Lab Week 13 : Example Exam Questions

Code

STAT5003

Dr. Justin Wishart

Semester 2, 2022

Below are some practice questions for the final exam. Some of these are taken from a previous year. Note in those exams the question could have a single correct or multiple correct response and it wasn't specified if there was more than one correct response For the exam this semester, the multiple choice section will always have two correct responses.

1 Multiple choice questions

1. Which of the following R code statements will compute the averages of the value variable for each level in group in the data.frame `dat` ?
    a. [**correct**] `vapply(split(dat$value, dat$group), mean, numeric(1L))`
    b. `mean(value ~ group, data = dat)`
    c. `lapply(value, mean, data = dat)`
    d. [**correct**] `aggregate(value ~ group, dat, mean)`
2. Which of the following are supervised learning techniques
    a. k-means
    b. [**correct**] Random Forest
    c. [**correct**] Linear Discriminant Analysis
    d. Density estimation
3. Which of the following are characteristics of a kernel function (one used in density estimation)?
    a. a frequency function from a histogram
    b. [**correct**] a symmetric function
    c. a function ranging from -1 to 1
    d. [**correct**] a function that integrates to 1 over its support.
4. Which of the following statement is TRUE for specificity and sensitivity?
    a. [**correct**] In the cancer prediction problem (cancer as positive and normal as negative), sensitivity is the number of true cancer cases that are captured by the predictive model divided by all cancer cases.
    b. Again, in cancer prediction, if a classifier predicts all samples as normal, it has 0% specificity and 100% sensitivity.

c. When evaluating classification models, specificity and sensitivity should be treated as equally important in all cases because they capture different aspects of model performance.

d. Overall classification accuracy is the best metric to summarise specificity and sensitivity.

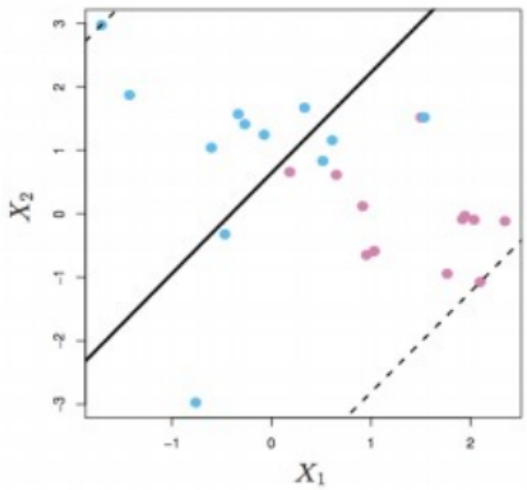
5. Which of the following support vector machine decision boundaries and margins correspond to the smallest C value of:

$$\max_{\beta_0,\beta_1,\ldots,\beta_p,\epsilon_1,\ldots,\epsilon_n} M \text{ such that } \sum_{j=1}^{p} \beta_j^2 = 1$$

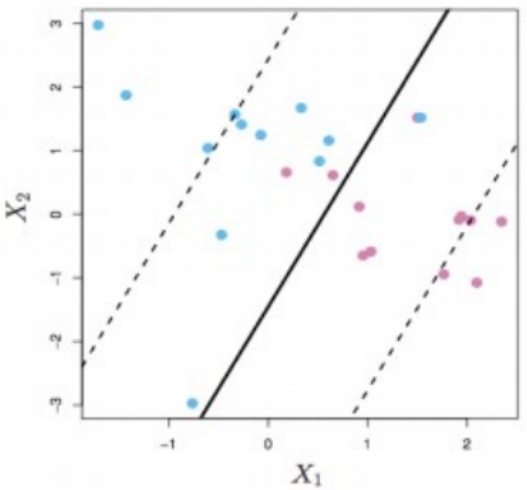
$$y_i(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}) \geq M(1 - \epsilon_i)$$

$$\epsilon_i \geq 0, \sum_{i=1}^{n} \epsilon_i \leq C.$$

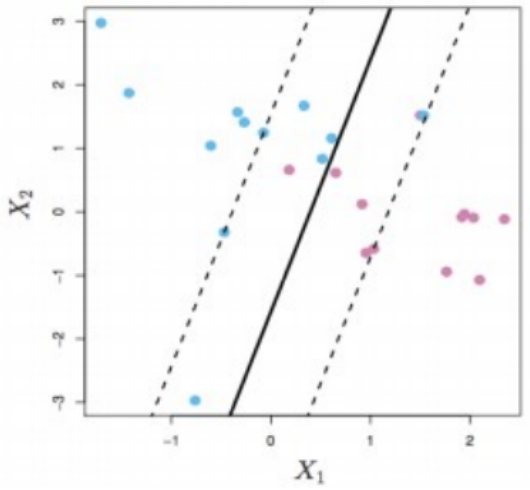
a.



