Exam in Statistical Machine Learning
Statistisk Maskininlärning (1RT700)

Date and time: March 16, 2021, 08.00–13.00 (plus 20 minutes for submitting)

Responsible teacher: David Sumpter

Number of problems: 5

Aiding material: Online and open book exam. You may not consult with anyone else during the exam.

Preliminary grades:

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

Some general instructions and information:

- Your solutions can be given in Swedish or in English.
- Typed or written answers are acceptable.
- Do not use a red pen.
- Submit the answers as 5 individual files or one combined file. **Do not submit more than one file per question.**
- For subproblems (a), (b), (c), . . . , it is usually possible to answer later subproblems independently of the earlier subproblems (for example, you can most often answer (b) without answering (a)).
- If you are enrolled at any other study program than a civilingenjörsprogram, you will *not* be allowed to take a later re-exam (to improve your grade) if you score grade 3 or higher on this exam. No exceptions will be made.

All your answers must be clearly motivated!
A correct answer without a proper motivation will score zero points!

Good luck!

Some relevant formulas

Pages 1–3 contain some expressions that may or may not be useful for solving the exam problems. *This is not a complete list of formulas used in the course*, but some of the problems may require knowledge about certain expressions not listed here. Furthermore, the formulas listed below *are not self-explanatory*, meaning that you need to be familiar with the expressions to be able to interpret them. They are possibly a support for solving the problems, but *not* a comprehensive summary of the course.

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

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

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

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

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

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

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

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

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. (a) Consider the following training data

i	$\mathbf{x}$		y
	x_1	x_2	
1	1	2	2
2	3	4	8
3	3	3	5
4	1	1	3

from which we want to learn a linear regression model

$$y = \theta_1 x_1 + \theta_2 x_2 + \varepsilon,$$

where we assume that ε has a Gaussian distribution. Use least squares to calculate $\widehat{\boldsymbol{\theta}} = [\widehat{\theta}_1 \ \widehat{\theta}_2]^T$. What value does your model predict for test data $x_1^* = 2$ and $x_2^* = 3$.

(4p)

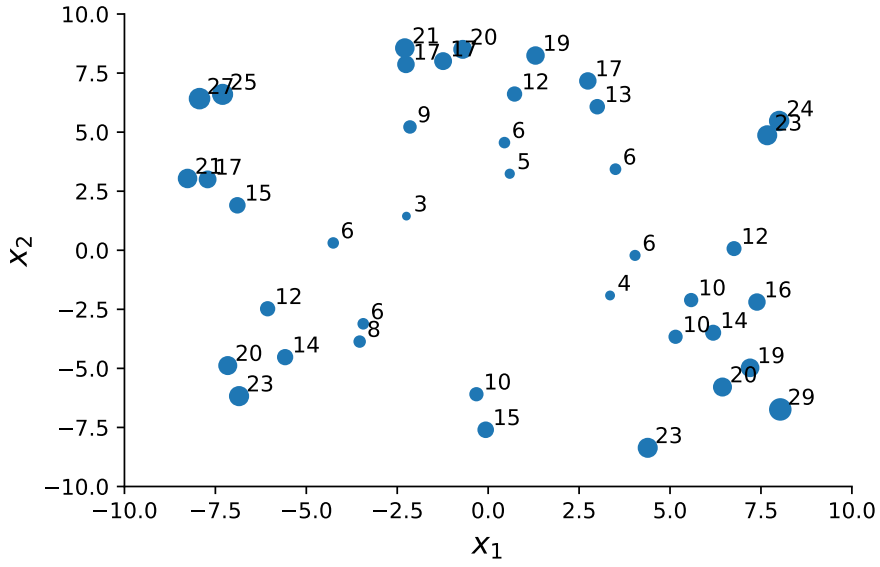
(b) If we used regularised linear regression (e.g. Ridge regression or LASSO) is the value of $\widehat{\boldsymbol{\theta}}$ obtained closer to the maximum likelihood estimate of $\boldsymbol{\theta}$ or further away? Explain in one sentence why.

(1p)

(c) Create a single split regression tree for the training data in (a) and use it to make a prediction for the test data $x_1^* = 2$ and $x_2^* = 3$.

(3p)

(d) Now consider the following training data x_1, x_2 and y.



The area of the circles is proportional to the value of y, i.e. larger values of y have larger circles. The numbers also show the value of y. By looking at the data, suggest input features for a linear regression model which you think might avoid overfitting the data. Justify your answer.

