EXAM IN STATISTICAL MACHINE LEARNING
STATISTISK MASKININLÄRNING

DATE AND TIME: June 8, 2022, 8.00–13.00

RESPONSIBLE TEACHER: Dave Zachariah

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!

Formula sheet for Statistical Machine Learning

Warning: This is not a complete list of formulas used in the course, some exam problems may require expressions not listed here. Furthermore, the formulas below are not self-explanatory, you need to be familiar with the expressions to interpret them.

The Gaussian distribution: The probability density function of the p-dimensional Gaussian distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ is

$$\mathcal{N}\left(\mathbf{x} \,|\, \boldsymbol{\mu},\, \boldsymbol{\Sigma}\right) = \frac{1}{(2\pi)^{p/2}\sqrt{\det \boldsymbol{\Sigma}}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right), \qquad \mathbf{x} \in \mathbb{R}^p.$$

Sum of identically distributed variables: For identically distributed random variables $\{z_i\}_{i=1}^n$ with mean μ, variance σ^2 and average correlation between distinct variables ρ, it holds that $\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n z_i\right] = \mu$ and $\text{Var}\left(\frac{1}{n}\sum_{i=1}^n z_i\right) = \frac{1-\rho}{n}\sigma^2 + \rho\sigma^2$.

Linear regression and regularization:

- The least-squares estimate of $\boldsymbol{\theta}$ in the linear regression model

$$y = \theta_0 + \sum_{j=1}^{p} \theta_j x_j + \epsilon$$

  is given by the solution $\widehat{\boldsymbol{\theta}}_{\text{LS}}$ to the normal equations $\mathbf{X}^{\mathsf{T}}\mathbf{X}\widehat{\boldsymbol{\theta}}_{\text{LS}} = \mathbf{X}^{\mathsf{T}}\mathbf{y}$, where

$$\mathbf{X} = \begin{bmatrix} 1 & -\mathbf{x}_1^{\mathsf{T}}- \\ 1 & -\mathbf{x}_2^{\mathsf{T}}- \\ \vdots & \vdots \\ 1 & -\mathbf{x}_n^{\mathsf{T}}- \end{bmatrix} \text{ and } \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \text{ from the training data } \mathcal{T} = \{\mathbf{x}_i, y_i\}_{i=1}^n$$

- Ridge regression uses the regularization term $\lambda\|\boldsymbol{\theta}\|_2^2 = \lambda\sum_{j=0}^p \theta_j^2$.
  The ridge regression estimate is $\widehat{\boldsymbol{\theta}}_{\text{RR}} = (\mathbf{X}^{\mathsf{T}}\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^{\mathsf{T}}\mathbf{y}$.
- LASSO uses the regularization term $\lambda\|\boldsymbol{\theta}\|_1 = \lambda\sum_{j=0}^p |\theta_j|$.

Maximum likelihood: The maximum likelihood estimate is given by

$$\widehat{\boldsymbol{\theta}}_{\text{ML}} = \arg\max_{\boldsymbol{\theta}} \ln \ell(\boldsymbol{\theta})$$

where $\ln \ell(\boldsymbol{\theta}) = \sum_{i=1}^{n} \ln p(y_i \,|\, \mathbf{x}_i; \boldsymbol{\theta})$ is the log-likelihood function (the last equality holds when the n training data points are modeled to be independent).

Logistic regression: The logistic regression combines linear regression with the logistic function to model the class probability

$$p(y = 1 \,|\, \mathbf{x}_i) = \frac{e^{\boldsymbol{\theta}^{\mathsf{T}}\mathbf{x}_i}}{1 + e^{\boldsymbol{\theta}^{\mathsf{T}}\mathbf{x}_i}}.$$

For multi-class logistic regression we use the *softmax* function and model

$$p(y = m \,|\, \mathbf{x}_i) = \frac{e^{\boldsymbol{\theta}_m^{\mathsf{T}}\mathbf{x}_i}}{\sum_{j=1}^{M} e^{\boldsymbol{\theta}_j^{\mathsf{T}}\mathbf{x}_i}}.$$

