UNIVERSITY OF EDINBURGH

COLLEGE OF SCIENCE AND ENGINEERING

SCHOOL OF INFORMATICS

INFR11007 DATA MINING AND EXPLORATION

May 12, 2017

14:30 to 16:30

INSTRUCTIONS TO CANDIDATES

Answer QUESTION 1 and ONE other question.

Question 1 is COMPULSORY. If both QUESTION 2 and QUESTION 3 are answered, only QUESTION 2 will be marked.

All questions carry equal weight.

CALCULATORS MAY BE USED IN THIS EXAMINATION

MSc Courses

Convener: F. Keller
External Examiners: A. Burns, P. Healey, M. Niranjan

THIS EXAMINATION WILL BE MARKED ANONYMOUSLY

1. THIS QUESTION IS COMPULSORY

(a) Let the data matrix $\boldsymbol{X}$ with $n = 9$ observations be

$$\boldsymbol{X} = \begin{pmatrix} 6 & 3 & 9 & 7 & 8 & 5 & 1 & 2 & 4. \end{pmatrix}$$

What is the median of the data? [*1 mark*]

(b) Assume that the first data point ("6") is wrongly measured and thus may be arbitrarily large.

i. How does this measurement error affect the value of the median? [*1 mark*]

ii. How does it affect the sample mean? [*1 mark*]

(c) You are using your friend's algorithm to compute the singular value decomposition of a centred data matrix $\boldsymbol{X}$ with $n = 4$ data points $\boldsymbol{x}_i \in \mathbb{R}^2$. The algorithm returns the following decomposition:

$$\boldsymbol{X} = 1.5577 \begin{pmatrix} -0.5208 \\ 0.8537 \end{pmatrix} \begin{pmatrix} 0.1442 \\ -0.5417 \\ 0.7495 \\ -0.3520 \end{pmatrix}^\top + 2.1899 \begin{pmatrix} -0.8537 \\ -0.5208 \end{pmatrix} \begin{pmatrix} -0.8456 \\ 0.1096 \\ 0.3999 \\ 0.3361 \end{pmatrix}^\top$$

where all vectors have unit norm.

i. What is the first principal component direction? [*2 marks*]

ii. What is the fraction of variance explained by the first principal component? [*3 marks*]

(d) Figure 1 shows a training loss in polynomial regression as a function of the degree of the polynomial model (solid line with circles). It further shows the prediction loss of the corresponding prediction function (dashed line with squares), and the average (expected) value of such prediction losses, when averaged over several learned prediction functions (solid line with diamonds).

i. Use Figure 1 to explain the concepts of under-fitting and over-fitting, and how under/over-fitting affects training performance and prediction performance. [*6 marks*]

ii. Both prediction loss and expected prediction loss measure generalisation performance. But they attain their minimal value for different degrees of the polynomial regression model (degree four versus degree three). Please explain how this apparent contradiction can be resolved. [*3 marks*]

QUESTION CONTINUES ON NEXT PAGE

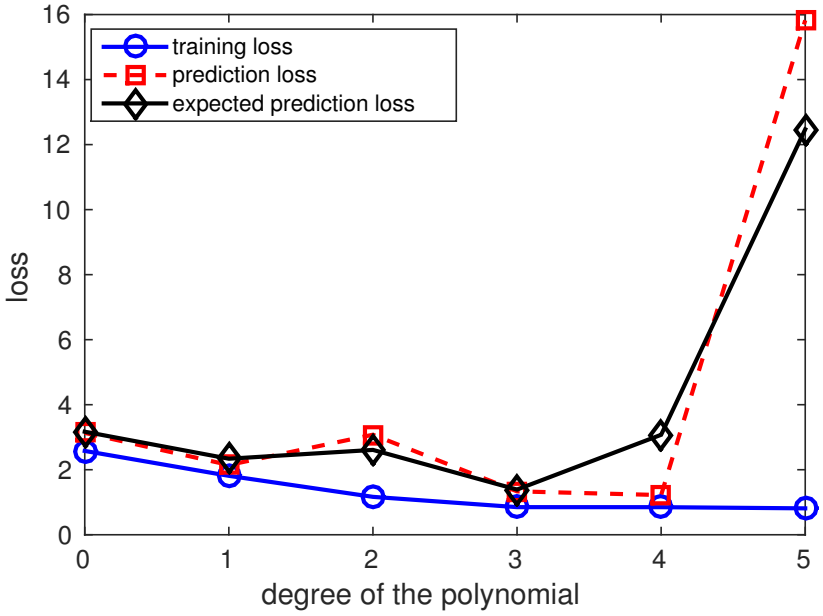

Figure 1: Polynomial regression with a square loss: training loss (solid line with circles), prediction loss (dashed line with squares), and expected prediction loss (solid line with diamonds). For Question 1d.

