Ludwig-Maximilians-Universitaet Muenchen 19.06.2020
Institute for Informatics
Prof. Dr. Volker Tresp
Rajat Koner
Andreas Lohrer
Sebastian Schmoll

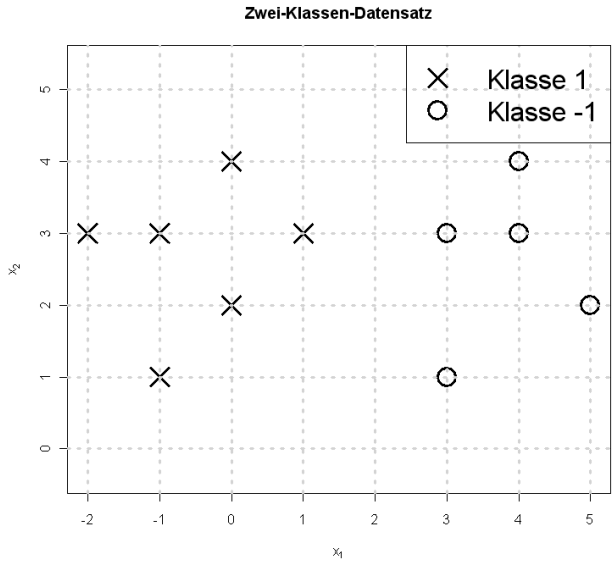
Machine Learning

Summer 2021

Exercise Sheet 9

Exercise 9-1 Optimal Seperating Hyperplane 1

Consider the following dataset consisting of points $\binom{x_1}{x_2}$ in $\mathbb{R}^2$. Using a hyperplane, points marked by $\times$ are to be mapped onto ≥ 1, points marked by $\bigcirc$ are to be mapped onto ≤ -1.



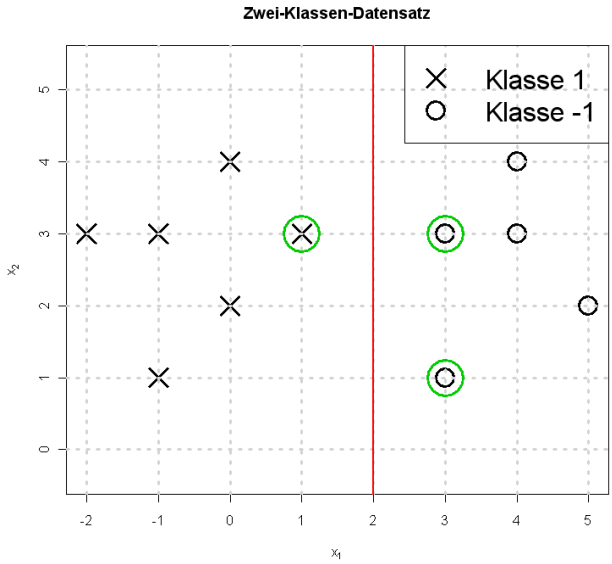
(a) Find the support vectors.

(b) Determine the equation of one separating hyperplane $h = \mathbf{x}^T\mathbf{w}$, optimize it and draw it within the figure.

(c) Compute the margin $\mathcal{C}$.

Possible Solution

Find support vectors that maximize the margin:



$\mathbf{x}_1 = \binom{1}{3}$ with $y = 1$ and $\mathbf{x}_2 = \binom{3}{1}$, and $\mathbf{x}_3 = \binom{3}{3}$ with $y = -1$.

Possible Solution

b)

$$s_1 = \begin{pmatrix}2\\2\end{pmatrix},\ s_2 = \begin{pmatrix}2\\3\end{pmatrix} \qquad \text{linear equation: } h = 0 = \mathbf{w}_0 + \mathbf{w}_1 \cdot \mathbf{x}_1 + \mathbf{w}_2 \cdot \mathbf{x}_2$$

$$\Rightarrow \text{I}): \quad \mathbf{w}_0 + \mathbf{w}_1 \cdot 2 + \mathbf{w}_2 \cdot 2 = 0$$

$$\text{II}): \quad \mathbf{w}_0 + \mathbf{w}_1 \cdot 2 + \mathbf{w}_2 \cdot 3 = 0$$

$$\text{II)}-\text{I}): \quad \mathbf{w}_2 = 0 \qquad \Rightarrow 2 \cdot \mathbf{w}_1 = -\mathbf{w}_0$$

$$\text{define } \mathbf{w}_1 = 1 \Rightarrow \mathbf{w}_0 = -2 \qquad \Rightarrow \mathbf{w} = \begin{pmatrix}-2\\1\\0\end{pmatrix}$$

