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!

1. i. True

   ii. False, a classifier is said to be linear if its decision boundary is linear. $\hat{G}$ takes values in a discrete set and can not be a linear function!

   iii. True

   iv. False, there is typically an irreducible error.

   v. False, the ensemble members are conditionally independent (given the training data set) and the correlation between any pair of ensemble members is independent of B.

   vi. False, misclassification loss yields a loss of 1 for any misclassified point, regardless of how far from the decision boundary it is.

   vii. True, the model becomes less flexible (= larger bias) as k increases. For large enough k the model will always predict according to the dominating class.

   viii. True

   ix. False, any mismatch between the postulated model and the true input-output relationship will result in a model bias which does not vanish as the number of data points becomes large.

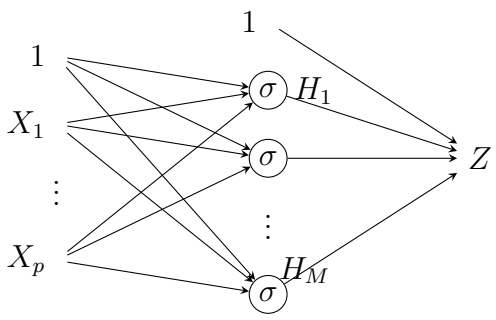
   x. True, in a probabilistic model the belief about unknown model parameters is represented using probability distributions.

2. (a) Classification and regression are both supervised learning problems, i.e. the task is to model the relationship between some input variables X and some output variable Y. The difference is that for classification problems Y is qualitative, whereas for regression problems Y is quantitative.

   (b) Let $\hat{f}$ denote the model of some true input-output relationship f, estimated from some training data set $\mathcal{T}$. The *model variance* tells us how much $\hat{f}$ would change if we were to estimate it using a different training data set $\mathcal{T}^\star$ independent of $\mathcal{T}$. Thus, if the model has high variance, then a small change in the training data can result in large changes in $\hat{f}$. The *model bias*, on the other hand, is the systematic error made by approximating the true function f (which is typically very complex) by some simpler class of models (e.g. some parametric model family). Specifically, the model bias is the expected error in $\hat{f}$ w.r.t. f where the expectation is taken over all possible training data sets. The trade-off between bias and variance comes from the fact that in order to attain low bias we need to use a flexible model, capable of capturing complex input-output relationships. However, a very flexible model will also be prone to overfitting to a given training data set $\mathcal{T}$, thus leading to high variance.

   (c) The inputs X_1 and X_2 could be correlated. In the extreme case, if X_2 is deterministically given by X_1 via the relationship $X_2 = (1.3 + 4.6X_1)/1.7$, then the models (P1) and (P2) would be equivalent. Even if the dependence between X_1 and X_2 is not this extreme, we can still obtain a similar effect on the regression model.

3. (a) A dense neural network for regression with one hidden layer can be illustrated as



Each link represents a multiplication of its incomming unit with a parameter. The parameters are different for each link. The number of links in the graph (=number of parameters in the model) is consequently $(p+1)\cdot M + 1 + M$.

(b) The mathematical model corresponding to the neural network above is

$$H_m = \sigma\left(\beta_{0m}^{(1)} + \sum_{j=1}^{p} \beta_{jm}^{(1)} X_j\right), \qquad m = 1, \ldots, M$$
$$Z = \beta_0^{(2)} + \sum_{m=1}^{M} \beta_m^{(2)} H_m$$

If we consider a linear activation function $\sigma(x) = x$ we get

$$\begin{aligned} Z &= \beta_0^{(2)} + \sum_{m=1}^{M} \beta_m^{(2)} \left(\beta_{0m}^{(1)} + \sum_{j=1}^{p} \beta_{jm}^{(1)} X_j\right) \\ &= \beta_0^{(2)} + \sum_{m=1}^{M} \beta_m^{(2)} \beta_{0m}^{(1)} + \sum_{j=1}^{p} \sum_{m=1}^{M} \beta_m^{(2)} \beta_{jm}^{(1)} X_j, \end{aligned}$$

which is a linear regression model

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

where $\beta_0 = \beta_0^{(2)} + \sum_{m=1}^{M} \beta_m^{(2)} \beta_{0m}^{(1)}$ and $\beta_j = \sum_{m=1}^{M} \beta_m^{(2)} \beta_{jm}^{(1)}$.

