Midterm

Fabiani Rafael

1. (22 points) Suppose that we wish to predict whether a given stock will issue a dividend this year (Y = 1 for "Yes" , Y = 0 for "No") based on X, last year's percent profit.We examine a large number of companies and discover that the mean value of X for companies that issued a dividend was $\bar{X} = 10$,while the mean for those that didn't was $\bar{X} = 0$. In addition, the variance of X for these two sets of companies was $\sigma^2 = 36$. Finally, 80% of companies issued dividends. Assume that X follows a normal distribution in each group, i.e. $X|Y = k\ N(\mu_k, \sigma^2)$. The density function for a normal distribution $N(\mu, \sigma^2)$ is

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

where μ is the mean and σ^2 is the variance.

(a) (4 points) According to Bayes' theorm, $P(Y = k|X = x) = \frac{\pi_k f_k(x)}{\pi_0 f_0(x)+\pi_1 f_1(x)}$, where $k = 0, 1$. What are the interpretations of π_0 and π_1? How to estimate them based on the given information ?

- $\pi_0\,\pi_1$ can be interpreted as representing the probabilities wherein the given company will not issue a dividend and will issue a dividend respectively. Now we can estimate them using the given information i.e given that 80% of the companies issued dividends, while 20% did not, we can say that $\pi_1 = 0.8$ and $\pi_0 = 0.2$.

(b) (4 ponts) What are the interpretations of σ^2, μ_0 and μ_1? How to estimate them based on the given information?

- Here $\sigma^2 = 36$ and $\sigma = 6$ can be interprered as the variance of the percentage profit of the companies, hence the standard deviation in reported % profit for all companies was 6%. Continuing μ_0 and μ_1 would correspond to the mean percentage profit of the companies that did not issue a dividend and the companies that did issue a dividend respectively. We can estimate them using the given information i.e $\bar{X} = 0$ for companies that did not issue a dividend and $\bar{X} = 10$ for companies that did issue a dividend.

(c) (4 points) We know that in logistic regression

$$\log \frac{P(Y = 1|X = x)}{P(Y = 0|X = x)} = \beta_0 + \beta_1 x$$

a linear function of x. Please show that this is also true here (the LDA framework)

- Now from the problem we have that

$$\begin{aligned} P(Y = 1|X = x) &= \frac{\pi_1 f_1(x)}{\pi_0 f_0(x) + \pi_1 f_1(x)} \\ &= \frac{\pi_1 \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu_1)^2}{2\sigma^2}\right)}{\pi_0 \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu_0)^2}{2\sigma^2}\right) + \pi_1 \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu_1)^2}{2\sigma^2}\right)} \end{aligned}$$

Now without loss of generality, $P(Y = 0 \mid X = x)$ can be written as the following

$$\frac{\pi_0 \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu_0)^2}{2\sigma^2}\right)}{\pi_0 \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu_0)^2}{2\sigma^2}\right) + \pi_1 \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu_1)^2}{2\sigma^2}\right)}$$

Now we have already estimated $\pi_0 = 0.2, \pi_1 = 0.8, \mu_0 = 0, \mu_1 = 10$ and were given that $\sigma^2 = 36$. We may now continue to use these values and yield the following

$$\begin{aligned}
P(Y=0|X=x) &= \frac{0.2 \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-10)^2}{2\sigma^2}\right)}{0.2 \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-10)^2}{2\sigma^2}\right) + 0.8 \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-10)^2}{2\sigma^2}\right)} \\
&= \frac{0.2 \exp\left(-\frac{x^2}{72}\right)}{0.2 \exp\left(-\frac{x^2}{72}\right) + 0.8 \exp\left(-\frac{(x-10)^2}{72}\right)} \\
P(Y=1|X=x) &= \frac{0.8 \exp\left(-\frac{(x-10)^2}{72}\right)}{0.2 \exp\left(-\frac{(x)^2}{72}\right) + 0.8 \exp\left(-\frac{(x-10)^2}{72}\right)} \\
\log \frac{P(Y=1|X=x)}{P(Y=0|X=x)} &= \log \frac{\frac{0.8 \exp\left(-\frac{(x-10)^2}{72}\right)}{0.2 \exp\left(-\frac{(x)^2}{72}\right) + 0.8 \exp\left(-\frac{(x-10)^2}{72}\right)}}{\frac{0.2 \exp\left(-\frac{x^2}{72}\right)}{0.2 \exp\left(-\frac{x^2}{72}\right) + 0.8 \exp\left(-\frac{(x-10)^2}{72}\right)}} \\
&= \log \frac{0.8 \exp\left(-\frac{(x-10)^2}{72}\right)}{0.2 \exp\left(-\frac{(x)^2}{72}\right) + 0.8 \exp\left(-\frac{(x-10)^2}{72}\right)} - \log \frac{0.2 \exp\left(-\frac{x^2}{72}\right)}{0.2 \exp\left(-\frac{x^2}{72}\right) + 0.8 \exp\left(-\frac{(x-10)^2}{72}\right)} \\
&= \log\left(0.8 \exp(-\frac{(x-10)^2}{72})\right) - \log\left(0.2 \exp(-\frac{x^2}{72}) + 0.8 \exp(-\frac{(x-10)^2}{72})\right) \\
&- \log\left(0.2 \exp(-\frac{(x)^2}{72})\right) + \log\left(0.2 \exp(-\frac{x^2}{72}) + 0.8 \exp(-\frac{(x-10)^2}{72})\right) \\
&= \log\left(0.8 \exp(-\frac{(x-10)^2}{72})\right) - \log\left(0.2 \exp(-\frac{(x)^2}{72})\right) \\
&= \log(0.8) - \frac{(x-10)^2}{72} - \log(0.2) + \frac{x^2}{72} \\
&= \left(2\ln(2) - \frac{25}{18}\right) + \frac{5}{18}x
\end{aligned}$$

