EXAM IN STATISTICAL MACHINE LEARNING
STATISTISK MASKININLÄRNING

DATE AND TIME: August 16, 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!

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

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

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

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

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

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. Regression models have quantitative outputs.

   ii. A classifier $\widehat{G}(X)$ is said to be linear if the function $\widehat{G}$, which maps each input to a predicted class, is a linear function of the model parameters.

   iii. LASSO regularization can be used as an input selection method.

   iv. The Bayes classifier can not be implemented in practice, but if it could it would always attain zero test error.

   v. The correlation between any pair of ensemble members of a bagged regression model

   $$\hat{f}^B_{\text{bag}}(x) = \frac{1}{B}\sum_{b=1}^{B} \hat{f}^{\star b}(x)$$

   tends to zero as the number of ensemble members B tends to infinity.

   vi. Misclassification loss is sensitive to outliers, i.e. incorrectly classified training data points far from the decision boundary.

   vii. The model bias of k-NN typically increases as k increases.

   viii. Quadratic discriminant analysis is a parametric model.

   ix. The model bias typically tends to zero as the number of training data points tends to infinity.

   x. Probabilistic models assign probability distributions to unknown model parameters. (10p)

2. (a) Explain briefly (a couple of sentences) the difference between classification and regression problems. (2p)

(b) Explain briefly ($\sim$ 0.5 page) the meaning of the bias-variance trade-off. I.e., what do we mean by model bias and model variance, and why is there a trade-off between the two? (5p)

(c) A friend of yours is faced with a regression problem with two possible inputs, X_1 and X_2. S/he considers two linear regression models:

(M1) $Y = \beta_0 + \beta_1 X_1 + \varepsilon$,

(M2) $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$.

Both models are fitted to a training data set $\mathcal{T} = \{(x_i, y_i)\}_{i=1}^{N}$ using least-squares, resulting in the two prediction models

(P1) $\hat{Y} = 12.9 + 3.2X_1$,

(P2) $\hat{Y} = 11.6 - 1.4X_1 + 1.7X_2$,

respectively. Your friend is puzzled by these results and comes to you for advice. S/he says:

> "In model (P1) a unit increase in X_1 results in an increase of the predicted output by 3.2 units, i.e. it is clear that Y is positively correlated with X_1. However, in model (P2) a unit increase in X_1 instead results in a *decrease* of 1.4 units in the predicted output, i.e. now X_1 appears to be negatively correlated with Y!"

Give a plausible explanation to your friend's dilemma. (3p)

3. (a) Draw the graph corresponding to a dense neural network for regression with p input variables $X_1, \dots, X_p$, one hidden layer with M units, activation function $\sigma : \mathbb{R} \mapsto \mathbb{R}$, and output $Z \in \mathbb{R}$. How many parameters does the model have (including offsets)? (3p)

   (b) Show that the model in 5(a) reduces to a linear regression model if $\sigma(x) = x$. Specifically, show how the parameters of the neural network relate to the parameters of the linear regression model

$$Z = \beta_0 + \sum_{j=1}^{p} \beta_j X_j.$$

(4p)

   (c) A boosted regression model can be written as

$$\widehat{f}^B_{\text{boost}}(X) = \sum_{b=1}^{B} \widehat{f}^b(X)$$

Assume that each ensemble member $\widehat{f}^b(X)$ is a *stump* (i.e. a regression tree with a single split). Show that the boosted model can be written as a so called *additive model*,

$$\widehat{f}^B_{\text{boost}}(X) = \sum_{j=1}^{p} \widehat{f}_j(X_j).$$

Note that the latter expression is a sum over the p input variables and each term of the sum depends on only one input variable. (3p)

4. Consider a regression problem with a two input variables X_1 and X_2, and one output Y. Based on the following training data

X_1	1.4	0.2	1.8	1.2	1.6	1.8	1.1	0.2
X_2	1.5	0.9	1.2	0.9	0.2	0.9	0.7	1.8
Y	0.5	0.2	0.7	0.7	0.1	0.7	0.6	0.1

Bob has constructed a regression tree shown in Figure 1 using recursive binary splitting.

