## Instructions

There are 8 questions worth a total of 100 marks. You should attempt them all.

### QUESTION 1

(a) True or false. The training error of the K Nearest neighbours (KNN) classifier with $K = 1$ is zero. Explain your answer.

[2 marks]

(b) True or false. The depth of an estimated decision tree can be larger than the number of training examples used to create the tree. Explain your answer.

[2 marks]

(c) Give one similarity and one difference between feature selection and principal component analysis (PCA).

[2 marks]

(d) Briefly explain why minimizing the training mean squared error (MSE) can lead to overfitting. Explain why adding a penalty term allows us to control the bias and variance tradeoff of our error estimation.

[2 marks]

(e) Suppose you apply clustering on a data set with 5 observations, explain if K-means or hierarchical clustering can produce the following three clusters: $C_1 = \{1, 3, 4\}$, $C_2 = \{2, 4\}$ and $C_3 = \{5\}$.

[2 marks]

(f) True or false. It is possible to produce a nonlinear regression fit using linear regression. Explain your answer.

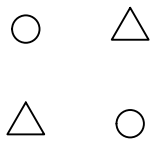
[2 marks]

**[Total: 12 marks]**

**— END OF QUESTION 1 —**

**QUESTION 2**

(a) Consider the following dataset:



Which classifiers will achieve zero training error on this data set? (i) logistic regression, (ii) SVM (with polynomial kernel of degree 2), (iii) 3-NN classifier.

[2 marks]

(b) Consider the following classification problem. We first choose the class $Y \sim \text{Bernouilli}(\frac{1}{2})$, which is class 1 with probability $\frac{1}{2}$. If $Y = 1$, then $X \sim \text{Bernouilli}(p)$; otherwise $X \sim \text{Bernouilli}(q)$. Assume that $p > q$.

(i) What is the Bayes optimal classifier $C^*(X)$?

[3 marks]

(ii) What is the Bayes error rate, i.e. $P(Y \neq C^*(X))$?

[2 marks]

(c) True or false. The Bayes error rate is the lowest possible error rate, and can never be zero. Explain your answer.

[2 marks]

(d) True or false. Let $Y$ be a binary response and $P(X) = \text{Probability}(Y = 1|X)$. If $P(X) = \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}}$, the log-odds are nonlinear in $X$. Prove your answer.

[2 marks]

(e) In the quadratic discriminant analysis (QDA) model, the observations within each class are drawn from a normal distribution with a class-specific mean vector and a class specific covariance matrix. We consider the case where there is only one feature, and $K$ classes. Prove that in this case, the Bayes' classifier is not linear. Argue that it is in fact quadratic.

[3 marks]

**[Total: 14 marks]**

**— END OF QUESTION 2 —**

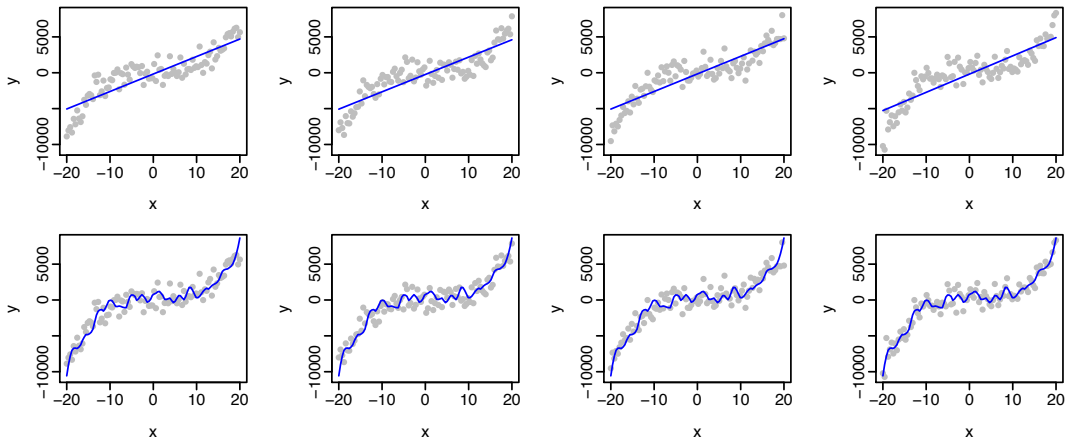
**QUESTION 3**

Consider the following data generating process:

$$y = x^3 - 2x^2 + 1.5x + \varepsilon, \tag{1}$$

where $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ and $\sigma = 1000$.

