匠人学院™
jiangren.com.au

DATA7001

Midterm 考前总结与试题分析

知识点框架

- Module 1: Design Thinking
- Module 2: Getting the Data I Need
- Module 3: Is my Data Fit for Use
- Module 4: Making Data Confess
- Module 5: Story Telling

Module 1 – Design thinking 概念

匠人学院
jiangren.com.au

Definition of Design Thinking: The **design thinking ideology** asserts that a **hands-on, user-centric approach** to problem solving can lead to innovation, and innovation can lead to differentiation and a competitive advantage.

Human Centered Design:

1. **Inspiration Phase:** learn directly from the people you`re designing for.
2. **Ideation Phase:** make sense of what you learned and prototype possible solutions.
3. **Implementation Phase:** bring your solution to life and to market.

Advantages of design thinking: --模拟题1

1. A user-centered process. （以用户为中心）
2. Leverages collective expertise and establishes a shared language amongst your team. （利用集体专业知识，增强团队协作能力）
3. Encourages innovation by exploring multiple avenues for the same problem. （探索多种解决问题的途径，鼓励创新）

Design Thinking Process (6 phases):

1. Understand: **Empathize**: Research about what your users do, say, think, and feel. **Define**: Combine all your research and observe where your users' problems exist, begin to highlight opportunities for innovation.
2. Explore: **ideate**: Brainstorm a range of crazy, creative ideas that address the unmet user needs identified in the define phase. **Prototype**: understand what components of your ideas work, and which do not.
3. Materialize: **test**: Return to your users for feedback. **Implement**: Put the vision into effect. Ensure that your solution is materialized and touches the lives of your end users.

Module 1 – Design thinking 实例

真题1：

Give two examples of human centered design thinking.

1. Beer and diaper.
    a) Understand: Daddies are always the one who come to buy diapers for the new born baby.
    b) Explore: Think about what daddies would also like to buy when they are tired with taking care of a baby? Beers!
    c) Materialize: Shop keeper put the beers together with diapers.
    d) Benefits: Customers bought what they want, and shop keeper had higher sales.
2. Airbnb.
    a) Understand: Travelers want cheap accommodation, residents have spare rooms.
    b) Explore: What if residents provide their spare rooms to travelers?
    c) Materialize: Create a platform for this need. Airbnb!
    d) Benefits: Travelers enjoy cheap accommodation while residents have income.

Module 1 – Design thinking

Characteristics of data：——模拟题3

- Volume：Terabytes，petabytes．Youtube．
- Velocity：Realtime，live stream．Taxi data in big city．
- Variety：Structured，Unstructured，multimedia．Facebook，Twitter．

Three generations of data：

- **Transactional**：directly derived as **a result of transactions**．**A reference data** describing the **time**，**place**，**prices**，**payment methods**，**discount values**，and **quantities** related to that particular transaction．It involves activities such as purchases，requests，insurance claims，deposits，withdraws．
- **Interactional**：the **real—time capture** of each procedural decision，complete with **data** and **time stamp**．Data gathered through **interaction with people**，including focus groups，questionnaires，and customer surveys．
- **Sensory**： or **machination data**，In addition to **large volumes** and **streaming nature**，such data typically have **high level of redundancy and low value density**．often found in sensor networks，GPS and RFID applications，vehicle on—board devices and medical monitoring devices．

Module 1 – Design thinking

(R)DBMS： —— 模拟题4

Advantages：

- Separation of data from applications (Millennium Bug)
- Separation of physical structures and logical structures.
- Relational model and theory.
- Non–procedural query language.
- Concurrency control and recovery (manage multi–user conflict).
- High performance query processing.

Disadvantage：

- Limited data types (Only structured).
- The Closed–World assumption.
- Extensions not well supported

Module 1 – Questions

QUESTION 1

Provide three benefits of approaching data science problem formulation through design thinking.

- A user–centered process.
- Leverages collective expertise and establishes a shared language amongst your team.
- Encourages innovation by exploring multiple avenues for the same problem.

Module 1 – Questions

匠人学院
jiangren.com.au

QUESTION 2

The data science process is a *sequence* of 5 steps

- ◯ True ⇦ 就是这节课的五个module
- ◯ False

Module 1 – Questions

匠人学院
jiangren.com.au

QUESTION 3

Three characteristics of big data are:

a. Volume, Variety, and Viscosity

b. Variety, Vanity, and Volume

c. Vastness, Variability, and Value

d. Volume, Variety, and Velocity

建议各准备一个例子

Module 1 – Questions

QUESTION 4

Which of the following statements is *not true* for relational database management system (RDBMS):

a. Mainly used for storing unstructured data ⟸ 只能存储 structured data

b. Separates data from applications

c. Separates physical and logical strucutres

d. Supports multi-user access

Module 2 – Getting the data I Need

Ethical and legal considerations:

1. Confidentiality, (business contract, medical prescription)
2. Consent, (articles, studies)
3. Potential Discrimination, (study on race and criminal rate)
4. Cultural diversity,
5. Ownership,
6. Commercialization.

Two principles of private data release:

1. The **privacy** of data owners is adequately protected.
2. The data is **useful** for its intended purpose to a desirable level.

Module 2 – Getting the data I Need

k–Anonymity：*k*–anonymity requires that each combination of quasi–identifiers (QI) is hidden in a group of at least size *k*.

Name	Age	Zipcode
Andy	20	1000
Bob	25	2000
Cathy	30	3000
Diane	35	4000

Adversary's knowledge:

Age	Zipcode	Grade
[20,29]	[1000,2000]	5
[20,29]	[1000,2000]	5
[30,39]	[2000,3000]	6
[30,39]	[2000,3000]	7

Other algorithms: 1. L-diversity: for a group there is at least L different sensitive attribute values (or combination), in addition to the basic k-anonymity algorithm. 2. Differential privacy

Module 2 – Getting the data I Need

Types of data:

1. **Structured** (RDBMS, SQL).
2. **Semi-Structured** (XML, <title> </title>).
3. **Unstructured** (Document, Short text).
4. **Spatial** (Location component, Vector Data, Raster Data).
5. **Time Series** (successive regularly spaced time intervals, Types: stock series & flow series).
6. **Graph** (include Nodes, Edges, Properties. Storage: Relational (SQL, limit: flexibility, scalability) & Key-Value (Neo4j, limit: integrity, adoption)).
7. **Multimedia** (text, images, graphics, audio and video).

Module 2 – Getting the data I Need

真题2：

Given two examples of analysis that will need two or more different types of data.

1. **Restaurant review website analysis.**

a) Customer's information including ID, sex, email, date of birth, etc. (structured)

b) Customer's reviews. (short text)

c) Restaurant's locations. (spatial data)

2. **Crop growth analysis.**

a) Crop growth situation including height, productivity, etc. (structured)

b) Temperature over time. (time series data, flow series)

c) Soil condition on different location. (spatial data)

Module 2 – Getting the data I Need

Meta—data：

1. Types of meta—data：

a) Descriptive. (author, title, etc. Enable searching)

b) Administrative. (file type, time created, access control, etc. Facilities management)

c) Structural. (table of content, frame index, version info, etc. Facilities navigation)

2. Storage types：

a) Embedded storage：Store metadata within the data file. markups, file header, or folders. Hard to disrupt the connection between data and metadata.

b) Centralized storage：Store in a centralized repository, such as searchable indices and databases. Easy to automate and standardize.

