Since $\hat{\theta}_i \cdot \hat{\theta}_j > 0$ we get from the vanishing gradient condition for the optimal parameter vector $\hat{\theta}$ that

$$\begin{aligned} 0 \;=\; \frac{\partial L(\hat{\theta})}{\partial \theta_i} \;&=\; x_i^\top(X^\top\hat{\theta} - y) + c\cdot\hat{\theta}_i + \lambda \mathsf{sign}(\hat{\theta}_i) \\ =\; \frac{\partial L(\hat{\theta})}{\partial \theta_j} \;&=\; x_j^\top(X^\top\hat{\theta} - y) + c\cdot\hat{\theta}_j + \lambda \mathsf{sign}(\hat{\theta}_j), \end{aligned}$$

where $x_i = \big(x_i^{(1)}, \ldots, x_i^{(m)}\big)^\top$ and $x_j = \big(x_j^{(1)}, \ldots, x_j^{(m)}\big)^\top$. By subtracting the second partial derivative from the first and using that $\mathsf{sign}(\hat{\theta}_i) = \mathsf{sign}(\hat{\theta}_j)$ we get

$$\big(x_j - x_i\big)^\top(y - X^\top\hat{\theta}) + c\cdot(\hat{\theta}_i - \hat{\theta}_j) \;=\; 0,$$

or equivalently,

$$\hat{\theta}_i - \hat{\theta}_j \;=\; \frac{1}{c}(x_i - x_j)^\top(X^\top\hat{\theta} - y).$$

Thus, it follows from the Cauchy-Schwarz inequality that

$$|\hat{\theta}_i - \hat{\theta}_j| \;\leq\; \frac{1}{c}\|x_i - x_j\|\cdot\|y - X^\top\hat{\theta}\| \;=\; \frac{1}{c}\sqrt{2\big(1-\rho_{ij}\big)}\cdot\|y - X^\top\hat{\theta}\|,$$

where we have used the definition of sample correlation coefficients and our assumption that the features are normalized Finally, by plugging the zero-vector into the objective function we get

$$\frac{1}{2}\|y - X^\top\hat{\theta}\|^2 + \frac{c}{2}\|\hat{\theta}\|^2 + \lambda\|\hat{\theta}\|_1 \;=\; L(\hat{\theta}) \;\leq\; L(0) \;=\; \frac{1}{2}\|y\|^2,$$

which implies $\|y - X^\top\hat{\theta}\|^2 \leq \|y\|^2$ and thus the claim of the theorem. □

### $\ell_2$-loss SVM

It turns out that both the LASSO and the elastic net problem are equivalent to an $\ell_2$-loss SVM without offset term. Here, we do not show the equivalence, but show only that the dual an $\ell_2$-loss SVM is also a polytope distance problem. The $\ell_2$-loss SVM with margin $\gamma$ and without offset term is the following problem

$$\begin{aligned} \min_{\theta\in\mathbb{R}^n,\xi\in\mathbb{R}^m,\gamma\in\mathbb{R}} \quad & \frac{1}{2}\|\theta\|^2 - \gamma + \frac{c}{2}\sum_{i=1}^m \xi_i^2 \\ \text{subject to} \quad & y^{(i)}\theta^\top x^{(i)} \geq \gamma - \xi_i,\ i \in [m], \end{aligned}$$

where $\left(x^{(i)}, y^{(i)}\right)_{i\in[m]}$ is a sequence of data points in $\mathcal{X} \times \mathcal{Y} = \mathbb{R}^n \times \{\pm 1\}$, and the $\xi_i$, $i \in [m]$ are slack variables. Note that, since slack is now penalized quadratically in the objective function, we no longer need a non-negativity constraint on the slack variables.
Let $z^{(i)} = y^{(i)}x^{(i)}$. The Lagrangian of the $\ell_2$-loss SVM is given as

$$L(\theta,\xi,\gamma,a) \;=\; \frac{1}{2}\|\theta\|^2 - \gamma + \frac{c}{2}\sum_{i=1}^m \xi_i^2 - \sum_{i=1}^m a_i(\theta^\top z^{(i)} - \gamma + \xi_i).$$

The gradients with respect to $\theta$ and $\xi$, respectively, need to vanish at a saddle point of the Lagrangian. Thus, we get from the saddle point condition that

$$\begin{aligned} \nabla_\theta L(\theta,\xi,\gamma,a) \;&=\; \theta - \sum_{i=1}^m a_i z^{(i)} \;=\; 0 \\ \nabla_\xi L(\theta,\xi,\gamma,a) \;&=\; c\xi - a \;=\; 0 \end{aligned}$$

and

$$\frac{dL(\theta,\xi,\gamma,a)}{d\gamma} \;=\; 1 - \sum_{i=1}^m a_i \;=\; 0.$$

Plugging these equations back into the Lagrangian gives

$$\begin{aligned} L(a) \;&=\; \frac{1}{2}a^\top Z^\top Za - \gamma + \frac{c}{2c^2}a^\top a - a^\top Z^\top Za + \gamma - \frac{1}{c}a^\top a \\ &=\; -\frac{1}{2}a^\top Z^\top Za - \frac{1}{2c}a^\top a \\ &=\; -\frac{1}{2}a^\top\Big(Z^\top Z + \frac{1}{c}\mathbb{1}_m\Big)a \\ &=\; -\frac{1}{2}\|Wa\|^2 \end{aligned}$$

where $Z \in \mathbb{R}^{n \times m}$ is the data matrix whose columns are the data points $z^{(i)}$, $i \in [m]$ and

$$W = \begin{pmatrix} Z \\ \frac{1}{\sqrt{c}}\mathbb{1}_m \end{pmatrix} \in \mathbb{R}^{(n+m)\times m}.$$

Hence, the dual $\ell_2$-loss SVM becomes, after negating the Lagrangian, the following polytope distance problem

$$\text{min}_{a\in\mathbb{R}^m} \frac{1}{2}\|Wa\|^2 \quad \text{subject to } \ a \in \Delta_{m-1}.$$

## Gilbert's algorithm

Gilbert's algorithm can be used for approximately solving the polytope distance problems that arise from the LASSO, the elastic net, and the $\ell_2$-loss SVM. It can be extended to handle also the standard hard and soft margin SVMs. Gilbert's algorithm builds on the duality of geometric margin maximization and polytope distance minimization.

