# STAT5003

## Week 13 : Sample questions

Jaslene Lin

*The University of Sydney*

# Example of Multiple Choice Questions

1. What is a good practice to avoid overfitting?

```
- What is overfitting?

- If a model is overfitting, what does this imply about the bias-variance
trade-off in the model performance?
```

A. Use a complicated model that includes all possible interaction terms and higher order terms of the covariates.
B. Using a two-part loss function which includes a regulariser to penalize model complexity.
C. Using a good optimizer to minimize error on training data.
D. Use cross-validation to monitor the generalisation performance.

2. What of the following statements about the linear discriminant analysis are correct?

A. The assumptions in linear discriminant analysis are that the features in each of the groups is a sample from an arbitrary multivariate distribution, and all of the populations have the same mean vector.

B. The assumptions in linear discriminant analysis are that the features in each of the groups is a sample from a multivariate normal, and all of the populations have the same covariance matrix.

C. Linear discriminant analysis directly models the probability of the label given the features.

D. Linear discriminant analysis requires features to be numeric.

# Example Multiple Choice Questions

3. Which of the following are supervised learning techniques?
   A. K-means clusters
   B. Random Forest
   C. Linear Discriminant Analysis
   D. Density estimation

4. Which of the following are characteristics of a kernel function (as used in density estimation)?
   A. a frequency function from a histogram
   B. a symmetric function
   C. a function ranging from $-1$ to $1$
   D. a function that integrates to $1$ over its support

# Example Multiple Choice Questions

5. Which of the following practices may overestimate the test performance?
   A. Using PCA to construct new independent features from the original features. B. Imputing missing values using the mean calculated from the entire dataset
   C. Using 10-fold cross-validation to assess model performance. D. To address class imbalance, reporting Cohen's kappa from the test set as the overall performance metric.

# Example Multiple Choice Questions

6. Which of the following statements about the support vector machine are correct?

A. SVM aims to find the hyperplane that maximises the margin between different classes.
B. SVMs can only be used for linear classification problems.
C. Changes in the position of the support vectors will not impact the decision boundary.
D. Increasing the value of $C$ in the SVM's optimisation function will lead to an increase in the bias but the model will generalises better to the unseen data.

# Example Multiple Choice Questions

7. Which of the following are indirect measures of the test error?

   a. $C_p = \frac{1}{n}\left(\mathrm{RSS} + 2d\widehat{\sigma}^2\right)$

   b. $\mathrm{RSS} = \sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2$

   c. $\mathrm{BIC} = \frac{1}{n}\left(\mathrm{RSS} + \log(n)d\widehat{\sigma}^2\right)$

   d. $F_1 = \frac{2\times\mathrm{Precision}\times\mathrm{Recall}}{\mathrm{Precision}+\mathrm{Recall}}$

   where in the above:
   $\widehat{Y}_i$ is the predicted response for the $i$th observation; $d$ is the number of features in the model, not including the intercept;

# Solutions of MCQ

1. B; D
2. B; D
3. B; D
4. B; D
5. A; B
6. A; D
7. A; D

# Example of Extended Answer Questions

Your friend recently started an internship as a data analyst at a university student support unit. The team is interested in building a model to predict whether a student is at risk of dropping out, using available academic and engagement data collected from the learning management system.
Your friend explains the dataset consists of 5,000 observations, where each data point is represented as $(\mathbf{x}_i, y_i)$ for $i = 1, \ldots, 5000$. $\mathbf{x}_i$ contains 80 features, including number of logins, average time spent per week on the platform, assignment grades, and forum participation. $y_i = 1$ means the student eventually dropped out, while $y_i = 0$ means the student successfully completed the semester. Around 2% of students in the dataset dropped out.
Here's the modeling workflow your friend followed:

1. They noticed a few missing values in some features and filled them using the mean of each variable.
2. Then, they applied feature selection on the imputed full dataset, selecting the top 20 features most correlated with the target variable.
3. Next, they randomly split the data into 75% training and 25% testing sets.