(a) The following Figure shows four samples with 100 observations drawn i.i.d. from (**??**), together with mean regression fitted values for two different models (in each row).




Describe the difference between the two models in terms of model complexity.

[2 marks]

(b) If we want to minimize expected out-of-sample squared errors, what is the optimal prediction $\hat{y}_*$ for $x_* = 10$? Explain your answer.

[2 marks]

(c) What is the expected out-of-sample squared errors of the optimal prediction at $x_* = 10$?

[2 marks]

(d) True or false. $\hat{f}(x) = x^3$ will have minimum expected out-of-sample squared errors for all $x$. Explain your answer.

[2 marks]

(e) Suppose we have generated a training sample of $n$ pairs $\boldsymbol{x}_i$, $y_i$ drawn i.i.d. from (**??**), where $\boldsymbol{x}_i = (x_i, x_i^2, x_i^3)$. We also generate a test sample of $n$ pairs $\boldsymbol{x}_i$, $y_i^{'}$ drawn i.i.d. from $y_i^{'} = x_i^3 - 2x_i^2 + 1.5x_i + \varepsilon_i^{'}$ where $\varepsilon_i^{'}$ and $\varepsilon_i$ are independent but identically distributed. Note that the training and test sample have the same predictors but different responses. Let $\hat{y}_i$ be the fitted values from linear regression (with intercept) applied on the training sample $\{(y_i, \boldsymbol{x}_i)\}_{i=1}^n$. We know that

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}(y_i^{'} - \hat{y}_i)^2\right] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2\right] + \frac{2}{n}\sum_{i=1}^{n}\mathrm{Cov}(\hat{y}_i, y_i).$$

In our scenario, if $n = 1000$, prove step by step that

$$\frac{2}{n}\sum_{i=1}^{n}\mathrm{Cov}(\hat{y}_i, y_i) = 8000.$$

[3 marks]

**[Total: 11 marks]**

**— END OF QUESTION 3 —**

**QUESTION 4**

Let $\boldsymbol{y} = (y_1, \ldots, y_n)^{'} \in \mathbb{R}^n$, $\boldsymbol{X}^{'} = [\boldsymbol{x}_1 \ \cdots \ \boldsymbol{x}_n] \in \mathbb{R}^{p \times n}$ with $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})' \in \mathbb{R}^p$, and consider the following model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$ and $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^{'} \sim \mathcal{N}_n(\boldsymbol{0}, \sigma^2 \boldsymbol{I})$. Given a realization of $\boldsymbol{y}$ and $\boldsymbol{X}$, we would like to estimate the coefficients $\boldsymbol{\beta}$.

(a) When the number of predictors $p$ is large, briefly explain why best subset selection is not computationally feasible.

[1 marks]

Suppose we estimate the coefficients $\boldsymbol{\beta}$ by solving the following optimization problem:

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta} \in \mathbb{R}^p}{\operatorname{argmin}} \left\{ \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda \sum_{j=1}^{p} \left[ (1-\alpha)\beta_j^2 + \alpha|\beta_j| \right] \right\} \tag{2}$$

where $\lambda \geq 0$, $\alpha \in \{0, 1\}$ and $\|\cdot\|_2$ is the $L_2$ norm.

(b) Assume $\boldsymbol{X}$ is an $n \times n$ identity matrix, and $p = n$.

(i) Write the solution $\hat{\boldsymbol{\beta}}$ when $\lambda = 0$.

