Stat 432 HW 03

Name: Your Name, netID: yournetID

Summer 2024

Include the R code for this HW.

```r
knitr::opts_chunk$set(echo = TRUE)
library(ISLR2)
library(GGally)
library(tibble)
library(dplyr)
library(knitr)
library(kableExtra)
```

Question 1 (Classification General)

Consider a categorical response Y which takes possible values 0 and 1 as well as a single numerical predictor X. Recall that $p(x) = P(Y = 1|X = x)$.

Consider the model

$$log\left(\frac{p(x)}{1-p(x)}\right) = \beta_0 + \beta_1 x$$

and estimated coefficients

- $\hat{\beta}_0 = 2$
- $\hat{\beta}_1 = 1$

(a) Provide a classification when $x = -1$ that attempts to minimize the classification error.

(b) Calculate an estimate of $P(Y = 1|X = -1)$.

(c) Find a value c that splits x into regions classified as 1 and 0.

Question 2 (Logistic Regression)

We will use data found in (wisc-trn.csv) and (wisc-tst.csv) [check the box folder for files] which contain train and test data respectively.

You should consider coercing the response (`class` variable) to be a factor variable.

(a) Create plots for EDA analysis with the variables: “class”, radius”, “texture”, “smoothness”, “concavity”, “symmetry”, and “fractal”.

(b) Our task is to conduct classification on `class`. Consider an additive logistic regression that uses “radius”, “texture”, “smoothness”, “concavity”, “symmetry”, and “fractal” as predictor variables. Use this model to estimate

$$p(x) = P(Y = \texttt{M} \mid X = x).$$

Report test sensitivity, test specificity, and test accuracy for the Bayes Classifier. We will consider `M` (malignant) to be the “positive” class when calculating sensitivity and specificity. Summarize these results.

(c) Repeat part (b) with two different classifiers, each using a different cutoff for predicted probability:

$$\hat{C}(x) = \begin{cases} M & \hat{p}(x) > c \\ B & \hat{p}(x) \le c \end{cases}$$

- $c = 0.3$
- $c = 0.8$

(d) Of the three classifiers proposed in (b) and (c), which one do you prefer? For the cancer study, we want to chose the model labels cancer as cancer, and we value this over labeling non-cancer as non-cancer. Which metric should you use?

Question 3 (LDA)

Consider the following estimates and information from data for a three-class classifications boundary.

Class A	Class B	Class C
$\hat{\mu}_A = 5$	$\hat{\mu}_B = 7$	$\hat{\mu}_C = 10$
$n_A = 100$	$n_B = 40$	$n_C = 60$

$$\hat{\sigma} = 1$$

(a) Use LDA to estimate the probability $P(Y = B|X = 7.5)$. Show your work.

(b) Provide a classification when $x = 5.5$. (Use the Bayes classifier.) Show your work.

(c) Calculate the LDA decision boundary. (Use the Bayes classifier.) Show your work.

Question 4 (QDA)

Consider the following estimates and information from data for a three-class classifications boundary.

Class A	Class B	Class C
$\hat{\mu}_A = 5$	$\hat{\mu}_B = 7$	$\hat{\mu}_C = 10$
$n_A = 100$	$n_B = 40$	$n_C = 60$
$\hat{\sigma}_A = 1$	$\hat{\sigma}_B = 2$	$\hat{\sigma}_C = 3$

(a) Use QDA to estimate the probability $P(Y = B|X = 7.5)$. Show your work.

(b) Provide a classification when $x = 5.5$. (Use the Bayes classifier.) Show your work.

(c) Calculate the QDA decision boundary. (Use the Bayes classifier.) Show your work.

Question 5 (LDA-2 dimension)

Consider the following estimates and information from data for a three-class classifications boundary with two feature variables.

Class A	Class B	Class C
$\hat{\mu}_A = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$	$\hat{\mu}_B = \begin{bmatrix} 3 \\ 1 \end{bmatrix}$	$\hat{\mu}_C = \begin{bmatrix} 2 \\ 4 \end{bmatrix}$
$n_A = 100$	$n_B = 40$	$n_C = 60$

$$\hat{\Sigma} = \begin{bmatrix} 1 & 0.3 \\ 0.3 & 1 \end{bmatrix}$$

(a) Provide a classification when $x = (2, 3)$. (Use the Bayes classifier.) Show your work.

(b) Use LDA to estimate the probability $P(Y = B|X = (2, 3))$. Show your work.