Module 2 – Getting the data I Need

Types of Data Sampling:

Simple Random Sampling:

- Each item has an **equal chance of appearing** in the sample.
- **Example**: A teachers puts students' names in a hat and chooses without looking to get a sample of students.

Weighted Random Sampling:

- Each item has a **weight**. The weight should capture data features of **particular interest**.
- Appears in sample proportional to weight.

Stratified Sampling:

- Distinct groups (strata) present in data.
- Maintain **representation** of all groups in the sample.

Ways of Data Sampling:

Sampling Without Replacement (WOR): Each time we add an item to the sample, it is excluded from being added again. No item is duplicated in the sample. Sampled items are **DEPENDENT.**

Sampling With Replacement (WR): Each time we add an item to the sample, it is NOT excluded from being added again. Items could be duplicated in the sample. Sampled items are **INDEPENENT.**

Reason: ONLY consider WR for its **Simplicity** and **Unbiasedness**.

Module 2 – Questions

QUESTION 5

Describe in detail how to take a stratified random sample of total size 12 from a set of 120 observations with three strata (20 observations in the first stratum, 30 observations in the second stratum and 70 in the other stratum).

$$12 \times \frac{20}{120} = 2, \quad 12 \times \frac{30}{120} = 3, \quad 12 \times \frac{70}{120} = 7$$

Module 2 – Questions

QUESTION 6

The following data has 2-anonymity with respect to the attributes 'Age', 'Gender', and 'State'. Explain how this data is still vulnerable. Suggest one alternative technique for data anonymization.

Age	Gender	State	Disease
[20 – 30]	F	NSW	Cancer
[20 – 30]	F	QLD	Viral Infection
[20 – 30]	F	NSW	TB
[10 – 20]	M	VIC	No illness
[10 – 20]	F	QLD	Heart
[10 – 20]	M	VIC	TB
[20 – 30]	F	QLD	Cancer
[10 – 20]	F	QLD	Heart

If the name of a female from QLD, aged between [10–20] is known, then we know she has heart disease.

Alternative: use 2–diversity to combine two sensitive attribute value.

For example: instead writing "Heart", we can write ["TB, Heart"].

5 minutes break

Module 3 – Is my Data Fit for Use

Data Quality Definition:

1. Degree to which data can be used for its **intended purpose**.
2. Degree to which data accurately **represent the real-world**.

Dimensions:

1. **Completeness** (Null value in a table, time period missing in climate dataset).
2. **Accuracy** (Wrong number in age, height, phone #, etc. 0483-656-100 vs 0483-656-110).
3. **Consistency** (Date, Name, Value format. 2020/05/01 vs May. 01. 2020).
4. **Freshness** (old phone number, change of name).

Module 3 – Is my Data Fit for Use

Variability in Data:

1. Group/Systematic Variability
   - Relationships between input/output variables,
   - can be represented by predictive model.
2. Natural Variability
   - intrinsic randomness,
   - same input and condition can have different output
3. Error
   - due to Sampling, Measurement, error in data collection methodology.
   - bias: "decimal data should round down but actually round up",
   - imprecision: "rounding, collect time in millisecond, but mechanism can measure up to second."
4. Residual Variability
   - Variability not accounted for

Module 3 – Is my Data Fit for Use

Variable Types:

1. Quantitative:

- Continuous: time, velocity.
- Discrete: Birth year, number of customers.

2. Qualitative:

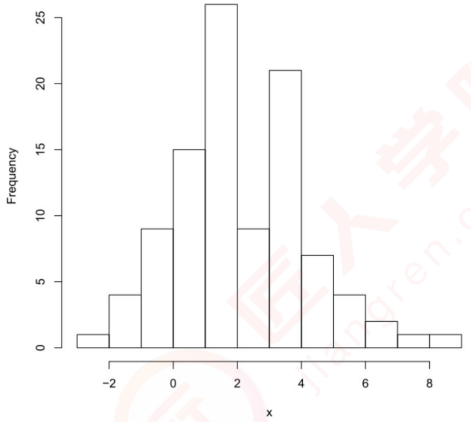
- Categorical: male/female, cities.
- Nominal: colors, animals. (no order)
- Ordinal: gold/silver/bronze, large/medium/small. (arbitrary/natural order)

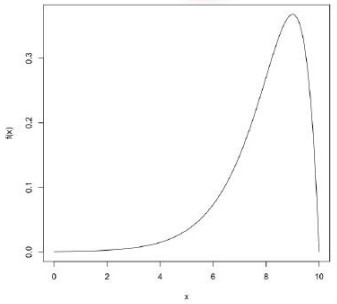
Module 3 – Is my Data Fit for Use

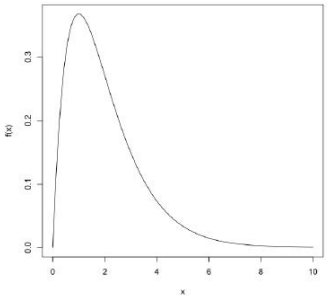
Visualizing Data

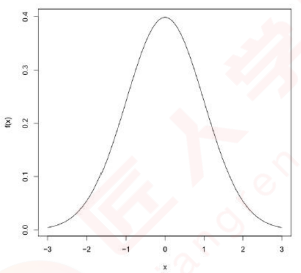
Basic visualisation techniques:

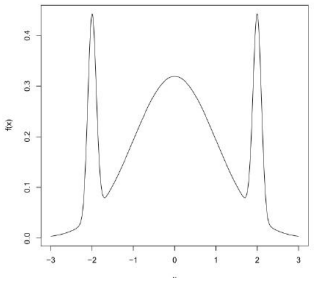
1. Univariate
   - Empirical CDF plot,
   - histogram (depend on choice of beam size),
   - box plot (heuristic decision about whether point should be **unusual**)
2. Bivariate
   - time series plot, scatter plot
3. **Symmetric** vs **left/right-Skewed** (the direction tail points)
4. **Unimodal** vs **Multimodal**











Module 3 – Is my Data Fit for Use

Rubin's missing data mechanisms

1. Missing Completely at Random (MCAR):

- Data missing at random, no relation with anything else.
- Does not introduce any bias.
- Complete case analysis.

2. Missing at Random (MAR):

- the chance of getting missing point in a column depend of the other value on that row, on other completely observed variables.
- Missing categorical data: add an extra category to indicate missingness.
- Example: Male are more likely to tell their weight than female.

