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]

TRUE. MSE = bias squared + variance + irreducible error. When bias = 0, and variance = 0, MSE = irreducible error.

(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.

TRUE. This can be explained geometrically. The L_1 ball has corners which encourage sparse solutions.

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

When the pairwise distance matrix uses Euclidean distances.

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

[2 marks]

The number of folds K control the bias and variance tradeoff in the cross-validation estimate. A large value decrease the variance but increase the bias, and vice versa. 10-fold cross-validation is often used in practice. However, 2-fold cross-validation can also be used with a large dataset.

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

[2 marks]

$$j^* = \text{argmax}_j \frac{1}{K} \sum_{i \in \mathcal{N}_0} \boldsymbol{I}(y_i = j)$$

where $\mathcal{N}_0$ contains the K nearest neighbors of x_0.

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

[2 marks]

TRUE. Suppose we have n_1 observations for class 1, and n_2 observations for class 2, with $\frac{n_1}{n_2} = 90\%$. Then classifying all observations as class 1 gives us a 90% error rate which is misleading since all class 2 observations have been missclassified.

[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]

$\hat{\boldsymbol{\beta}} = (\boldsymbol{X'X} + \lambda \boldsymbol{I})^{-1} X'y$. In the limit, when $\lambda = \infty$, the variance is zero, and the bias is very large. The other extreme is when $\lambda = 0$, which reduce to the least square estimate, with small bias and high variance. By using a value of λ between these two extremes, we can control the bias and variance tradeoff.

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

[2 marks]

If we do not apply standardization, the predictors with the largest variance will dominate the penalty term.

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

A lot of zero coefficients provides better interpretability of the final model. Possibly better predictions due to the lower variance. Finally, the best on sparsity principle.

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

Although it shrinks the estimate towards zero, it does not provide sparse estimate, expect when $\lambda = \infty$ where all the coefficients are zero which is not a useful solution in general. To obtain sparse estimate, we need to consider L_q norm with $q \leq 1$.

[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.

Random forests. It allows to build multiple trees and average the results to reduce the variance.

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

- Sample with replacement the i.i.d. observations
- For each bootstrap sample, compute the statistic
- The distribution of these bootstrap statistics is an estimate of the sampling distribution of the statistic.

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

No. Time series data do not satistify the i.i.d assumption required by the bootstrap procedure. It is possible to use another bootstrap procedure for time series, called block bootstrap for example.

[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.

The sum of the squared euclidean distance between the observations in a cluster is equal to two times the sum of the squared eunclidean distance between each observation and the centroid. The fact that K-means assign observations to the closest centroid, it is guaranteed that the sum of the squared euclidean distance will decrease.

(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.

FALSE. The K-means algorithms converges to a local optimum.

(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.

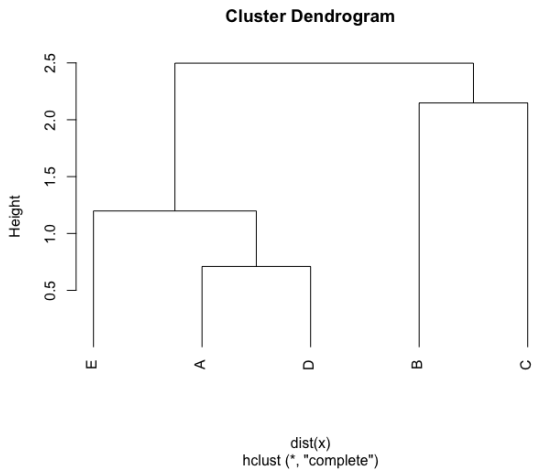
[2 marks]

0.98

(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]

A, D



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

[2 marks]

