Tut 1: Basics of Statistical Learning

June 2024

Business Understanding

1. (Final Exam Jan 2024, Jan 2023 Sem, Q1(a), 3 marks) State the phases/processes involved in the CRISP-DM (Cross Industry Standard Process for Data Mining).

   Solution. The phases in CRISP-DM are:

   - Business understanding ..... [0.5 mark]
   - Data understanding ..... [0.5 mark]
   - Data preparation ..... [0.5 mark]
   - Modelling ..... [0.5 mark]
   - Evaluation ..... [0.5 mark]
   - Deployment ..... [0.5 mark]

   Average: 2.52 / 3 marks in Jan 2024; 14.55% below 1.5 marks.

   Average: 2.52 / 3 marks in Jan 2023; 13% below 1.5 marks. □

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

   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 Understanding

Look at the available data (may be from computer files or from company database) to **form a table** of features and response and understand the statistical correlation between each feature and response.

3. (Final Exam Jan 2023 Sem, Q1(b), 4 marks) State two examples of unstructured data and then state the techniques for us to transform the two examples of unstructured data to tabular data for predictive modelling.

Solution. Two examples of unstructure data:

- Texts from Email, Word document or PDF document ..... [1 mark]
- Images of possibly different sizes ..... [1 mark]

Techniques to transform the two unstructured data suitable for predictive modelling are:

- Text can be converted to a tabular data using the **bag of word** representation or the **TF-IDF (Term Frequency-Inverse Document Frequency)** representation. [1 mark]
- We need to convert (rescale) the images to 3D matrices or 2D arrays of a fixed shape. [1 mark]

Average: 1.29 / 4 marks in Jan 2023; 54% below 2 marks.

Reason: Many didn't bother to read the lecture notes to understand the complexity of things we need to classify in real-world problems. □

Data Preparation

Once we understand the structure of the data (text, image, SQL, NoSQL, etc.), we need to 'collect' the data (relevant to the objective) by downloading them or writing SQL/NoSQL queries to get them. After that, they may be processed for use in predictive models. E.g. for some predictive models (e.g. LR), the missing values cannot be handled in the mathematical formulation and we may need to impute them. For some predictive models (e.g. kNN and LDA), the standard deviation of numeric features should not be too large and need to be rescaled.

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

(b) Hot deck

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

5. 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
   Total^* = 16182kg − 580kg + 58kg = 15660kg
   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
   Total^* = 16182kg − 580kg = 15602kg
   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:
   $\text{weight}^* = 0.39(160) - 6.8 = 55.6\text{kg}$
   Recalculate new mean:
   $\text{mean}^* = \frac{15602\text{kg}+55.6\text{kg}}{290} \approx 53.99\text{kg}$ □

6. (Final Exam Jan 2023 Sem, Q1(c), 5 marks) Given the training data in Table 1.1 and the new data in Table 1.2, write down the formula for min-max scaling and then apply the min-max scaling to the new data in Table 1.2.

Obs.	x_1	x_2
1	44	3.87
2	69	1.35
3	46	1.81
4	13	2.94
5	73	1.96
6	35	4.51
7	12	3.82
8	50	2.70

Table 1.1: Training Data

Obs.	x_1	x_2
1	65	2.65
2	53	4.53
3	36	4.05
4	96	3.02
5	42	3.57

Table 1.2: New Data

Solution.

The formulas for min-max scaling on the two columns are

$$s_1(x_1) = \frac{x_1 - 12}{73 - 12} = \frac{x_1 - 12}{61}$$
$$s_2(x_2) = \frac{x_2 - 1.35}{4.51 - 1.35} = \frac{x_2 - 1.35}{3.16}$$

.....[2 marks]

After min-max scaling, the new table (Table 1.2) becomes

Obs.	x_1'	x_2'
1	0.8688525	0.4113924
2	0.6721311	1.0063291
3	0.3934426	0.8544304
4	1.3770492	0.5284810
5	0.4918033	0.7025316

.....[3 marks]

Average: 2.18 / 5 marks in Jan 2023; 67% below 2.5 marks. □

7. (Final Exam May 2023 Sem, Q1(b), 5 marks) Given the training data in Table 1.1 and the new data in Table 1.2, write down the formula for standardisation and then apply the standardisation to the new data in Table 1.2.

Obs.	x_1	x_2
1	3.05	27
2	2.76	64
3	4.41	77
4	3.63	16
5	6.73	53
6	3.80	29
7	2.33	49
8	3.66	72

Table 1.1: Training Data

Obs.	x_1	x_2
1	9.60	62
2	3.83	35
3	5.56	105
4	2.88	71
5	5.80	74

Table 1.2: New Data

Solution.

