EXAM IN STATISTICAL MACHINE LEARNING
STATISTISK MASKININLÄRNING

DATE AND TIME: March 10, 2022, 8.00–13.00

RESPONSIBLE TEACHER: Jens Sjölund

NUMBER OF PROBLEMS: 5

AIDING MATERIAL: Calculator, mathematical handbook

PRELIMINARY GRADES:

grade 3	23 points
grade 4	33 points
grade 5	43 points

Some general instructions and information:

- Your solutions should be given in English.
- Only write on one page of the paper.
- Write your exam code and a page number on all pages.
- Do not use a red pen.
- Use separate sheets of paper for the different problems (i.e. the numbered problems, 1–5).

With the exception of Problem 1, all your answers must be clearly motivated! A correct answer without a proper motivation will score zero points!

Good luck!

1. i. False. (See Figure 4.9 (a) in the draft (April 30, 2021) SML book. The bias is approximately constant as the number of training examples increases.)

   ii. False. (Because the logistic loss is convex. Reference in the draft (April 30, 2021) SML book on page 96: "Examples of convex functions are the cost functions for logistic regression, linear regression and L^1-regularized linear regression.")

   iii. True. Reference in the draft (April 30, 2021) SML book on page 94: "Whereas L^2 regularization pushes all parameters towards small values (but not necessarily exactly zero), L^1 tends to favor so-called sparse solutions where only a few of the parameters are non-zero, and the rest are exactly zero."

   iv. True (Because of the sparse interactions and parameter sharing in convolutional layers. Reference in the draft (April 30, 2021) SML book on page 126: "Furthermore, a convolutional layer uses significantly fewer parameters compared to the corresponding dense layer.")

   v. False. (Reference in the draft (April 30, 2021) SML book on page 140: "It is important to understand that by the construction of bagging, more ensemble members does not make the resulting model more flexible, but only reduces the variance.")

   vi. True. (Reference in the draft (April 30, 2021) SML book on page 141: "When using bagging, it turns out that there is a way to estimate the expected new data error E_{new} without using cross-validation.")

   vii. False. (Reference in the draft (April 30, 2021) SML book on page 151: "Another unfortunate aspect of the sequential nature of boosting is that it is not possible to parallelize the learning.")

   viii. True. (Reference in the draft (April 30, 2021) SML book on page 240: "For models that are trained iteratively we can reduce $E_{\text{train}}$by training longer.")

   ix. False. (Reference in the draft (April 30, 2021) SML book on page 49: "linear regression is a model which is linear in its parameters")

   x. True. (Reference in the draft (April 30, 2021) SML book on page 29)

2. (a) Liz is correct given a Gaussian noise assumption. Using the squared error loss is equivalent to assuming a Gaussian noise distribution in the maximum likelihood formulation.

   (b) The decision boundary of a logistic regression classifier is given by $\Theta^T \tilde{\mathbf{x}} = 0$. This gives us the following decision boundary:

$$\Theta^T \tilde{\mathbf{x}} = -6.75 + 0.5m + v = 0 \iff v = 6.75 - 0.5m.$$

   For each sample (m_*, v_*) in the test dataset, the expression for the decision boundary can be used to make a prediction according to:

$$\begin{cases} 1 & \text{for } -6.75 + 0.5m_* + v_* > 0 \\ -1 & \text{for } -6.75 + 0.5m_* + v_* < 0 \end{cases}$$

   Comparing the predicted class to the true class for each sample in the test dataset gives a misclassification rate of $\frac{1}{6} \approx 17\%$.

   (c) Liz is not correct. The material is a qualitative input feature that can be encoded using dummy variables. One suggestion is to use the dummy variables x_1, x_2 and x_3 according to:

$$x_1 = \begin{cases} 1 & \text{if plastic} \\ 0 & \text{else} \end{cases}, \quad x_2 = \begin{cases} 1 & \text{if wood} \\ 0 & \text{else} \end{cases}, \quad x_3 = \begin{cases} 1 & \text{if metal} \\ 0 & \text{else} \end{cases}.$$

   (d) The general criterion is that the data has to be (approximately) separable by a linear decision boundary in the space spanned by all inputs, including transformed inputs (1p).

   Logistic regression can be used for all of the datasets depicted (1p per dataset; inputs x_1, x_2 do not have to be mentioned explicitly):

   (i) The decision boundary already is linear, no transformed inputs are required.

