EXAM IN STATISTICAL MACHINE LEARNING
STATISTISK MASKININLÄRNING

DATE AND TIME: January 10, 2023, 14.00-19.00

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 side 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

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

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 $\mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^n z_i\right) = \frac{1-\rho}{n}\sigma^2 + \rho\sigma^2$.

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 $\widehat{\boldsymbol{\theta}}_{\mathrm{LS}}$ to the normal equations $\mathbf{X}^{\mathsf{T}}\mathbf{X}\widehat{\boldsymbol{\theta}}_{\mathrm{LS}} = \mathbf{X}^{\mathsf{T}}\mathbf{y}$, 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 $\widehat{\boldsymbol{\theta}}_{\mathrm{RR}} = (\mathbf{X}^{\mathsf{T}}\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^{\mathsf{T}}\mathbf{y}$.
- 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

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

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

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

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:

$$\text{Misclassification error:} \quad Q_\ell = 1 - \max_m \widehat{\pi}_{\ell m}$$

$$\text{Gini index:} \quad Q_\ell = \sum_{m=1}^{M} \widehat{\pi}_{\ell m}(1 - \widehat{\pi}_{\ell m})$$

$$\text{Entropy/deviance:} \quad Q_\ell = -\sum_{m=1}^{M} \widehat{\pi}_{\ell m} \log \widehat{\pi}_{\ell m}$$

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 $\widehat{y}(\mathbf{x}) = \text{sign}\{C(\mathbf{x})\}$, 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,

$$\text{Exponential loss:} \quad L(y, c) = \exp(-yc).$$

$$\text{Hinge loss:} \quad L(y, c) = \begin{cases} 1 - yc & \text{for } yc < 1, \\ 0 & \text{otherwise.} \end{cases}$$

$$\text{Binomial deviance:} \quad L(y, c) = \log(1 + \exp(-yc)).$$

$$\text{Huber-like loss:} \quad L(y, c) = \begin{cases} -yc & \text{for } yc < -1, \\ \frac{1}{4}(1 - yc)^2 & \text{for } -1 \leq yc \leq 0, \\ 0 & \text{otherwise.} \end{cases}$$

$$\text{Misclassification loss:} \quad L(y, c) = \begin{cases} 1 & \text{for } yc < 0, \\ 0 & \text{otherwise.} \end{cases}$$

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. In classification problems, the input variables are always categorical.

   ii. Learning a classifier with the logistic loss always produces a linear decision boundary in input space.

   iii. The model sensitivity of k-NN typically decreases as k increases.

   iv. The quadratic discriminant analysis classifier is only applicable to inputs that follow a Gaussian distribution.

   v. The squared-error performance of a learning method for regression can be decomposed into three terms:

   $$\mathbb{E}[(f_0(\mathbf{x}_\star) - \bar{f}(\mathbf{x}_\star))^2] + \mathbb{E}[(f(\mathbf{x}_\star; \widehat{\boldsymbol{\theta}}(\mathcal{T})) - \bar{f}(\mathbf{x}_\star))^2] + \sigma^2,$$

   where the first term can be reduced by using a more flexible family of models.

   vi. If the second term in the sum above is dominant, we call this 'underfitting'.

   vii. Ridge regression is typically used as an input selection method.

   viii. When assessing the out-of-sample performance of a given learned model $f(\mathbf{x}; \widehat{\boldsymbol{\theta}})$, its new expected error can be estimated using k-fold cross-validation.

   ix. Consider learning a linear classification model $f(\mathbf{x}; \boldsymbol{\theta}) = \text{sign}(\mathbf{x}^\top \boldsymbol{\theta})$ using the missclassification loss. The gradient descent method will find a model that has the minimum average missclassification.

   x. Stochastic gradient descent can be faster than ordinary gradient descent even though it requires more search steps.