3. Missing not at Random (MNAR):

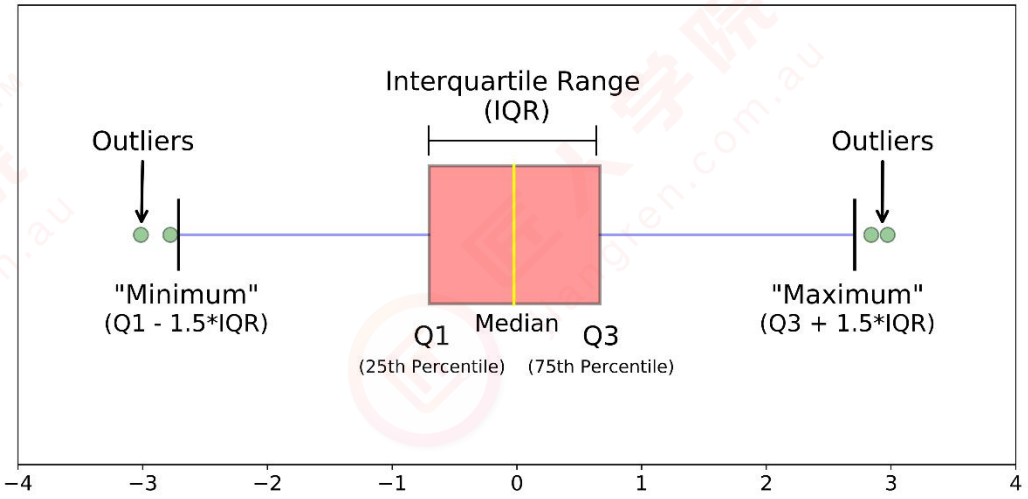
- cannot tell the chance of missing points base on the data you've got.
- Reason: The chance of the value is missing depends on the value of missing thing itself or depends on other valuable not measured.
- Example: Sensor returns error when temperature goes above 40 degree.

Module 3 – Is my Data Fit for Use

匠人学院
jiangren.com.au

Unusual Data

1. **Outliers**:
   - Deviate significantly from statistical model assumptions.
   - **Erroneous Data**: Height in metres vs millimetres
   - **Natural Outliers**: May indicate inappropriate model assumptions.
2. **Influential observations**:
   - inclusion/exclusion significantly affects statistical analyses
   - Influential observations may or may not be outliers.



Basic Error Model Assumption: Random variability (noise) is normally distributed.

Basic Outlier Detection:

1. For normally distributed data, slightly less than 1% of probability density lies outside the range: (Q1−1.5 × IQR, Q3+1.5 × IQR)
2. A frequently applied rule is to flag any observations outside this range as outliers.

** Q1: First Quartile; Q3: Third Quartile; IQR: Inter−Quartile Range

Module 3 – Is my Data Fit for Use

匠人学院
jiangren.com.au

Data Cleaning

1. Cleaning from rules:

- adult heights less than 1m and more than 2.5m tall, call them missing value.
- phone numbers have amount of digits not equal to 10 are missing values.

2. Cleaning from filter:

- subset to certain amount of years.
- subset customers of age between 20 to 50 years old for behavior analysis.

3. Cleaning from source:

- spreadsheet copied too many times, find out clean from the reliable source.

Module 3 – Is my Data Fit for Use

Data Imputation

What is data imputation:

NOT to recreate missing data, rather, to obtain statistically valid inferences from incomplete data.

Why use data imputation:

1. after eliminating cases with missing value, the data may not be enough to perform the analysis.
2. the analysis might run but the results may not be statistically significant because of the small amount of input data.
3. results may be misleading if the cases you analyze are not a random sample of all cases.

Module 3 – Is my Data Fit for Use

Data Imputation

Types of imputation:

1. Mean Imputation:

- Process: replace each missing value by the mean of the non-missing values.
- Flaw: Severely distort the distribution of the variable: Variability in data and estimates "too low"; Weakens relationships between variables; – Biases estimates other than the mean

2. Simple Random Imputation:

- Process: Fill in missing value with a simple random sample from of the non-missing values.
- Flaw: Ignores information contained in other variables: Weakens relationships between variables; Variability in estimates "too low".

3. (Deterministic) Regression Imputation:

- Process: Fit a regression model to non-missing values and replace missing value by its predicted value.
- Flaw: Distorts distribution of data: Strengthens relationships between predictor and response; Variability in data and estimates ``too low".

4. (Stochastic) Regression Imputation: (right direction)

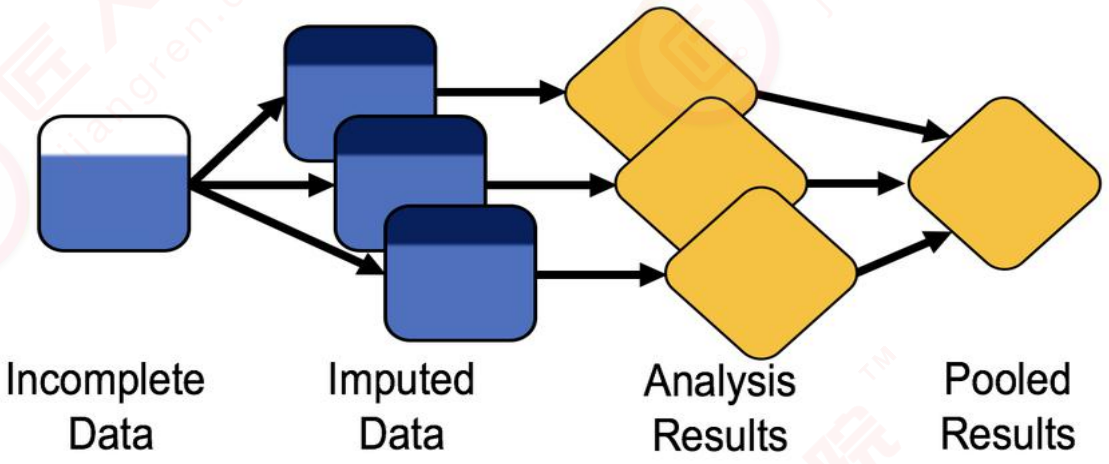
- Process: Fit a regression model to non-missing values and replace missing value by its predicted value 'plus noise'. 'Noise' distribution estimated from non-missing residuals. Unbiassed estimates of regression coefficients and correlation under MAR.
- Flaw: Variability in estimates ``too low"

Module 3 – Is my Data Fit for Use

Data Imputation

Multiple Imputation:

1. Address the problem of variability in estimates ``too low''.
2. Creates m complete datasets (missing data drawn from distribution).
3. Each dataset analyzed using standard techniques.
4. The m results are pooled into a final point estimate, together with an estimate of its variance.



Module 3 – Is my Data Fit for Use

Data Integration

1. Schema Level:

- Structural differences

  Example: different table col names, build global schema

- Semantic differences

  Example: two time serial with different template

2. Instance Level:

- Field linking/matching

  Example: identify same cols

- Record linking

  Example: identify same rows

- Entity linking

  Example: identify same pseudo-identifiers, Tiger Woods, Tiger Balm

Module 3 – Is my Data Fit for Use

Edit (Levenshtein) Distance:

One distance: delete, insert or replace one character.

Advantage:

1. Two strings are considered as the same if their ED is less than a **pre-defined threshold**.
2. Suitable for **common typing mistakes**.