The formulas for min-max scaling on the two columns are

$$s_1(x_1) = \frac{x_1 - 3.79625}{1.352795}$$
$$s_2(x_2) = \frac{x_2 - 48.375}{22.43682}$$

.....[2 marks]

After the standardisation, the new table (Table 1.2) becomes

Obs.	x_1'	x_2'
1	4.29019056	0.6072609
2	0.02494834	-0.5961184
3	1.30378180	2.5237538
4	-0.67730125	1.0083873
5	1.48119222	1.1420961

.....[3 marks]

□

Modelling

8. (Final Exam Jan 2021 Sem, 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] □

9. (Final Exam Jan 2022 Sem, 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] □

10. (Final Exam May 2023 Sem, Q1(a)) Assuming the inputs of a data are all numeric and its output is binary.

    (a) State an example of supervised learning model for the data which is a parametric discriminative model. (2 marks)

    Solution. logistic regression model (other choices: linear SVM, ANN) . [2 marks] □

    (b) State two examples of supervised learning models for the data which are nonparametric discriminative models. (3 marks)

    Solution. kNN model and decision tree model (alternatives: Random Forest, nonlinear/kernel SVM) ..... [$1.5 \times 2 = 3$ marks] □

11. 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. □

12. 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 Malaysia
    $p = 3$; 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

13. (Final Exam Jan 2023 Sem, Q1(d)) Given the predictions from a regressor in Table 1.3.

X	Actual Y	Prediction $\hat{Y}$
38.62	36.58	25.95
24.14	6.10	12.83
28.97	25.60	17.21
38.62	21.34	25.95
25.75	9.75	14.29
38.62	28.04	25.95
32.19	14.63	20.13
30.58	14.02	18.67
11.27	6.71	1.17
32.19	19.51	20.13

Table 1.3: Comparison of actual output against predicted output.

Calculate

(a) the mean squared error; (3 marks)

(b) the coefficient of determination, R^2. (3 marks)

Solution. (a) $$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \frac{(36.58 - 25.95)^2 + \cdots + (19.51 - 20.13)^2}{10}$$
$$= \frac{357.8622}{10} = 35.78622$$
.....[3 marks]

Average: 2.28 / 3 marks in Jan 2023; 20% below 1.5 marks.

(b) $R^2 = 1 - \dfrac{\sum_{i=1}^{n}(y_i - \hat{y}_i))^2}{\sum_{i=1}^{n}(y_i - \frac{1}{n}\sum_{i=1}^{n} y_i)^2} = 1 - \dfrac{357.8622}{881.0278} = 0.5938128$ .....[3 marks]

Average: 1.99 / 3 marks in Jan 2023; 29% below 1.5 marks.

Remark: Need to be good with calculators.

□

14. 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. □

15. (Final Exam Jan 2022 Sem, 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] □

16. (Final Exam Jan 2023 Sem, Q1(e)) Given the confusion matrix in Table 1.4 with "No" as positive.

	Actual	
Prediction	No	Yes
No	31	14
Yes	6	29

Table 1.4: Confusion matrix.

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

(a) Sensitivity (2 marks)

Solution. $\text{TPR} = \dfrac{\text{TP}}{\text{FN} + \text{TP}} = \dfrac{31}{31+6} = 0.8378378$ ..... [2 marks]

Average: 1.89 / 2 marks in Jan 2023; 6% below 1 mark. □

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

Solution. $\text{NPV} = \dfrac{\text{TN}}{\text{TN} + \text{FN}} = \dfrac{29}{29+6} = 0.8285714$ ..... [2 marks]

Average: 1.85 / 3 marks in Jan 2023; 4% below 1 mark. □

(c) Accuracy and Kappa Statistic

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

where $\text{RandomAccuracy} = \frac{(\text{FN}+\text{TP})\times(\text{FP}+\text{TP})+(\text{TN}+\text{FP})\times(\text{TN}+\text{FN})}{(\text{Total Number of Test Data})^2}$. (3 marks)

Solution. $\text{Accuracy} = \dfrac{31+29}{31+29+6+14} = 0.75$ .....[1 mark]

$\text{RandomAccuracy} = \dfrac{(31+6)(31+14)+(29+6)(29+14)}{80^2} = 0.4953125$

.....[1 mark]

$\text{Kappa statistic} = \dfrac{0.75 - 0.4953125}{1 - 0.4953125} = 0.504644$ .....[1 mark]

Average: 2.82 / 3 marks in Jan 2023; 3% below 1.5 marks. □

17. (Final Exam May 2023 Sem, Q1(c)) Retaining customers has become a top problem for financial institutions in today's cutthroat banking sector. Customer churn, the phenomena where customers switch banks, has a big impact on banks' profitability, market share, and long-term viability. Suppose the logistic regression model is trained on a bank customer churn dataset with the response 0 and 1 which represent "not a churn" and "is a churn" respectively and the confusion matrix on the training data is given in Table 1.3.

