EXAM IN STATISTICAL MACHINE LEARNING
STATISTISK MASKININLÄRNING

DATE: January 9, 2024

RESPONSIBLE TEACHER: Dave Zachariah

NUMBER OF PROBLEMS: 5

AIDING MATERIAL: Calculator, mathematical handbook

GRADES: grade 3 21 points
grade 4 32 points
grade 5 42 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, kept in order).

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

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

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:

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

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

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

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.

   Hint: It is often better to only answer statements where you are confident. You do not need to classify all statements.

   i. Bagging techniques inject randomization to lower the bias.

   ii. Minimizing
   $$J(\boldsymbol{\theta}) = \sum_{i=1}^{n} L(\mathbf{x}_i, y_i; \boldsymbol{\theta})$$
   where $L(\mathbf{x}_i, y_i; \boldsymbol{\theta}) = \ln p(y_i|\mathbf{x}_i; \boldsymbol{\theta})$ corresponds to the maximum likelihood method.

   iii. Using a deep regression tree increases the risk of a high bias.

   iv. For image recognition problems, convolutional neural networks try to approximate rotational invariance in image.

   v. Consider a regression problem, where we use the loss function $L(\mathbf{x}, y) = |y - f(\mathbf{x})|^2$. The regression function that minimizes the new expected loss is $f_0(\mathbf{x}) = \mathbb{E}[y|\mathbf{x}]$.

   vi. Suppose we use a Gaussian model $p(y|\mathbf{x}; \boldsymbol{\theta})$ with mean $\mathbf{x}^T\boldsymbol{\theta}$ and variance σ^2. Using this model to approximate $f_0(\mathbf{x})$ leads to a linear regression model.

   vii. Lasso regression penalizes the learning of models with many inputs.

   viii. Decision trees provide interpretable models.

   ix. The cross-validation method can be used for learning or tuning hyperparameters.

   x. Learning a regression model that yields high variance in a problem indicates that the model family is not flexible enough to approximate the best regression model well.

   (10p)

2. Given an individual feature x, we want to predict the incubation time y of an infected person. We have collected dataset as shown in Table 1.

i	x	y
1	199.1	6.60
2	228.5	9.14
3	-23.8	0.77
4	210.9	10.99
5	171.6	5.57

Table 1: Incubation time data.

a) Use the two first data points ($n = 2$) to train a linear regression model

$$f(\mathbf{x}; \boldsymbol{\theta}) = \mathbf{x}^T \boldsymbol{\theta}$$

using the squared-error loss, where $\mathbf{x} = [1\ x]^T$ (5p)

b) Use the learned model $f(\mathbf{x}; \widehat{\boldsymbol{\theta}})$ to predict the incubation time of an infected person with feature $x = 50$. (1p)

b) Estimate the new expected error E_{new} of the learned model $f(\mathbf{x}; \widehat{\boldsymbol{\theta}})$. (4p)

3. We consider again the problem of predicting the incubation times of infected persons from Problem 2. However, this time we will incorporate some additional model structure. Specifically, we know that the incubation time is positive, $y > 0$ and that the log-normal distribution is a reasonable model for its conditional distribution:

$$p(y|\mathbf{x};\boldsymbol{\theta},v) = \frac{1}{y\sqrt{2\pi v}}\exp\left(-\frac{(\ln y - \mathbf{x}^T\boldsymbol{\theta})^2}{2v}\right),$$

which has mean

$$\exp(\mathbf{x}^T\boldsymbol{\theta} + v/2)$$

a) Consider approximating the best regression model $f_0(\mathbf{x}) = \mathbb{E}[y|\mathbf{x}]$ using the model above. That is,

$$f(\mathbf{x};\widehat{\boldsymbol{\theta}}) = \exp(\mathbf{x}^T\widehat{\boldsymbol{\theta}} + \widehat{v}/2)$$

where the parameters are learned by minimizing

$$J(\boldsymbol{\theta},v) = \sum_{i=1}^{n} -\ln p(y|\mathbf{x};\boldsymbol{\theta},v).$$

Show that the learned parameters are given by

$$\widehat{\boldsymbol{\theta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\widetilde{\mathbf{y}}, \quad \text{where } \widetilde{\mathbf{y}} = \begin{bmatrix}\ln y_1\\ \vdots \\ \ln y_n\end{bmatrix},$$

and

$$\widehat{v} = \frac{1}{n}\|\widetilde{\mathbf{y}} - \mathbf{X}\boldsymbol{\theta}\|^2$$

(6p)

b) Use the two first data points ($n = 2$) from Table 1 to learn the model $f(\mathbf{x};\widehat{\boldsymbol{\theta}},\widehat{v})$ and then predict the incubation time of an infected person with feature $x = 50$. (1p)

c) Use the remaining data to estimate to new expected error E_{new} of the model using the squared-error loss. (2p)

d) From this sample, can we conclude that the new regression model of incubation times has a better new expected error E_{new} than the learned linear regression model? Why/why not? (1p)

