Homework 4 (Nov. 16th)

Deadline: Wednesday, November 30th, at 11:59pm.

Submission: Read the submission instruction carefully! There are 5 questions in this assignment. You need to submit two files through Quercus for this assignment.

- The first file should be a PDF file titled `hw4_writeup.pdf` containing your answers to Questions 1 – 5, as well as R code and R outputs requested for Questions 4 and 5. You can produce the file however you like (e.g. LaTeX, Microsoft Word, scanner), as long as it is readable.

- The second file should be your completed R code, named as `discriminant_analysis.R`. You need to ensure that this file has the exact name as indicated. DO NOT set or modify the working directory within this file.

Neatness Point: You will be deducted one point if we have a hard time reading your solutions or understanding the structure of your code.

Late Submission: 10% of the total possible marks will be deducted for each day late, up to a maximum of 3 days. After that, no submissions will be accepted.

- **Problem 1 (6 pts)**
  In this question, you will derive the maximum likelihood estimates for Gaussian Naïve Bayes in which a random discrete class label $Y \in [K] := \{1, 2, \ldots, K\}$ and a random feature $X \in \mathbb{R}^D$ satisfy

$$\mathbb{P}(Y = k) = \pi_k, \qquad X \mid Y = k \sim N_p(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k), \qquad \forall k \in [K]. \tag{0.1}$$

  Here $\pi_1, \ldots, \pi_K$ are the priors of the class label Y, and conditioning on $Y = k$ for any $k \in [K]$, the feature vector $X \in \mathbb{R}^p$ has a p-dimensional Gaussian density with mean $\boldsymbol{\mu}_k \in \mathbb{R}^p$ and a diagonal covariance matrix

$$\boldsymbol{\mu}_k = \begin{bmatrix} \mu_{k1} \\ \vdots \\ \mu_{kp} \end{bmatrix}, \qquad \boldsymbol{\Sigma}_k = \begin{bmatrix} \sigma_{k1}^2 & 0 & \cdots & 0 \\ 0 & \sigma_{k2}^2 & \cdots & 0 \\ & & \ddots & \\ 0 & 0 & \cdots & \sigma_{kp}^2 \end{bmatrix}.$$

  Let $(y_1, \mathbf{x}_1), \ldots, (y_n, \mathbf{x}_n)$ be n i.i.d. realizations of (Y, X).

  1. **(3 pts)** Write down the log-likelihood function of $(y_1, \mathbf{x}_1), \ldots, (y_n, \mathbf{x}_n)$.
     Hint: Let Z be a categorical variable taking values from $\{1, \ldots, K\}$ with corresponding probabilities $\theta_1, \ldots, \theta_K$ with $\theta_k \geq 0$ and $\sum_k \theta_k = 1$. Its probability mass function at any $Z = z$ is

$$\mathbb{P}(Z = z) = \prod_{k=1}^{K} \theta_k^{1\{z=k\}}.$$

  2. **(3 pts)** Derive the maximum likelihood estimators of π_k, $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$ for all $k \in [K]$. You may assume $\sum_{i=1}^{n} 1\{y_i = k\} > 0$ for all $k \in [K]$.

- **Problem 2 (4 pts)**
  It was mentioned in the lecture that a cubic regression spline with one knot at ξ can be obtained as
  $$f(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \beta_4 (x - \xi)_+^3$$
  where
  $$(x-\xi)_+^3 = \begin{cases} (x-\xi)^3 & \text{if } x > \xi \\ 0 & \text{otherwise} \end{cases}.$$
  We now verify parts (1) – (4) below to conclude that $f(x)$ is indeed a cubic regression spline.

  1. **(1 pt)** Verify that $f(x)$ is a piecewise cubic polynomials. That is, show that $f(x)$ can be written as two cubic polynomials for $x > \xi$ and $x \leq \xi$.
  2. **(1 pt)** Denote the two polynomials as $f_1(x)$ and $f_2(x)$. Show that $f_1(\xi) = f_2(\xi)$, that is, $f(x)$ is continuous at ξ.
  3. **(1 pt)** Show that $f_1'(\xi) = f_2'(\xi)$, that is, the first order derivative $f'(x)$ is continuous at ξ.
  4. **(1 pt)** Show that $f_1''(\xi) = f_2''(\xi)$, that is, the second order derivative $f''(x)$ is continuous at ξ.