   (ii) Use additional transformed input $x_3 = x_1^2$.

   (iii) Use transformed input $x_3 = r = \sqrt{x_1^2 + x_2^2}$.

3. a) **k-NN**

For the first missing point we have

$$||x_1 - x_A||_2 = \sqrt{(-2)^2 + (-8)^2} = \sqrt{68} = 2\sqrt{17} \approx 8.246$$

For the second missing point we have

$$||x_3 - x_A||_2 = \sqrt{(-7)^2 + (-7)^2} = \sqrt{98} = 7\sqrt{2} \approx 9.899$$

For our two cases we use the majority vote within the neighbor-

Runner id	x		y	$\|x - x_A\|_2$
1	38	82	1	$\sqrt{68} \approx 8.246$
2	32	85	1	9.4
3	33	83	-1	$\sqrt{98} \approx 9.899$
4	28	85	-1	13.0
5	44	93	-1	5.0
6	57	97	-1	18.4

hood R, to assign x_A to class m with $\sum_{i \in R} 1\{y_i = m\}$. Using this we have:

i $k = 1$, we simply consider the class of the closest distance which is x_5.
Hence, our prediction is $y = -1$.

ii $k = 3$, we take the three closest points x_5, x_1 and x_2. The sum of the classes for $m = 1$ is 2 and for $m = -1$ is 1.
Hence our prediction is $y = 1$.

b) In a QDA model the prior/marginal probability $p(y = 1)$ is estimated by:

$$\hat{\pi}_1 = \frac{n_1}{n} = \frac{2}{6} = \frac{1}{3}$$

where n is the total number of data points in the dataset and n_1 the number of data points with label $y = 1$.

The likelihood $p(x|y = 1)$ is estimated as a normal distribution with the parameters $\hat{\mu}_1$ and $\hat{\Sigma}_1$ as following:

The mean of the distribution is given by:

$$\begin{aligned}\hat{\mu}_1 &= \frac{1}{n_1}\sum_{i:y_i=1} x_i \\ &= \frac{1}{2}(x_1 + x_2) \\ &= 0.5(\begin{bmatrix}38 & 82\end{bmatrix}^\mathsf{T} + \begin{bmatrix}32 & 85\end{bmatrix}^\mathsf{T}) \\ &= \begin{bmatrix}35 & 83.5\end{bmatrix}^\mathsf{T}\end{aligned}$$

The estimation of the covariance matrix for the distribution in QDA is given by:

$$\begin{aligned}\hat{\Sigma}_1 &= \frac{1}{n_1 - 1}\sum_{i:y_i=1} (x_i - \hat{\mu}_1)(x_i - \hat{\mu}_1)^\mathsf{T} \\ &= \frac{1}{2-1}[(x_1 - \hat{\mu}_1)(x_1 - \hat{\mu}_1)^\mathsf{T} + (x_2 - \hat{\mu}_1)(x_2 - \hat{\mu}_1)^\mathsf{T}] \\ &= \begin{bmatrix}9 & -4.5 \\ -4.5 & 2.25\end{bmatrix} + \begin{bmatrix}9 & -4.5 \\ -4.5 & 2.25\end{bmatrix} \\ &= \begin{bmatrix}18 & -9 \\ -9 & 4.5\end{bmatrix}\end{aligned}$$

The likelihood is estimated by $p(x|y=1) = \mathcal{N}(x|\hat{\mu}_1, \hat{\Sigma}_1)$.

c) For runner A, we have that

$$\begin{aligned}p(y=1|x_A) &= \frac{\mathcal{N}(x_A|\hat{\mu}_1, \hat{\Sigma}) \cdot \hat{\pi}_1}{\mathcal{N}(x_A|\hat{\mu}_1, \hat{\Sigma}) \cdot \hat{\pi}_1 + \mathcal{N}(x_A|\hat{\mu}_{-1}, \hat{\Sigma}) \cdot \hat{\pi}_{-1}} \\ &= \frac{0.0020 \cdot 0.31}{0.0020 \cdot 0.31 + 0.0030 \cdot 0.69} \\ &= 0.23\end{aligned}$$

Thus, we obtain $p(y=-1|x_A) = 1 - p(y=1|x_A) = 1 - 0.23 = 0.77$.