Let $P \subset \mathbb{R}^n$ a finite point set and conv$(P)$ its convex hull. In the polytope distance problem we want to compute the point in conv$(P)$ that is closest to the origin. For $x \in$ conv$(P)$ let $f(x) = \|x\|$ be its distance to the origin, and

$$w(x) \;=\; \text{min}_{p\in P}\, \frac{x^\top p}{\|x\|}$$

the geometric margin of the separating hyperplane whose normal is $x$, i.e., the distance between the hyperplane with normal $x$ that passes through the origin and the hyperplane with normal $x$ that touches conv$(P)$ and has the smallest distance to the origin. Note that we always have $f(x) \geq w(x)$. We call

$$g(x) \;=\; f(x) - w(x)$$

the duality gap at $x$. Given $\varepsilon > 0$, we call $x \in$ conv$(P)$ an $\varepsilon$-approximate solution to the polytope distance problem, if $g(x) \leq \varepsilon f(x)$. Gilbert's algorithm can now be stated as follows:

Within the while-loop, Gilbert's algorithm needs to compute

$$x := \text{argmin}_{z\in qx}\; \|z\|$$

the point in the line segment $qx$, i.e., in the convex hull of $q$ and $x$, that is closest to the origin. Note that this is a much simpler one-dimensional polytope distance problem that can be solved analytically.

**Algorithm 1** Gilbert's algorithm for the polytope distance problem

```
Require: finite point set $P \subset \mathbb{R}^n$, $\varepsilon > 0$
Ensure: output is an $\varepsilon$-approximate solution $x \in \text{conv}(P)$
  $x := \text{argmin}_{p \in P}\ \|p\|$
  while $g(x) > \varepsilon f(x)$ do
      $q := \text{argmin}_{p \in P}\ \frac{p^\top x}{\|x\|}$
      $x := \text{argmin}_{z \in xq}\ \|z\|$
  end while
  return $x$
```

**Theorem 17.** *Gilbert's algorithm returns an $\varepsilon$-approximate solution after at most $2\lceil \frac{4E}{\varepsilon} \rceil$ iterations, where $E = d^2/2\rho^2$ is the eccentricity of the polytope conv$(P)$ with diameter $d$ and distance to the origin $\rho$.*

*Proof.* Let $x^{(i)}$ be the solution after the $i$-th iteration of Gilbert's algorithm and let

$$q^{(i)} \;=\; \text{argmin}_{p \in P}\ \frac{p^\top x^{(i)}}{\|x^{(i)}\|} \;\in, P$$

which is computed in the first line of the while-loop during the $i$-th iteration. Note that $q^{(i)} \neq x^{(i-1)}$, because otherwise the duality gap at $x^{(i-1)}$ is zero and thus $x^{(i-1)}$ is the optimal solution to the polytope distance problem and the algorithm stops before the $i$-th iteration. Hence, we also have $x^{(i)} \neq x^{(i-1)}$, because otherwise the projection of the non-trivial segment $x^{(i-1)}q^{(i)}$ onto the segment from $x^{(i-1)}$ to the origin would be empty, in contradiction to the choice of $q^{(i)}$. Note that this also implies that $\|x^{(i)}\| < \|x^{(i-1)}\|$. Finally, since $x^{(0)} = \text{argmin}_{p \in P}\ \|p\|$ by the initialization in the first line of the algorithm, we have $x^{(i)} \neq q^{(i)} \in P$, because otherwise we have a contradiction to the minimality of $x^{(0)} \in P$. Hence, $x^{(i)}$ is a point in the interior of the segment $x^{(i-1)}q^{(i)}$ and thus the vector $x^{(i)}$ is orthogonal to the segment $x^{(i-1)}q^{(i)}$. From these considerations we can conclude that the triangle whose vertices are $x^{(i-1)}, x^{(i)}$ and the origin has a right angle at $x^{(i)}$.
Let $f_i = f(x^{(i)}), w_i = w(x^{(i)})$ and $g_i = f_i - w_i$ be the duality gap. By construction, we have $f_i = f_{i-1}\cos\alpha$, where $\alpha$ is the angle between $x^{(i-1)}$ and $x^{(i)}$ since the triangle whose vertices are $x^{(i-1)}, x^{(i)}$ and the origin has a right angle at $x^{(i)}$. It follows that

$$\begin{aligned} f_{i-1} - f_i \;&=\; f_{i-1}(1 - \cos\alpha) \;\geq\; \frac{f_{i-1}}{2}(1 + \cos\alpha)(1 - \cos\alpha) \\ &=\; \frac{f_{i-1}}{2}(1 - \cos^2\alpha) \;=\; \frac{f_{i-1}}{2}\sin^2\alpha \;=\; \frac{f_{i-1}}{2}\frac{g_{i-1}^2}{\|q^{(i)} - x^{(i-1)}\|^2}, \end{aligned}$$

where we have used a simple geometric argument that shows

$$\sin\alpha \;=\; \frac{f_{i-1} - w_{i-1}}{\|q^{(i)} - x^{(i-1)}\|} \;=\; \frac{g_{i-1}}{\|q^{(i)} - x^{(i-1)}\|}.$$