Disadvantage:

1. May be costly operation for large strings.
2. Problematic for specific domains:

- AT&T Corporation vs AT&T Corp
- IBM Corporation vs AT&T Corporation.
- 500 vs {499, 800}

X=dva y=dave
find Levenstein distance using dynamic programming.

		y0	y1	y2	y3	y4
			d	**a**	**v**	**e**
x0		0	1	2	3	4
x1	**d**	1	0	1		
x2	**v**	2				
x3	**a**	3				

		y0	y1	y2	y3	y4
			d	**a**	**v**	**e**
x0		0	1	2	3	4
x1	**d**	1	0	1	2	3
x2	**v**	2	1	1	1	2
x3	**a**	3	2	1	2	2

x = d – v a
| | | |
y = d a v e

substitute a with e
insert a (after d)

Module 3 – Is my Data Fit for Use

匠人学院
jiangren.com.au

Edit Distance Practice

1. shoe → show(ED = 1)
- replace "e" with "w".

2. broad → brand (ED = 2)
- delete "o"
- insert "n"

3. kitten → sitting (ED = 3)
- replace "k" with "s"
- replace "e" with "i"
- delete "g"

Module 3 – Questions

QUESTION 7

With the help of an example, explain the difference between cleaning from rules and cleaning from filter.

1. **Cleaning from rules**:
   - adult heights less than 1m and more than 2.5m tall, call them missing value.
   - phone numbers have amount of digits not equal to 10 are missing values.
2. **Cleaning from filter**:
   - subset to certain amount of years.
   - subset customers of age between 20 to 50 years old for behavior analysis.

Module 3 – Questions

QUESTION 8

Describe how deterministic regression imputation works and describe two problems arise from using deterministic imputation?

(Deterministic) Regression Imputation:

- Process: Fit a regression model to non–missing values and replace missing value by its predicted value.
- Flaw: Distorts distribution of data: Strengthens relationships between predictor and response; Variability in data and estimates ``too low''.

Module 4 – Making Data Confess

Multi–dimensional model:

- Consists of a **Fact Table (result/data we want)** and **multiple Dimension Tables (factors).**
- Each **key** is a **dimension**. Dimensions have **hierarchical organization**.
- Dimensions are organized by **dimension tables**.

fact table

Key			Facts
Day	Product	Store	Sales (AUD)
9.2.04	Milk	Toowong	3412
10.2.04	Milk	Toowong	2918
9.2.04	Bread	Toowong	2918
10.2.04	Bread	Toowong	3445
9.2.04	Milk	Sunnybank	5440
10.2.04	Milk	Sunnybank	4992
9.2.04	Bread	Sunnybank	2918
10.2.04	Bread	Sunnybank	3067

dimension tables

Day	Month	Qtr	Year
9.2.04	Feb	1	2004
10.2.04	Feb	1	2004

Store	District	Region
Toowong	North	Brisbane
Sunnybank	South	Brisbane

Product	Kind	Type	Class
Milk	Dairy	Perishable	Food
Bread	Bakery	Perishable	Food

Module 4 – Making Data Confess

Machine Learning

Purpose: find a functional relationship between the input and the output, without finding out the distribution generating the data.

Approaches:

1. Supervised Learning:

- collect a set of input–output pairs (x1, y1), (x2, y2)...,(xn, yn), and use it to fit a model.

  Algorithm: Linear regression

- Classification: Y is qualitative (categorical, Example: handwritten digit recognition).

  Algorithm: Logistic regression, K–nearest neighbor, Linear/Quadratic discriminant analysis, Neural Networks, SVM.

- Regression: Y is quantitative (numerical, Example: stock price prediction).

2. Semi–supervised learning: collect a set of input–output pairs and a set of unlabeled inputs x'1, x'2, ..., x'm, and use them to fit a model.

- Useful when it is expensive to label the inputs.
- Example: in handwritten digit recognition, we use both labelled and unlabeled digit images

3. Unsupervised Learning: only collect the inputs, but not the outputs.

- Example: Clustering, Density estimation, Principal component analysis (PCA).

  Algorithm: K–means

Module 4 – Making Data Confess

匠人学院
jiangren.com.au

Simple Linear Regression:

1. Only a single predictor, The output is assumed to be a linear function of the predictor.
2. $Y = f(X) + \varepsilon$, where $f(X) = \beta_0 + \beta_1 X$.
3. Errors **independent** and **identically distributed (iid)** according to a Normal distribution with mean zero and constant variance .
4. Estimating the standard deviation of the error term using **residual standard error**, which is the square root of SSE/(n−2).

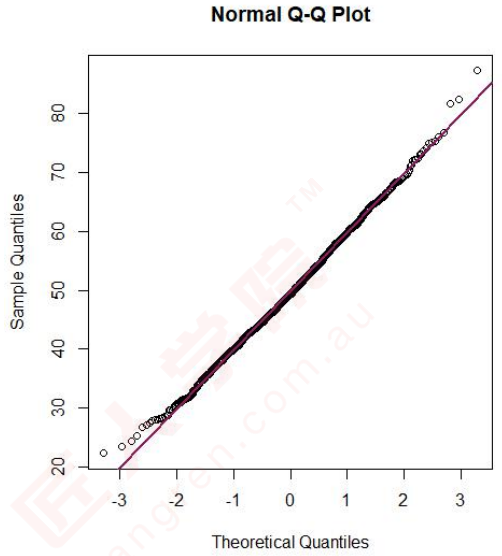
Goodness of Fit:

R^2 statistic. the proportion of variance explained by the model (0~1), explain how well the model fit the training data.

Model Validation:

1. Plotting the residuals themselves.
2. Constructing a quantile–quantile (qq) plot.
3. Examining summary statistics of the residuals.

$$N = \frac{X - \mu}{\sigma}$$

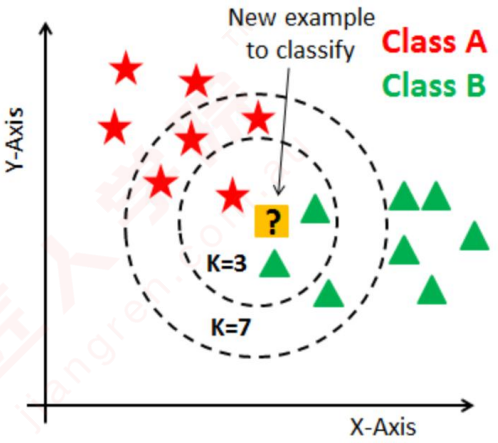


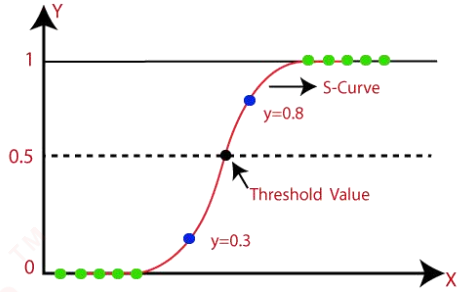
Module 4 – Making Data Confess

Classification:

we want to find out the **relationship** between the **predictors** and a **categorical output**.