For runner B, we obtain similarly

$$\begin{aligned}p(y=1|x_B) &= \frac{\mathcal{N}(x_B|\hat{\mu}_1, \hat{\Sigma}) \cdot \hat{\pi}_1}{\mathcal{N}(x_B|\hat{\mu}_1, \hat{\Sigma}) \cdot \hat{\pi}_1 + \mathcal{N}(x_B|\hat{\mu}_{-1}, \hat{\Sigma}) \cdot \hat{\pi}_{-1}} \\ &= \frac{0.0014 \cdot 0.31}{0.0014 \cdot 0.31 + 0.00019 \cdot 0.69} \\ &= 0.768\end{aligned}$$

For the threshold $r = 0.5$, we predict $\hat{y}(x_A) = -1$ and $\hat{y}(x_B) = 1$. Thus, the average misclassification error is 0.5.

d) The figure can be drawn e.g. like the one in Figure 1. QDA

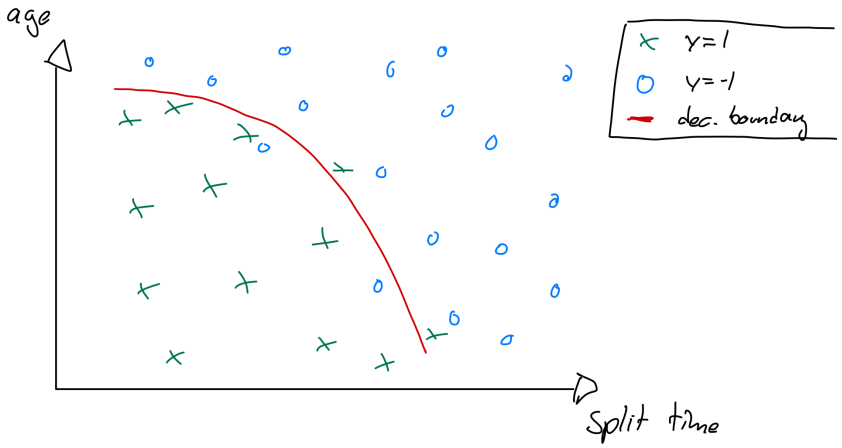

Figure 1: Possible solution for LDA/QDA scatter plot

would work better in this setting since the true decision boundary is quadratic. LDA would just define a linear decision boundary which would then have a higher error rate than a quadratic one.

4. (a) No, quite the opposite. A robust loss is insensitive to a limited amount of large errors (outliers), whereas the minimax loss picks out the largest errors.

   (b) The prediction for each leaf is given by the average (the minimizer of the squared error). Based on this, we have to compute the minimax loss for each cutpoint $s \in \{2, 4, 6\}$ and select the one that minimizes the loss. As shown by the comparison in Table 1, $s = 2$ minimizes the loss. The resulting prediction is plotted in Figure 2.

x	1	3	5	7	$L(\mathbf{y}, \hat{\mathbf{y}})$
y	0	5	2	1	
$\hat{y}_{s=2}$	0	8/3	8/3	8/3	**7/3**
$\hat{y}_{s=4}$	5/2	5/2	3/2	3/2	5/2
$\hat{y}_{s=6}$	7/3	7/3	7/3	1	8/3

Table 1: Predictions of a stump with different cutpoints.

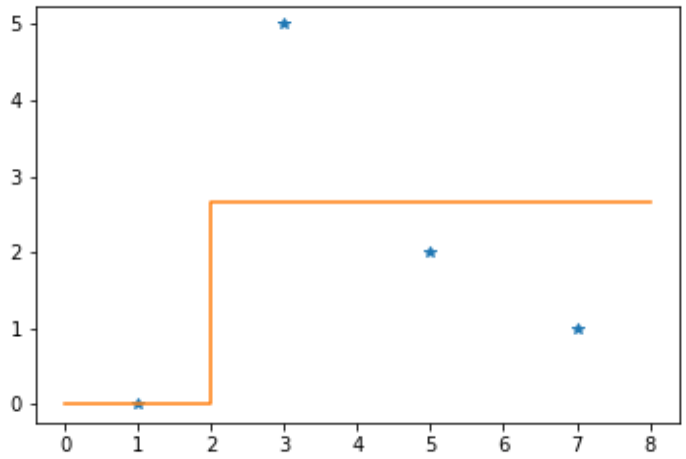

Figure 2: Prediction from stump.

