UNIVERSITY OF EDINBURGH

COLLEGE OF SCIENCE AND ENGINEERING

SCHOOL OF INFORMATICS

INFR11007 DATA MINING AND EXPLORATION

May 21, 2018

09:30 to 11:30

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 NOT BE USED IN THIS EXAMINATION

MSc Courses

Convener: G. Sanguinetti
External Examiners: W. Knottenbelt, M. Dunlop, M. Niranjan, E. Vasilaki

THIS EXAMINATION WILL BE MARKED ANONYMOUSLY

1. THIS QUESTION IS COMPULSORY

(a) List at least three key points that exploratory data analysis emphasises. [*3 marks*]

(b) i. Name one robust numerical measure each to characterise location, scale and shape of a dataset. [*3 marks*]

ii. What do robust numerical measures and robust plots guard against? [*1 mark*]

(c) i. Consider the following data matrix $\boldsymbol{X}$ containing 10 observations:

$$\boldsymbol{X} = \begin{pmatrix} 5 & 2 & 8 & 7 & 6 & 7 & 1 & 0 & 11 & 4 \end{pmatrix}$$

What is the interquartile range (IQR) of these data? [*1 mark*]

ii. Due to a faulty measurement device, observation "11" is replaced by "255". Which of the following measures are unaffected by this fault? [*1 mark*]
sample kurtosis, sample variance, sample median, sample mean, sample skewness, mode, Galton's measure of skewness, median absolute deviation

(d) For a large data matrix $\boldsymbol{X}$ containing 3-dimensional centred observations, you are given the eigendecomposition of its sample covariance matrix $\boldsymbol{\Sigma} = \mathrm{cov}(\boldsymbol{X})$:

$$\boldsymbol{\Sigma} = \begin{pmatrix} 2/3 & -2/3 & 1/3 \\ 1/3 & 2/3 & 2/3 \\ 2/3 & 1/3 & -2/3 \end{pmatrix} \begin{pmatrix} 2 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 4 \end{pmatrix} \begin{pmatrix} 2/3 & -2/3 & 1/3 \\ 1/3 & 2/3 & 2/3 \\ 2/3 & 1/3 & -2/3 \end{pmatrix}^\top$$

You want to represent the following two observations from $\boldsymbol{X}$ in two dimensions: $\boldsymbol{x}_1 = (9, 9, -18)^\top$ and $\boldsymbol{x}_2 = (18, 9, 9)^\top$

i. What are the first two principal component directions? [*2 marks*]

ii. Now calculate the principal component scores of $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ corresponding to the first two principal component directions. [*2 marks*]

iii. What is the projection matrix for projecting the data into the 2-dimensional subspace with smallest expected mean square error? Note that this is a $\mathbb{R}^3 \to \mathbb{R}^3$ transformation. [*2 marks*]

iv. Calculate the data approximations in this subspace. [*2 marks*]

v. Calculate the expected mean square approximation error. [*1 mark*]

QUESTION CONTINUES ON NEXT PAGE

(e) You are given a dataset $\mathcal{D}$ containing 1000 observations and a model family $h_{\boldsymbol{\lambda}}$ indexed by hyperparameters $\boldsymbol{\lambda}$. We want to choose the hyperparameters $\boldsymbol{\lambda}$ using 3-fold cross-validation scores and estimate the generalisation performance of the resulting prediction function using a hold-out test set.

i. Into how many parts do you split the dataset $\mathcal{D}$? [*1 mark*]

ii. Describe the precise steps you take to choose the hyperparameters $\boldsymbol{\lambda}$. You can assume that we know the set of possible $\boldsymbol{\lambda}$ and that this set is small enough to iterate over. [*5 marks*]

iii. How do you estimate generalisation performance of your chosen model? [*1 mark*]

2. ANSWER EITHER THIS QUESTION OR QUESTION 3

(a) We are considering the skewness defined as the third standardised moment of a random variable and its corresponding sample estimator: the sample skewness (as in the lecture).