(e) Among others, the data analysis process generally involves the following steps:

- Getting (raw) data
- Exploratory data analysis
- Preparing the data for further analysis
- Building and fitting models

Explain the general goal of each of the steps above as well as potential issues that one needs to consider. (2 marks per step) [*8 marks*]

2. ANSWER EITHER THIS QUESTION OR QUESTION 3

Let $\boldsymbol{\Delta}$ be the $n \times n$ distance matrix with the squared Euclidean distances between all, possibly uncentred, data points $\tilde{\boldsymbol{x}}_i \in \mathbb{R}^d$ $(d > 2)$.

(a) What is the corresponding distance matrix for the centred data points? Justify your answer. [*1 mark*]

(b) The lecture introduced the centring matrix

$$\boldsymbol{H}_n = \boldsymbol{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\top.$$

i. Let $\boldsymbol{a} = (a_1, \ldots, a_n)^\top$ be a n dimensional column vector. Show that $\boldsymbol{H}_n\boldsymbol{a}$ removes the average value

$$\bar{a} = \frac{1}{n}\sum_{i=1}^{n} a_i$$

from each element of $\boldsymbol{a}$. [*3 marks*]

ii. Let $\boldsymbol{b} = (b_1, \ldots, b_n)$ be a n dimensional row vector. Using the result for the column vector, show that $\boldsymbol{b}\boldsymbol{H}_n$ removes the average value from $\boldsymbol{b}$. [*1 mark*]

(c) We showed that the Gram matrix $\boldsymbol{G}$ with all inner products between the centred data points $\boldsymbol{x}_i$ can be computed as

$$\boldsymbol{G} = -\frac{1}{2}\boldsymbol{H}_n\boldsymbol{\Delta}\boldsymbol{H}_n.$$

How are the eigenvectors and eigenvalues of $\boldsymbol{G}$ related to the singular vectors and singular values of the centred $d \times n$ data matrix $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$? [*5 marks*]

(d) Explain step-by-step, using concrete formulae, how PCA can be used to visualise data points in the two-dimensional plane given their distance matrix $\boldsymbol{\Delta}$ only. [*7 marks*]

(e) Isomap is a dimensionality reduction method that is based on the geodesic distance.

i. Use a simple two-dimensional example to explain the difference between the Euclidean and the geodesic distance. [*2 marks*]

ii. How do we compute the geodesic distances between the data points $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ in practice? [*2 marks*]

iii. What is classical multidimensional scaling (MDS) and how does it relate to dimensionality reduction by PCA? [*2 marks*]

iv. How is classical MDS used in Isomap? [*2 marks*]

3. ANSWER EITHER THIS QUESTION OR QUESTION 2

(a) The fourth standardised moment of a random variable x is called its kurtosis.

i. Provide a formula for the kurtosis. [*1 mark*]

ii. What properties of a random variable does the kurtosis assess? [*1 mark*]

iii. Figure 2 shows the probability density functions of two random variables. Which one has a larger kurtosis? Justify your answer. [*2 marks*]

(b) Figure 3 visualises the performance of four binary classifiers (labelled A to D).

i. Specify a classification rule that corresponds to classifier A. [*1 mark*]

ii. Do you prefer classifier B or C? Justify your answer. [*2 marks*]

iii. Assume classifiers C and D were trained to detect a rare heart condition whose treatment requires a risky surgery. Would you generally deploy classifier C or classifier D? Justify your answer. [*2 marks*]

(c) What is the misclassification rate of classifier D in Figure 3 if the positive labels ("1") occur twice as often as the negative labels ("-1")? [*7 marks*]