   (c) Following a similar procedure as in part (b), the models for each bootstrapped dataset can be found as shown in Tables 2 and 3:

   The prediction of the bagged model is the average of the predictions from those two models (both of which happen to have the

x	3	5	5	$L(\mathbf{y}, \hat{\mathbf{y}})$
y	5	2	2	
$\hat{y}_{s=4}$	5	2	2	**0**
$\hat{y}_{s=6}$	3	3	3	2

Table 2: Stump trained on $\mathcal{T}^{(1)}$.

x	1	5	7	$L(\mathbf{y}, \hat{\mathbf{y}})$
y	0	2	1	
$\hat{y}_{s=4}$	0	3/2	3/2	**1/2**
$\hat{y}_{s=6}$	1	1	1	1

Table 3: Stump trained on $\mathcal{T}^{(2)}$.

x	1	3	5	7	$L(\mathbf{y}, \hat{\mathbf{y}})$
y	0	5	2	1	
$\hat{y}_{\text{bagging}}$	5/2	5/2	7/4	7/4	5/2

Table 4: Predictions of the bagged model.

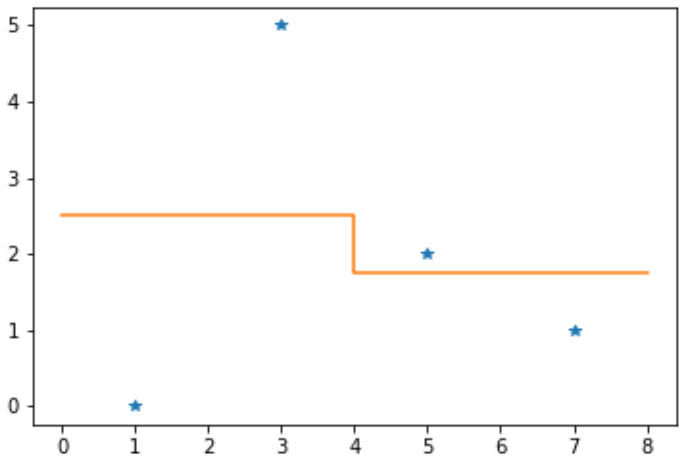

Figure 3: Prediction from bagged model.

same cutpoint, $s = 4$), as shown in Table 4 and plotted in Figure 3.

(d) The leaf $s < 2$ has only one element, thus $\hat{y}_{\text{minimax}}(1) = y(0) = 0$ attains the lower-bound on the loss: 0. The leaf $s \geq 2$ has three

elements, and the prediction is the solution to the optimization problem:

$$\hat{y}(x \geq 2)_{\text{minimax}} = \arg\min_{y} \left(\max\left(|5-y|, |2-y|, |1-y|\right)\right) = 3.$$

This solution can be found, for instance, by plotting each component separately and finding the minimum of the pointwise max. Table 5 summarizes the predictions.

x	1	3	5	7	$L(\mathbf{y}, \hat{\mathbf{y}})$
y	0	5	2	1	
$\hat{y}_{\text{minimax}}$	0	3	3	3	2

Table 5: Predictions of the minimax model.

(e) Bagging is a method for reducing variance, which can improve the performance of base models with low bias and high variance such as *deep* regression trees. In contrast, a shallow regression tree, such as a stump, has high bias and lower variance. Thus, bagging is not likely to improve the performance.

Boosting, on the other hand, is a method for reducing bias that is, in fact, often used with shallow regression trees.

5. (a)

$$\begin{aligned}1-\sigma(z) &= 1-\frac{1}{1+\exp(-z)} = \frac{1+\exp(-z)}{1+\exp(-z)} - \frac{1}{1+\exp(-z)} \\ &= \frac{1+\exp(-z)-1}{1+\exp(-z)} = \frac{\exp(-z)}{1+\exp(-z)} \\ &= \frac{\frac{\exp(z)}{\exp(z)}}{\exp(z)+\frac{\exp(z)}{\exp(z)}} = \frac{1}{\exp(z)+1} = \sigma(-z)\end{aligned}$$

(2p)

