Tut 1: Basics of Statistical Learning

May/June 2022

Cross industry standard **process** of data mining (CRISP-DM):

- Business understanding
- Data understanding (Prob & Stat I)
- Data preparation
- Modelling
- Evaluation
- Deployment

Business Understanding

1. Describe the things that predictive analytics can help tackle in real-world business problems. (Come out in 2022 Jan 2022 Semester Final Exam. 4 marks)

Data Preparation

2. You are given the following data.

Candidate	Project	Experience	Major	Hired (Class)
1	Y	H	CS	Y
2	N	H	SE	Y
3	Y	M	CE	Y
4	N	L	AS	N
5	Y	L	AM	N
6	Y	M	CE	Y
7	Y	L	FM	N
8		H	SE	Y
9	Y	H	AM	Y
10	N	L	AS	N

Use the following method to replace the missing value (of a categorical data)

(a) Mode

(b) Hot deck

3. There are 290 customers in ABC company. Given that the mean customer weight from ABC company database is 55.8kg. It is found that a customer's weight was incorrectly recorded as 580kg. Recalculate the mean if

(a) The correct weight is 58kg.

(b) The error is replaced by mean.

(c) The error is replaced by regression. Note that the height of this customer is 160cm and from overall data and the regression line of weight, y, against the height of the customer, x, is

$$y = 0.39x - 6.8$$

Modelling

4. (Jan 2021 Final Q1(a)) Describe the classification of supervised models using

   (a) the Bayesian approach. (1 mark)

   (b) the output's type. (1 mark)

5. (Jan 2022 Final Q1(b)) Assuming the inputs of the data are all numeric and the output is binary. Give two examples of supervised learning models for each of the following class.

   (a) parametric discriminative models (2 marks)

   (b) nonparametric discriminative models (2 marks)

   (c) generative models (2 marks)

6. For each parts (a) through (d), indicate whether we would generally expect the performance of a flexible statistical learning method to be better or worse than an inflexible method. Justify your answer.

   (a) The sample size n is extremely large, and the number of predictors p is small.

   (b) The number of predictors p is extremely large, and the sample size n is small.

   (c) The relationship between the predictors and response is highly non-linear.

   (d) The variance of the error terms $\sigma^2 = var(\epsilon)$ is extremely high.

7. Explain whether each scenario is a classification or regression problem, and indicate whether we are most interested in inference or prediction. Finally, provide n and p.

   (a) We collect a set of data on the top 500 firms in Malaysia. For each firm we record profit, number of employees, industry and the CEO salary. We are interested in understanding which factors affect CEO salary.

   (b) We are considering launching a new product and wish to know whether it will be a success or a failure. We collect data on 20 similar products that were previously launched. For each product we have recorded whether it was a success or failure, price charged for the product, marketing budget, competition price, and ten other variables.

   (c) We are interested in predicting the percentage change in MYR in relation to the weekly changes in the world stock markets. Hence, we collect weekly data for all of 2015. For each week we record the percentage change in MYR, the percentage change in KLSE, the percentage change in NASDAQ and the percentage change in Nikkei 225.

Evaluation

8. Table below shows a confusion matrix for a binary classification problem after applying Model A.

	True +	True -
Predicted +	114	16
Predicted -	72	125

   (a) Calculate the following accuracy measures.

      (i) Sensitivity

      (ii) Specificity

      (iii) Accuracy

(iv) Positive predictive value

(v) Negative predictive value

(b) Compare the recall and precision for both classes (positive and negative). Interpret your results with refer to the performance of Model A.

9. (Jan 2022 Final Q1(c)) Given the confusion matrix of a 1002 training data for a predictive model of the prostate cancer diagnostic with a response variable *Result* of values "B" (positive, an abbreviation for benign) and "M" (negative, an abbreviation for malignant) in Table 1.1.

Table 1.1: Confusion matrix.

	Actual	
Prediction	B	M
B	507	131
M	104	260

Calculate the following statistical measures for evaluating the performance of the predictive model.

(a) Accuracy (ACR) (2 marks)

(b) Sensitivity (2 marks)

(c) Specificitiy (2 marks)

(d) Negative Predictive Value (NPV) (2 marks)

(e) Kappa Statistic

$$\text{Kappa} = \frac{\text{Accuracy} - \text{RandomAccuracy}}{1 - \text{RandomAccuracy}}$$

where

$$\text{RandomAccuracy} = \frac{(\text{TN+FP}) \times (\text{TN+FN}) + (\text{FN+TP}) \times (\text{FP+TP})}{(\text{Total Number of Test Data})^2}.$$

The Kappa statistic compares the accuracy of the system to the accuracy of a random system. The accuracy of the system is an observational probability of agreement and the random accuracy is a hypothetical expected probability of agreement under an appropriate set of baseline constraints. (2 marks)

Deployment

10. (Jan 2021 Final Q1(b)) Write down two applications of supervised learning. In the two applications, state the target variables. (2 marks)