Thus we have that $\beta_0 = \left(2\ln(2) - \frac{25}{18}\right), \beta_1 = \frac{5}{18}$, hence we have shown that the LDA framework is also a linear function of x i.e $\log \frac{P(Y=1|X=x)}{P(Y=0|X=x)} = \beta_0 + \beta_1 x$.

(d) (5 points) Predict the probability that a company will issue a dividend this year given that its percentage profit was $X = 4$ last year. *Hint: plug in normal densities and calculate* $P(Y = k|X = 4)$

- Using the results obtained previously we have

$$P(Y=1|X=x) = \frac{0.8 \exp\left(-\frac{(x-10)^2}{72}\right)}{0.2 \exp\left(-\frac{(x)^2}{72}\right) + 0.8 \exp\left(-\frac{(x-10)^2}{72}\right)}$$

Now we can plug in the values $x = 4$ and calculate the probability that a company will issue a dividend this year given that its percentage profit was $X = 4$ last year yielding

$$P(Y=1|X=4) = \frac{0.8\exp\left(-\frac{(4-10)^2}{72}\right)}{0.2\exp\left(-\frac{(4)^2}{72}\right) + 0.8\exp\left(-\frac{(4-10)^2}{72}\right)} \approx 0.75$$

So we predict that the probability that a company will issue a dividend this year given that its percentage profit was $X = 4$ last year is approximately 0.75 or 75%.

(e) (5 points) Write down the explicit discriminant functions $\delta_k(x) = x\frac{\mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2} + log(\pi^k)$ and evaluate at $x = 4$. Make predictions using discriminant functions.

- $$\begin{aligned}\delta_0(x) &= x\frac{\mu_0}{\sigma^2} - \frac{\mu_0^2}{2\sigma^2} + \log(\pi_0)\\ &= 4\frac{0}{36} - \frac{0^2}{2*36} + \log(0.2)\\ &= log(0.2)\\ \delta_1(x) &= x\frac{\mu_1}{\sigma^2} - \frac{\mu_1^2}{2\sigma^2} + \log(\pi_1)\\ &= 4\frac{10}{36} - \frac{10^2}{2\cdot 36} + \log(0.8)\end{aligned}$$

  so we then have $\delta_0(4) \approx -1.6, \delta_1(4) \approx -0.5$ and since $\delta_1(4) > \delta_0(4)$ we predict that the company will issue a dividend this year given that its percentage profit was $X = 4$ last year moreover this alligns with the previous prediction made in part (d).

2. (20 points) The table below provides a training data set containing six observations, three predictors, and one qualitative response variable. Suppose we want to use this data set to

Obs.	$\mathbf{X_1}$	$\mathbf{X_2}$	$\mathbf{X_3}$	**Y**
1	2	-1	0	Red
2	1	1	0	Green
3	-1	1	3	Red
4	-2	1	0	Green
5	1	0	1	Green
6	2	3	1	Red

make a prediction for Y when $X_1 = 1, X_2 = -1, X_3 = 0$ using K-nearest neighbors.

(a) (10 points) Compute the Euclidean distance between each observation and the test point, $X_1 = 1, X_2 = -1, X_3 = 0$.