(2p)

2. A medical test for asthma is the peak flow test and is measured by a patient blowing in to a tube and seeing how many millimetres a plastic marker moves. You are asked to develop a model to help in the first stage of identifying children with asthma. Children without asthma have an average test score of 100mm and a standard deviation of 100mm. Those with asthma have an average score of 81mm and a standard deviation of 81mm.

   (a) You decide to use QDA to classify the children in terms of an output variable $y \in \{\text{Asthma}, \text{No Asthma}\}$. 10% of the general population have asthma. If a child with no previous symptoms gets a score of 300mm when she blows in the tube, what would you calculate as the probability that the child has asthma? i.e. calculate $p(y = \text{Asthma} \,|\, X = 300)$.

   (4p)

   (b) State one limitation of assuming scores are Normally distributed. (1p)

   (c) You plot the data from tube blowing tests and find that the probabilities are exponentially distributed, so that the probability that a child with asthma blows less than x mm is

   $$p(X < x) = 1 - \exp(-x/81)$$

   Similarly, the probability that a child without asthma blows less than x mm

   $$p(X < x) = 1 - \exp(-x/100)$$

   Show that the probability that a child with score x has asthma is given by

   $$p(y = \text{Asthma} \,|\, X = x) = \frac{1}{1 + \frac{9^3}{100} \exp\left(\frac{19x}{8100}\right)}$$

   (3p)

   (d) Use the equation directly above to calculate $p(y = \text{Asthma} \,|\, X = 50)$ and $p(y = \text{Asthma} \,|\, X = 300)$. Comment on the usefulness of the test in practice.

   (2p)

3. (a) Consider a binary classification problem with $Y \in \{0, 1\}$ where the training data consists of 600 data points of each class. Assume that a classification tree is to be fitted to this data. At the root node there are two possible splits:

   A: The **left branch** contains 500 data points of class $Y = 0$ and 100 data points of class $Y = 1$.

   B: The **left branch** contains 200 data points of class $Y = 0$ and 600 data points of class $Y = 1$.

   Compute the cost for the two splits, A and B, using both misclassification error and the Gini index impurity measures. For which, if any, of the splits are these measures minimised?

   (4p)

   (b) Which of the splits A or B will give the best conditions for further splitting? On this basis, which impurity measure is a better choice?

   (2p)

   (c) Explain briefly how the process of bagging can be applied to the data set above? What effect will it have on variance and bias?

   (2p)

   (d) Now consider how to treat split A above if it is the first iteration of a process of boosting. For which points should the weights increase in the second iteration?

   (2p)

4. Consider a dataset with 10'000 color images of pets, each consisting of 20×20 scalar pixels and three channels representing the colors red, green and blue. Each image is labeled with one of the four classes $\mathbf{y} \in \{\text{cat}, \text{dog}, \text{rabit}, \text{hamster}\}$. You put 90% of the images in the training data and 10% in the validation data.

   For this data you design a small convolutional neural network with one convolutional layer and one dense layer. The convolutional layer is parameterized with a weight tensor $\mathbf{W}^{(1)}$ and a offset vector $\mathbf{b}^{(1)}$ producing a hidden layer $\mathbf{q}$. The convolutional layer has the following design

Number of filters/output channels	24
Filter rows and columns	(5×5)
Stride	2

   The stride = 2 means that the filter is moving by two steps (both row- and column-wise) during the convolution using zero-padding such that the hidden layer $\mathbf{q}$ has half as many rows and columns as the input image.

   The dense layer is parameterized with the weight matrix $\mathbf{W}^{(2)}$ and offset vector $\mathbf{b}^{(2)}$ producing the logits $\mathbf{z}$. The logits $\mathbf{z}$ are pushed through a softmax function to produce the four class probabilities.

   (a) What are the sizes of the weight tensor $\mathbf{W}^{(1)}$, offset vector $\mathbf{b}^{(1)}$ and the hidden layer $\mathbf{q}$? (2p)

   (b) What are the sizes of the weight tensor $\mathbf{W}^{(2)}$, offset vector $\mathbf{b}^{(2)}$ and the logits $\mathbf{z}$? (2p)

   (c) What is the total number of parameters used to paramaterize this network? (1p)

   (d) In the context of neural networks, describe, using a few sentences, the difference between a dense layer and a convolutional layer. (2p)

   Note: This question can be answered independently of question 4a-4c

   (e) You train the model with stochastic gradient decent for 10 epochs. After you have trained the model you evaluate it on all data points in the training data as well as on all data points in the validation data. On the training data the model predicts 100 images incorrectly and also on the validation data the model predicts 100 images incorrectly. Suggest three adjustment in the model and/or the training that you think could decrease the misclassification rate on the validation data and motivate your suggestions. (3p)

   Note: This question can be answered independently of question 4a-4d

5. You are working for a company that works with building security and are developing a face recognition software.

   (a) Your colleagues have created a machine learning algorithm to operate an automatic door opener. It opens the door automatically when it recognises a face, otherwise the person has to use a card to access the building.

