ETC3250

Practice exam 2017

Instructions

- In the actual exam, there are 8 questions worth a total of 100 marks. You should attempt them all.
- Open book
- You can use the approved calculator

This practice exam has a range of questions that show possible range of topics that are examined. They come from last year's exam and practice exam, and there are a few additional new questions.

QUESTION 1

(a) *The minimum value of the mean squared error is the irreducible error term of the bias and variance decomposition.*

[2 marks]

(b) True or false. By constraining the L_1 norm of the vector of coefficients, it is possible to obtain sparse estimate for the coefficients. Explain your answer.

(c) When is Multidimensional scaling (MDS) is equivalent to Principal Component Analysis (PCA)?

(d) What is the difference between 2-fold cross-validation and 10-fold cross-validation? Which one would you use in practice?

[2 marks]

(e) Write the formula to compute the prediction of a K-NN classifier for the new data point x_0.

[2 marks]

(f) *Using the error rate to measure classification accuracy can be misleading if the dataset is unbalanced.*

[2 marks]

[Total: 8 marks]

— END OF QUESTION 1 —

QUESTION 2

Suppose you estimate the coefficients of a linear regression by solving the following optimization problem:

$$\underset{\boldsymbol{\beta}}{\text{minimize}} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

where $\lambda \geq 0$ is a **tuning parameter**.

(a) Briefly explain how λ affects the bias and variance tradeoff of your estimate $\hat{\boldsymbol{\beta}}$.

[2 marks]

(b) Explain why is it important to standardize the predictors in this case

[2 marks]

(c) Sparsity is very useful in high-dimensional regression. Explain why.

(d) Does the previous optimization allows sparse estimates? Explain your answer.

[Total: 4 marks]

— END OF QUESTION 2 —

QUESTION 3

This question is about bootstrapping.

(a) Give an example of algorithm that uses bootstrapping and why.

(b) Bootstrapping can be used to estimate the sampling distribution of a statistic. Explain this procedure.

(c) Can we use the previous bootstrap procedure when the data is a time series? Explain.

[Total: 0 marks]

— END OF QUESTION 3 —

QUESTION 4

(a) Briefly explain why the K-means algorithm is guaranteed to decrease the value of the objective at each step.

(b) True or false. For any starting values of the assignment of the observations, the K-means algorithm will always converge to the same solution.

(c) For the following data.

	X1	X2	X3	X4
A	-1.02	0.27	-0.81	-0.34
B	-0.61	0.97	0.76	0.71
C	0.70	-0.38	0.88	-0.32
D	-0.82	0.48	-0.71	-0.98
E	-0.72	0.97	-0.33	0.04

and the associated distance matrix:

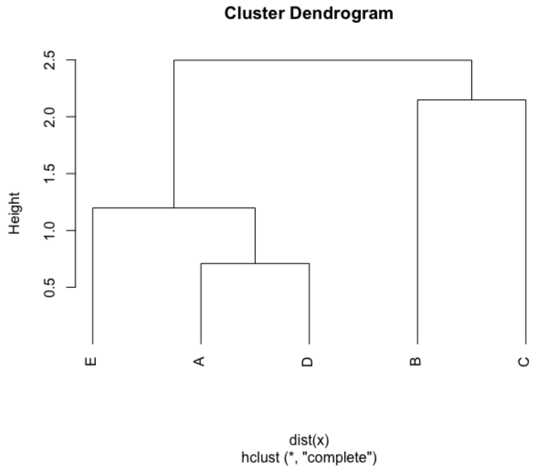
	A	B	C	D	E
A	0.00	2.06	2.50	0.71	0.98
B	2.06	0.00	2.15	2.30	1.28
C	2.50	2.15	0.00	2.45	2.33
D	0.71	2.30	2.45	0.00	1.20
E	0.98	1.28	2.33	1.20	0.00

(i) Compute the Euclidean distance between observations A and E.