Let $h_i = f_i - \rho$, then we have $g_i \geq h_i$, because $\rho \geq w_i$. We can compute

$$\begin{aligned} h_{i-1} - h_i \;=\; f_{i-1} - f_i \;\geq\; & \frac{f_{i-1}}{2}\frac{g_{i-1}^2}{\|q^{(i)} - x^{(i-1)}\|^2} \\ \geq\; & \frac{\rho}{2}\frac{g_{i-1}^2}{d^2} \;=\; \frac{1}{4E\rho}g_{i-1}^2 \;\geq\; \frac{1}{4E\rho}h_{i-1}^2. \end{aligned}$$

By setting $\bar{h}_i = \frac{h_i}{4E\rho}$ and $\bar{g}_i = \frac{h_i}{4E\rho}$ we thus get the following inequality

$$\bar{h}_{i-1} - \bar{h}_i \;\geq\; \bar{g}_{i-1}^2 \;\geq\; \bar{h}_{i-1}^2$$

that we are going to exploit in the following.
The first observation is that the inequality implies

$$\bar{h}_i \;\leq\; \bar{h}_{i-1} - \bar{h}_{i-1}^2 \;=\; \bar{h}_{i-1}(1 - \bar{h}_{i-1}) \;\leq\; \frac{\bar{h}_{i-1}}{1 + \bar{h}_{i-1}} \;=\; \left(1 + \frac{1}{\bar{h}_{i-1}}\right)^{-1}$$

If we can show that $\bar{h}_0 \leq \frac{1}{2}$, then we get inductively that

$$\bar{h}_i \;\leq\; \frac{1}{i+2} \qquad \text{and thus} \qquad \bar{h}_i \leq \varepsilon \;\text{ for }\; i \geq \lceil 1/\varepsilon \rceil.$$

For showing that $\bar{h}_0 \leq \frac{1}{2}$ we use again that the triangle whose vertices are $x^{(0)}, x^{(1)}$ and the origin has a right angle at $x^{(1)}$. Thus, we can use Pythagoras' Theorem and get

$$f_0^2 \;=\; \|x^{(0)}\|^2 \;=\; \|x^{(0)} - x^{(1)}\|^2 + \|x^{(1)}\|^2 \;=\; \|x^{(0)} - x^{(1)}\|^2 + f_1^2,$$

or equivalently $f_0^2 - f_1^2 = \|x^{(0)} - x^{(1)}\|^2$. Using this and the inequality from above we can conclude that

$$\begin{aligned} \bar{h}_0^2 \leq \bar{h}_0 - \bar{h}_1 \;=\; & \frac{f_0 - f_1}{4E\rho} \;=\; \frac{1}{4E\rho}\frac{(f_0 - f_1)(f_0 + f_1)}{f_0 + f_1} \;=\; \frac{1}{4E\rho}\frac{f_0^2 - f_1^2}{f_0 + f_1} \\ =\; & \frac{1}{4E\rho}\frac{\|x^{(0)} - x^{(1)}\|^2}{f_0 + f_1} \;\leq\; \frac{1}{4E\rho}\frac{d^2}{2\rho} \;=\; \frac{1}{4E}\frac{d^2}{2\rho^2} \;=\; \frac{E}{4E} \;=\; \frac{1}{4}. \end{aligned}$$

Our next claim, that also follows from the inequality $\bar{h}_i - \bar{h}_{i+1} \geq \bar{g}_i^2$ (for the $(i+1)$-th iteration, see above), is that

$$\bar{g}_i \;\leq\; \varepsilon \qquad \text{for some } \;\lceil 1/\varepsilon \rceil \;\leq\; i \;\leq\; 2\lceil 1/\varepsilon \rceil.$$

For a contradiction, assume that this is not the case. We get

$$\begin{aligned}\bar{h}_{\lceil 1/\varepsilon\rceil} - \bar{h}_{2\lceil 1/\varepsilon\rceil+1} &= \sum_{i=\lceil 1/\varepsilon\rceil}^{2\lceil 1/\varepsilon\rceil} \bar{h}_i - \bar{h}_{i+1} \geq \sum_{i=\lceil 1/\varepsilon\rceil}^{2\lceil 1/\varepsilon\rceil} \bar{g}_i^2 \\ &> \sum_{i=\lceil 1/\varepsilon\rceil}^{2\lceil 1/\varepsilon\rceil} \varepsilon^2 = \lceil 1/\varepsilon\rceil \cdot \varepsilon^2 \geq \varepsilon,\end{aligned}$$

which is a contradiction, because $\bar{h}_{\lceil 1/\varepsilon\rceil} \leq \varepsilon$ and $\bar{h}_{2\lceil 1/\varepsilon\rceil+1} \geq 0$.
The assertion of the theorem, namely

$$g_i \leq \varepsilon\rho \leq \varepsilon f_i$$

for some $i \leq 2\lceil \frac{4E}{\varepsilon}\rceil$, now follows immediately, because the first inequality is equivalent to

$$\bar{g}_i \leq \frac{\varepsilon\rho}{4E\rho} = \frac{\varepsilon}{4E},$$

which is satisfied for some $i \leq 2\lceil \frac{4E}{\varepsilon}\rceil$, and the second inequality is just $f_i \geq \rho$. □

*Remark:* The $\varepsilon$-approximate solution computed by Gilbert's algorithm is a convex combination of at most $2\lceil \frac{4E}{\varepsilon}\rceil \leq |P|$ points in $P$, since the algorithm considers just one additional point from $P$ in each iteration. In the high-dimensional case, when $|P| < n$, then the sparsity of the solution, i.e., the number of non-zero convex coefficients, is at most $2\lceil \frac{4E}{\varepsilon}\rceil \leq |P|$.