(10p)

2. (a) Consider a regression problem in which we observe the following training data:

x_1	2	4	6
y	4	7	11

Let $\mathbf{x} = [1 \; x_1]^\mathsf{T}$ and show that the linear model learned using the least-squares method equals:

$$f(\mathbf{x}; \widehat{\boldsymbol{\theta}}) = \mathbf{x}^\mathsf{T}\widehat{\boldsymbol{\theta}}, \quad \text{where} \quad \widehat{\boldsymbol{\theta}} = \begin{bmatrix} \widehat{\theta}_0 \\ \widehat{\theta}_1 \end{bmatrix} = \begin{bmatrix} \frac{1}{3} \\ \frac{7}{3} \end{bmatrix}$$

(6p)

(b) The data is drawn from a distribution $p(\mathbf{x}, y)$ such that the expected value of x_1 is zero, i.e. $\mathbb{E}[x_1] = 0$ and the output can be expressed as

$$y = \alpha x_1 + \varepsilon,$$

where ε is independent of x_1 and

$$\mathbb{E}[\varepsilon] = 0 \qquad \mathbb{E}[\varepsilon^2] = \sigma^2$$

Show the expected new squared error can be expressed as:

$$\mathbb{E}[(y - f(\mathbf{x}; \widehat{\boldsymbol{\theta}}))^2] = \left((\alpha - \widehat{\theta}_1)^2 \mathbb{E}[x_1^2] + \widehat{\theta}_0^2\right) + \sigma^2$$

and provide an interpretation of the two terms in the expression. (3p)

(c) Suppose we manage to obtain additional information in the form of a second input, with which we extend our linear model.

x_2	3	4	5

However when learning the least-squares linear model, we encounter a problem. What is the problem? (1p)

3. a) Consider a binary classification problem with training data set $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$ and labels $y_i \in \{-1, 1\}$ (the n data points are assumed to be independent). Logistic regression is a parametric model, defined according to,

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

Write down the corresponding expression for $p(y = -1|\mathbf{x}; \boldsymbol{\theta})$, and then show that the maximum likelihood estimate $\hat{\boldsymbol{\theta}}_{\mathrm{ML}}$ is given by,

$$\hat{\boldsymbol{\theta}}_{\mathrm{ML}} = \arg\min_{\boldsymbol{\theta}} \sum_{i=1}^{n} \ln(1 + e^{-y_i\boldsymbol{\theta}^\mathsf{T}\mathbf{x}_i}).$$

(5p)

b) Show that the decision boundary of the logistic regression model is a linear hyperplane. Based on this decision boundary, show that hard predictions (positive or negative class) of the model can be written as $f(\mathbf{x}_*; \widehat{\boldsymbol{\theta}}) = \mathrm{sign}(\widehat{\boldsymbol{\theta}}^\mathsf{T}\mathbf{x}_*)$. (3p)

c) If the optimal decision boundary for the classification problem is linear, can we expect logistic regression to perform better than k-NN (with $k = 1$) on the *test* set? How about on the *training* set? Justify your answer. (2p)

4. We consider the following binary classification problem: The presence of a medication, encoded by $y \in \{-1, 1\}$, leads to a reaction that changes the concentration of a chemical compound as measured by a numerical input x (one-dimensional biomarker).

   a) Suppose probability of a sample being subject to medication is $p(y = 1) = \frac{1}{3}$ and that under medication the distribution of biomarkers is described by $\mathcal{N}(x; 2, 4)$. When no medication is applied, the distribution of biomarkers is $\mathcal{N}(x; 0, 2)$

   Sketch the shapes and locations of the two conditional distributions $p(x|y = 1)$ and $p(x|y = -1)$, respectively, along the x-axis. (2p)

   b) Show that the best classifier (with respect to the missclassification loss) is given by:

$$f_0(x) = \begin{cases} 1, & -\frac{(x-2)^2}{2} + x^2 > \ln 32 \\ -1, & \text{otherwise} \end{cases}$$

   Hint: Use the chain rule $p(x, y) = p(y|x)p(x) = p(x|y)p(y)$

   (7p)

   c) Suppose we observe a biomarker $x_\star = 1$, determine the predicted output $f_0(x_\star)$. (1p)

5. Consider a classification problem with two input variables x_1 and x_2, and one output y. Based on the training data

x_1	x_2	y
2.4	2.6	red
3.2	2.6	red
2.0	3.0	blue
0.3	0.2	red
3.4	3.5	blue
2.8	3.1	blue
2.0	1.6	blue
1.5	2.2	blue

Alice has constructed the classification tree in Figure 1 using recursive binary splitting.