- Towards the intent of computing the Euclidean distance for every obeservation and the test point $X_1 = 1, X_2 = -1, X_3 = 0$ we use the formula :

$$\sqrt{(1 - X_1)^2 + (-1 - X_2)^2 + (0 - X_3)^2}$$

The resulting distances are then shown in the following table :

Obs.	**Distance**	**Color**
1	1	Red
2	2	Green
3	4.1231	Red
4	3.6056	Green
5	1.4142	Green
6	4.2426	Red

(b) (3 points) What is out prediction with K =1? Why?

- With $K = 1$ we would predict that the test point $X_1 = 1, X_2 = -1, X_3 = 0$ would be classified as Red since the nearest point to the test point is observation 1 which is red .

(c) (3 points) What is out prediction with K =3? Why?

- With $K = 3$ we would predict that the test point $X_1 = 1, X_2 = -1, X_3 = 0$ would be classified as Green since the three nearest points to the test point are observations 1, 5 and 2 which are Red, Green followed by Green again respectively. Evidently the majority is green hence the prediction itself is green.

(d) (4 points) If the Bayes decision boundary in this problem is highly non-linear, then would we expect the best value for K to be large or small? Why?

- If it were that the Bayes decision boundary was of a highly non-linear nature, then a smaller K would bee preferable in K-nearest neighbors since, a smaller K would mean less smoothing of the boundary. Consequently this would allow for the model itself to closely follow non-linear contours corresponding to the true decision boundary. Contrastly, a larger value for K would average over more points and yield a smoother, more linear boundary but the fault would be in the failure to as adequately capture the underlying non-linear nature.

3. (12 points) Consider the fitted values that result from performing simple linear regression (one predictor) without an intercept, i.e., the model is $Y_i = \beta X_i + \epsilon_i, i = 1, ..., n$. By minimizing the RSS find the estimated coefficient $\hat{\beta}$. (the least square estimator)

- The residual sum of squares (RSS) is defined by $RSS = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$ wherein $\hat{Y}_i = \hat{\beta}X_i$. We can then substitute $\hat{Y}_i$ into the RSS equation to yield

$$RSS = \sum_{i=1}^{n}(Y_i - \hat{\beta}X_i)^2$$

Now we can take the derivative of the RSS with respect to $\hat{\beta}$ and set it equal to zero to find the least square estimator $\hat{\beta}$.

$$\begin{aligned}\frac{d}{d\hat{\beta}}RSS &= \frac{d}{d\hat{\beta}}\sum_{i=1}^{n}(Y_i - \hat{\beta}X_i)^2 \\ &= \sum_{i=1}^{n}2(Y_i - \hat{\beta}X_i)(-X_i) \\ &= -2\sum_{i=1}^{n}X_i(Y_i - \hat{\beta}X_i) \\ &= -2\sum_{i=1}^{n}X_iY_i + 2\hat{\beta}\sum_{i=1}^{n}X_i^2\end{aligned}$$

Setting the $\frac{d}{d\hat{\beta}}RSS = 0$ then yields,

$$\begin{aligned}0 &= -2\sum_{i=1}^{n}X_iY_i + 2\hat{\beta}\sum_{i=1}^{n}X_i^2 \\ 2\sum_{i=1}^{n}X_iY_i &= 2\hat{\beta}\sum_{i=1}^{n}X_i^2 \\ \hat{\beta} &= \frac{\sum_{i=1}^{n}X_iY_i}{\sum_{i=1}^{n}X_i^2}\end{aligned}$$

Thus we have that the least square estimator $\hat{\beta}$ is given by $\hat{\beta} = \frac{\sum_{i=1}^{n}X_iY_i}{\sum_{i=1}^{n}X_i^2}$.

4. (24 points) Recall the Boston data set, which record medv (median house value in thousands of dollars, eg.,medv=50 mean $50,000)for 506 neighborhoods around Boston. We fitted a linear regression model using two predictors lstat (percent of households with low socioeconomic status, eg., lstat=1 means 1%), rm (average number of rooms per dwelling) and their intereaction. The output is provided below.

```
Call:
lm(formula = medv ~ lstat + rm + rm * lstat, data = Boston)

Residuals:
    Min      1Q  Median      3Q     Max
-23.2349 -2.6897 -0.6158  1.9663 31.6141

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -29.12452     3.34250  -8.713  < 2e-16 ***
lstat         2.19398     0.20570  10.666  < 2e-16 ***
rm            9.70126     0.50023  19.393  < 2e-16 ***
lstat:rm     -0.48494     0.03459 -14.018  < 2e-16 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.701 on 502 degrees of freedom
Multiple R-squared:  0.7402,   Adjusted R-squared:  0.7387
F-statistic: 476.9 on 3 and 502 DF,  p-value: < 2.2e-16
```

Use the output and the **context** to answer the following questions.