[2 marks]

(ii) The dendrogram shows hierarchical clustering with complete linkage. Which two points were fused at the first step of hierarchical clustering with complete linkage?

[2 marks]



(iiI) What is the intercluster distance (linkage) values between the new cluster and the remaining three points?

[2 marks]

[Total: 6 marks]

— END OF QUESTION 4 —

QUESTION 5

A principal component analysis is conducted on a subset of Mexico City data, health and pollution where missing values have been imputed using regression methods. There are five variables in the subset: deaths (number of deaths each day), temp_mean (average temperature), humidity, NOX (nitrogen oxide, pollutant), O3 (ozone, pollutant).

```
> mexico.pca1 <- prcomp(mexico[,c("deaths", "temp_mean",
    "humidity", "NOX", "O3")], scale=T, retx=T)
> mexico.pca1
Standard deviations:
[1] 1.37 1.27 0.86 0.64 0.62

Rotation:
            PC1    PC2    PC3    PC4    PC5
deaths    -0.30  0.573 -0.471 -0.597  0.077
temp_mean -0.33 -0.578  0.338 -0.631  0.213
humidity   0.64 -0.074  0.003 -0.468 -0.608
NOX       -0.25  0.510  0.766 -0.028 -0.300
O3        -0.58 -0.268 -0.279  0.161 -0.699

> summary(mexico.pca1)
Importance of components:
                         PC1   PC2   PC3    PC4   PC5
Standard deviation     1.367 1.267 0.856 0.6362 0.624
Proportion of Variance 0.374 0.321 0.146 0.0809 0.078
Cumulative Proportion  0.374 0.695 0.841 0.9220 1.000
```

(a) Compute the total variance.

[1 marks]

(b) Make a sketch of the scree plot. Label your axes.

[2 marks]

(c) The PCA was conducted on the correlation matrix. Why do you think that this was necessary?

[1 marks]

(d) What proportion of variance is explained by four PCs?

[1 marks]

(e) Interpret the first principal component.

[2 marks]

(f) How many principal components would you use to summarize the variation of this data? Why?

[2 marks]

[Total: 9 marks]

— END OF QUESTION 5 —

QUESTION 6

(a) Select the propositions that are true, and briefly explain your answer:

[4 marks]

- □ The test mean squared error can be smaller than the training mean squared error.
- □ The model is underfitting when the training error is very large.
- □ Increasing the number of neighbours in the K-nearest neighbours algorithm will increase the flexibility of the model.
- □ Using cross-validation for model selection will necessarily provide models with better prediction accuracy compared to models selected by AIC and BIC.

(b) Consider a simple classification procedure applied to a two-class dataset with 5000 predictors and 50 samples:

Step 1. Find the 100 predictors having the largest correlation with the class labels.

Step 2. Apply logistic regression using only these 100 predictors .

How do we estimate the test set performance of this classification procedure?

[2 marks]

Can we use cross-validation? If yes, describe the step-by-step cross-validation procedure.

[3 marks]

(c) If we have n data points, what is the probability that a given data point does not appear in a bootstrap sample?

[3 marks]

[Total: 12 marks]

— END OF QUESTION 6 —

QUESTION 7

(a) Suppose we have a data set with five predictors, X_1 = GPA, X_2 = IQ, $X3$ = Gender (1 for Female and 0 for Male), X_4 = GPA×IQ, and X_5 = GPA×Gender. The response Y is starting salary after graduation (in thousands of dollars). Suppose we use least squares to fit the model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + e$$

where e is a random error term, and we obtain estimates $\hat{\beta}_0 = 50$, $\hat{\beta}_1 = 20$, $\hat{\beta}_2 = 0.07$, $\hat{\beta}_3 = 35$, $\hat{\beta}_4 = 0.01$, $\hat{\beta}_5 = -10$.

Which answer is correct, and why?

[4 marks]

(a) For a fixed value of IQ and GPA, males earn more on average than females.