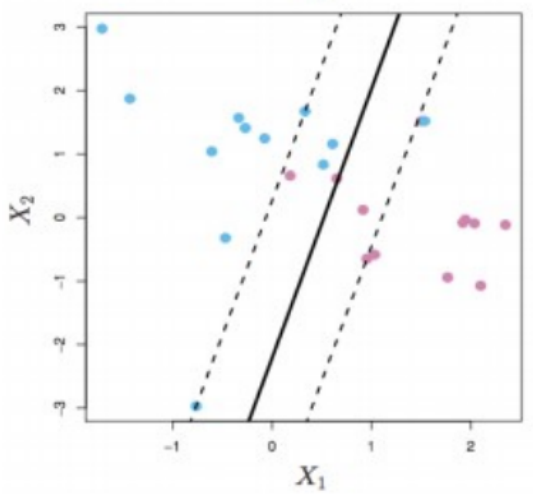
b.



c.



d. [**correct**]



6. Which of the following are indirect measures of the test error?
   a. [**correct**] $C_p = \frac{1}{n}\left(RSS + 2d\hat{\sigma}^2\right)$
   b. $RSS = \sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2$
   c. [**correct**] $BIC = \frac{1}{n}\left(RSS + \log(n)d\hat{\sigma}^2\right)$
   d. $F_1 = \frac{2\text{Precision}\times\text{Recall}}{\text{Precision}+\text{Recall}}$

where in the above

$\widehat{Y}_i$ is the predicted response for the ith observation. d is the number of features in the model, not including the intercept. Precision is TP/(TP + FP) with TP, FP and FN being a True Positive, False Positive and False Negative respectively. Recall is TP/(TP + FN) with TP, FP and FN defined above

2 Short answer example

2.1 Logistic regression 2.2 Distance matrix

Suppose we collect data for a group of students in a statistics class with variables X_1 = hours studied, X_2 = undergraduate weighted average mark, X_3 = years of previous programming experience and Y = receive a High Distinction (HD). We decide to use logistic regression to solve this classification problem. Below is the R code and results:

Hide

```
summary(lr.results)
```

```
##
## Call:
## glm(formula = Grade ~ ., family = binomial, data = students)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.02914  -0.23784  -0.05285  -0.00832   2.59620
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -22.26464    8.28123  -2.689  0.00718 **
## UgradMark     0.02621    0.11756   0.223  0.82361
## HrsStudied    0.74146    0.15142   4.897 9.74e-07 ***
## YearsProg     0.53504    0.42993   1.244  0.21332
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 140.453  on 129  degrees of freedom
## Residual deviance:  45.944  on 126  degrees of freedom
## AIC: 53.944
##
## Number of Fisher Scoring iterations: 7
```

a. Which predictor variable best explains the grade response variable? Justify your answer briefly.
b. For a student with undergraduate mark of 65, 20 hours of studying, and no programming experience, what is the probability the student will obtain a HD result?
c. Briefly describe how to interpret the coefficients of the logistic regression.

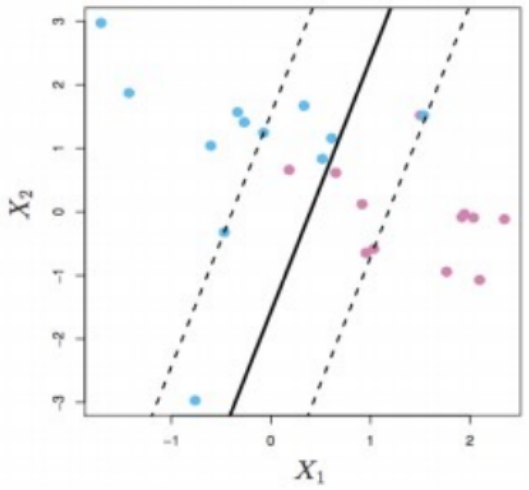
Solution

a. *HrsStudied* best explains the grade response variable since it is the coefficient that is the most statistically significant conditional on all other features being in the model with a p-value of 9.74e-7
b. The linear predictor gives the value `-22.266464 + 65 * 0.02621 + 20 * 0.74146 + 0 * 0.53504 =` -5.733614. Using the logistic function to convert this to a predicted probability,

