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

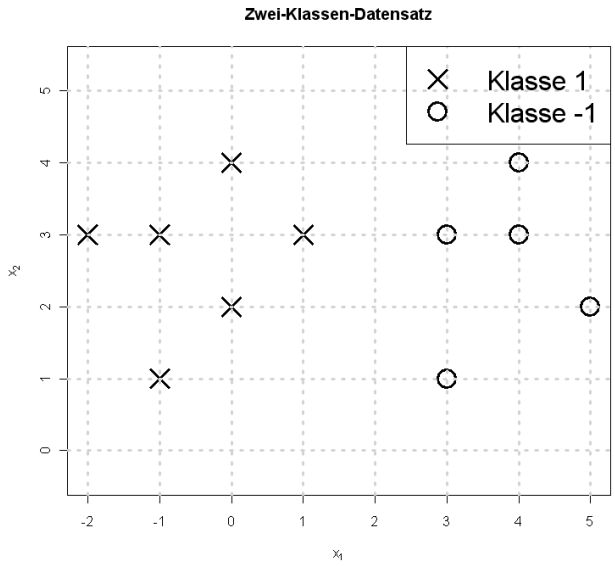
# Machine Learning

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 $\geq 1$, points marked by $\bigcirc$ are to be mapped onto $\leq -1$.




(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 $\mathcal{C}$.

**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 $-1$.

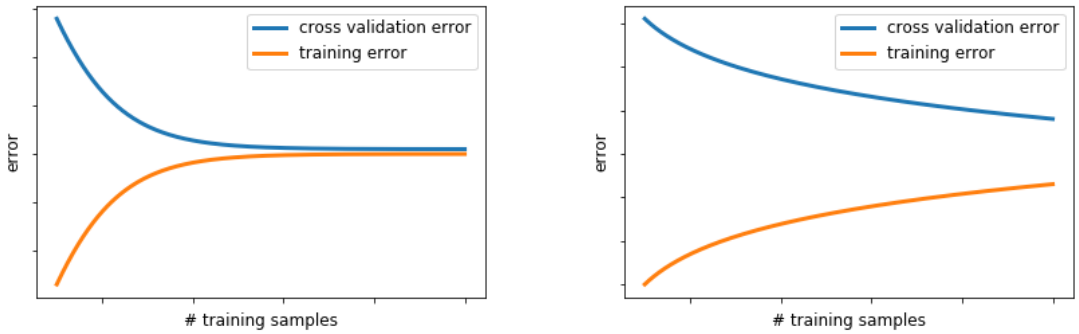
Name the support vectors, compute the optimal separating hyperplane and visualize the result. How wide is the margin?

**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 $\frac{1}{2}$.

**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?




**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: $\mathbb{E}[X \cdot Y] = \mathbb{E}[X] \cdot \mathbb{E}[Y]$.

(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.

(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?