(b) For a fixed value of IQ and GPA, females earn more on average than males.

(c) For a fixed value of IQ and GPA, males earn more on average than females provided that the GPA is high enough.

(d) For a fixed value of IQ and GPA, females earn more on average than males provided that the GPA is high enough.

(b) Predict the salary of a female with IQ of 110 and a GPA of 4.0.

[2 marks]

(c) True or false: Since the coefficient for the GPA×IQ interaction term is very small, there is very little evidence of an interaction effect. Justify your answer.

[3 marks]

(d) Suppose that the true relationship between starting salary and IQ is nonlinear for fixed values of GPA and Gender. You wish to compare the cubic model (including X_2, X_2^2 and X_2^3) to the linear model (including X_2 but not the quadratic or cubic terms). You split the available data into two forming a training set and test set, and you estimate both models using only the training data. Which of the two models would you expect to have larger mean squared error computed on the training data, or is there not enough information to say? Which of the two models would you expect to have larger mean squared error computed on the test data, or is there not enough information to say? Justify your answer.

[4 marks]

(e) You need to decide whether to add the quadratic and cubic terms to the model. Explain how you would do this using a test set, using cross-validation, and using the AIC. Comment on which of these three methods you would prefer and why.

[3 marks]

[Total: 16 marks]

— END OF QUESTION 7 —

QUESTION 8

This question is about ensemble methods.

(a) Suppose B is the number of trees in a random forest, σ^2 is the variance of each tree, and $\rho > 0$ is the correlation between trees in a random forest. Explain how these three components affect the variance of the forest, and how to calibrate them to reduce the variance without affecting too much the bias.

[2 marks]

(b) Suppose you are not allowed to use a validation set or cross-validation, how can you estimate the performance of the random forest algorithm for unseen observations?

[2 marks]

(c) When using bagging, explain why it is often recommended to consider highly flexible model such as trees instead of simple linear models.

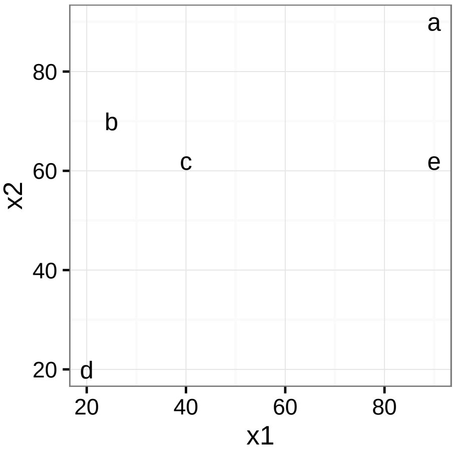
[2 marks]

[Total: 6 marks]

— END OF QUESTION 8 —

QUESTION 9

We consider the following 5 data points: $a(90,90), b(25,70), c(40,62), d(20,20), e(90,62)$ given in the Figure below.



(a) Compute the Euclidean distance between points a and b.

[2 marks]

(b) We will perform $k-$means clustering with $k = 2$. The algorithm is at this stage, cluster 1 contains one point, $\{a\}$, and cluster 2 contains 4 points, $\{b, c, d, e\}$. Compute the means for the two clusters.

[2 marks]

(c) At the next stage in k-means would point e be grouped with cluster 1 or 2? Why.

[2 marks]

(d) We are going to perform hierarchical clustering using Euclidean distances using *complete linkage*. The interpoint distance matrix is below. Which two points would be joined at the first step?

[2 marks]

	a	b	c	d	e
a	0.00	68.01	57.31	98.99	28.00
b	68.01	0.00	17.00	50.25	65.49
c	57.31	17.00	0.00	46.52	50.00
d	98.99	50.25	46.52	0.00	81.63
e	28.00	65.49	50.00	81.63	0.00

(e) With numerical data, it is a good idea to standardize the variables before computing the interpoint clusters. Why?

[2 marks]

[Total: 10 marks]

— END OF QUESTION 9 —