AD-B = 2.30, AD-C=2.45, AD-E=1.20

[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
```

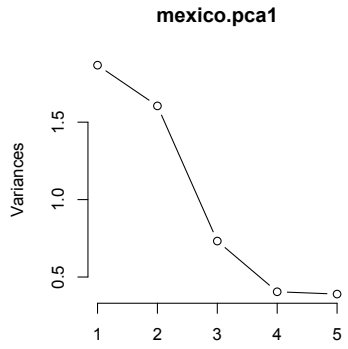
(a) Compute the total variance.

[1 marks]

5

(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]

The variables were recorded in different units.

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

[1 marks]

0.922

(e) Interpret the first principal component.

[2 marks]

The first PC is a combination of all of the variables, but mostly a contrast between humidity and O3. If humidity is high O3 tends to be low.

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

[2 marks]

Scree plot really suggests 4. Even two PCs explains a lot of the variation, though. Definitely need two, they explain different things in the variables. Coefficients of third PC contradicts the second, deaths and NOX have opposite signs, although it might be interpreted as only NOX. Similarly for PC4. Given that the eigenvalue for PC 3 is so much smaller, and not a lot more variation is explained, I'd go with two rather than 4.

[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. TRUE
- ☐ The model is underfitting when the training error is very large. TRUE
- ☐ Increasing the number of neighbours in the K-nearest neighbours algorithm will increase the flexibility of the model. FALSE
- ☐ Using cross-validation for model selection will necessarily provide models with better prediction accuracy compared to models selected by AIC and BIC. FALSE

(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]

Split into training and testing. Not forgetting to include step 1 in the procedure.

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

[3 marks]

Yes. A typical cross-validation procedure, but check if step 1 has also been considered in the cv procedure.

(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]

$$(1 - \frac{1}{n})^n$$

[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.
False

(b) For a fixed value of IQ and GPA, females earn more on average than males.
True, if GPA < 3.5, otherwise false.

(c) For a fixed value of IQ and GPA, males earn more on average than females provided that the GPA is high enough.
True. Once GPA > 3.5

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

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

[2 marks]

137.1

(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]

This is true. The coefficient is small but the magnitude of values large, so the effect is small. In contrast, the coefficient for GPA×Gender is large, and this is where the interaction does have an effect on salary.

(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]

The MSE for the linear model should be higher than the more complicated model. Adding more terms reduces the error term. It may be hard to say which has the bigger error for the test data. Errors should be higher onthe test data anyway, but because the model was fitted to the training set, it is possible to have larger error with the more complicated model because the training data was overfitted.

(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]

It is always good to have the training/test set because it is clear what the model is fitted on, and how it performs on the new data. However, if this approach is used, you really need to have training, validation and test sets, so that you have a set of data for prediction that was not used at all in the model building. The drawback is that you have one slice of the pie and all inference is depending on the training test split.

Cross-validation should give a better handle on error with future data, because it guards against odd sampling effects. You could use cross-validation on training and validation sets, to decide whether the cubic model gets substantial reduction in test error enough to use that model. Then you would need to fit the model with the training set, and finally report error for the test set.

AIC is useful for model selection. The linear vs cubic models are nested, which is the appropriate situation for comparing AIC values. You can still work with a training set of data, and use AIC to select the model. It should help to prevent overfitting.

[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]

The total variance is given by $\rho\sigma^2 + \sigma^2\frac{1-\rho}{B}$. If the regression function is highly nonlinear, in order to have a small bias, σ^2 will be high. To reduce the total variance, we want to reduce ρ, and increase B.

(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]

Trees are repeatedly fit to bootstrapped subsets of the observations. We have shown that on average, each bagged tree makes use of around two-thirds of the observations. The remaining one-third of the observations not used to fit a given bagged tree are referred to as the out-of-bag (OOB) observations. We can use them as test samples.

(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]

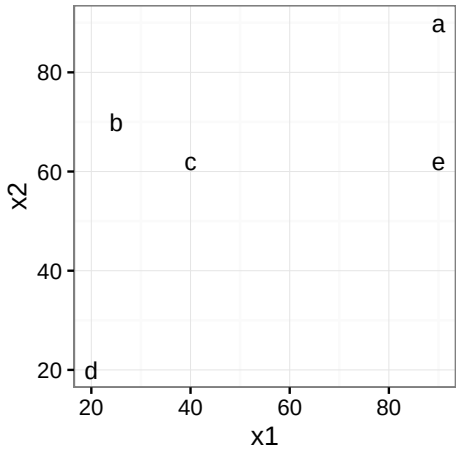
By using highly flexible models, we can assume that the bias is almost zero. The bagging procedure will average a bunch of high variance estimators, which will (under certain conditions) reduce the total variance.

[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]

68.00735

(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 centroids for the two clusters.

[2 marks]

$c_1 = (90, 90)$ and $c_2 = (43.75, 53.50)$

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

[2 marks]

Cluster 2 since it is closer to the centroid 2.

(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

b and c

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

[2 marks]

Yes. For example, if the variables have different scales and we are using Euclidean distances

[Total: 10 marks]

— END OF QUESTION 9 —