Test if the hyperplane $\mathbf{w} = (-2, 1, 0)^T$ classifies correctly:
$x_1 : y_1(\mathbf{x}_1^T \mathbf{w}) = 1 \cdot (-2 + 1 \cdot 1 + 0 \cdot 3) = -1 \Rightarrow$ not correct.
Change the signs:

$$\mathbf{w} = \begin{pmatrix}2\\-1\\0\end{pmatrix} \qquad \Rightarrow y_1(\mathbf{x}_1^T \mathbf{w}) = 1 \cdot (2 \cdot 1 - 1 \cdot 1) = 1 \qquad \text{The solution is correct.}$$

Possible Solution

c) $\mathcal{C} = \frac{1}{\|\tilde{\mathbf{w}}_{\text{opt}}\|}$, with $\mathbf{w}_{\text{opt}}$ beeing the optimale $\mathbf{w}$.
Since $x_1 : y_1(\mathbf{x}_1^T \mathbf{w}) = 1$ we already have the optimal $\mathbf{w}_{opt}$.
$\tilde{\mathbf{w}}_{\text{opt}}$ is $\mathbf{w}_{\text{opt}}$ without the bias dimension, therefore $\tilde{\mathbf{w}}_{\text{opt}} = \binom{\mathbf{w}_{\text{opt}_1}}{\mathbf{w}_{\text{opt}_2}}$. Thus the margin is:

$$\mathcal{C} = \frac{1}{\|\tilde{\mathbf{w}}_{\text{opt}}\|} = \frac{1}{\sqrt{(-1)^2 + (0)^2}} = 1, \qquad with \tilde{\mathbf{w}}_{\text{opt}} = \begin{pmatrix}\mathbf{w}_1\\\mathbf{w}_2\end{pmatrix} = \begin{pmatrix}-1\\0\end{pmatrix}$$

Exercise 9-2 Optimal Separating Hyperplane 2

Determine the optimal separating hyperplane of the following dataset, partitioned into two classes A and B:

$$A = \left\{ p_1 = \begin{pmatrix} 2 \\ 4 \end{pmatrix},\ p_2 = \begin{pmatrix} 3 \\ -1 \end{pmatrix},\ p_3 = \begin{pmatrix} 1 \\ 0.5 \end{pmatrix},\ p_4 = \begin{pmatrix} 2.5 \\ 3 \end{pmatrix},\ p_5 = \begin{pmatrix} 2 \\ 2 \end{pmatrix} \right\},$$
$$B = \left\{ p_6 = \begin{pmatrix} 0.5 \\ 1.5 \end{pmatrix},\ p_7 = \begin{pmatrix} -1 \\ 3 \end{pmatrix},\ p_8 = \begin{pmatrix} 0 \\ 0.5 \end{pmatrix} \right\}$$

Instances of class A shall be labeled with 1, instances of class B with -1.

Name the support vectors, compute the optimal separating hyperplane and visualize the result. How wide is the margin?

Possible Solution

Visual solution: $\{p_1, p_3, p_6\}$ are the support vektors, thus:

$$s_1 = \begin{pmatrix} 1.25 \\ 2.75 \end{pmatrix},\ s_2 = \begin{pmatrix} 0.75 \\ 1 \end{pmatrix} \qquad \text{linear equation: } h = 0 = \mathbf{w}_0 + \mathbf{w}_1 \cdot \mathbf{x}_1 + \mathbf{w}_2 \cdot \mathbf{x}_2$$