4. (a) In her master thesis, Taylor is exploring the use of machine learning for various football-related applications. As an illustrative example, she has constructed a logistic regression classifier that predicts whether a team A will win ($y = 1$) or loose ($y = -1$) an ongoing game against a team B based on the feature vector $\mathbf{x}^T = [1, x_1, x_2]^T$, where $x_1 = N_A - N_B$ is the difference between the number of successful passes within each team in the first half, and $x_2 = K_A - K_B$ is the difference between the number of free kicks awarded to each team in the fist half.
The classifier $f(\mathbf{x}; \boldsymbol{\theta})$ uses the model

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

with learned parameters $\hat{\boldsymbol{\theta}} = [\hat{\theta}_0, \hat{\theta}_1, \hat{\theta}_2]^T = [-2, 0.008, 0.9]^T$. She records the following predictions using the learned model:

i	x_1	x_2	f
1	55	5	1
2	90	1	-1
3	-20	0	-1
4	100	2	-1
5	-103	4	1

Suppose the classifier $f(\mathbf{x}; \boldsymbol{\theta})$ uses a threshold r on the conditional probability. Determine the largest possible interval $a < r < b$ for this threshold r that could have been used to obtain the predictions above. That is, the smallest a and largest b. (2p)

(b) Now consider the model $p(y|\mathbf{x}; \boldsymbol{\theta})$. What is the interpretation of the intercept term θ_0? Given this interpretation, is the learned model reasonable? (2p)
Hint: Consider the model output for a datapoint suggesting that the teams performed equally well in the first half.

(c) Now, assume that instead of modeling which team wins, we are interested in predicting how many goals team A scored in the game. This can be achieved by modeling the goals scored by a team as a Poisson regression problem where the probability that y goals are scored is modelled as

$$p(y|\mathbf{x}; \boldsymbol{\theta}) = \frac{(\lambda(\boldsymbol{\theta}))^y}{y!} e^{-\lambda(\boldsymbol{\theta})}$$

and the mean of the distribution is given by

$$\lambda(\boldsymbol{\theta}) = e^{\boldsymbol{\theta}^T \mathbf{x}}$$

Note that this is a discrete distribution over $\mathbb{N} = \{0, 1, 2, \dots\}$, since the goals scored by a team are a non-negative real-number.

Evaluate the model probability of scoring less than two goals, $y < 2$, when the input features $\mathbf{x}$ correspond to those in dtapoint $i = 1$ from the table in task (a). Use the model parameters $\boldsymbol{\theta} = [-17, 0.11, 2.35]^T$ (2p)

(d) In order to learn the parameters of the model using techniques from her machine learning course, Taylor chooses a maximum likelihood approach. Show that maximizing the likelihood of the iid sampled data $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$ is equivalent to minimizing

$$J(\boldsymbol{\theta}) = \sum_{i=1}^{n} -y_i \boldsymbol{\theta}^T \mathbf{x}_i + e^{\boldsymbol{\theta}^T \mathbf{x}_i}$$

(4p)

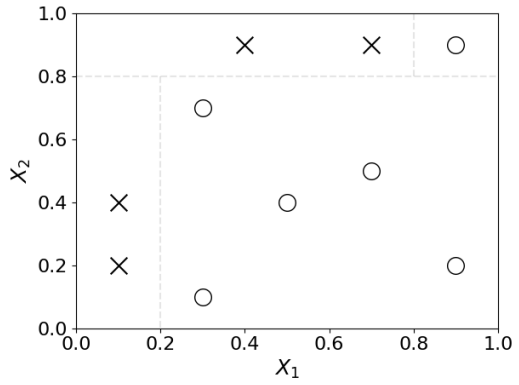

Figure 1: A dataset with two-dimensional feature space.

5. (a) Consider a classification problem with two input variables x_1 and x_2, and two class labels $\circ$ and $\times$ as shown in Figure 1. Based on the dataset in Figure 1, sketch a binary classification tree with *depth=2*.

(2p)

(b) How do bias and variance change as we increase the depth of the decision tree? Bagging and Boosting are two ensemble algorithms that can be applied on top of decision trees. Which one is more suitable for shallow trees? Explain why!

(2p)

(c) In Bagging we use averaging of B individual predictions $\{z_b\}_{b=1}^{B}$ which are identically distributed variables. Let us assume $\mathbb{E}[z_b] = \mu$ and $\mathrm{Var}[z_b] = \sigma^2$ with correlation between z_i and z_j of $\rho = \frac{1}{\sigma^2}\mathbb{E}[(z_i - \mu)(z_j - \mu)] > 0$ for $i \neq j$.

Compute the mean of the averaged prediction, i.e. $\mathbb{E}[\frac{1}{B}\sum_{b=1}^{B} z_b]$ and its variance $\mathrm{Var}[\frac{1}{B}\sum_{b=1}^{B} z_b]$.

Hint 1: Recall that the variance in terms of moments is given by $\mathbb{E}[Z^2] = \mathrm{Var}[Z] + \mathbb{E}[Z]^2$.
Hint 2: use the fact that $(\sum_{i=1}^{B} z_i)^2 = \sum_{i,j=1}^{B} z_i z_j$.
Hint 3: rewrite the correlation to obtain an expression for $\mathbb{E}[z_i z_j]$.

(5p)

(d) Show from the equation for the variance, that increasing the number of predictors B reduces the variance if $\rho < 1$.

(1p)