(a) (4 points) What is the interpretation of $R^2 = 0.7402$? What is the interpretation of the F-test results (F-statistic = 476.9 on 3 and 502 DF, p-value < 2.2e-16)?

- The $R^2 = 0.7402$ can be interpreted as the proportion of the variance in the response variable that can be explained by the predictors. In this case, 74.02% of the variance in median house value can be explained by the predictors lstat, rm and their interaction ($lstat \times rm$). and the F-test results can be interpreted as the overall significance of the model. It tests the null hypothesis that all of the coefficients in the model are equal to zero. In this case, the F-statistic is 476.9 with a p-value less than 2.2e-16, which indicates that the model is statistically significant and so the null hypothesis that all of the coefficients in the model are equal to zero is rejected.

(b) (4 points) What predictors are significant? Why?

- The predictors lstat, rm, and the interaction term lstat:rm are all significant moreover we know this by the p-values associated with each coefficient being less than 0.05. The p-values being less than 0.05, indicates that they are statistically significant predictors of median house value.

(c) (6 points) What is the interpretation of the interaction effect? (Please interprete two ways: when *rm* is held constant, when *lstat* is held constant)

- The interaction effect between lstat and rm can be interpreted as the effect of the interaction between the percentage of households with lstat and the rm on the median house value. When rm is held constant, the interaction effect represents the change in the median house value for a one-unit increase in lstat, and vice versa. On the other hand when lstat is held constant, similarly the interaction term represents the change in the median house value for a one-unit increase in rm, and again vice versa.

(d) (4 points) Given a new neighbourhood's with 10% low socioeconomic households and 6 years as the average number of rooms per dwelling. What is the predicted median value of houses in this neighborhood?

- The predicted median value of houses in this neighborhood can be calculated by plugging in the values of lstat = 10, rm = 6 into the regression equation. The regression equation is then given by

  $$\text{medv} = -29.12452 + 2.19398 \times 10 + 9.70126 \times 6 - 0.48494 \times 10 \times 6 = 21.92644$$

  Thus we find the predicted median value of houses in this neighborhood to be $21.92644 \approx \$21,926$.

(e) (6 points) For the neighborhoods with 10% low socioeconomic households and 6 years as the average number of rooms per dwelling, we can calculate the confidence interval is (21.36483, 22.4876) and prediction interval is (12.67235, 31.18008). What are their interpretations?

- The confidence interval (21.36483, 22.4876) can be interpreted as the range of values for which we are 95% confident that the true median house value for neighborhoods with 10% low socioeconomic households and 6 years as the average number of rooms per dwelling lies. The prediction interval (12.67235, 31.18008) can be interpreted as the range of values within which we would expect the median house value for a new neighborhood to fall with 95% confidence. In essense both can be seen as conveying the extent to which we are sure about the median house value for the neighborhoods.

5. (12) points) Please answer the following conceptual questions about classification. Assume we have two classes and one predictor, i.e., $p = 1$ and $K = 2$.

   (a) (2 points) What is Bayesian Classifier? How does it make prediction?

   - The Bayesian Classifier calculates the probability of an observation belonging to a class using Bayes' theorem. The classifier then continues to assign the observation to the class having the highest probability.

   (b) (2 points) What is the general idea of logisitic regression? How does it make prediction?

   - The general idea behind logistic regression is to calculate the probability of an observation belonging to a class using a logistic function and then assigning the observation to the class with the highest probability.

   (c) (2 points) What is the general idea of LDA? How does it make prediction?

   - The general idea of LDA is to go off the assumption that observations come from normal distributions sharing the same covariance matrix. From these assumptions together with Bayes' theorem LDA then defines a linear decision boundary to classify the observations.

   (d) (2 points) What is the general idea of QDA? How does it make prediction?

   - The general idea of QDA is comparable to LDA however QDA allows eaach class its own covariance matrix consequently covariances differ from class to class leading to decision boundaries pertubing to produce quadratic like decisions boundary, consequently it can provide a more flexible decision boundary for when the decision boundary in question is quadratic in nature.

   (e) (2 points) What is the general idea of KNN? How does it make prediction?

   - KNN assigns a label by looking at the K available training points which lie closts to the new observation. It peerforms classification by assigning the most common label among the K nearest training points. For $K = 2$ as given in the problem it looks to the two nearest training points and assigns the label that is most common among the two. If it were that they were equally common a measure that can be taken is looking to the weight of the test data as a whole to be used as a determining factor for classification.

   (f) (2 points) What is the general idea of Naive Bayes? How does it make prediction?

   - Naive Bayes assumes that the predictors are each independent with respect to one another. It calculates the probability of an observation belonging to a given class it then proceeds to assign the observation to the class with the highest probability.