$$\Rightarrow \text{I):} \quad \mathbf{w}_0 + \mathbf{w}_1 \cdot 1.25 + \mathbf{w}_2 \cdot 2.75 = 0$$
$$\text{II):} \quad \mathbf{w}_0 + \mathbf{w}_1 \cdot 0.75 + \mathbf{w}_2 \cdot 1 = 0$$
$$\text{I)}-\text{II):} \quad \mathbf{w}_1 \cdot 0.5 + \mathbf{w}_2 \cdot 1.75 = 0 \qquad \Rightarrow \mathbf{w}_1 = -3.5 \cdot \mathbf{w}_2$$
$$\text{define } \mathbf{w}_2 = 1 \Rightarrow \mathbf{w}_1 = -3.5 \overset{\text{in I)}}{\Rightarrow} \mathbf{w}_0 = 3.5 \cdot 1.25 - 1 \cdot 2.75 = 1.625$$

Margin-condition: $y_i(\mathbf{x}_i^T \mathbf{w}_{\text{opt}}) \geq 1$. Test for correct classification:
$y_i \cdot \sum_{j=0}^{2} \mathbf{w}_j \mathbf{x}_{i,j} \overset{p_1}{=} 1 \cdot (1.625 - 3.5 \cdot 2 + 1 \cdot 4) = -1.375 \not\geq 1.$

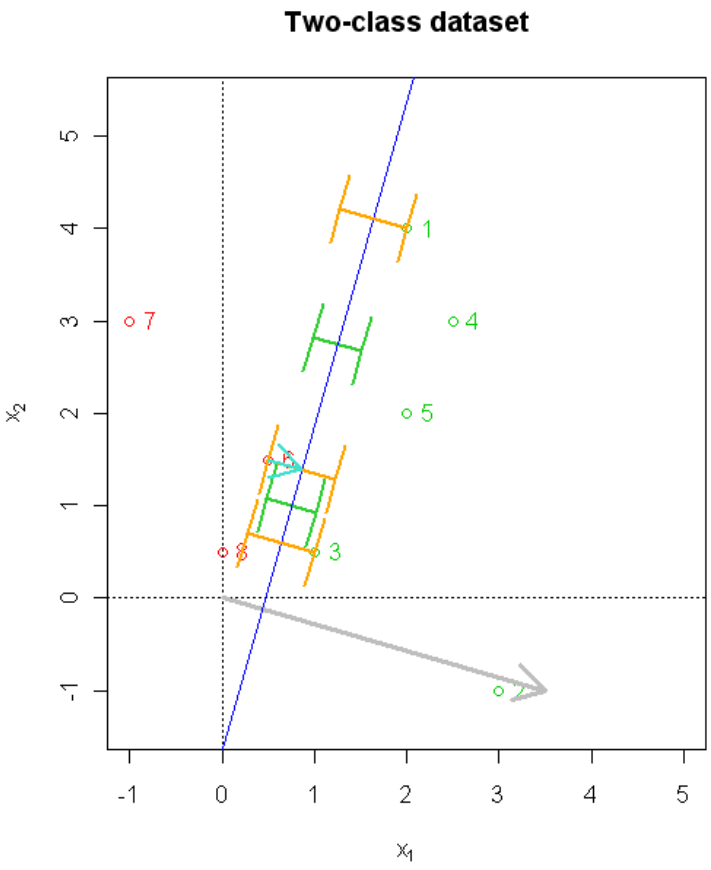
Thus the signs of the vector have to be multiplied with $(-1) \Rightarrow \mathbf{w} = \begin{pmatrix} -1.625 \\ 3.5 \\ -1 \end{pmatrix}$.

Now we want to get the minimal $\mathbf{w}$, since the margin $\mathcal{C} = \frac{1}{\|\tilde{\mathbf{w}}_{\text{opt}}\|}$ is supposed to be maximal. Furthermore the margin-condition has to hold and $y_i(\mathbf{x}_i^T \mathbf{w}_{\text{opt}}) \overset{!}{=} 1$ for the support vectors. So far $\min\limits_{p_1,p_3,p_6} y_i(\mathbf{x}_i^T \mathbf{w}) = 1.375$, thus it's enough to divide by 1.375 to get the optimal separating hyperplane:

$$\Rightarrow \mathbf{w}_{\text{opt}} = \frac{\mathbf{w}}{y_i(\mathbf{x}_i^T \mathbf{w}_{\text{opt}})} = \frac{\mathbf{w}}{1.375} = \begin{pmatrix} -1.\overline{18} \\ 2.\overline{54} \\ -0.\overline{72} \end{pmatrix}$$