Males	Don't work in building	Work in building
Don't admit	800	10
Admit to building	200	90

Females	Don't work in building	Work in building
Don't admit	800	10
Admit to building	200	190

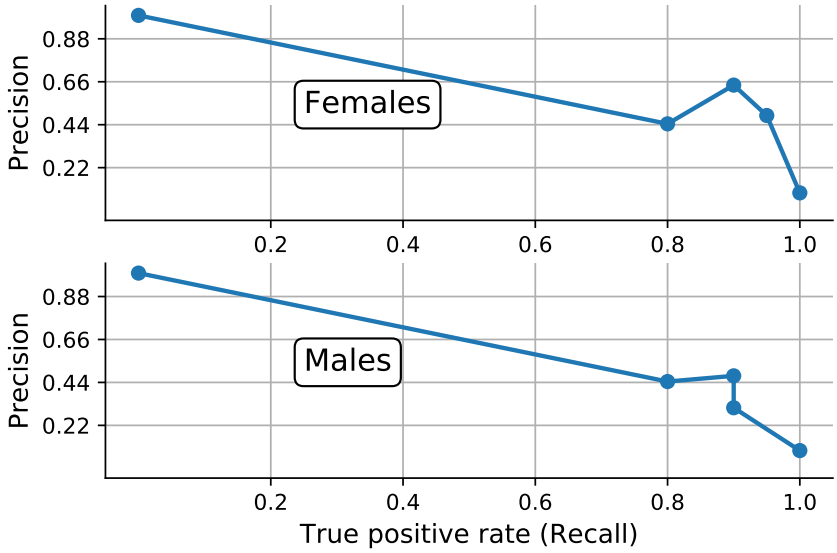
   We denote work in/admit to building as the positive class. Calculate the false positive rate, the true positive rate and the precision for the model for both groups.

   (3p)

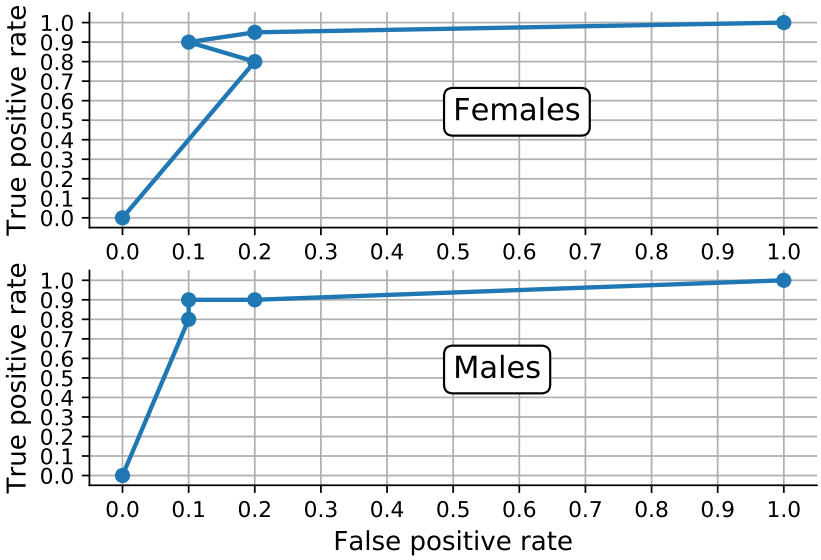
   (b) For each of the three performance measures say whether it is 'fair', 'unfair' or 'neither fair or unfair'. For those you consider unfair, say (in your view) who it is unfair against and write one or two sentences explaining why.

   (2p)

   (c) Your employer asks you to set equal levels of precision for men and women. The precision-recall curves for the method is shown below.



And the ROC-AUC curve for the same models is.



Each version of the model is represented by a dot in the Precision-Recall and ROC curves. What are the true positive and false positive rates for men and women when precision is equal for both groups and better than that achieved by guessing at random? Do you think your employers request for equal levels of precision is fair? Justify your answer.

(3p)

(d) Your employers decide to use the method from the table above, which (according to the ROC curve) has the highest true positive rate. They claim it is 95% accurate for women and 90% accurate for men. Write a three sentence message to them explaining any reservations you have about their claims and contrast it to the model (shown in the AUC curve) for which both true and false positive rates are 90%.

(2p)