i. What is the formula to calculate sample skewness? [*1 mark*]

ii. After computing sample skewness on your dataset, you realise that you neglected an offset. Hence, you transform your data by adding a constant scalar value m to each observation. How will this affect the sample skewness? [*1 mark*]

iii. You measure a sample skewness of 0 for a unimodal data distribution. Does this mean that the data distribution is symmetric around its mean? Justify your answer. [*2 marks*]

(b) In the probabilistic model underlying probabilistic PCA:

i. What parameters are specified? Give also the sizes of the parameters. [*2 marks*]

ii. What variables are there, how are they distributed and how are they related? [*3 marks*]

iii. How can we change the probabilistic PCA noise to recover the PCA projection from the maximum likelihood probabilistic PCA projection? [*1 mark*]

QUESTION CONTINUES ON NEXT PAGE

(c) Consider a probabilistic model similar to the one underlying probabilistic PCA but with a joint distribution of the following multivariate normal form:

$$p(z, \boldsymbol{x}) = \frac{1}{\text{const}} \exp\left(-\frac{1}{2}\left[zmz + \boldsymbol{x}^\top(\boldsymbol{I}_d + \boldsymbol{b}m\boldsymbol{b}^\top)\boldsymbol{x} - \boldsymbol{x}^\top\boldsymbol{b}mz - zm\boldsymbol{b}^\top\boldsymbol{x}\right]\right),$$

where $m = (2 - \boldsymbol{b}^\top\boldsymbol{b})^{-1}$, $\boldsymbol{I}_d$ denotes the d-dimensional identity matrix, $\boldsymbol{b}$ is a constant vector of length d and "const" denotes terms that are independent of the scalar variable z and the d-dimensional variable $\boldsymbol{x}$. $\boldsymbol{b}$ is thus the only parameter of the model.
In the following, you can use that for matrix partitions $\boldsymbol{A}$, $\boldsymbol{B}$, $\boldsymbol{C}$, $\boldsymbol{D}$, the following identity for the inverse of a partitioned matrix holds:

$$\begin{pmatrix}\boldsymbol{A} & \boldsymbol{B}\\ \boldsymbol{C} & \boldsymbol{D}\end{pmatrix}^{-1} = \begin{pmatrix}\boldsymbol{M} & -\boldsymbol{M}\boldsymbol{B}\boldsymbol{D}^{-1}\\ -\boldsymbol{D}^{-1}\boldsymbol{C}\boldsymbol{M} & \boldsymbol{D}^{-1} + \boldsymbol{D}^{-1}\boldsymbol{C}\boldsymbol{M}\boldsymbol{B}\boldsymbol{D}^{-1}\end{pmatrix},$$

where $\boldsymbol{M} = (\boldsymbol{A} - \boldsymbol{B}\boldsymbol{D}^{-1}\boldsymbol{C})^{-1}$.

i. Calculate the covariance matrix of the joint distribution $p(z, \boldsymbol{x})$ where you treat the concatenation of z and $\boldsymbol{x}$ as a single vector variable $\begin{pmatrix} z \\ \boldsymbol{x}\end{pmatrix}$. *[3 marks]*

ii. Calculate the covariance matrix of the conditional distribution $p(\boldsymbol{x}|z)$. *[2 marks]*

iii. Calculate the mean of the conditional distribution $p(\boldsymbol{x}|z)$. *[3 marks]*

(d) Consider a binary classifier $\hat{y}(\boldsymbol{x}) = \text{sign}(h(\boldsymbol{x}; \boldsymbol{\theta}))$, where $h(\boldsymbol{x}; \boldsymbol{\theta})$ is a real-valued function with tuning parameters $\boldsymbol{\theta}$ and $\text{sign}(a)$ returns the sign of a (-1 or 1).

