Exam in Statistical Machine Learning
Statistisk Maskininlärning (1RT700)

Date and time: January 11, 2019, 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 $\boldsymbol{\beta}$ in the linear regression model

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

  is given by the solution $\widehat{\boldsymbol{\beta}}_{\text{LS}}$ to the normal equations $\mathbf{X}^{\mathsf{T}}\mathbf{X}\widehat{\boldsymbol{\beta}}_{\text{LS}} = \mathbf{X}^{\mathsf{T}}\mathbf{y}$, 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 $\lambda\|\boldsymbol{\beta}\|_2^2 = \lambda\sum_{j=0}^p \theta_j^2$.
  The ridge regression estimate is $\widehat{\boldsymbol{\beta}}_{\text{RR}} = (\mathbf{X}^{\mathsf{T}}\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^{\mathsf{T}}\mathbf{y}$.
- LASSO uses the regularization term $\lambda\|\boldsymbol{\beta}\|_1 = \lambda\sum_{j=0}^p |\theta_j|$.

Maximum likelihood: The maximum likelihood estimate is given by

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

where $\ln \ell(\boldsymbol{\beta}) = \sum_{i=1}^{n} \ln p(y_i \,|\, \mathbf{x}_i; \boldsymbol{\beta})$ 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

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

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

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

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. Consider the following $n = 5$ training data points for a binary classification problem ($y = \{-1, 1\}$ here) with one input variable:

x_i	6	10	8	11	7
y_i	1	-1	1	-1	1

(a) A logistic regression classifier is constructed using the function

$$f(x) = \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)}$$

Calculate the log-likelihood of the model, $P(\mathbf{y}|\mathbf{x}, \beta_0, \beta_1)$, given parameters $\beta_0 = 9$ and $\beta_1 = -1$.

(3p)

(b) You create a classification function

$$\hat{y}(x_\star) = \begin{cases} 1 & \text{if } f(x_\star) > r, \\ -1 & \text{otherwise,} \end{cases}$$

where $0 \leq r \leq 1$ is a threshold. Using the same parameter values ($\beta_0 = 9$ and $\beta_1 = -1$) as above, for what range of values of r do you get zero misclassification error in your training data? In other words, give the smallest and largest values of r for which your model is gives perfect predictions on the training data.

(2p)

(c) Now consider the following $n = 5$ test data points:

x_i	8	9	11	7	12
y_i	1	-1	1	1	-1

Calculate false positive rate, true positive rate and misclassification error for $r = 0.3$ and $r = 0.6$

(2p)

(d) Sketch an ROC curve using the values you calculated in (c) together with values corresponding to $r = 0.0$ and $r = 1.0$.

(2p)

(e) Explain, with reference to the example above, why logistic regression is a linear classifier?

(1p)

2. Consider the following training data

i	1	2	3	4	5	6	7	8
x_1	3.0	9.0	6.0	4.0	3.0	4.0	4.0	5.0
x_2	7.0	4.0	7.0	8.0	2.0	5.0	2.0	3.0
y	1	1	1	1	0	0	0	0

where x_1 and x_2 are the input variables, y is the output and i is the data point index.

(a) Illustrate the training data points in a graph with x_1 and x_2 on the two axes. Represent the points belonging to class $y = 0$ with a circle and those belonging to class $y = 1$ with a cross.

(2p)

(b) Based on the training data we want to construct bagging classification trees with three ensemble members. For this we draw three new datasets by bootstrapping the training data (sampling with replacement). The following data points' indices have been drawn for each of the three bootstrapped datasets

	Data point indices i							
Dataset 1	2	3	3	5	6	7	8	8
Dataset 2	1	3	4	5	6	6	7	8
Dataset 3	2	2	3	4	4	5	6	7

Construct three classification trees, one for each of the three bootstrapped datasets. Each tree shall consist of one single binary split that minimizes the entropy/deviance loss.

(5p)

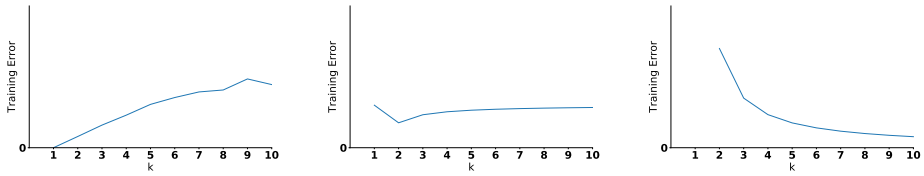
(c) The final classifier predicts according to the majority vote of the three classification trees. Sketch the decision boundary of the final classifier.

(2p)

(d) Is there a linear classifier which gives zero misclassification error for this dataset? Justify your answer.

(1p)

3. (a) Your lecturer sends you three graphs detailing the relationship between the training error of a model and a parameter k. They are shown below.