Possible Solution

The margin is defined as
$\mathcal{C} = \frac{1}{\|\tilde{\mathbf{w}}_{\text{opt}}\|} = \frac{1}{\|(\mathbf{w}_{\text{opt},1}, \mathbf{w}_{\text{opt},2})^T\|} =$
$= \frac{1}{\sqrt{\frac{3.5}{1.375}^2 + \frac{1}{1.375}^2}} \approx$
$\approx 1/7.0082 \approx 0.142689.$



Exercise 9-3 Failure of k-fold cross validation

Consider a case in that the label is chosen at random according to $P[y = 1] = P[y = 0] = \frac{1}{2}$. Consider a learning algorithm that outputs the constant predictor $h(x) = 1$ if the parity of the labels on the training set is 1 and otherwise the algorithm outputs the constant predictor $h(x) = 0$. Prove that the difference between the leave-one-out estimate and the true error in such a case is always $\frac{1}{2}$.

Possible Solution

Let S be an i.i.d sample. Let h be the output of the described learning algorithm. It is given that $Loss_D(h) = \frac{1}{2}$, where D denotes the probability distribution for the data generating process..

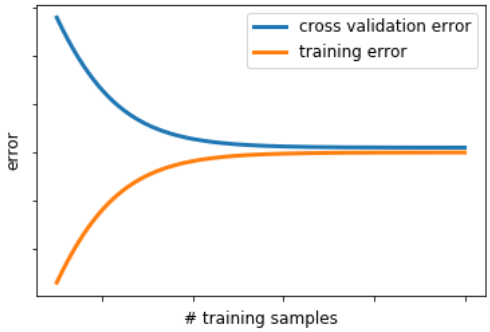
Let us calculate the estimate $Loss_{Val}(h)$. Assume that the parity of S is 1. Fix some fold $\{(x, y)\} \subseteq S$. We distinguish between two cases:

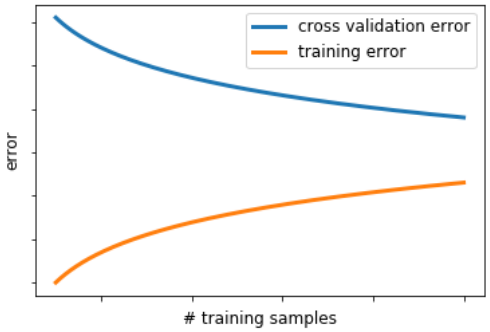
- The parity of $S \setminus \{x\}$ is 1. It follows that $y = 0$. When being trained using $S \setminus \{x\}$, the algorithm outputs the constant predictor $h(x) = 1$. Hence, the leave-one-out estimate using this fold is 1.
- The parity of $S \setminus \{x\}$ is 0. It follows that $y = 1$. When being trained using $S \setminus \{x\}$, the algorithm outputs the constant predictor $h(x) = 0$. Hence, the leave-one-out estimate using this fold is 1.

Averaging over the folds, the estimate of the error of h is 1. Consequently, the difference between the estimate and the true error is $\frac{1}{2}$. The case in which the parity of S is 0 is analyzed analogously.

Exercise 9-4 Bias vs. Variance - General

Consider the following learning curves showing training and cross-validation error for two different models when the respective model is trained on an increasing amount of data. For both cases, discuss bias and variance. What are indicators for a high variance or a high bias problem?