# Chapter 21

# Conjoint analysis

Support vector machines can be used also for solving a seemingly unrelated problem in the area of preference learning. Assume that we are given a set $A$ of *options*. In a preference elicitation task we provide respondents with two options $a, b \in A$ and ask them to point out their preferred option. From a sequence of such *choice tasks* we want to learn a *value function*

$$v : A \to \mathbb{R}$$

that we want to use for predicting the outcome of choice tasks as follows

$$a \succeq b, \text{ if } v(a) \geq v(b),$$

where $\preceq$ symbolizes the preference relation. The value function provides us with metric information that can be interpreted as a degree of confidence in our prediction, i.e., we are more confident in our prediction, if the difference $|v(a) - v(b)|$ is large compared to the difference of the value function on other pairs of choice options.

In *conjoint analysis* we make two additional assumptions. First, we assume that the option space $A$ is a Cartesian product of finite sets $A_i, \in i \in [n]$, i.e.,

$$A = A_1 \times \ldots \times A_n.$$

Second, we assume that the value function can be decomposed linearly, i.e.,

$$v = \sum_{i=1}^{n} v_i, \text{ with } v_i : A_i \to \mathbb{R}.$$

The functions $v_i,\, i \in [n]$ are called *partworth functions*. Since the sets $A_i,\, i \in [n]$ are finite, the problem of estimating the value function $v$ from a sequence of choice data boils down to estimating a vector $v \in \mathbb{R}^k$, where

$$k \;=\; \sum_{i=1}^{k} k_i \;\text{ and }\; k_i \;=\; |A_i|.$$

For $a = (a_1, \ldots, a_n) \in A$ the values $v_i(a_i),\, i \in [n]$ are called the *partworth values*. The name conjoint analysis stresses the fact the partworth values are measured *jointly* in the choice experiments.

The vector $v \in \mathbb{R}^k$ is just the vector of all partworth values. The entries in $v$ are ordered according to the natural order on $[n]$ and arbitrary but fixed orders on each of the sets $A_i$. Using this ordering, an element $a \in A$ can be represented by a characteristic vector $\mathbf{1}_a \in \{0,1\}^k$ that has exactly $n$ entries whose value is $1$, namely one entry for every set $A_i, i \in [n]$, and $k - n$ entries that are $0$.

Assume now that we are given choice data of the form

$$\big((a^{(1)}, b^{(1)}), y^{(1)}\big), \ldots, \big((a^{(m)}, b^{(m)}), y^{(m)}\big),$$

where the $a^{(i)}, b^{(i)} \in A$ are two choice options, and $y^{(i)} = 1$, if $a^{(i)}$ is preferred over $b^{(i)}$, and $y^{(i)} = -1$ otherwise. The partworth value vector $v \in \mathbb{R}^k$ should satisfy the following constraints

$$y^{(i)}\big(v(a^{(i)}) - v(b^{(i)})\big) \;\geq\; 0.$$

Using the characteristic vectors $\mathbf{1}_a,\, a \in A$ we can rewrite these constraints as

$$y^{(i)}\big(\mathbf{1}_{a^{(i)}} - \mathbf{1}_{b^{(i)}}\big)^\top v \;\geq\; 0,$$

of, if we define $n^{(i)} = \mathbf{1}_{a^{(i)}} - \mathbf{1}_{b^{(i)}} \in \{-1,0,1\}^k,\, i \in [n]$, as

$$y^{(i)} {n^{(i)}}^\top v \;\geq\; 0.$$

The $n^{(i)},\, i \in [m]$ can be considered as normal vectors of hyperplanes that pass through the origin. Hence, the data for a conjoint analysis problem are labeled hyperplanes

$$(n^{(1)}, y^{(1)}), \ldots, (n^{(m)}, y^{(m)}) \;\in\; \{-1,0,1\}^k \times \{\pm 1\}$$

and the sought for model is given by a partworth value vector $v \in \mathbb{R}^k$.

There is a striking similarity with binary linear classification problems, where the data are labeled points and the sought for model is a hyperplane. We can summarize this situation in the following table

| | binary linear classification | conjoint analysis |
|---|---|---|
| input | labeled points | labeled hyperplanes |
| output | hyperplane | point |

That is, the roles played by points and hyperplanes, respectively, are reversed in the two problems. It turns out that there is a geometric duality between points and hyperplanes that establishes a duality between binary linear classification problems (without offset) and conjoint analysis problems.