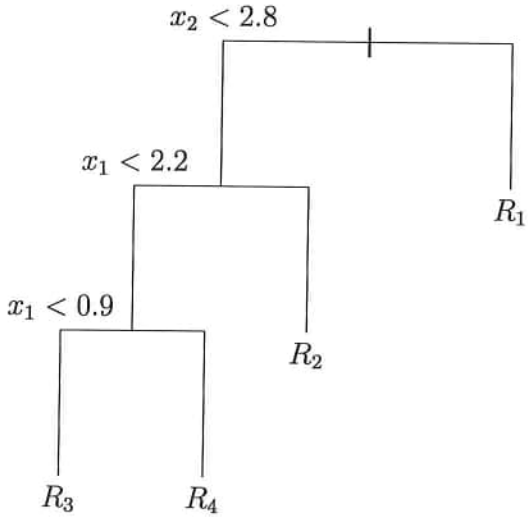

Figure 1: Alice's decision tree of the input space for Problem 5.

(a) Draw the corresponding input partitioning to this tree. Mark the regions with the names of the leaf nodes, $R_1, \ldots, R_4$. (1p)

(b) Use the classification tree to predict the output of the test input $x_\star = \begin{bmatrix} x_1^\star & x_2^\star \end{bmatrix}^\mathsf{T} = \begin{bmatrix} 1.6 & 1.8 \end{bmatrix}^\mathsf{T}$. (1p)

Based on the same training data but using recursive binary splitting with a different impurity measure, Bob has constructed the decision tree as shown in Figure 2.

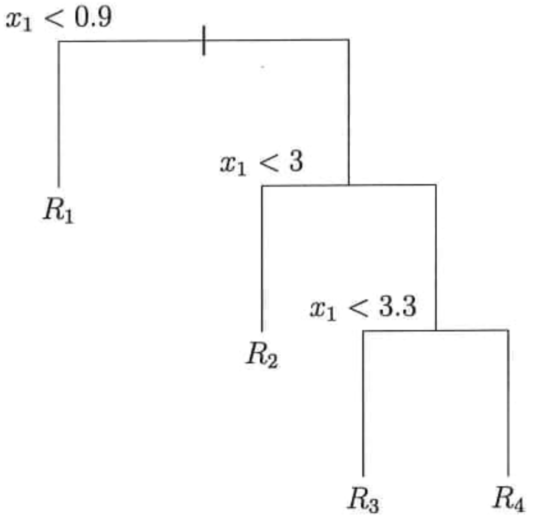

Figure 2: Bob's decision tree of the input space for Problem 5.

(c) We know that either Alice or Bob used the Gini index to construct their tree, and that the other one used the classification error rate. Explain with which of the two measures Alice constructed the tree in Figure 1 and with which measure Bob constructed the tree in Figure 2. (2p)

Alice is the leader of a team in an international machine learning competition where each team is tasked with solving a classification problem. The teams have access to a static training data set, and each team is required to submit a well-chosen model to be tested on a top-secret test set by a jury. The winning team is the team achieving the highest accuracy on this test set.

After exploring a number of different model families, Alice's team decides to move forward with decision trees.

(d) To improve the predictive performance, the team analyzes the model performance on the validation set using a bagging classifier with decision trees as base models. Explain briefly what bagging is and why it can help improve performance on the test set. (2p)

(e) It turns out that bagging does not improve the validation performance as much as the team expects. Give a short mathematical explanation for this, and describe why this can be the case for bagging with decision tree base models in particular. (2p)

(f) You suggest that the team tries to improve the performance using a random forest classifier. When learning a random forest classifier, each split is optimized based on a random subset consisting of $q \leq p$ out of the original p input features. Explain why this can improve the performance on the test set compared to standard bootstrap aggregation. (1p)

(g) How can the team choose the size q of the random subset of input features in a systematical way? (1p)