Tut 1: Basics of Statistical Learning

Jan 2023

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)

Solution. (a) Detecting outliers/anomalies/fraud: Strange / criminal behaviours usually have different patterns from regular patterns. As cybersecurity becomes a growing concern, the examination of all actions on a network in real time to spot abnormalities that may indicate spam, fraud, zero-day vulnerabilities and advanced persistent threats.

(b) Optimising marketing campaigns: Predictive analytics are used to determine customer responses or purchases, as well as promote **cross-sell** opportunities. The effectiveness of marketing campaigns can be better assessed. E.g. recommendation engines are widely used for online shopping recommendations as predictions are made from using customers' prior purchasing and browsing behaviour.

(c) Improving operations: Many companies use predictive models to forecast **inventory** and manage **resources**. Airlines use predictive analytics to set ticket prices. Hotels try to predict the number of guests for any given night to maximise occupancy and increase revenue. Engineering uses it to estimate component/part replacement and maintenance.

(d) Reducing risk: Credit scores are used to assess a buyer's likelihood of default for purchases. A credit score is a number generated by a predictive model that incorporates all data relevant to a person's creditworthiness. Other risk-related uses include insurance claims and collections.

□

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

Solution. $\mathrm{P}(\text{Project} = \text{Y}) = \frac{2}{3}$;
$\mathrm{P}(\text{Project} = \text{N}) = \frac{1}{3}$;
Mode = Y;
Hence, Project = Y □

(b) Hot deck

Solution. Experience = H; Major = SE, similar to Candidate 2, which Project = Y.
Hence, Project = Y □

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.

Solution. Replace the error weight from total:
Total = 290 * 55.8kg = 16182kg
$\text{Total}^* = 16182\text{kg} - 580\text{kg} + 58\text{kg} = 15660\text{kg}$
Recalculate new mean:
$\text{mean}^* = \frac{15660\text{kg}}{290} = 54.0\text{kg}$ □

(b) The error is replaced by mean.

Solution. Exclude the error weight from total:
Total = 290 * 55.8kg = 16182kg
$\text{Total}^* = 16182\text{kg} - 580\text{kg} = 15602\text{kg}$
Recalculate new mean (exclude the error):
$\text{mean}^* = \frac{15602\text{kg}}{289} \approx 53.99\text{kg}$
□

(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

$$y = 0.39x - 6.8$$

Solution. Exclude the error weight from total:
Total = 290 * 55.8kg = 16182kg
Total* = 16182kg − 580kg = 15602kg
Estimate the new weight with regression:
weight* = 0.39(160) − 6.8 = 55.6kg
Recalculate new mean:
mean* = $\frac{15602\text{kg}+55.6\text{kg}}{290} \approx 53.99\text{kg}$ □

Modelling

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

   (a) the Bayesian approach. (1 mark)

   Solution. discriminative models vs generative models ..... [1 mark] □

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

   Solution. classifiers vs regressors ..... [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)

   Solution. logistic regression model and artificial neural network model (alternative: linear SVM) ..... [2 marks] □

   (b) nonparametric discriminative models (2 marks)

   Solution. kNN model and decision tree model (alternatives: Random Forest, nonlinear/kernel SVM) ..... [2 marks] □

   (c) generative models (2 marks)

   Solution. naive bayes model and linear discriminant analysis model ... [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.

   Solution. Better — a flexible approach will fit the data closer and with the large sample size a better fit than an inflexible approach would be obtained. □

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

   Solution. Worse — a flexible method would overfit the small number of observations. □

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

   Solution. Better — with more degrees of freedom, a flexible model would obtain a better fit. □

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

Solution. Worse — a flexible method fit to the noise in the error terms and increase variance. □

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.

   Solution. Regression; inference — quantitative output of CEO salary based on CEO firm's features
   $n = 500$ firms in the US
   $p = 3$; profit, number of employees, industry □

   (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.

   Solution. Classification; prediction — predicting new product's success or failure
   $n = 20$ similar products previously launched
   $p = 13$; price charged, marketing budget, comp. price, 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.

   Solution. Regression; prediction — quantitative output of percentage change in MYR
   $n = 52$ weeks of 2015 weekly data
   $p = 3$; change in KLSE, percentage change in NASDAQ, 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

   Solution. $TPR = \dfrac{114}{114 + 72} = 0.6129 = 61.29\%$ □

   (ii) Specificity

   Solution. $TNR = \dfrac{125}{16 + 125} = 0.8865 = 88.65\%$ □

   (iii) Accuracy

Solution. $ACR = \dfrac{114+125}{114+72+16+125} = 0.7309 = 73.09\%$ □

(iv) Positive predictive value

Solution. $PPV = \dfrac{114}{114+16} = 0.8769 = 87.69\%$ □

(v) Negative predictive value

Solution. $NPV = \dfrac{125}{72+125} = 0.6345 = 63.45\%$ □

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

Solution. For positive (+) class, **recall** (sensitivity) is 61.29% and precision (PPV) is 87.69%. The positive (+) class has a low recall and high precision. This means that Model A casts a small but highly specialised model — does not capture a lot of prediction on positive (+) class, but mostly the prediction for positive (+) class is correct. For negative (-) class, recall is 88.65% and precision is 63.45%. The negative (-) class has a high recall and low precision. This means that Model A casts a wide but generalised model — captures a lot of prediction on negative (-) class, but the prediction for negative (-) class might be incorrect. □

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)

Solution. $\text{ACR} = \dfrac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{FN} + \text{TN}} = \dfrac{507+260}{507+131+104+260} = 0.765469$ [2 marks] □

(b) Sensitivity (2 marks)

Solution. $\text{TPR} = \dfrac{\text{TP}}{\text{TP} + \text{FN}} = \dfrac{507}{507+104} = 0.829787$ .....[2 marks] □

(c) Specificitiy (2 marks)

Solution. $\text{TNR} = \dfrac{\text{TN}}{\text{FP} + \text{TN}} = \dfrac{260}{131+260} = 0.664962$ ..... [2 marks] □

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

Solution. $\text{NPV} = \dfrac{\text{TN}}{\text{TN} + \text{FN}} = \dfrac{260}{104 + 260} = 0.714286$ ..... [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)

Solution.

$$\text{RandomAccuracy} = \frac{(260 + 131)(260 + 104) + (507 + 104)(507 + 131)}{1002^2}$$
$$= 0.5300198$$
$$\text{Kappa} = \frac{0.765469 - 0.5300198}{1 - 0.5300198} = 0.5009768$$

..... [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)

Solution. Anyone of the following or any reasonable answer will be accepted. A minimum of 0.5 mark will be deducted if the targets are not mentioned.

- Spam filter. Target: The type which allows for the decision to move the email, SMS text, etc. to the spam folder.
- OCR / Handwriting recognition. Target: characters and words.
- Object recognition in computer vision. Target: the term associated with the object type.
- Speech recognition. Target: sentences.
- Database marketing. Target: potential customer.

However, the listing of predictive models will not be accepted. ..... [2 marks] □