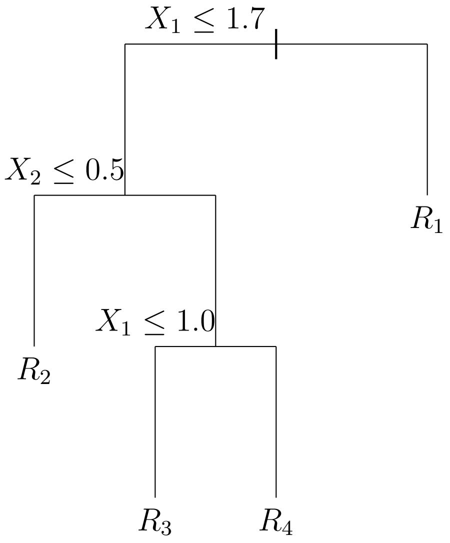

Figure 1: Decision tree of the input space for Problem 4

(a) Draw the corresponding input partitioning to this tree. Mark the regions with the names of the leaf nodes $R_1, \ldots, R_4$. (2p)

(b) Use the regression tree to predict the output of the test input $X^\star = [X_1^\star,\ X_2^\star]^\mathsf{T} = [1.5\ 1.8]^\mathsf{T}$ (3p)

(c) Continue to grow the tree in Figure 1 such that there are at most two data points in each region by minimizing the mean-square-error. Which region(s) do you split where? (there are multiple possible splits that are equally good) (3p)

(d) Explain briefly (a couple of sentences) the disadvantage of growing a decision tree too deep. (2p)

5. (a) Consider the scatter plots in Figure 2 which depict three different training data sets for three binary classification problems. In which dataset(s) could the classes be well separated by...

- ...an LDA classifier?
- ...a QDA classifier?

(In both cases we assume that X_1 and X_2 are the only inputs to the classifiers.)

(2p)

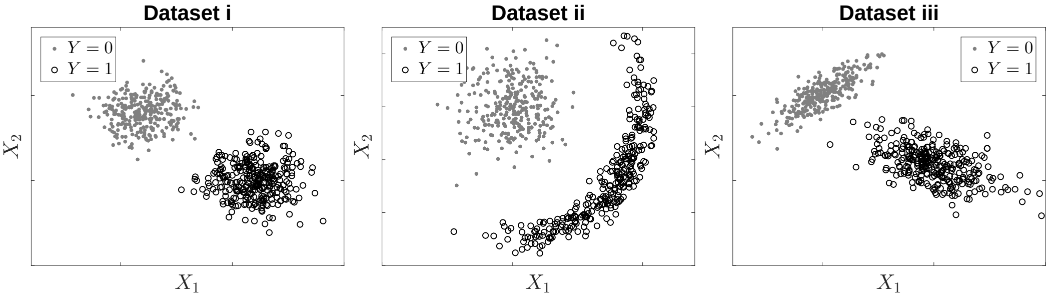

Figure 2: Scatter plots of training data for Problems 5a and 5b.

(b) Consider again the scatter plots in Figure 2. The LDA and QDA classifiers are based on different *assumptions* about the properties of the data. Which dataset(s) in Figure 2 appear to correspond well to the assumptions made by LDA and QDA, respectively? (4p)

(c) Suppose we want to predict whether or not a student will pass an exam, based on the time spent studying. Historical data shows that the average study time of the students who passed the exam was $\bar{X}_{\text{pass}} = 40$ (in some unspecified unit of time). For the students who failed, the average study time was $\bar{X}_{\text{fail}} = 25$. Furthermore, the variances within these two groups were $\hat{\sigma}^2_{\text{pass}} = 10^2$ and $\hat{\sigma}^2_{\text{fail}} = 7^2$, respectively. Finally, 60% of the students passed the exam. Construct a QDA classifier for predicting `pass` or `fail` based on the time spent studying X. Specifically, what is the decision boundary of the QDA classifier? What is the prediction (fail or pass) for a student who has studied 33 time units? (4p)

Note: This question can be answered independently of 3a and 3b