[1 marks]

(ii) Calculate the bias and estimation variance of the estimator from (i).

[1 marks]

(iii) Write the solution $\hat{\boldsymbol{\beta}}$ when $\alpha = 0$.

[1 marks]

(iv) Calculate the bias and variance of the estimator from (iii).

[1 marks]

(v) Compare the results of the bias and variance that you obtained in (ii) and (iv) and provide an explanation.

[1 marks]

(c) Briefly explain the difference between ridge regression and principal component regression (PCR).

[2 marks]

(d) What does "sparse coefficients" mean? Which values of $\alpha$ and $\lambda$ provide "sparse coefficients"?

[1 marks]

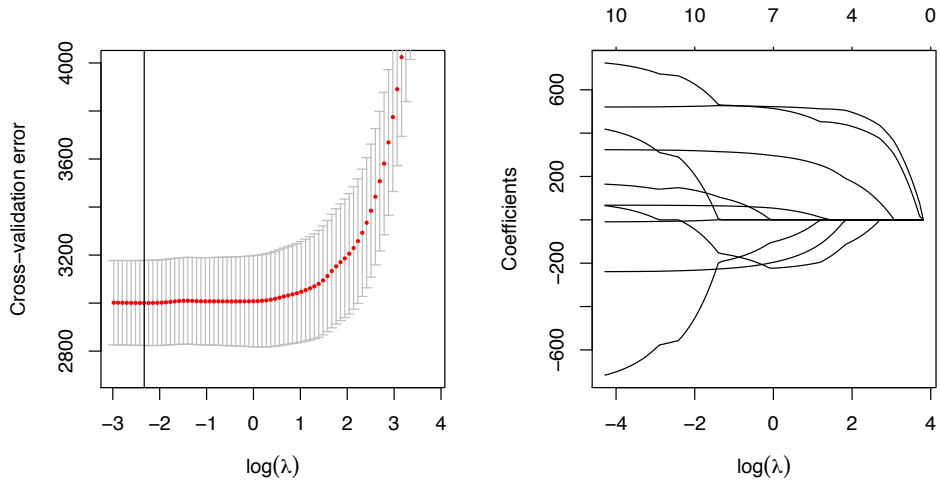
(e) In the Figure below, the left and right panels shows the cross-validation error and the regularization path as a function of the regularisation parameter $\lambda$ (in log scale), respectively. The vertical line in the left panel shows the minimum value of the cross-validation error.

(i) Which value of the parameter $\lambda$ will be selected if you apply the one-standard error rule?

[2 marks]

(ii) Which value of $\alpha \in \{0, 1\}$ has been used to produce the plot in the right panel? Explain your answer.

[2 marks]




**[Total: 13 marks]**

**— END OF QUESTION 4 —**

**QUESTION 5**

(a) For the following dataset, which classifier has larger Leave-One-Out Cross validation error: (a) 1-NN classifier, (b) 3-NN classifier?

[2 marks]

(b) True or false. 10-fold cross-validation is always better than using a validation set. Explain your answer.

[3 marks]

(c) If we have $n$ data points, derive the probability that the $j$th data point does not appear in a bootstrap sample?

[3 marks]

(d) Consider a simple classification procedure applied to a two-class dataset with 2000 predictors and 100 samples:

- Step 1. Find the 2 predictors having the largest correlation with the response
- Step 2. Fit a logistic regression model using only these 2 predictors

We estimate the performance of this classification procedure on new samples using the following algorithm: (1) apply the classification procedure to multiple bootstrap samples, (2) for each classifier estimated using a bootstrap sample, predict the original dataset, and (3) average the classification error for all the bootstrap samples.

(i) Using your answer in (c), explain why this algorithm will provide a bad estimate of the classification performance on new samples.

[2 marks]

(ii) Modify this algorithm and show that it will provide a better estimate of the classification performance on new samples.