1. **Bayes optimal classifier:**
   - classifier with minimum classification error is the highest probability
   - error rate is 1 – classifier,
   - **flaw**: we can't compute it because we don't know the true class distribution
2. **KNN (classification and regression):**
   - Find the k nearest neighbors of x, assign x to the most commonly–occurring class of these neighbors
   - If the predictors are on the same scale. normalize the predictors first.
   - 1–NN is self doesn't generalize on new examples
   - **flaw**: In **HIGH DIMENSIONAL DATA**, kNN is computationally very demanding.
3. **Logistic regression**: Parameters found by **Maximum Likelihood**
4. **Support vector machines (SVM)**: When two classes in a dataset are linearly separable, we have many separating hyperplanes, we prefer **the hyperplane farthest away from the nearest example**

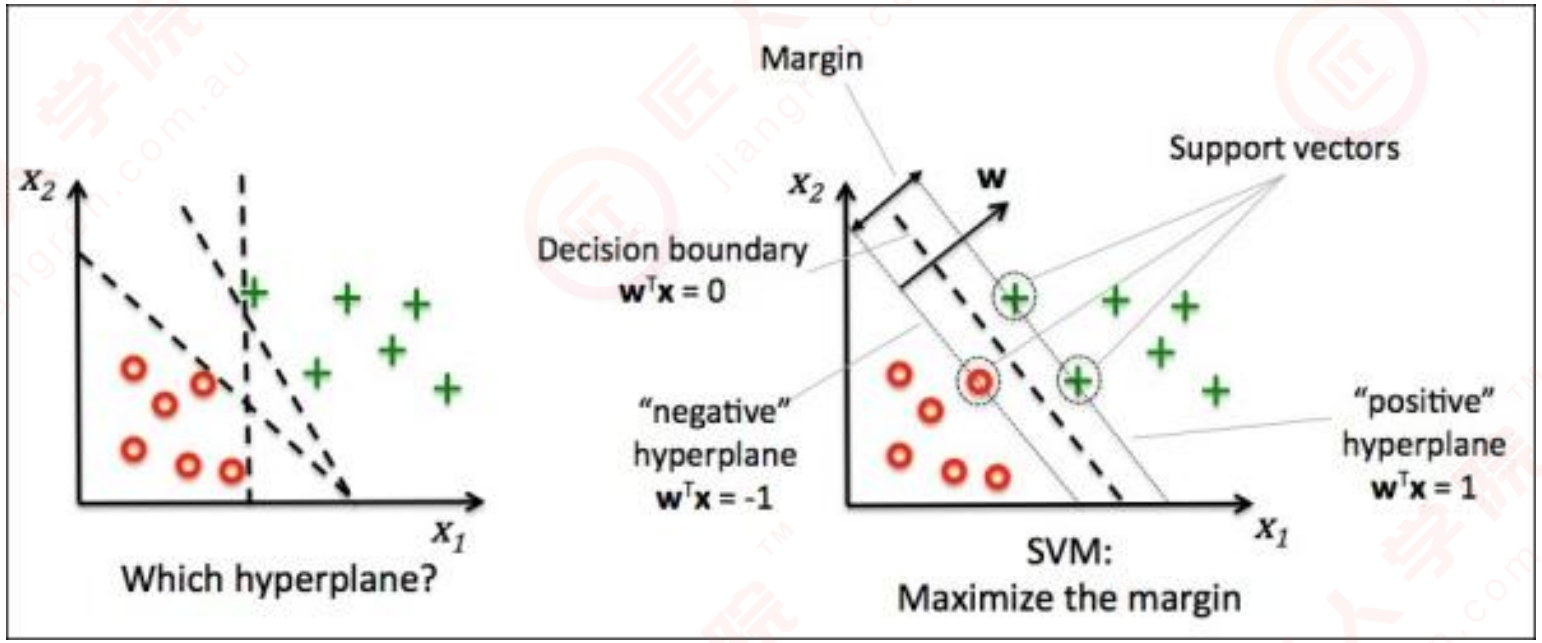




Module 4 – Making Data Confess

Support Vector Machines (SVMs):

1. Consider a two–class dataset (red and blue).
2. When the dataset is linearly separable (that is, we can draw a hyperplane separating the two classes), we have many separating hyperplanes.
3. In SVM, we prefer the hyperplane farthest away from the nearest example.



Module 4 – Making Data Confess

Parametric Method:

1. Assume functional form, e.g. linear.
2. Model fitting / training through parameter fitting.
3. Flaw: Limited by the assumption on the functional form, but generally efficient.

Nonparametric Method:

1. No assumptions on functional form.
2. Functional form is often data-dependent (e.g. k-NN).
3. Flaw: Capable of learning very complex functional relationship, but often computationally expensive.

Module 4 – Making Data Confess

Confusion matrix

Error rate=(FP+FN)/(TP+TN+FP+FN);

Accuracy=(TP+TN)/(TP+TN+FP+FN);

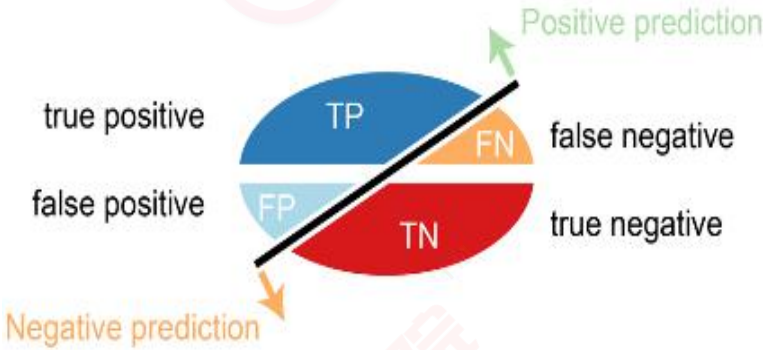
Sensitivity (Recall or True Positive Rate)=TP/(TP+FN);

Specificity (True Negative Rate) =TN/(TN+FP);

Precision=TP/(TP+FP)

False Positive Rate=FP/(TN+FP) =1− Specificity.

		True label	
		Positive	Negative
Predicted label	Positive	True Positive (TP)	False Positive (FP)
	Negative	False Negative (FN)	True Negative (TN)



Why not use:

Not perform well when estimating a very biased data like the probability of winning a lottery (0.0001 winning probability)