Discriminant Analysis: The linear discriminant analysis (LDA) classifier models $p(y \,|\, \mathbf{x})$ using Bayes' theorem and the following assumptions

$$p(y = m \,|\, \mathbf{x}) = \frac{p(\mathbf{x} \,|\, m) p(y = m)}{\sum_{j=1}^{M} p(\mathbf{x} \,|\, j) p(y = j)} = \frac{\mathcal{N}\left(\mathbf{x} \,|\, \widehat{\boldsymbol{\mu}}_m, \widehat{\boldsymbol{\Sigma}}\right) \widehat{\pi}_m}{\sum_{j=1}^{M} \mathcal{N}\left(\mathbf{x} \,|\, \widehat{\boldsymbol{\mu}}_j, \widehat{\boldsymbol{\Sigma}}\right) \widehat{\pi}_j},$$

where

$$\begin{aligned}
\widehat{\pi}_m &= n_m/n \text{ for } m = 1, \dots, M \\
\widehat{\boldsymbol{\mu}}_m &= \frac{1}{n_m} \sum_{i:y_i=m} \mathbf{x}_i \text{ for } m = 1, \dots, M \\
\widehat{\boldsymbol{\Sigma}} &= \frac{1}{n - M} \sum_{m=1}^{M} \sum_{i:y_i=m} (\mathbf{x}_i - \widehat{\boldsymbol{\mu}}_m)(\mathbf{x}_i - \widehat{\boldsymbol{\mu}}_m)^{\mathsf{T}}.
\end{aligned}$$

For quadratic discriminant analysis (QDA), the model is

$$p(y = m \,|\, \mathbf{x}) = \frac{\mathcal{N}\left(\mathbf{x} \,|\, \widehat{\boldsymbol{\mu}}_m, \widehat{\boldsymbol{\Sigma}}_m\right) \widehat{\pi}_m}{\sum_{j=1}^{M} \mathcal{N}\left(\mathbf{x} \,|\, \widehat{\boldsymbol{\mu}}_j, \widehat{\boldsymbol{\Sigma}}_j\right) \widehat{\pi}_j},$$

where $\widehat{\boldsymbol{\mu}}_m$ and $\widehat{\pi}_m$ are as for LDA, and

$$\widehat{\boldsymbol{\Sigma}}_m = \frac{1}{n_m - 1} \sum_{i:y_i=m} (\mathbf{x}_i - \widehat{\boldsymbol{\mu}}_m)(\mathbf{x}_i - \widehat{\boldsymbol{\mu}}_m)^{\mathsf{T}}.$$

Classification trees: The cost function for tree splitting is $\sum_{\ell=1}^{|T|} n_\ell Q_\ell$ where T is the tree, $|T|$ the number of terminal nodes, n_ℓ the number of training data points falling in node ℓ, and Q_ℓ the impurity of node ℓ. Three common impurity measures for splitting classification trees are:

$$\begin{aligned} \text{Misclassification error:} \qquad & Q_\ell = 1 - \max_m \widehat{\pi}_{\ell m} \\ \text{Gini index:} \qquad & Q_\ell = \sum_{m=1}^{M} \widehat{\pi}_{\ell m}(1 - \widehat{\pi}_{\ell m}) \\ \text{Entropy/deviance:} \qquad & Q_\ell = -\sum_{m=1}^{M} \widehat{\pi}_{\ell m} \log \widehat{\pi}_{\ell m} \end{aligned}$$

where $\widehat{\pi}_{\ell m} = \frac{1}{n_\ell} \sum_{i:\, \mathbf{x}_i \in R_\ell} \mathbb{I}(y_i = m)$