[2 marks]

**[Total: 12 marks]**

**— END OF QUESTION 5 —**

## QUESTION 6

(a) Consider 3 observations $x_1$, $x_2$ and $x_3$ where $x_i \in \mathbb{R}^p$, $i = 1, 2, 3$. The pairwise Euclidean distance matrix is given below.

| | $x_1$ | $x_2$ | $x_3$ |
|---|---|---|---|
| $x_1$ | 0 | 10.296 | 30.265 |
| $x_2$ | 10.296 | 0 | 22.847 |
| $x_3$ | 30.265 | 22.847 | 0 |

(i) Compute the $K = 2$ clusters $C_1$ and $C_2$ that are solutions to the following optimization problem:

$$\underset{C_1,\ldots,C_K}{\text{minimize}} \left\{ \sum_{k=1}^{K} \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2 \right\}$$

Explain how you computed it.

[2 marks]

(ii) What is the value of the objective function for the clusters you computed in (i)?

[2 marks]

(b) Let $\{x_1, \ldots, x_n\}$ be a set of points where $x_i \in \mathbb{R}^p$, and $C_1, C_2, \ldots, C_K$, a set of $K$ clusters. Prove that the total variation in the data set $T$ is equal to the sum of the within-cluster variation $W_K$ and the between-cluster variation $B_K$. In other words, prove that

$$T = W_K + C_K,$$

with

$$T = \sum_{i=1}^{n} \|x_i - \bar{x}\|_2^2,$$

$$W_K = \sum_{k=1}^{K} \sum_{i \in C_k} \|x_i - \bar{x}_k\|_2^2,$$

$$C_K = \sum_{k=1}^{K} n_k \|\bar{x}_k - \bar{x}\|_2^2,$$

where $\bar{x}$ is the overall average, i.e. $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, and $\bar{x}_k$ is the average of points in cluster $k$, i.e. $\bar{x}_k = \frac{1}{|C_k|}\sum_{i \in C_k} x_i$. (Hint 1: you need to start from the total variation. Hint 2: you need to add and subtract the cluster averages $\bar{x}_k$.)

[4 marks]

(c) Suppose that we have five observations, for which we compute a dissimilarity matrix, given by

| | a | b | c | d | e |
|---|---|---|---|---|---|
| a | 0 | 17.804 | 20.616 | 44.045 | 33.121 |
| b | 17.804 | 0 | 36.359 | 61.555 | 50 |
| c | 20.616 | 36.359 | 0 | 26.926 | 14.213 |
| d | 44.045 | 61.555 | 26.926 | 0 | 13 |
| e | 33.121 | 50 | 14.213 | 13 | 0 |

On the basis of this dissimilarity matrix, sketch the dendrogram that results from hierarchically clustering these five observations using *complete linkage*. Be sure to indicate on the plot the height at which each fusion occurs, as well as the observations corresponding to each leaf in the dendrogram

[3 marks]

(d) Repeat (c), this time using *single linkage* clustering.

[2 marks]

**[Total: 13 marks]**

**— END OF QUESTION 6 —**

**QUESTION 7**

We consider a data set which contains statistics, in arrests per 100,000 residents for assault, murder, and rape in each of the 50 US states in 1973. Also given is the percent of the population living in urban areas. Below is a principal component analysis (PCA) of this data set after centering and scaling each column.

| | PC1 | PC2 | PC3 | PC4 |
|---|---|---|---|---|
| Murder | -0.536 | 0.418 | -0.341 | 0.649 |
| Assault | -0.583 | 0.188 | -0.268 | -0.743 |
| UrbanPop | -0.278 | -0.873 | -0.378 | 0.134 |
| Rape | -0.543 | -0.167 | 0.818 | 0.089 |
| Variance | 2.480 | 0.990 | ? | 0.173 |
| Cum. Prop. | 0.620 | ? | ? | ? |

(a) Complete the missing entry in the row "Variance" for PC3. Explain your answer.