(c) Since each ensemble member $\widehat{f}^b(X)$ is a stump, it only depends on one input variable

$$\widehat{f}^b(X) = \widehat{f}^b(X_{j_b}),$$

where j_b is the index of the input variable that is splitted in the bth ensemble member. We then have

$$\begin{aligned}
\widehat{f}^B_{\text{boost}}(X) &= \sum_{b=1}^{B} \widehat{f}^b(X) \\
&= \sum_{b=1}^{B} \widehat{f}^b(X_{j_b}) \\
&= \sum_{b=1}^{B} \sum_{j=1}^{p} \widehat{f}^b(X_j) I(j = j_b) \\
&= \sum_{j=1}^{p} \sum_{b=1}^{B} \widehat{f}^b(X_j) I(j = j_b) \\
&= \sum_{j=1}^{p} \widehat{f}_j(X_j),
\end{aligned}$$

where we define

$$\widehat{f}_j(X_j) \stackrel{\text{def}}{=} \sum_{b=1}^{B} \widehat{f}^b(X_j) I(j = j_b).$$

4. (a) The first split divides X_1 two half-spaces, the region $X_1 > 1.7$ correspond to leaf node R_1. The second split divides the region $X_1 \leq 1.7$ at $X_2 = 0.5$ where the region $X_2 \leq 0.5$ corresponds to node R_2. Finally, the third split divides the region $X_2 > 0.5$ and $X_1 \leq 1.7$ at $X_1 = 1.0$ resulting in two regions where R_3 corresponds to $X_1 \leq 1.0$ and R_4 corresponds to $X_1 > 1.0$. The partitioning of the input space is thus as follows

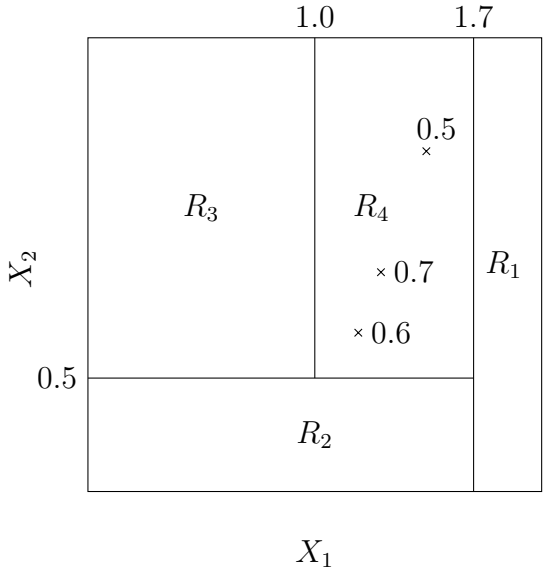


Figure 1: Partitioning of the input space for Problem 4a.

(b) Since $X_1^* = 1.5 < 1.7$, $X_2^* = 1.8 > 0.5$ and $X_1^* = 1.5 > 1.0$, the test point belongs to region R_4. To compute the predicted output we also need to know which regions the training data points fall into. We do this for all eight data points and get

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
Region	R_4	R_3	R_1	R_4	R_2	R_1	R_4	R_3

Thus, we have three data points in region R_4 and we take the mean of these to compute the predicted output, which is (0.5+0.7+0.6)/3 = 0.6.

(c) All regions already have two or less data points, except for region R_4 which has three. Therefore, we need to make one additional

split in that region, where one of the resulting regions will have two data points and the other region one data point.