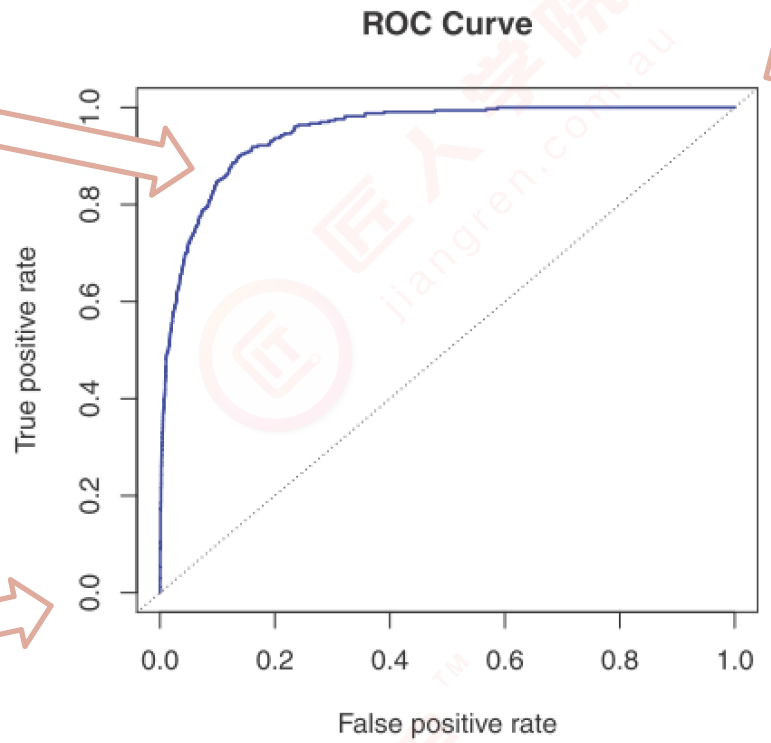
Module 4 – Making Data Confess

ROC：X~ False Positive Rate，Y~ True Positive Rate.

AUC is area under ROC，a quantitative measure of classifier performance.

Threshold = 0

Best Threshold

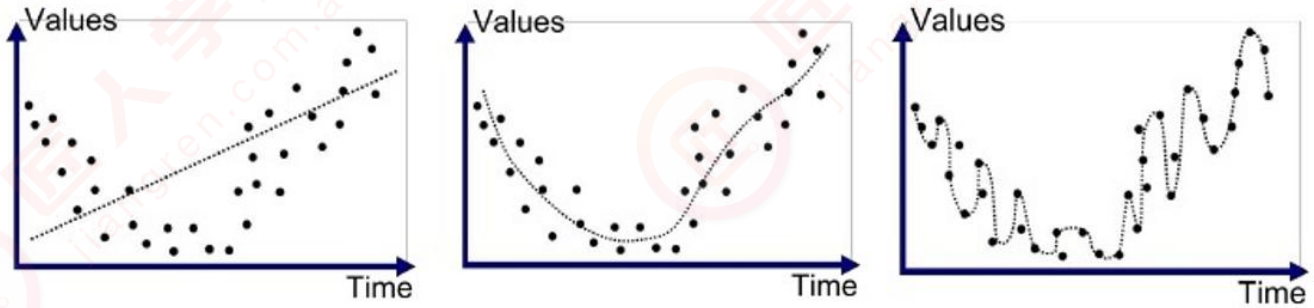


Threshold = 1

Module 4 – Making Data Confess

Overfitting:

Model learns not only the pattern but also the noise. An overfit model has "memorized" the training data, and will make larger errors on new data. In general, a model that achieves 0 training set error does not necessarily achieve 0 error on new examples.



Underfitted | Good Fit/Robust | Overfitted

Module 4 – Making Data Confess

Solution:

Lots of training data: Train–Validation Split

1. Split it into **a training set** and **a validation**,
2. Train candidate models on the training set.
3. Choose the one with best predictive performance on the validation set.
4. The validation set performance provides a much better estimate for a model's predictive performance on the test set.

Don't have lots data: Cross Validation (CV)

1. Create multiple training and test splits for the training set.
2. For each split: a. Fit a model on its training set. b. Compute the model's prediction error on the test set.
3. Compute the average error using the above estimates.

Module 4 – Making Data Confess

K folder–CV vs. Leave one out CV:

k–fold CV is has larger bias but smaller variance as compared to LOOCV.

Choosing between k–fold CV and LOOCV:

For large (>hundreds) datasets, k–fold CV is often recommended.

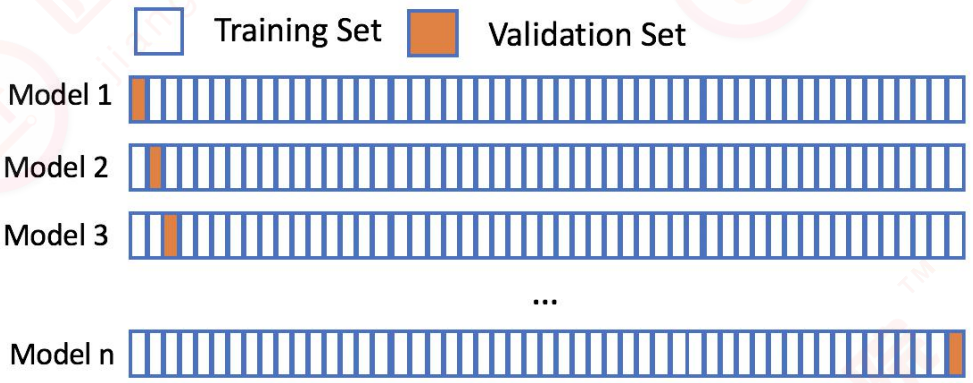
For small datasets, LOOCV may be a better option.

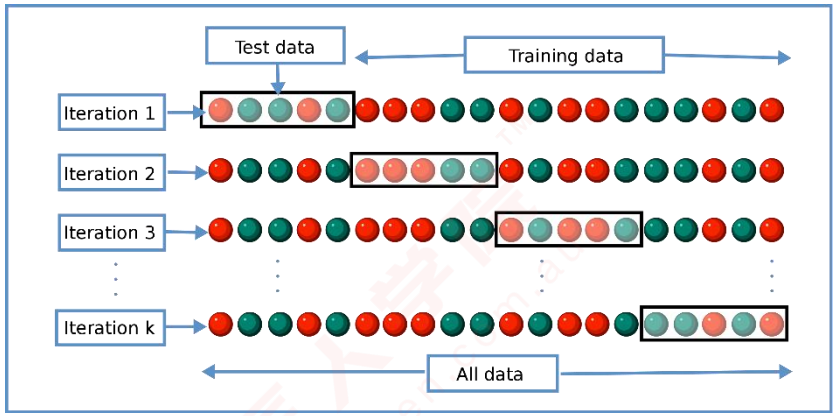
LOOCV:

$$\mathrm{CV}_{(n)} = \frac{1}{n}\sum_{i=1}^{n}\mathrm{MSE}_i.$$

K–fold:

$$\mathrm{CV}_{(k)} = \frac{1}{k}\sum_{i=1}^{k}\mathrm{MSE}_i.$$





Module 4 – Questions

QUESTION 9

Match the description to one of the following:

a. Supervised learning
b. Density estimation
c. Clustering

Unsupervised, Density estimation → - : Given a sample drawn from a Gaussian distribution, determine the parameters of the Gaussian distribution.

Supervised learning, Classification → - : Given a collection of emails labelled as SPAM or NOT SPAM, build a model to check whether an email is a spam or not.

Supervised learning, Linear Regression → - : Given a dataset containing the education level, gender, age, and salary for a group of people, build a model to predict the salary of a person given the person's education level, gender, and age.

Unsupervised, Clustering → - : Given a collection of articles about science and sports, divide them into groups about these respective topics.

Module 4 – Questions

QUESTION 10

For a linear regression model, what are the assumptions on the residuals? Why are these assumptions made?