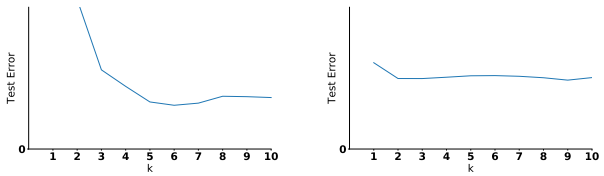


The problem is that the lecturer used k to denote three different types of parameters during the course! He used it for

- The number of variables, k, in the training data $(x_1, x_2...x_k)$ when explaining logistic regression.
- The number of k nearest neighbours in k-NN algorithm.
- The number of batches, k , in k-fold cross validation of QDA. For $k = 1$ all data is used for fitting the model and no cross-validation is done.

Explain which of the figures (left, middle, right) corresponds to which method. Justify your answer while referring to as many of the properties of the graphs as possible. (3p)

(b) Now your lecturer sends you two new graphs detailing the relationship between the test error of a model (applied to a new production data set, not used in training) and the parameter k. These are shown below.



Explain which possible uses of k could be consistent with the left and right figures. Note that for each figure it can be 0, 1, 2 or 3 different uses of k. (2p)

(c) Explain the main advantage of random forests over bagging in (at most) 3 sentences. (2p)

(d) Explain the key ideas of using boosting in less than 3 sentences, using the terms 'weak models', 'weights' and 'classifier' in your explanation. (3p)

4. Three of your colleagues (A, B and C) are working on a regression problem involving two input variables and an output variable. They use the same dataset but don't communicate with each other during their work, but do consult with you. Please answer the following questions in at most one or two sentences each.

   (a) Colleague A performs linear regression and obtains the following model:

   $$\text{(A1) } \widehat{y} = 6.8 + 0.03x_1 - 0.003x_1^2 + 0.18x_2 - 0.015x_2^2 - 0.004x_1x_2$$

   Explain in words how inputs x_1 and x_2 affect the output y (assume x_1 and x_2 take positive values only).

   (2p)

   (b) Colleague A then performs ridge regression and finds a new model such that

   $$\text{(A2) } \widehat{y} = 6.8 + 0.02x_1 - 0.0007x_1^2 + 0.18x_2 - 0.014x_2^2 - 0.0001x_1x_2$$

   She looks at the sum of squares error function of the new model given by,

   $$\|\mathbf{X}\boldsymbol{\beta} - \mathbf{y}\|_2^2$$

   Is this value bigger or smaller for this model than the model in question (a)? Explain in one sentence how you know the answer.

   (1p)

   (c) Colleague A now wants to find a good value of the hyperparameter λ for ridge regression. Explain to her, referring to the bias/variance trade-off, a method to find a good value for λ.

   (3p)

   (d) Colleague B separates the data in to two sets: training (90%) and test (10%) data. He decides (to keep things simple) to use at most three parameters to fit the models He trains three models on the training data and finds.

   (B1) $\widehat{y} = 6.8 + 0.1x_1 - 0.01x_1^2$
   (B2) $\widehat{y} = 6.4 + 0.2x_2 - 0.015x_2^2$
   (B3) $\widehat{y} = 6.5 - 0.05x_1 + 0.2x_2$

   Colleague B is puzzled and says: "When x_1 increases in (B1) it leads to an increase in y, but in (B3) an increase in x_1 leads to a decrease in y". Explain to him why this might have happened.

   (2p)

(e) Colleagues A and B now meet to compare their final models (A2 and B2). They both perform equally well (have very similar values to the sum of squares error function) on the test data which colleague B saved. They both claim that their model should be used. Which model do you think is most reliable and (in one or two sentences) why?

(1p)

(f) Colleague C decides to find out what the two variables represent. It turns out y is number of confirmed cases of a disease per day, x_1 is the number of journeys on public transport in the last week and x_2 is number of calls to a national health hotline. She decides to discard the x_2 data set, does a 90%/10% training/test split and presents model B1 as her final model. Colleague B disagrees, saying that he found that model B2 was a much better fit to data. Give one reason colleague C might have made the correct choice?

(2p)

5. You are working for a company that helps people get started in machine learning. You are asked if you can develop a machine learning algorithm to help them find potential students using their LinkedIn profiles.

   (a) List the type of input features and output data which you might use for this task. (You are recommended to search online in answering this question).

   (2p)

   (b) You are asked to lead a group of 4 data scientists, all of who have read the Uppsala course in statistical machine learning. Describe the methods you would use, why you would use them and how you would organise their work in order to maximally utilise their skills. (Write at most 250 words)

   (4p)

   (c) Your colleagues have created a machine learning algorithm to find people who might be interested in studying in Uppsala. Their algorithm either recommends or doesn't recommend Uppsala course to users. They have tested it on two different groups of people (600 non-Swedes and 1200 Swedes) all of whom would be eligible for the course and have given permission for their data to be used. They first applied the algorithm then asked the potential students whether or not they would be interested. The confusion matrix for the respective groups are:

Non-Swedes	Recommended course	Not recommended
Interested	100	100
Not interested	100	300

Swedes	Recommended course	Not recommended
Interested	400	50
Not interested	350	400

   Your colleagues tell you that their algorithm has the same misclassification error for the two groups and want to go ahead and put it in to production. First check whether they are correct about the misclassification error. Then discuss, with reference to the confusion matrix, whether you agree that the method can be put in to production.

   (4p)