[2 marks]

(b) What proportion of the total variance does the fourth principal component explain?

[2 marks]

(c) Complete the bottom line of the previous Table which gives the cumulative proportions of total variance explained.

[2 marks]

(d) How many principal component directions would we need to explain at least 60% of the variance?

[2 marks]

(e) Interpret the first principal component.

[2 marks]

(f) Let $\boldsymbol{\phi}_1 \in \mathbb{R}^4$ and $\boldsymbol{\phi}_2 \in \mathbb{R}^4$ be the first two loading vectors. What is the value of $\boldsymbol{\phi}_1'\boldsymbol{\phi}_2$? Explain your answer.

[2 marks]

**[Total: 12 marks]**

**— END OF QUESTION 7 —**

## QUESTION 8

(a) True or false. Regression trees can only model constant functions. Explain your answer.

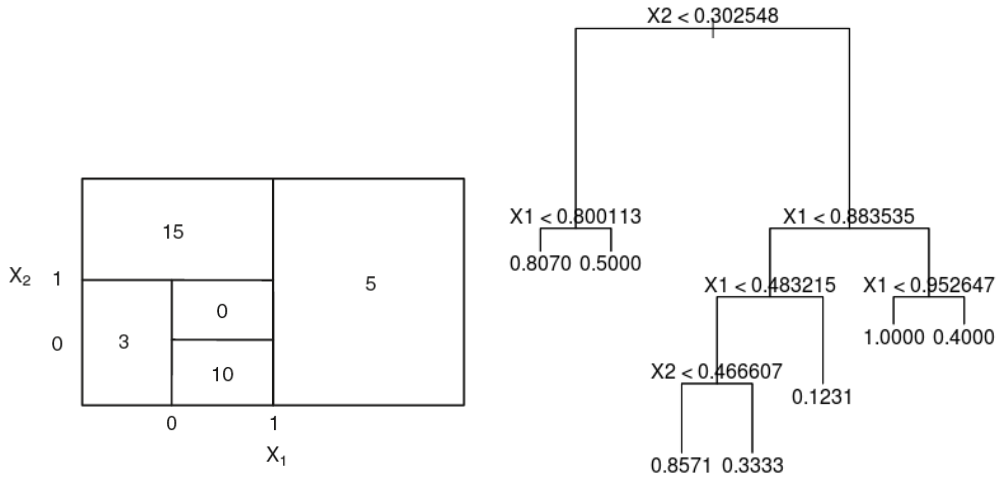
[2 marks]

(b) Suppose we have a sample of $n$ pairs $x_i$, $y_i$ drawn i.i.d. from $y_i = f(x_i) + \varepsilon_i$, where $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$, $f(x) = \sum_{j=1}^{J} c_j \, I(x \in R_j)$ with $R_1; R_2; \ldots; R_J$ being $J$ partitions of the input space. What are the optimal values of $c_1, \ldots, c_J$ that minimize $\sum_{i=1}^{n} (y_i - f(x_i))^2$? Give a proof of your answer.

[3 marks]

(c) Create a diagram similar to the left-hand panel of the following Figure, using the tree illustrated in the right-hand panel of the same figure. You should divide up the predictor space into the correct regions, and indicate the mean for each region.

[3 marks]




(d) Let $Z_1, \ldots, Z_B$ be a set of $B$ i.d. (identically distributed, but not necessarily independent) random variables, each with variance $\sigma^2$, with positive pairwise correlation $\rho$ ($\rho > 0$), and $\bar{Z} = \frac{1}{B} \sum_{b=1}^{B} Z_b$.

(i) Prove that

$$\mathrm{Var}(\bar{Z}) = \rho\sigma^2 + \sigma^2 \frac{1-\rho}{B}. \tag{3}$$

[3 marks]

(ii) Which algorithm discussed in the class exploit expression (**??**)? Briefly explain how the algorithm exploits it.

[2 marks]

**[Total: 13 marks]**

**— END OF QUESTION 8 —**