Errors **independent** and **identically distributed (iid)** according to a **Normal distribution** with **mean zero** and **constant variance** .
Because linear regression model can only be applied when the residuals are normally distributed, so that we can construct a confidence interval for the predicted value. Mean zero implies there is no relationship between errors and the independent variables (no information leaked into the error term).

5 minutes break

Module 5 – Story Telling

匠人学院
jiangren.com.au

What is Story telling:

Who is your audience, What do they need to know, How can you effectively communicate.

Storytelling methods:

1. Change over time.
- Uses a chronology to illustrate a trend.

2. Drill down.
- Sets context so that your audience better understands what's going on in a particular category.

3. Zoom out.
- Describes how something your audience cares about relates to the bigger picture.

4. Contrast.
- Shows how two or more subjects differ.

5. Intersections.
- Highlights important shifts when one category overtakes another.

6. Factors.
- Explains a subject by dividing it into types or categories.

7. Outliers
- Shows anomalies or where things are exceptionally different.

Module 5 – Story Telling

匠人学院
jiangren.com.au

Story telling practice:

Assume that you have access to sale data including: property address (street, suburb, state), property feature such as no. of bedroom, lot size, car spaces etc.; sale history (data and price). Tell stories using different method respectively.

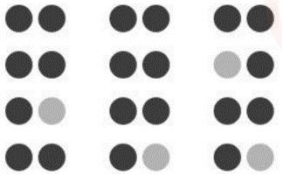
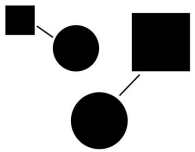
1. **Change over time:**
   - Show the average price change of the properties in each state over 10 years.
2. **Drill down:**
   - Show the average property price of each state in the whole country, and focus on one state with highest price, then the suburb and street with highest price.
3. **Zoom out:**
   - Show the street with highest average property price in the current month and zoom out to the whole state and year.
4. **Contrast:**
   - Compare the conditions (no. of bedroom, car spaces, etc.) of the house in same lot size with the highest average property price with the house with the lowest in the same suburb.
5. **Intersection:**
   - Show that in the past a larger lot space is the main indicator of a higher price, but now the location is a more important factor.
6. **Factors:**
   - Show how each property feature affect the property price.
7. **Outlier:**
   - Find out the property of extremely high price in a state and study why.

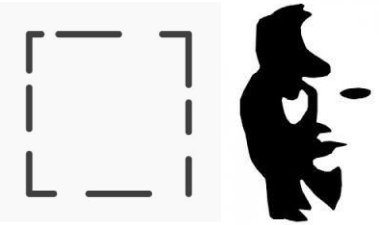
Module 5 – Story Telling

匠人学院
jiangren.com.au

Gestalt laws：

1. Proximity：相关的元素靠近，不相关的元素分开。
2. Similarity：相关的元素用相同的表现方法，不相关的元素用不同的表现方法。
3. Connectedness：相连的元素更相关。
4. Continuity：在一条直线或曲线上的元素比不在一条线上的更相关。
5. Closure：缺失元素是容易被补全的，不会引起错误脑补。
6. Enclosure：相关的元素置于框内，不相关的元素在框外。

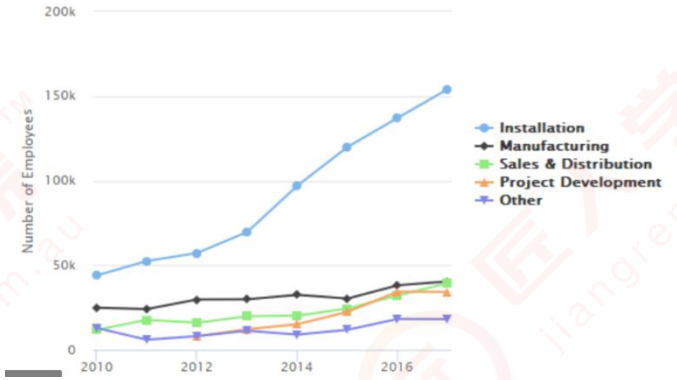






Module 5 – Story Telling

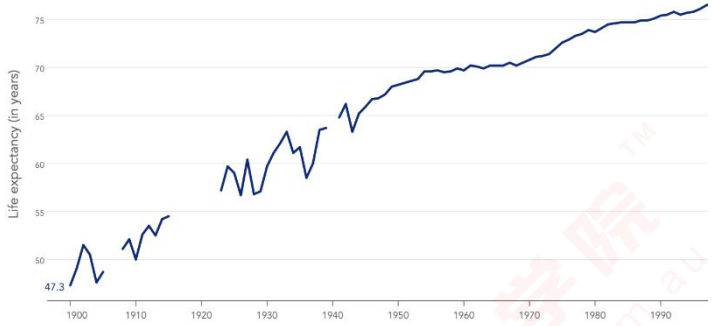
Which Gestalt's principles are used?

Proximity: data of same quarter are put together
Similarity: data of same product are in same color



Connectedness: points of same group are connected
Similarity: points of same group are in same color

Continuity: put bars in a rank from big to small shows a line, indicating there is a relationship.



Closure: although there are missing values in the graph, the overall pattern is clearly presented.

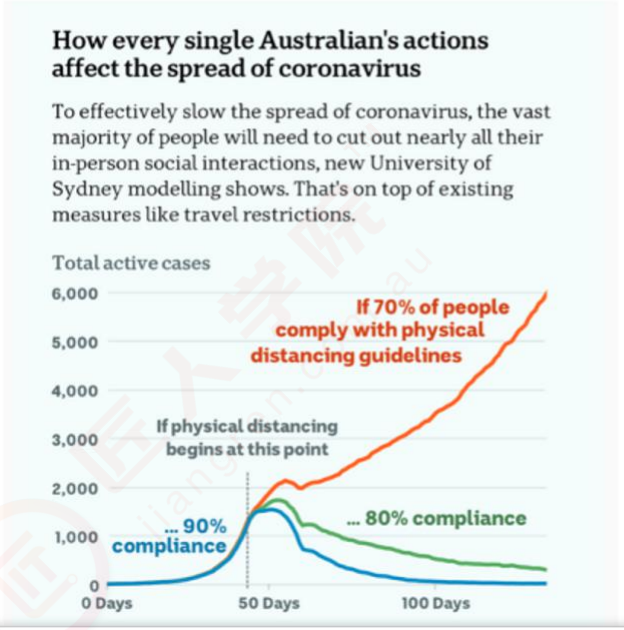
Module 5 – Questions

QUESTION 11

Recall the importance of understanding the context and consider the following visual. Does this visual persent an effective way of understanding the context ((Y/N)? If so ...

Who is the expected audience?

What action is expected from the audience?



Yes.

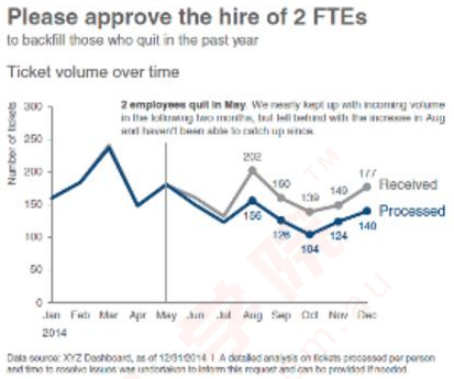
The expected audience is every Australian who is facing the risk of coronavirus.