(d) Write pseudo-code that uses cross-validation with $K = 5$ folds to choose between two competing classification models (e.g. neural networks versus support vector machines), and that returns the trained selected model. You can assume that you have access to a function `train` that takes data and a given model-class as input and returns a trained classifier as output. While the function `train` minimises the log-loss, we are ultimately interested in a classifier with a small classification error. [*7 marks*]

(e) Would it generally be a good idea to use the cross-validation score of the selected model as an estimate of its generalisation performance? Justify your answer. [*2 marks*]

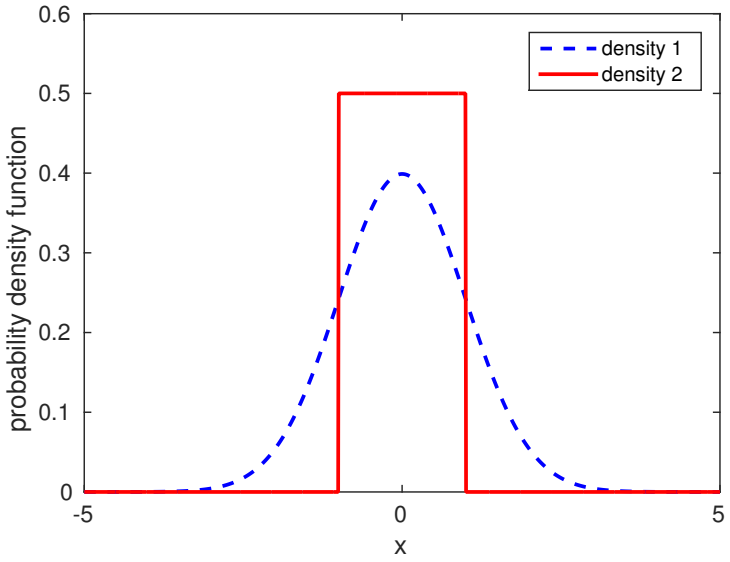

Figure 2: The probability density functions of two random variables. For Question 3a.

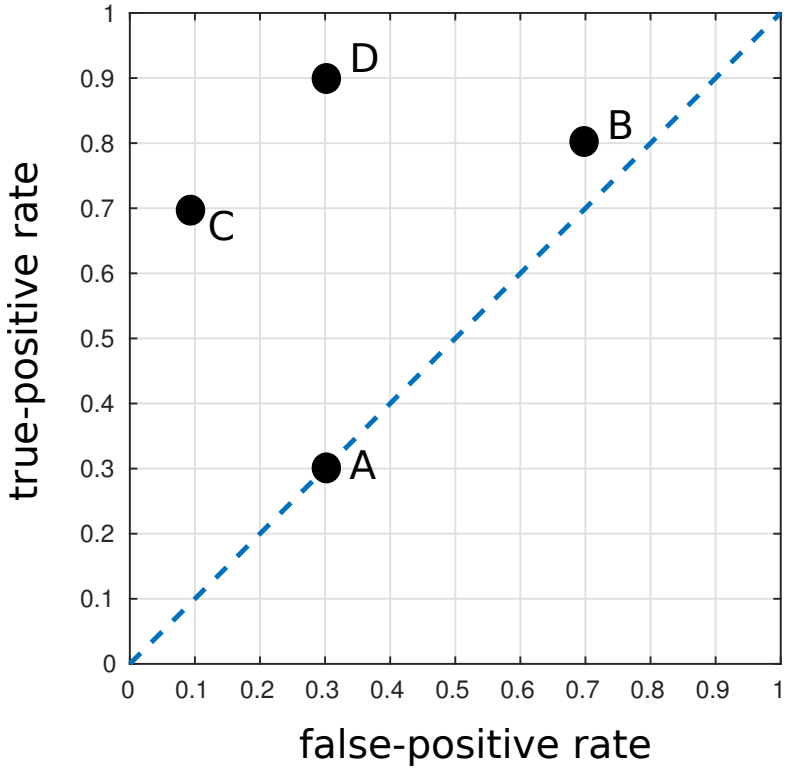

Figure 3: The performance of different classifiers in the ROC space. For Questions 3b and 3c.