i. What do you plot on the axes of the receiver-operating-characteristic (ROC)? *[1 mark]*

ii. Why is it not a good strategy to minimise the false-negative rate alone? Give an example of a function h that minimises the false-negative rate. *[2 marks]*

iii. You are evaluating training loss and prediction loss of a family of functions $h(\boldsymbol{x}; \boldsymbol{\theta})$, where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)^\top$. You increase n and find that the training loss is decreasing whereas the prediction loss starts increasing. What is the name of this phenomenon? We assume that you obtained accurate estimates of training loss and prediction loss. Would you further increase n? Justify your answer. *[2 marks]*

iv. Suppose that you already added a regularisation term $\lambda_{\text{reg}} R(\boldsymbol{\theta})$ to your training loss function, where $R(\boldsymbol{\theta})$ is a non-negative penalty function. How should you change λ_{reg} if you have a low training loss but a large prediction loss? Justify your answer. *[2 marks]*

3. ANSWER EITHER THIS QUESTION OR QUESTION 2

(a) You are given three 2-dimensional data points:
$\boldsymbol{x}_1 = (1, 3)^\top, \boldsymbol{x}_2 = (2, 1)^\top, \boldsymbol{x}_3 = (4, 2)^\top$.

i. Which point pairs are concordant? Which pairs are discordant? [*2 marks*]

ii. Calculate Kendall's tau for the data points. [*1 mark*]

iii. The data points are transformed using the function:

$$(a, b)^\top \mapsto (\log(a), \log(b))^\top$$

Calculate Kendall's tau for the transformed data points. [*1 mark*]

(b) Consider a data matrix $\boldsymbol{X} = \begin{pmatrix} \boldsymbol{x}_1 & \boldsymbol{x}_2 & \dots & \boldsymbol{x}_n \end{pmatrix}$ consisting of n d-dimensional observations that have not been centred.

i. Suppose we are using a k-dimensional space for dimensionality reduction, $k < d$. Describe all the precise steps necessary to do PCA dimensionality reduction from the inner products $\boldsymbol{X}^\top \boldsymbol{X}$. You can use the centring matrix $\boldsymbol{C}_n = \boldsymbol{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\top$. [*4 marks*]

ii. Show that the eigenvalues of the Gram matrix are non-negative. [*4 marks*]

iii. How can we calculate the fraction of variance explained from the eigenvalues $\lambda_1, \dots, \lambda_d$ of the sample covariance matrix of $\boldsymbol{X}$? [*1 mark*]

iv. Show that the principal components (i.e. the random variables corresponding to the principal component scores, not the principal component directions) are uncorrelated. [*4 marks*]

v. Now we want to apply kernel PCA with the polynomial kernel $k(\boldsymbol{x}, \boldsymbol{x}') = (\boldsymbol{x}^\top \boldsymbol{x}')^a$ where $a = 2$. Calculate the transformed Gram matrix for the (very small) data matrix $\boldsymbol{X} = \begin{pmatrix} \boldsymbol{x}_1 & \boldsymbol{x}_2 \end{pmatrix} = \begin{pmatrix} 1 & -1 \\ 1 & 0 \end{pmatrix}$. [*3 marks*]

QUESTION CONTINUES ON NEXT PAGE

(c) We want to apply classical multidimensional scaling, i.e. we want to find lower-dimensional vectors $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n$ that represent given dissimilarities δ_{ij} of a dissimilarity matrix $\boldsymbol{\Delta}$.

i. The eigendecomposition of the hypothetical Gram matrix $\boldsymbol{G}'$ takes the form

$$\boldsymbol{G}' = \boldsymbol{V} \begin{pmatrix} 0 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & -0.2 \end{pmatrix} \boldsymbol{V}^\top,$$

where $\boldsymbol{V}$ denotes an orthogonal matrix. Please ignore the low number of observations (the size of the matrix means that we have 3 observations only). What dimensionality do you pick for your lower dimensional vectors $\boldsymbol{z}_i$ to obtain a meaningful configuration representing the dissimilarities? Justify your answer. [*2 marks*]

ii. You realise that your data lie on a complicated manifold, so you want to apply isometric feature mapping (Isomap) to visualise it. How is Isomap related to classical MDS? [*1 mark*]

iii. Describe one way to obtain the neighbourhood graph for Isomap. [*1 mark*]

iv. Given the neighbourhood graph, how can we calculate the Isomap dissimilarity between a pair of data points? [*1 mark*]