In Figure 1 the three data points are depicted. The two possible splits are (i) to put 0.5 in one region and 0.7 and 0.6 in the other region, or (ii) to put 0.6 in one region and 0.5 and 0.7 in the other region. The MSE for these two options will be

(i) $\text{MSE} = (0.5 - 0.5)^2 + (0.7 - 0.65)^2 + (0.6 - 0.65)^2 = 2 * 0.05^2$

(ii) $\text{MSE} = (0.6 - 0.6)^2 + (0.7 - 0.6)^2 + (0.5 - 0.6)^2 = 2 * 0.1^2$

Clearly, option (i) will give a smaller MSE. This split could be realized, for example with the split $X_2 \leq 1.2$.

(d) The disadvantage with growing a decision tree too deep is overfitting. The performance on an unseen test data set and no generalization is achieved if there are too few data points in each node.

5. (a) The LDA classifier is linear and will therefore separate the two classes with a straight line in 2D. Looking at the scatter plots in Figure 2 it is clear that dataset i and iii can be separated using a straight line, but dataset ii can not. The QDA classifier is non-linear and will hence result in a non-linear (quadratic) decision boundary. Looking at the scatter plots in Figure 2 it is clear that in all three plots the two classes can be separated by a quadratic curve.

   (b) Both LDA and QDA assume the inputs corresponding to each class have a Gaussian distribution with some mean value μ_k for each class k. The difference between LDA and QDA is that LDA assumes that the covariance matrix Σ is the same for all classes k whereas QDA assumes that each class k has a unique covariance matrix Σ_k.

   Looking at dataset i it seems as if both classes have a Gaussian distribution and there is no clear correlation (pos or neg) in any of the classes, and their spread seems to be the same. Hence the assumptions of LDA (same covariance matrix for both classes) are fulfilled by dataset i. (Since QDA is a generalization of LDA, we could also argue that dataset i satisfies the assumptions of QDA as well.)

   Looking at dataset iii it seems as if both classes have a Gaussian distribution, but there is also some correlation in both classes. Since the correlation is positive for $Y = 0$ and negative for $Y = 1$ the two classes have different covariance matrices. Hence the assumptions of QDA (but not LDA) are fulfilled by dataset iii.

   Looking at dataset ii, class $Y = 0$ seems to have a Gaussian distribution whereas the class $Y = 1$ does not have a Gaussian distribution. Hence dataset ii does not fulfill the assumptions of either LDA or QDA.

   (Note that even though dataset ii does not fulfill the assumptions of QDA, a QDA decision boundary can still separate the classes in this case! The same holds for LDA and dataset iii. This shows that the LDA/QDA classifier *can* perform well in practice, even if the Gaussian assumptions are not satisfied.)

(c) We have the following parameters: $\hat{\mu}_F = 25$, $\hat{\mu}_P = 40$, $\hat{\Sigma}_F = \hat{\sigma}_F^2 = 7^2$, $\hat{\Sigma}_P = \hat{\sigma}_P^2 = 10^2$. Since 60 % passed the exam $\hat{\pi}_F = 0.4$ and $\hat{\pi}_P = 0.6$. The discriminant function for QDA is:

$$\hat{\delta}_k(x) = -\frac{1}{2}\log|\hat{\Sigma}_k| - \frac{1}{2}(x-\hat{\mu}_k)^T\hat{\Sigma}_k^{-1}(x-\hat{\mu}_k) + \log\hat{\pi}_k$$

Hence, after inserting parameter values we get:

$$\hat{\delta}_F(x) = -\frac{1}{98}x^2 + \frac{25}{49}x - 9.240$$

$$\hat{\delta}_P(x) = -\frac{1}{200}x^2 + \frac{40}{100}x - 10.814$$

The decision boundary is given by $\hat{\delta}_F = \hat{\delta}_P$ which gives

$$-\frac{1}{98}x^2 + \frac{50}{98}x - 9.240 = -\frac{1}{200}x^2 + \frac{2}{5}x - 10.814 \quad \Leftrightarrow \quad x \approx 31$$

Note that there are two solutions but one is negative and thus disregarded (you can't study for a negative number of time units). The prediction for a student who has studied $X = 33$ time units hence becomes pass, since it is on the 'pass side' of the decision boundary (> 31).