Loss functions for classification: For a binary classifier expressed as $\widehat{y}(\mathbf{x}) = \operatorname{sign}\{C(\mathbf{x})\}$, for some real-valued function $C(\mathbf{x})$, the margin is defined as $y \cdot C(\mathbf{x})$ (note the convention $y \in \{-1, 1\}$ here). A few common loss functions expressed in terms of the margin, $L(y, C(\mathbf{x}))$ are,

$$\begin{aligned} \text{Exponential loss:} \qquad & L(y,c) = \exp(-yc). \\ \text{Hinge loss:} \qquad & L(y,c) = \begin{cases} 1 - yc & \text{for } yc < 1, \\ 0 & \text{otherwise.} \end{cases} \\ \text{Binomial deviance:} \qquad & L(y,c) = \log(1 + \exp(-yc)). \\ \text{Huber-like loss:} \qquad & L(y,c) = \begin{cases} -yc & \text{for } yc < -1, \\ \frac{1}{4}(1 - yc)^2 & \text{for } -1 \leq yc \leq 0, \\ 0 & \text{otherwise.} \end{cases} \\ \text{Misclassification loss:} \qquad & L(y,c) = \begin{cases} 1 & \text{for } yc < 0, \\ 0 & \text{otherwise.} \end{cases} \end{aligned}$$

1. This problem is composed of 10 true-or-false statements. You only have to classify these as either `true` or `false`. For this problem *(only!)* no motivation is required. Each correct answer scores 1 point and each incorrect answer scores -1 point (capped at 0 for the whole problem). Answers left blank score 0 points.

   i. QDA is a non-linear classification algorithm and always has a quadratic decision surface.

   ii. An underfitted model has high bias and low variance. Therefore, it shows a low accuracy on trainset and high accuracy on testset.

   iii. Both CART and K-Nearest Neighbor are non-parametric.

   iv. In A Linear Regression model with Gaussian noise, MLE and MSE always give the same result.

   v. Dropout is a regularization technique which prevents overfitting and generalizes the model.

   vi. Compared to Bagging, Random Forest on the same trainset performs better by decreasing the bias and increasing the variance.

   vii. When using bagging, an out-of-bag estimate of the expected new data error E_{new} is computationally much cheaper than a k-fold cross-validation.

   viii. Like Bagging technique, cross-validation helps to reduce the flexibility of model.

   ix. In neural networks, sigmoid activation function is a common choice for the last layer in a multi-class classification problem.

   x. Standard Gradient Descent is guaranteed to converge and find the global minimum.

(10p)

2. Consider a small training dataset with only $N = 3$ data points with x as input and y as output

x	-1.0	0.0	1.0
y	0.0	0.0	-1.0

We want to model the output y based on the input x. We opt between two regression models to fit the data.

$$y = \alpha_0 + \epsilon \tag{1}$$

$$y = \beta_1 x + \beta_0 + \epsilon \tag{2}$$

where ϵ is a measurement error.

(a) For both of the two models (1) and (2), find the parameters $\hat{\alpha}_0$, and $\hat{\beta} = [\hat{\beta}_0 \ \hat{\beta}_1]^\mathsf{T}$ that minimizes the mean-square errors (MSE) on the training data above. (3p)

(b) The training MSE for the model (2) will always be equal or smaller than the training MSE for model (1) regardless of the training data. Explain why! (2p)

(c) Evaluate the two linear regression models (1) and (2) by performing leave-one-out cross-validation based on MSE on each of the two models. Which model performs the best?

Hint: leave-one-out cross-validation is the same as k-fold cross-validation where k=n. (5p)

3. Consider a binary classification problem $y \in \{0, 1\}$ with one input variable x_1 and the following $N = 4$ training data points:

x_1	-3	-1	1	7.2
y	0	1	1	0

With this training data, Alice has trained in total six different classifiers, using (i) linear discriminant analysis, (ii) quadratic discriminant analysis, (iii) logistic regression, (iv) k-Nearest Neigbour with $k = 1$, (v) k-Nearest Neigbour with $k = 3$, and (vi) a classification tree with a single binary split based on missclassification error, respectively.

(a) To evaluate the performance, she computed the missclassification error on the training data for each of the six classifiers and got the number 50% for two of the classifiers, 25% for two other classifiers and 0% for the remaining two classifiers. Which missclassification error belongs to which classifier? (6p)

Hint: It is possible to solve the problem without any complicated calculations.

A correct pairing without a proper motivation scores zero points!