## Point-hyperplane duality

For every point $p \in \mathbb{R} \setminus \{0\}$ we can define a *dual hyperplane*

$$p^* \,=\, \{x \in \mathbb{R}^n \,:\, p^\top x = 1\},$$

i.e., the hyperplane whose normal points in the direction of $p$ and whose distance to the origin is $1/\|p\| > 0$. Similarly, we can define a dual point for every origin avoiding hyperplane

$$H \,=\, \{x \in \mathbb{R}^n \,:\, \theta^\top x = \theta_0\},$$

where $\theta \in \mathbb{R}^n \setminus \{0\}$ defines the direction of the normal of $H$ and $\theta_0 > 0$ defines the distance $\theta_0/\|\theta\|$ of $H$ to the origin. The *dual point* of $H$ is defined as

$$H^* \,=\, \left(\frac{\theta_1}{\theta_0}, \ldots, \frac{\theta_n}{\theta_0}\right)^\top \in \mathbb{R}^n \setminus \{0\}.$$

Obviously, we have

$$p^{**} \,=\, p \quad \text{and} \quad H^{**} \,=\, H.$$

There are more possible choices for defining a hyperplane from a point and a point from a hyperplance. The definitions above are special in the sense that they preserve incidences. Before we can state this property more formally we need the definition of *positive* and *negative halfspaces*. The closed, positive halfspace associated with the hyperplane

$$H \,=\, \{x \in \mathbb{R}^n \,:\, \theta^\top x = \theta_0\}$$

is the set

$$H_+ \,=\, \{x \in \mathbb{R}^n \,:\, \theta^\top x \geq \theta_0\},$$

and the closed, negative halfspace is the set

$$H_- \,=\, \{x \in \mathbb{R}^n \,:\, \theta^\top x \leq \theta_0\}.$$

**Lemma 10.** *We have*

$$p \,\in\, \begin{cases} H_+ \\ H_- \\ H \end{cases} \quad \textit{if and only if} \quad H^* \,\in\, \begin{cases} p^*_+ \\ p^*_- \\ p^* \end{cases}$$

*Proof.* Assume that $p \in H_+$, then we have

$$\begin{aligned} p \,\in\, H_+ \quad & \text{if and only if} \quad \theta^\top p \,\geq\, \theta_0 \\ & \text{if and only if} \quad \frac{1}{\theta_0}\theta^\top p \,\geq\, 1 \\ & \text{if and only if} \quad H^* \,\in\, p^*_+. \end{aligned}$$

The remaining two cases can be proved analogously. □

Lemma 10 implies that if $n+1$ points in $\mathbb{R}^n$ lie on a common, origin avoiding hyperplane, then their dual hyperplanes meet in a common point. Note that $n$ points in $\mathbb{R}^n$ always lie on a common hyperplane, but $n+1$ points on a common hyperplane is a special, non-generic situation.

## Duality in projective spaces

So far our discussion of duality is restricted to points that are not the origin and to origin avoiding hyperplanes. These singularities at the origin can be removed in *projective geometry*.

The points of the projective space $\mathbb{R}P^n$ are given as

$$\mathbb{R}P^n \,=\, \Big\{\{ax \,:\, a \in \mathbb{R}\setminus\{0\}\} \,:\, x \in \mathbb{R}^{n+1}\setminus\{0\}\Big\},$$

i.e., a point in projective space is an equivalence class $\{ax \,:\, a \in \mathbb{R}\}$. In projective geometry we distinguish between finite points and points at infinity. For finite points we have $x_{n+1} \neq 0$ and for points at infinity we have $x_{n+1} = 0$. Note that the distinction between *finite* points and *points at infinity* does not depend on the representative of the equivalence class. It is common to use *homogeneous coordinates* for representing the equivalence class of finite points. The *canonical representative* of a *finite* point in $\mathbb{R}P^n$ is given in homogeneous coordinates as

$$\left(\frac{x_1}{x_{n+1}}, \ldots, \frac{x_n}{x_{n+1}}, 1\right)^\top, \ \text{ if } x_{n+1} \,\neq\, 0.$$

A hyperplane in projective geometry is also defined by an equivalence class

$$H \,=\, \big\{x \in \mathbb{R}P^n \,:\, \theta^\top x = 0\big\},$$

where $\theta \in \mathbb{R}^{n+1} \setminus \{0\}$ is again a representative of the equivalence class since scaling $\theta$ by $a \in \mathbb{R} \setminus \{0\}$ does not change the set $H$. Here, $\theta^\top x = 0$ means that the inner product vanishes for the whole equivalence classes that are represented by $x$ and $\theta$, respectively. The equivalence class of $(0, \theta_0)^\top \in \mathbb{R}^{n+1}$, where $\theta_0 \neq 0$, is called the *hyperplane at infinity*.

Duality now simply means interpreting a point $p \in \mathbb{R}^{n+1} \setminus \{0\}$ either as a point or a hyperplane in the projective space $\mathbb{R}P^n$. Note that the dual of the origin, whose homogeneous coordinates are $(0, \ldots, 0, 1)^\top$, becomes the hyperplane at infinity, and the dual of a hyperplane with homogeneous coordinates $(\theta, 0)^\top \in \mathbb{R}^{n+1} \setminus \{0\}$, i.e., a hyperplane that contains the origin, becomes a point at infinity. Hence, by considering duality in the more general setting of projective geometry that includes the objects at infinity, the singularities of the point-hyperplane duality in Euclidean space are removed.

## Conjoint analysis revisited

For our application in conjoint analysis it is necessary to define the embedding of a partworth value vector $v \in \mathbb{R}^k$ into the projective space $\mathbb{R}P^k$ and the extraction of partworth values from a point in $\mathbb{R}P^k$. A natural embedding of a partworth value vector $v$ is the equivalence class of the vector whose homogeneous coordinates are $(v, 1)^\top \in \mathbb{R}^{k+1}$, and we only consider finite points in projective space as valid partworth value vectors. Note that all *positively* scaled partworth value vectors are equivalent with respect to predicting the outcome of choice experiments. But since the scaling in the definition of the points in $\mathbb{R}P^k$ is not restricted to positive scaling we cannot take just any representative of the equivalence class and remove the last coordinate for extracting the partworth values, because this would allow also for negative scaling of the partworth values. Since the scaling in the definition of points in $\mathbb{R}P^k$ involves also the last coordinate we can simply eliminate the effect of any scaling on the partworth values by using

$$\left( \frac{v_1}{v_{k+1}}, \ldots, \frac{v_k}{v_{k+1}} \right)^\top$$

for the partworth values given a representative $(v_1, \ldots, v_{k+1})^\top$ of a finite point in $\mathbb{R}P^k$. That is, we extract the partworth values from the canonical representative of a point in $\mathbb{R}P^k$ by projecting it onto its first $k$ coordinates.