(b) i. We obtain $\begin{pmatrix}13 & -16 & -1\\ 20 & 28 & 22\end{pmatrix}$,
where for example $4\cdot 2+3\cdot 4+0\cdot -1=20$. (2p)

ii. If the vectorization is done row-wise, i.e.,

$$\mathbf{x} = \text{vec}\left[\begin{pmatrix}x_1 & x_2 & x_3\\ x_4 & x_5 & x_6\\ x_7 & x_8 & x_9\\ x_{10} & x_{11} & x_{12}\end{pmatrix}\right] = \begin{pmatrix}x_1\\ x_2\\ x_3\\ \vdots\\ x_{11}\\ x_{12}\end{pmatrix}$$

we obtain

$$\mathbf{wx} = \mathbf{s}$$

$$\begin{pmatrix}f_1 & 0 & 0 & f_2 & 0 & 0 & f_3 & 0 & 0 & 0 & 0 & 0\\ 0 & f_1 & 0 & 0 & f_2 & 0 & 0 & f_3 & 0 & 0 & 0 & 0\\ 0 & 0 & f_1 & 0 & 0 & f_2 & 0 & 0 & f_3 & 0 & 0 & 0\\ 0 & 0 & 0 & f_1 & 0 & 0 & f_2 & 0 & 0 & f_3 & 0 & 0\\ 0 & 0 & 0 & 0 & f_1 & 0 & 0 & f_2 & 0 & 0 & f_3 & 0\\ 0 & 0 & 0 & 0 & 0 & f_1 & 0 & 0 & f_2 & 0 & 0 & f_3\end{pmatrix}\begin{pmatrix}x_1\\ x_2\\ x_3\\ \vdots\\ x_{11}\\ x_{12}\end{pmatrix} = \begin{pmatrix}s_1\\ s_2\\ s_3\\ s_4\\ s_5\\ s_6\end{pmatrix}$$

In case of a column-wise vectorization, i.e.,

$$\mathbf{x} = \text{vec}\left[\begin{pmatrix}x_1 & x_2 & x_3\\ x_4 & x_5 & x_6\\ x_7 & x_8 & x_9\\ x_{10} & x_{11} & x_{12}\end{pmatrix}\right] = \begin{pmatrix}x_1\\ x_4\\ x_7\\ x_{10}\\ \vdots\\ x_9\\ x_{12}\end{pmatrix}$$

we get

$$\mathbf{wx} = \mathbf{s}$$

$$\begin{pmatrix} f_1 & f_2 & f_3 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & f_1 & f_2 & f_3 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & f_1 & f_2 & f_3 & 0 \\ 0 & f_1 & f_2 & f_3 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & f_1 & f_2 & f_3 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & f_1 & f_2 & f_3 \end{pmatrix} \begin{pmatrix} x_1 \\ x_4 \\ x_7 \\ x_{10} \\ \vdots \\ x_9 \\ x_{12} \end{pmatrix} = \begin{pmatrix} s_1 \\ s_2 \\ s_3 \\ s_4 \\ s_5 \\ s_6 \end{pmatrix}$$

(3p)

(c) For $i = j$ we have

$$\begin{aligned} \frac{\partial p_j}{\partial x_i} &= \frac{\exp(x_j)\left(\sum_k \exp(x_k)\right) - \exp(x_j)\exp(x_i)}{\left(\sum_k \exp(x_k)\right)^2} && \text{(derivative of a rational)} \\ &= \frac{\exp(x_j)}{\sum_k \exp(x_k)} \cdot \frac{\left(\sum_k \exp(x_k)\right) - \exp(x_i)}{\sum_k \exp(x_k)} && \text{(separate } \exp(x_j)\text{)} \\ &= \frac{\exp(x_j)}{\sum_k \exp(x_k)} \cdot \left(\frac{\left(\sum_k \exp(x_k)\right)}{\sum_k \exp(x_k)} - \frac{\exp(x_i)}{\sum_k \exp(x_k)} \right) \\ &= p_j \cdot (1 - p_i) \end{aligned}$$

and for $i \neq j$ we get

$$\begin{aligned} \frac{\partial p_j}{\partial x_i} &= \frac{0 - \exp(x_j)\exp(x_i)}{\left(\sum_k \exp(x_k)\right)^2} \\ &= \frac{\exp(x_j)}{\sum_k \exp(x_k)} \cdot \frac{-\exp(x_i)}{\sum_k \exp(x_k)} \\ &= p_j \cdot (0 - p_i). \end{aligned}$$

Combined, that yields

$$\frac{\partial p_j}{\partial x_i} = p_j(\delta_{ij} - p_i),$$

where δ_{ij} is 1 if $i = j$ and 0 otherwise.

(3p)