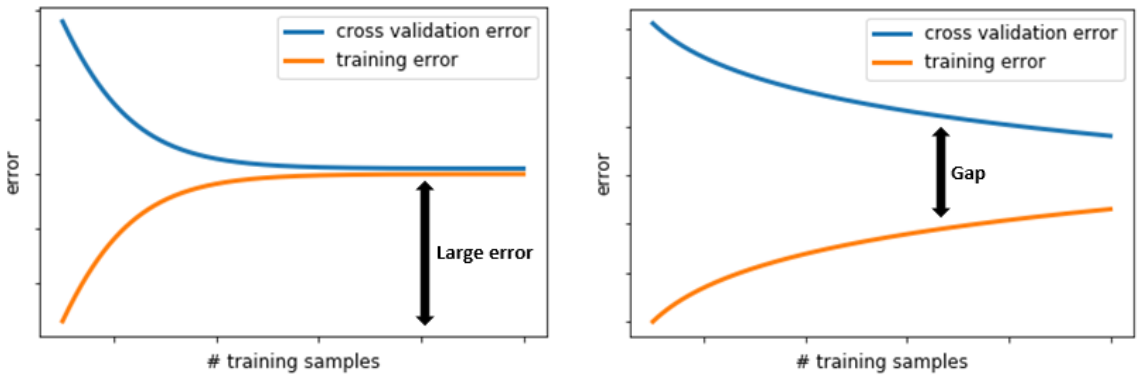




Possible Solution

The general behavior of the curves can be explained as follows: When you have only a small amount of training samples, fitting them is relatively easy, so you get a small training error. With a growing number of data points, it gets increasingly difficult to fit all of the points well with your model, so the training error can be expected to increase. The cross-validation error on the other hand will be relatively large if you have only a small amount of data available, since the model is not able to generalize well. This behavior gets better when more data is available. The given plots show two idealized settings indicating special cases w.r.t. bias and variance:

- First case: The learning curves in this case indicate a *high bias*. High bias means that you make strong assumptions on how your data looks like and select a small, restricted class of models, e.g. the class of linear functions. In a sense, you are biased towards a particular class of models. If the underlying data distribution is more complex than what your model is able to represent, your model will never be able to achieve a small error, even if you sample more and more data. For instance, you can never accurately model data sampled from a higher-order polynomial with a linear function. In this case, you are *underfitting* your data.
- Second case: In the second plot, the contrary case of *high variance* is indicated. High variance implies small bias, i.e. you make very weak assumptions on your data and end up selecting a fairly large class of models, e.g. high-order polynomials. High variance means that your model will end up very differently when trained on different portions of the data, since it strongly adapts to the training data. In this sense, the model *overfits* the training data. Correspondingly, your training error with stay pretty low when the number of data points increases. The cross-validation error on the other hand will be quite large, since the model is not able to generalize well to parts of the data space it has not seen during training. This behavior is indicated by the gap between the two curves. However, with more data available, the model will be able to cover a larger part of the data space and be able to generalize better, so that the two curves will converge. In contrast to the first case, more data is likely to help.



Exercise 9-5 Bias-Variance Decomposition

Assume a fixed distribution $P(x)$ over x and a dependent variable $y = f(x) + \epsilon$ with $\epsilon \sim \mathcal{N}(0, \sigma^2)$. We try to model this relationship by a function $\hat{f}(x, w)$ whose parameter w we learn. For short, we write $f(x) = f$ and $\hat{f}(x, w) = \hat{f}$. Consider the expected square loss $\mathbb{E}[L] = \mathbb{E}[(\hat{f} - y)^2]$.

(a) Show that: $\mathbb{E}[(\hat{f} - y)^2] = \mathbb{E}[(\hat{f} - f)^2] + \mathbb{E}[(f - y)^2]$
with $\mathbb{E}[(f - y)^2] = \mathrm{Var}[y] = \sigma^2$ being the intrinsic noise of the data.
Hint: Make use of the calculation rules for the expected value: For two random variables X and Y and a constant c it holds that: $\mathbb{E}[cX + Y] = c\mathbb{E}[X] + \mathbb{E}[Y]$. Moreover, if X and Y are independent: $\mathbb{E}[X \cdot Y] = \mathbb{E}[X] \cdot \mathbb{E}[Y]$.