Note that we have encountered the restriction to positive scaling before when we were considering *orientations* of hyperplanes for defining the halfspaces above and below a hyperplane. The orientation of a hyperplane was fixed by the positivity constraint on $\theta_0$. That is, also for hyperplanes we agreed on a choice of canonical representatives, namely the ones with $\theta_0 \geq 0$. Again, a natural choice for a single representative for an origin avoiding hyperplane would be the one with $\theta_0 = 1$.

Let us come back now to the conjoint analysis problem. The dual hyperplane of the sought for partworth value vector $v \in \mathbb{R}^k \setminus \{0\}$ is still defined as

$$v^* \;=\; \{x \in \mathbb{R}^k \;:\; v^\top x = 1\},$$

i.e., as $(v, 1)^\top \in \mathbb{R}^{k+1}$ in canonical, homogeneous coordinates. All the input hyperplanes

$$H^{(i)} \;=\; \left\{x \in \mathbb{R}^k \;:\; {n^{(i)}}^\top x = 0\right\}$$

are passing through the origin. Hence, their dual points are the points at infinity whose canonical, homogeneous coordinates are given as $(n^{(i)}, 0)^\top$, $i \in [m]$. A point $(n^{(i)}, 0)^\top$ lies on the hyperplane $(v, 1)$ if

$$\begin{pmatrix} v \\ 1 \end{pmatrix}^\top \begin{pmatrix} n^{(i)} \\ 0 \end{pmatrix} \;=\; v^\top n^{(i)} \;=\; 0,$$

it lies above the hyperplane if $v^\top n^{(i)} > 0$, and it lies below the hyperplane if $v^\top n^{(i)} < 0$.

We are now in the situation that we can employ any binary linear classification method for conjoint analysis simply by applying it to the binary linear classification problem whose input data is the sequence of labeled points in homogeneous coordinates

$$\left((n^{(i)}, 0)^\top, y^{(i)}\right)_{i \in [m]}.$$

and the sought for hyperplane is of the form $(v, 1)^\top$ in homogeneous coordinates.

For instance, we can use a primal soft-margin SVM for the binary linear classification problem given by the labeled data points $\left((n^{(i)}, 0)^\top, y^{(i)}\right)_{i \in [m]}$. This primal soft-margin SVM reads as

$$\begin{aligned} &\text{min}_{v \in \mathbb{R}^k,\, \xi \in \mathbb{R}^m} \quad \frac{1}{2}\|(v, 1)\|^2 + c \sum_{i=1}^m \xi_i \\ &\text{subject to} \quad y^{(i)} \begin{pmatrix} v \\ 1 \end{pmatrix}^\top \begin{pmatrix} n^{(i)} \\ 0 \end{pmatrix} \geq 1 - \xi_i \text{ and } \xi_i \;\geq\; 0,\; i \in [m]. \end{aligned}$$

Note that the last coordinate, that has the value zero for all the data points, does not matter for the constraints and thus can be removed. Also, $\|(v,1)\|^2 = \|v\|^2 + 1$. Since the plus one does not affect the solution of the SVM it also can be removed. Hence, the SVM from above becomes the following standard primal soft-margin SVM (without intercept)

$$\begin{aligned} &\text{min}_{v\in\mathbb{R}^k,\,\xi\in\mathbb{R}^m} \quad \frac{1}{2}\|v\|^2 + c\sum_{i=1}^{m}\xi_i \\ &\text{subject to} \quad y^{(i)}v^\top n^{(i)} \geq 1 - \xi_i \text{ and } \xi_i \geq 0,\ i \in [m]. \end{aligned}$$

If $\hat{v}$ is an optimal solution of this SVM, i.e., the hyperplane $\{x \in \mathbb{R}^k : \hat{v}^\top x = 1\}$, then its dual is the finite point $(\hat{v}, 1)^\top \in \mathbb{R}P^k$. Thus, the optimal partworth value vector is simply $\hat{v}$ by our extraction rule, i.e., the projecting of $(\hat{v}, 1)^\top$ onto its first $k$ coordinates.

# Chapter 22

# Exercises

## Chapter 20

**Exercise 1.** Let
$$H = \{x \in \mathbb{R}^d \,:\, \theta^\top x + \theta_0 = 0\}$$
be a hyperplane and let $x \in \mathbb{R}^d$ be a point. Derive a formula for the closest point to $x$ in $H$.

**Exercise 2.** Let
$$H = \{x \in \mathbb{R}^d \,:\, \theta^\top x + \theta_0 = 0\}$$
be a hyperplane and let $x \in \mathbb{R}^d$ be point. Derive a formula for the distance of $x$ to $H$.

**Exercise 3.** Show that maximizing the functional margin $\gamma$ for the hypothesis class of linear classifiers is equivalent to the primal, hard-margin SVM. That is, show that the following optimization problem

$$\begin{aligned} &\text{min}_{\gamma,\theta,\theta_0} \quad \gamma \\ &\text{subject to} \quad (\theta^\top x^{(i)} + \theta_0) \geq \gamma,\ i \in [m] \ \text{ and } \ \|\theta\| = 1 \end{aligned}$$

is equivalent to the primal, hard-margin SVM.