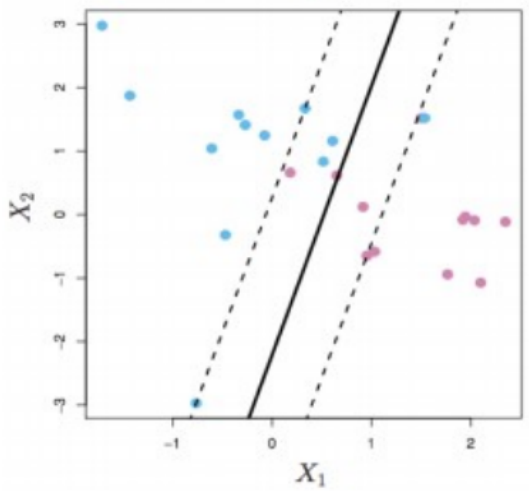
$$\widehat{p} = 1/(1 + exp(- - 5.733614)) = 0.0032249$$

c. (longer than it necessary) Each coefficient represents the increase (or decrease) in the log odds (log (p/(1-p)) for each unit increase in the predictor. Or looking at the odds ratio. p/(1-p) = exp(X beta) and the impact of the multiplicative increase (or decrease) on the odds. E.g. Assume Xa and Xb have the same values in each feature except Xa has one unit larger in X1. Then we can compute the ratio of odds ratios. pa/(1 - pa)/(pb/(1-pb)) = exp(beta1). Or using the example above, suppose there are two students (A and B) that are identical except student A studied for 1 hour longer than the other. Then we would have, pa/(1 - pa) = pb/(1 - pb) exp(0.74146) = pb/(1 - pb) * 2.099. That is, for each extra hour of study, the odds of getting a HD almost doubles.

c.



d. [**correct**]



6. Which of the following are indirect measures of the test error?
   a. [**correct**] $C_p = \frac{1}{n}\left(RSS + 2d\hat{\sigma}^2\right)$
   b. $RSS = \sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2$
   c. [**correct**] $BIC = \frac{1}{n}\left(RSS + \log(n)d\hat{\sigma}^2\right)$
   d. $F_1 = \frac{2\text{Precision}\times\text{Recall}}{\text{Precision}+\text{Recall}}$

where in the above

$\widehat{Y}_i$ is the predicted response for the ith observation. d is the number of features in the model, not including the intercept. Precision is TP/(TP + FP) with TP, FP and FN being a True Positive, False Positive and False Negative respectively. Recall is TP/(TP + FN) with TP, FP and FN defined above

2 Short answer example

2.1 Logistic regression 2.2 Distance matrix

Suppose we have been hired by a leading supermarket chain to analyse the profiles of their customers. They provided us with the following shopping basket data for six customers.

Compute the distance matrix for the customers below (use *Jaccard distance* as the distance measure between samples and complete distance as distance measure between clusters).

	Pasta	Tomato	Onion	Carrot	Nappies	Tissues	Soap	Catfood	Chips	Beer
Katie	1	1	1	1	1	0	0	0	0	0
John	0	0	0	0	1	1	1	0	1	1

3 Long answer question

3.1 Retirement problem 3.2 Monty Hall problem

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.

> 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%. Assume you retire at 65 and will live until you are 100, compute the chance that your investment will support your lifestyle until you die.

No need to calculate a number, just describe the process you would go through.

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
        1. If not, stop the algorithm, the money is exhausted.
        2. If yes, proceed to 2.
2. Multiply the capital by the return rate (1 + rnorm with mean 0.09 and 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 0-3 a number of times (say 1000), to get the number of years each retirement simulation survived.
5. Estimate the tail probability function.

	Pasta	Tomato	Onion	Carrot	Nappies	Tissues	Soap	Catfood	Chips	Beer
Jean	1	1	1	1	0	0	0	0	0	0
Pual	0	1	1	1	0	0	0	1	0	0
Jenny	1	1	1	1	0	0	1	0	0	0
Robert	1	0	0	0	0	0	1	0	1	1

Jaccard distance definition:

$$d(A,B) = 1 - \frac{|A \cap B|}{|A \cup B|}$$

where

a. $|A \cap B|$ = size of the intersection between sets A and B.
b. $|A \cup B|$ = size of the union between sets A and B.
c. Example: Jaccard distance $d(\text{Katie, John}) = 1 - 1/9 = 8/9$.

Solution:

Table for $d(A,B)$ given below

	Katie	John	Jean	Pual	Jenny	Robert
Katie	NA	0.8888889	0.2	0.5	0.3333333	0.8750000
John	NA	NA	1.0	1.0	0.8888889	0.5000000
Jean	NA	NA	NA	0.4	0.2000000	0.8571429
Pual	NA	NA	NA	NA	0.5000000	1.0000000
Jenny	NA	NA	NA	NA	NA	0.7142857
Robert	NA	NA	NA	NA	NA	NA

3 Long answer question

3.1 Retirement problem 3.2 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.

Solution

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. E.g. (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 x.
2. Simulate a single random integer from 1, 2, 3 to represent the user choice (1. in the problem statement)
3. Check which elements of x that the user hasnt chosen are worthless and remove it from 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 1-5 complete, check the proportion of outcome values that were valuable.