4. Finally, they trained a SVM classifier and evaluated it using test set accuracy, achieving 92% accuracy.

Based on your understanding of statistical learning and good modeling practice, identify **three** problematic aspects of your friend's modeling workflow. For each issue: briefly explain why it is problematic, and suggest an alternative.

# Example of Extended Answer Questions

## Retirement problem

> You are planning your retirement and decide that you will retire with $1,000,000 invested in an index fund. During retirement you plan to withdraw $50,000 each year from your investment with the remaining money being invested in an index fund. Assume the index fund has an average return rate of 9% and a standard deviation of 15% (normally distributed). Assume you retire at 65 and will live until you are 100, and the CPI adjustment is 104% each year. Compute the probability that your investment will support your lifestyle until you die.

Your friend uses Monte Carlo to study your retirement plan and proposes the pseudo code to solve this problem. Evaluate this code and fill in <1> to <4>.

# Pseudo Code

```
# Set initial parameters
initial_investment ← 1000000
annual_withdrawal ← 50000
mean_return ← 0.09
sd_return ← 0.15
cpi ← 1.04
years ← 35
n_sim ← <1>
success_count ← 0

# Start Monte Carlo simulation
for (sim in 1:n_sim) {
  investment ← initial_investment
  withdrawal ← annual_withdrawal

  for (year in 1:years) {
    # Simulate annual return from normal distribution
    annual_return ← random value from <2>

    # Update investment value
    investment ← investment * (1 + annual_return)
    investment ← investment - withdrawal
```

Write down an expression for <1> to <4> respectively.

# Solution with explanations

## Solution

0. Initialize the situation
    a. retirement capital at $1,000,000.
    b. Initialize counter for number of years past 65.
1. Withdraw from the capital $50,000 multiplied by the CPI adjustment (104%) to the power of the number of years past 65.
    a. Check if the capital is positive
        - If not, stop the algorithm: the money is exhausted.
        - If yes, proceed to 2.
2. Multiply the capital by the return rate (`1 + rnorm(1, mean = 0.09, sd = 0.15)`)
3. Check if the capital is positive
    a. If not, stop the algorithm: the money is exhausted.

   b. If yes, increment the year counter and if less then 100 and goto step 1, otherwise stop.

4. Repeat the steps 0 to 3 a number of times (say 1,000) to get the number of years each retirement simulation survived.
5. Estimate the tail probability function.

# Monty Hall problem

The Monty Hall Problem: You are a contestant on a game show and the game is to pick one of three doors. Behind one of the doors is a valuable prize, while the other two doors have a worthless prize. You cannot see what is behind the door; only the game host can open the doors. The game has the following rules.

You are to choose a single door of the three available. The game host then opens one of the other doors to reveal there is a worthless prize behind it. Leaving two unopended doors remaining. Your door chosen in step 1 and the last door. The host gives you the choice to keep the door you chose at step one or change to the other unopened door. Determine what are the chances of winning if you keep the same door or switch doors.

Describe in your own words how you would solve the following problem using Monte Carlo simulation. You may use pseudo code as part of your answer.

## Solution with explanations

To solve this, consider applying the same strategy (keep same door or switch doors) repeatedly and check the chance of obtaining the valuable prize. The strategy below uses the optimal strategy which is to switch.
Step 0. Consider values to represent the valuable and worthless prizes behind each door. Say for example, 0 for worthless and 1 for valuable. Then a vector consisting of two zeros and a single one would represent the game scenario. For example, (0, 0, 1) to represent worthless prizes in doors 1 and 2 and the valuable prize in door 3.
The algorithm to determine the chance of winning with a strategy consists of repeating the steps below a large number of times ($M = 1000$ say) and recording the percentage of times the strategy was successful.

1. Simulate a random vector as per details in Step 0; call this vector $\boldsymbol{x}$.
2. Simulate a single random integer from 1, 2, 3 to represent the user choice.
3. Check which elements of $\boldsymbol{x}$ that the user hasn't chosen are worthless and randomly remove one from $\boldsymbol{x}$.

4. Employ the switch strategy: change the user selection to the other one.
5. Check if the user got the valuable prize and record the outcome.

Final Step: once $M$ iterations of Steps 1 to 5 complete, check the proportion of outcome values that were valuable.