**Exercise 4.** Let $P = \{x^{(1)} = (0,0), x^{(2)} = (1,0), x^{(3)} = (0,1), x^{(4)} = (1,1)\}$. Draw the following three sets:

1.
$$\text{conv}(P) = \left\{ \sum_{i=1}^{4} a_i x^{(i)} \in \mathbb{R}^2 \,:\, \sum_{i=1}^{4} a_i = 1 \wedge \forall i : a_i \geq 0 \right\}.$$

2.

$$\mathsf{conv}_{3/4}(P) \;=\; \left\{\sum_{i=1}^{4} a_i x^{(i)} \in \mathbb{R}^2 \;:\; \sum_{i=1}^{4} a_i = 1 \,\wedge\, \forall i:\, 0 \leq a_i \leq \frac{3}{4}\right\}.$$

3.

$$\mathsf{conv}_{1/2}(P) \;=\; \left\{\sum_{i=1}^{4} a_i x^{(i)} \in \mathbb{R}^2 \;:\; \sum_{i=1}^{4} a_i = 1 \,\wedge\, \forall i:\, 0 \leq a_i \leq \frac{1}{2}\right\}.$$

## Chapter 21

**Exercise 1.** Let $P = \{x^{(1)}, \ldots, x^{(m)}\} \subset \mathbb{R}^n$ be a finite point set. The smallest enclosing ball problem asks to compute the center $c \in \mathbb{R}^n$ and radius $r > 0$ of the smallest ball that contains $P$. Formally, the smallest enclosing ball problem can be written as

$$\begin{array}{ll} \mathsf{min}_{c\in\mathbb{R}^n, r\in\mathbb{R}} & r \\ \text{subject to} & \|x^{(i)} - c\|^2 \leq r. \end{array}$$

Derive the Lagrangian dual of this problem.

## Chapter 22

**Exercise 1.** Derive an analytical solution to the following simple polytope distance problem

$$\hat{x} \;=\; \mathsf{argmin}_{x\in pq}\; \|x\|$$

that has to be solved within Gilbert's algorithm. Here, $p, q \in \mathbb{R}^n$ and $pq$ is the line segment

$$pq \;=\; \big\{\lambda p + (1 - \lambda q) \;:\; \lambda \in [0,1]\big\}.$$

# Chapter 23

# Supplemental material

### McDiarmid's Inequality

McDiarmid's Inequality is an extension of Hoeffding's Inequality. Its proof makes use of an extension of Hoeffding's Lemma.

**Lemma 11.** *Let $U$ and $X$ be real valued random variables with $E[U|X] = 0$ with probability $1$ and assume that there exists a function $g : \mathbb{R} \to \mathbb{R}$ and a constant $c > 0$ such that*

$$g(X) \; \leq \; U \; \leq \; g(X) + c.$$

*Then we have for every $s \geq 0$ that*

$$E\big[\exp(sU)|X\big] \; \leq \; \exp\big(s^2c^2/8\big).$$

*Proof.* The proof follows the lines of the proof of Hoeffding's Lemma. □

**Theorem 18. [McDiarmid]** *Let $X_i$, $i \in [n]$ be random variables that take values in $\mathcal{X}$, and let $f : \mathcal{X}^n = \prod_{i=1}^n \mathcal{X} \to \mathbb{R}$ be a function that satisfies the bounded difference assumption*

$$sup_{(x_i)_{i\in[n]}, x_i'} \big|f(x_1, \ldots, x_n) - f(x_1, \ldots, x_{i-1}, x_i' x_{i+1}, \ldots, x_n)\big| \; \leq \; c_i$$

*for $i \in [n]$. Then we have for all $t > 0$ that*

$$P\big[f(X_1, \ldots, X_n) - E[f(X_1, \ldots, X_n)] \geq t\big] \; \leq \; \exp\left(-\frac{2t^2}{\sum_{i=1}^n c_i^2}\right)$$

*and*

$$P\big[E[f(X_1, \ldots, X_n)] - f(X_1, \ldots, X_n) \geq t\big] \; \leq \; \exp\left(-\frac{2t^2}{\sum_{i=1}^n c_i^2}\right).$$

*Proof.* Let $V_0 = \mathsf{E}\big[f(X_1,\dots,X_n)\big]$ and

$V_i(x_1,\dots,x_i) \;=\; \mathsf{E}_{i+1\dots n}\big[f(X_1,\dots,X_n)|X_1 = x_1,\dots,X_i = x_i\big] \quad$ for $\;i \in [n]$,

then the sequence $\big(V_i(X_1,\dots,X_i)\big)_{i\in[n]}$ of random variables has the following *martingale property*

$$\mathsf{E}\big[V_i(X_1,\dots,X_i)|X_1 = x_1,\dots,X_{i-1} = x_{i-1}\big] \;=\; V_{i-1}(x_1,\dots,x_{i-1}).$$

We also have

$$\begin{aligned} f(X_1,\dots,X_n) - \mathsf{E}\big[f(X_1,\dots,X_n)\big] \;&=\; V_n(X_1,\dots,X_n) - V_0 \\ = \sum_{i=1}^{n}\Big(V_i(X_1,&\dots,X_i) - V_{i-1}(X_1,\dots,X_{i-1})\Big). \end{aligned}$$