(b) Consider the same training data as above, but with an additional input variable x_2

x_2	0	0	0	10

The classifiers have been retrained using the two-dimensional input $\mathbf{x} = [x_1,\ x_2]^\mathsf{T}$. Does the missclassification error on the training data for the classifiers (iii), (v) and (vi) improve with this additional input variable? (3p)

(c) How would classifier (v) classify the test point $\mathbf{x}^* = [x_1^*,\ x_2^*]^\mathsf{T} = [-2,\ 10]^\mathsf{T}$. (1p)

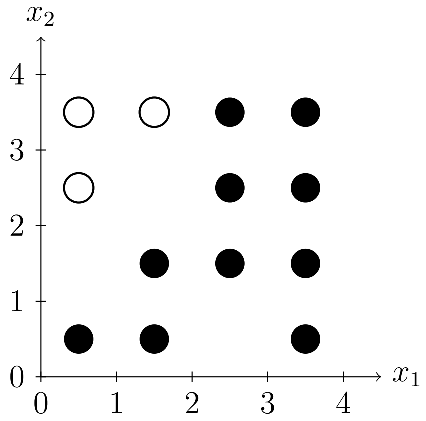

Figure 1: Training data point with two inputs x_1 and x_2 and two possible classes, illustrated by $y =$ ● and $y =$ ○, respectively.

4. Classification trees are constructed greedily using recursive binary splitting. Each split is typically made in order to minimize the cost function,

$$C(T) = \sum_{m=1}^{|T|} N_m Q_m$$

where T is the tree object, $|T|$ is the total number of terminal nodes in the tree, N_m is the number of training data points in the region corresponding to the terminal node m, and Q_m is a node impurity measure of the same region. Two common measures are

$$\text{Misclassification error: } Q_m = 1 - \max_k \hat{p}_{mk},$$

$$\text{Gini index: } Q_m = \sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk}),$$

where $\hat{p}_{mk}$ is the proportion of training data points with label k in the region corresponding to node m. Since the classification for any data point within a certain region is decided with a 'majority vote', the proportion of correct training data classifications in region m can thus be expressed as $\max_k \hat{p}_{mk}$, and the proportion of misclassifications as $1 - \max_k \hat{p}_{mk}$.

(a) Consider a binary classification problem with two inputs and a training data set illustrated in Figure 1. Give an optimal classification tree with two split points (three terminal nodes) for minimizing the training misclassification loss. An intuitive motivation for the optimality of the proposed solution is sufficient. (2p)

(b) For the training data in Figure 1, what is the resulting classification tree with two split points obtained by recursive binary splitting, using misclassification error as node impurity measure? Explain the suboptimality of the solution. (5p)

(c) Consider the first split point. Compute the cost $C(T)$ for the two candidate splits:

    i. $R_1 = \{\mathbf{x} : x_1 \leq 1\}$, $R_2 = \{\mathbf{x} : x_1 > 1\}$, and

    ii. $R_1 = \{\mathbf{x} : x_1 \leq 2\}$, $R_2 = \{\mathbf{x} : x_1 > 2\}$,

    respectively, if we instead *use the Gini index as impurity measure.* Based on the result, why might the Gini index be preferable over misclassification error when growing the classification tree?

    (3p)

5. (a) Briefly describe (max $\sim \frac{1}{2}$ page) how K-fold cross-validation works and what it is used for. (3p)

(b) Consider a regression problem $y = f(x) + \epsilon$. Assume (as is often the case in practice) that we have access to a family of models with different levels of flexibility. Make a sketch of the typical shapes of the following curves, as the model flexibility goes from low to high:

i. The squared model bias.

ii. The model variance.

iii. The irreducible error, $\mathrm{Var}(\epsilon)$.

iv. The training mean-squared error.

v. The expected test mean-squared error.

Sketch all five curves in the empty figure axes handed out together with the exam, marked **”Problem 5b: Sketch the curves here”**. Clearly mark all curves in the plot with **i–v**. You should also explain and motivate the shape of each curve, as well as how they relate to each other. (7p)

Don’t forget to attach the sketch when handing in the solution!