Possible Solution

$$
\begin{aligned}
\mathbb{E}[(\hat{f} - y)^2] &= \mathbb{E}[(\hat{f} - f - \epsilon)^2] \\
&= \mathbb{E}[(\hat{f} - f)^2 + \epsilon^2 - 2(\hat{f} - f)\epsilon] \\
&= \mathbb{E}[(\hat{f} - f)^2] + \mathbb{E}[\epsilon^2] - 2\mathbb{E}[\hat{f}\epsilon] + 2\mathbb{E}[f\epsilon] \\
&= \mathbb{E}[(\hat{f} - f)^2] + \mathrm{Var}[\epsilon] - 2\mathbb{E}[\hat{f}] \cdot \mathbb{E}[\epsilon] + 2f \cdot \mathbb{E}[\epsilon] && \text{since } \epsilon \perp \hat{f} \text{ and } f \text{ is deterministic} \\
&= \mathbb{E}[(\hat{f} - f)^2] + \sigma^2 && \text{since } \mathbb{E}[\epsilon] = 0
\end{aligned}
$$

In the 4th step we used that $\mathbb{E}[\epsilon^2] = \mathrm{Var}[\epsilon] + (\mathbb{E}[\epsilon])^2 = \sigma^2 + 0$.
Also note that $\mathbb{E}[\epsilon^2] = \mathbb{E}[(y - f)^2] = \mathbb{E}[(y - \mathbb{E}[y])^2] = \mathrm{Var}[y]$

(b) Now show that: $\mathbb{E}[(\hat{f} - f)^2] = \mathbb{E}[(\hat{f} - \mathbb{E}[\hat{f}])^2] + (\mathbb{E}[\hat{f}] - f)^2$
where $\mathbb{E}[(\hat{f} - \mathbb{E}[\hat{f}])^2]$ is the variance and $(\mathbb{E}[\hat{f}] - f)^2$ the squared bias of our estimation $\hat{f}$.
Hint: Substract and add $\mathbb{E}[\hat{f}]$ to the squared term in the loss and simplify.

Possible Solution

$$
\begin{aligned}
\mathbb{E}[(\hat{f} - f)^2] &= \mathbb{E}[(\hat{f} - \mathbb{E}[\hat{f}] + \mathbb{E}[\hat{f}] - f)^2] \\
&= \mathbb{E}[(\hat{f} - \mathbb{E}[\hat{f}])^2] + \mathbb{E}[(\mathbb{E}[\hat{f}] - f)^2] + 2\mathbb{E}[(\hat{f} - \mathbb{E}[\hat{f}])(\mathbb{E}[\hat{f}] - f)] \\
&= \mathrm{Var}[\hat{f}] + (\mathbb{E}[\hat{f}] - f)^2 + 2(\mathbb{E}[\hat{f}] - f) \cdot \mathbb{E}[(\hat{f} - \mathbb{E}[\hat{f}])] && \text{since } (\mathbb{E}[\hat{f}] - f) \text{ is deterministic} \\
&= \mathrm{Var}[\hat{f}] + (\mathrm{Bias}(\hat{f}))^2 + 2(\mathbb{E}[\hat{f}] - f) \cdot (\mathbb{E}[\hat{f}] - \mathbb{E}[\hat{f}]) && \text{since } \mathbb{E}[\mathbb{E}[X]] = \mathbb{E}[X] \\
&= \mathrm{Var}[\hat{f}] + (\mathrm{Bias}(\hat{f}))^2
\end{aligned}
$$

(c) With that we have:

$$\begin{aligned}\mathbb{E}[L] = \mathbb{E}[(\hat{f} - y)^2] &= \mathbb{E}[(\hat{f} - \mathbb{E}[\hat{f}])^2] + (\mathbb{E}[\hat{f}] - f)^2 + \mathbb{E}[(f - y)^2] \\ &= \mathrm{Var}[\hat{f}] + (\mathrm{Bias}(\hat{f}))^2 + \sigma^2.\end{aligned}$$

What is the minimum of the expected loss?

Possible Solution

For the best possible model we would have $\hat{f}(x) = f(x)$ and thus

$$\mathbb{E}[L] = \mathbb{E}[(f - f)^2] + \mathbb{E}[(f - y)^2] = 0 + \sigma^2 = \sigma^2.$$

In this case our model has zero bias and variance and only the intrinsic noise remains as loss.