Let

$$W_i(X_1,\dots,X_{i-1}) \;=\; \mathsf{sup}_{\bar{x}\in\mathcal{X}_i}\, V_i(X_1,\dots,X_{i-1},\bar{x}) - V_{i-1}(X_1,\dots,X_{i-1})$$

and

$$Z_i(X_1,\dots,X_{i-1}) \;=\; \mathsf{inf}_{\bar{x}\in\mathcal{X}_i}\, V_i(X_1,\dots,X_{i-1},\bar{x}) - V_{i-1}(X_1,\dots,X_{i-1}).$$

Then

$$\begin{aligned} Z_i(X_1,\dots,X_{i-1}) \;&\le\; V_i(X_1,\dots,X_i) - V_{i-1}(X_1,\dots,X_{i-1}) \\ &\le\; W_i(X_1,\dots,X_{i-1}) \end{aligned}$$

and

$$\begin{aligned} &W_i(X_1,\dots,X_{i-1}) - Z_i(X_1,\dots,X_{i-1}) \\ &\quad= \sup_{x\in\mathcal{X}_i}\,\sup_{\bar{x}\in\mathcal{X}_i}\, V_i(X_1,\dots,X_{i-1},x) - V_i(X_1,\dots,X_{i-1},\bar{x}) \;\le\; c_i. \end{aligned}$$

by the bounded difference assumption. Let

$$U_i(X_1,\dots,X_i) \;=\; V_i(X_1,\dots,X_i) - V_{i-1}(X_1,\dots,X_{i-1}).$$

Then we have from above that

$$\begin{aligned} f(X_1,\dots,X_n) - \mathsf{E}\big[f(X_1,\dots,X_n)\big] \;&=\; \sum_{i=1}^{n} U_i(X_1,\dots,X_i), \\ Z_i(X_1,\dots,X_{i-1}) \;\le\; U_i(X_1,\dots,X_i) \;&\le\; Z_i(X_1,\dots,X_{i-1}) + c_i, \end{aligned}$$

and by linearity of expectation and the martingale property

$$\begin{aligned}
&\mathsf{E}\big[U_i(X_1,\ldots,X_i)|X_1 = x_1,\ldots,X_{i-1} = x_{i-1}\big] \\
&= \mathsf{E}\big[V_i(X_1,\ldots,X_i) - V_{i-1}(X_1,\ldots,X_{i-1})|X_1 = x_1,\ldots,X_{i-1} = x_{i-1}\big] \\
&= \mathsf{E}\big[V_i(X_1,\ldots,X_i)|X_1 = x_1,\ldots,X_{i-1} = x_{i-1}\big] - V_{i-1}(x_1,\ldots,x_{i-1}) \\
&= V_{i-1}(x_1,\ldots,x_{i-1}) - V_{i-1}(x_1,\ldots,x_{i-1}) \;=\; 0.
\end{aligned}$$

Hence, we can apply Lemma 11 and get

$$\mathsf{E}\big[\exp(sU_i(X_1,\ldots,X_i)|X_1 = x_1 \ldots, X_{i-1} = x_{i-1}\big] \;\le\; \exp\left(s^2 c_i^2/8\right)$$

for every $(x_1,\ldots,x_{i-1}) \in \prod_{j=1}^{i-1}\mathcal{X}_j$, $s \ge 0$ and $i \in [n]$. Plugging this into Chernoff's bounding method gives

$$\begin{aligned}
&\mathsf{P}\big[f(X_1,\ldots,X_n) - \mathsf{E}[f(X_1,\ldots,X_n)] \ge t\big] \\
&= \mathsf{P}\left[\sum_{i=1}^{n} U_i(X_1,\ldots,X_n) \ge t\right] \\
&\le \exp(-st)\,\mathsf{E}\left[\exp\left(s\sum_{i=1}^{n} U_i(X_1,\ldots,X_i)\right)\right] \\
&= \frac{\mathsf{E}\left[\exp\left(s\sum_{i=1}^{n-1} U_i(X_1,\ldots,X_i)\right)\exp\Big(sU_n(X_1,\ldots,X_n|X_1,\ldots,X_{n-1}\Big)\right]}{\exp(st)} \\
&= \frac{\mathsf{E}_{1\ldots n-1}\left[\exp\left(s\sum_{i=1}^{n-1} U_i(X_1,\ldots,X_i)\right)\mathsf{E}_n\left[\exp\Big(sU_n(X_1,\ldots,X_n|X_1,\ldots,X_{n-1}\Big)\right]\right]}{\exp(st)} \\
&= \exp(-st)\,\mathsf{E}\left[\exp\left(s\sum_{i=1}^{n-1} U_i(X_1,\ldots,X_i)\right)\exp\left(s^2c_i^2/8\right)\right] \\
&\le \exp\left(\frac{s^2c_n^2}{8st}\right)\mathsf{E}\left[\exp\left(s\sum_{i=1}^{n-1} U_i(X_1,\ldots,X_i)\right)\right] \\
&\le \ldots \le \exp\left(\frac{s^2\sum_{i=1}^{n} c_i^2}{8st}\right).
\end{aligned}$$

Plugging

$$s \;=\; \frac{4t}{\sum_{i=1}^{n} c_i^2}$$

into this bound gives the first inequality. The proof of the second inequality follows similarly by using

$$\mathsf{E}\big[f(X_1,\ldots,X_n)\big] - f(X_1,\ldots,X_n) \;=\; -\sum_{i=1}^{n} U_i(X_1,\ldots,X_i) \quad \text{and}$$

$$-\,W_i(X_1,\ldots,X_{i-1}) \;\le\; -U_i(X_1,\ldots,X_i) \;\le\; -W_i(X_1,\ldots,X_{i-1}) + c_i.$$

□