	Actual	
Prediction	0	1
0	7876	1830
1	87	207

Table 1.3: Confusion matrix on the training data. 0 is positive.

(a) Find the accuracy and kappa statistic for the data characterised by Table 1.3.

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

where $\text{RandomAccuracy} = \frac{(\text{TP}+\text{FN})\times(\text{TP}+\text{FP})+(\text{TN}+\text{FP})\times(\text{TN}+\text{FN})}{(\text{Total Number of Data})^2}$. (5 marks)

Solution. Accuracy $= \dfrac{7876+207}{7876+1830+87+207} = \dfrac{8083}{10000} = 0.8083$ ..... [2 marks]

$$\text{RandomAccuracy} = \frac{(7876+87)(7876+1830)+(207+1830)(207+87)}{10000^2}$$
$$= 0.77887756 \quad \text{[1 mark]}$$

$$\text{Kappa statistic} = \frac{0.8083-0.77887756}{1-0.77887756} = 0.133059 \quad \text{..... [2 marks]}$$ □

(b) The F_1 score is defined as the harmonic mean of the sensitivity (TPR) and positive predictive value (PPV), i.e. $F_1 = \left[\dfrac{\frac{1}{\text{TPR}}+\frac{1}{\text{PPV}}}{2}\right]^{-1}$. Calculate the sensitivity, the PPV and the F_1 score.

(3 marks)

Solution.
- sensitivity (TPR) $= \dfrac{7876}{7876+87} = 0.989074$ ..... [1 mark]
- PPV $= \dfrac{7876}{7876+1830} = 0.811457$ ..... [1 mark]
- $F_1 = \dfrac{2}{\frac{1}{0.989074}+\frac{1}{0.811457}} = 0.891505$ ..... [1 mark]

□

(c) Are the high accuracy in part (i) and high F_1 score in part (ii) sufficient to support that the statement that the logistic regression model is good at capturing the churns? Justify your answer. (2 marks)

Solution. No. ..... [1 mark]
Since the negative response 1 represents "is a churn", the accuracy and F_1 score only indicates the estimate of the majority (not a churn) in an imbalanced data is fine but the specificity is very low indicating the predictive model has bad recall for a churn. [1 mark] □

Deployment

18. (Final Jan 2021 Sem, 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] □

19. (Final Jan 2024 Sem, Q1(d)) Write down **four** major categories of **unsupervised learning methods** and provide a concrete application for each category. (6 marks)

Solution. The four major categories of unsupervised learning methods are

(a) Dimensionality reduction ..... [0.5 mark]

(b) Cluster analysis ..... [0.5 mark]

(c) Finding association rules ..... [0.5 mark]

(d) Anomaly detection / Visualisation / feature extraction, etc. ..... [0.5 mark]

Biologists employ dimensionality reduction to determine the relations between various living species based on the (complete or fragment of) genetic information. ..... [1 mark]

Clustering is the process of grouping the given data into different clusters or groups. E-commerce websites like Amazon use clustering algorithms to implement the user-specific recommendation system. ..... [1 mark]

(a) Market segmentation divides the consumers of the market into some groups. In a group, consumers will be similar to each other based on some predefined set of characteristics. If two customers are not similar based on these characteristics, they are in different groups. Companies use this clustered data and the features of the customers to decide their market strategies, like which group of customer they should target or which group of customers needs more advertising etc. etc.

(b) There are millions of people in social networking websites and analysing their behaviours sounds really fun. Clustering plays the role here. This idea of social networks analysis can be extended to real life social scenarios.

(c) Search engines and many other websites use clustering to group similar web pages, videos, songs etc. and improve results for their users.

Finding Association Rules is the process of finding associations between different parameters in the available data. It discovers the probability of the co-occurrence of items in a collection, such as people that buy X also tend to buy Y. It is used in supermarket item placement (association rules) and logistics. ..... [1 mark]

Anomaly detection: The identification of rare items, events or observations which brings suspicions by differing significantly from the normal data. ..... [1 mark]

Average: 1.11 / 6 marks in Jan 2024; 90.91% below 3 marks. □