The audience is expected to comply with social distancing guidelines in order to prevent the spread of coronavirus.

Module 5 – Questions

QUESTION 12

Recall Gestalt's principles of visual perception: Proximity, Similarity, Enclosure, Closure, Continuity, and Connectedness. Which of these are evident in the visual below? Briefly explain your answer.



Similarity: The data points of "Received" are grey, while those of "Processed" are blue.
Connectedness: The data points of the same group are connected with lines.

Module 5 – Questions

QUESTION 13

Assume that you have access to sale data including: property address (street, suburb, state), property features such as no. of bedrooms, lot size, car spaces etc; sale history (date and price). Describe how a data story using *Drill Down* can help understand the sales data.

Drill down:

- Show the average property price of each state in the whole country, and focus on one state with highest price, then the suburb, finally study on the street with highest price.

真题讲解

A famous scientist claims that he has invented a new COVID-19 test with true positive rate of 0.95, true negative rate of 0.96, and an accuracy of 0.94. However, a data scientist insists that the scientist's claim is incorrect. Which of the following statement is correct?

(i) The scientist is right.

(ii) The data scientist is right.

(iii) There is insufficient information to tell who is right.

Justify your answer.

		True label	
		Positive	Negative
Predicted label	Positive	True Positive (TP)	False Positive (FP)
	Negative	False Negative (FN)	True Negative (TN)

TP = 95	FP = 4
FN = 5	TN = 96

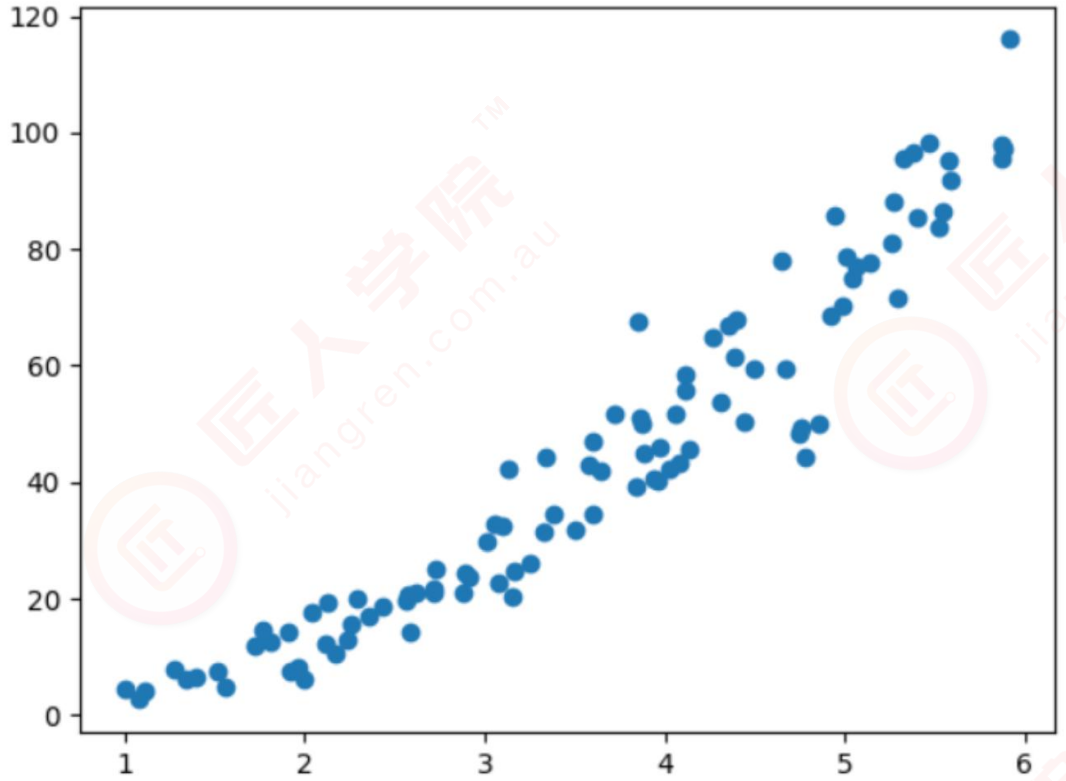
Accuracy=(TP+TN)/(TP+TN+FP+FN);

Sensitivity (Recall or True Positive Rate)=TP/(TP+FN);

Specificity (True Negative Rate) =TN/(TN+FP);

真题讲解

The following is a plot of the response y against a single predictor x. Is linear regression suitable for this dataset? Justify your answer.



1. Variance of residuals increases with x—axis;
2. The relation between x and y does not follow a straight line. (not linearly related)

真题讲解

Match the description to the right term.

In a survey, all respondents provide their income, but people with high income often do not provide their address. The address variable is

Answers: Missing completely at random

Missing not at random

Missing at Random

真题讲解

In a set of experiment records, some values are missing because the handwriting is sometimes not legible. There are no specific characters or digits that are consistently illegible though. The missing values are

Answers:
- Missing completely at random
- Missing not at random
- Missing at Random

真题讲解

A survey was conducted to investigate how factors like education, exercise and diet affects one's health in a city. The city has 3 towns A, B, C with population size 10k, 30k, and 40k respectively, and 500, 300, 1000 residents were interviewed from these towns respectively. A data analyst was asked to construct a subsample of size 160 and perform analysis on the subsample. Describe an appropriate way to construct such a subsample.

160 x 1/8 = 20
160 x 3/8 = 60
160 x 4/8 = 80

What if we need a subsample of size 1600?

真题讲解

Consider the following points on the real axis: -4, -2, -1, 1, 4, 8. Use k-means algorithm to cluster them into three groups, assuming that initially they are grouped as follows: {-4, 8}, {-2, 4}, {-1, 1}. Show your working. What are the centers of the final clusters?

Iteration 1:

- Centroids: 2, 1, 0
- Clusters: 2:{4, 8}, 1:{1}, 0:{−4, −2, −1}

Iteration 2:

- Centroids: 6, 1, −7/3
- Clusters: 6:{4, 8}, 1:{1}, −7/3{−4, −2, −1}

真题讲解

Bias due to sampling may arise when the sampled data does not reflect the actual population of interest. Given an example scenario which gives rise to bias due to sampling

1. Study workload and anxiety in university students.
- Only the students with low workload and anxiety level will participate the survey;
- Cause biased study result.

2. Study average people's opinion on technology with online survey.
- Only the ones who is pro—technology will participate the survey;
- Cause biased study result.

真题讲解

Match the description to the right term:

Given a sample drawn from an exponential distribution, determine the parameters of the exponential distribution.

Answers: Regression

Density Estimation

Classification

Clustering

真题讲解

Match the description to the right term:

Given a sample drawn from an exponential distribution, determine the parameters of the exponential distribution.

Answers: Regression

Density Estimation

Classification

Clustering

匠人学院
jiangren.com.au

Question time