- **Problem 3 (4 pts)**
A 1-dimensional binary classification training set $\{(x_i, y_i)\}_{i=1}^n$ with $x_i \in \mathbb{R}$ and $y_i \in \{0, 1\}$ is *linear separable* if there exists a threshold $t \in \mathbb{R}$ such that

$$\begin{aligned} x_i &< t, \quad \text{for all } y_i = 0 \\ x_i &\geq t, \quad \text{for all } y_i = 1. \end{aligned}$$

1. **(2 pts)** Suppose we have the following 1-D dataset for binary classification:

x_i	y_i
-1	1
1	0
3	1

Argue briefly (at most a few sentences) that this dataset is not linearly separable.

2. **(2 pts)** Now suppose we map the 1-dimensional feature into a 2-dimensional space

$$h(x) = \begin{bmatrix} h_1(x) \\ h_2(x) \end{bmatrix} = \begin{bmatrix} x \\ x^2 \end{bmatrix}.$$

Is the new data set $(h(x_1), y_1), (h(x_2), y_2), (h(x_3), y_3)$ linear separable? That is, does there exist pairs of (t_1, t_2) such that

$$\begin{aligned} t_1 h_1(x_i) + t_2 h_2(x_i) &< 1, \quad \text{for all } y_i = 0 \\ t_1 h_1(x_i) + t_2 h_2(x_i) &\geq 1, \quad \text{for all } y_i = 1. \end{aligned}$$

- **Problem 4 (12 pts)**
  For this question you will build classifiers to label images of handwritten digits. Each image is 8 by 8 pixels and is represented as a vector of dimension 64 by listing all the pixel values in raster scan order. The images are grayscale and the pixel values are between 0 and 1. The labels y are $\{0, 1, 2, \ldots, 9\}$ corresponding to which character was written in the image. There are 700 training points and 400 test points for each digit; they can be found in `digits_train.txt` and `digits_test.txt`. These data sets can be loaded by using the helper function in `utils.R`.

  You will implement both linear discriminant analysis (LDA) and quadratic discriminant analysis (QDA) to classify these images. Recall that conditioning on each class $k \in \{0, 1, \ldots, 9\}$, the feature $X \mid Y = k$ follows a multivariate Gaussian distribution, that is,

$$\mathbb{P}(X = \mathbf{x} \,|\, Y = k) = (2\pi)^{-p/2}|\boldsymbol{\Sigma}_k|^{-1/2} \exp\left\{-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_k)^\top \boldsymbol{\Sigma}_k^{-1}(\mathbf{x} - \boldsymbol{\mu}_k)\right\} \tag{0.2}$$

  where $\boldsymbol{\mu}_k \in \mathbb{R}^p$ is the conditional mean and $\boldsymbol{\Sigma}_k \in \mathbb{R}^{p\times p}$ is the conditional covariance matrix. For LDA, $\boldsymbol{\Sigma}_k$ is assumed to be the same across classes. The priors are

$$\pi_k = \mathbb{P}(Y = k), \qquad \text{for all} \quad k \in \{0, 1, \ldots, 9\}.$$

  You will compute the maximum likelihood estimators of the priors π_k, the conditional means $\boldsymbol{\mu}_k$ and the conditional covariance matrices $\boldsymbol{\Sigma}_k$ for $k \in \{0, 1, \ldots, 9\}$, and use the estimators to construct classifiers.

  Read carefully the structure of `discriminant_analysis.R`. Include your code for all sub-questions.

  1. **(4 pts)** Complete the functions `Comp_priors`, `Comp_cond_means` and `Comp_cond_covs` in the file `discriminant_analysis.R`.
  2. **(2 pts)** Complete the functions `Predict_posterior` and `Predict_labels` in the file `discriminant_analysis.R`.
  3. **(2 pts)** Use LDA to classify the test data by completing part a in `hw4_starter.R`. Report the misclassification error of LDA.
  4. **(2 pts)** Use QDA to classify the test data by completing part b in `hw4_starter.R`. Report the misclassification error of QDA.
  5. **(2 pts)** Complete part c in `hw4_starter.R`, i.e. perform LDA and QDA by using the built-in `lda` and `qda` functions and compare with your implementation in terms of both misclassification rates and computational speed.

- **Problem 5 (11 pts)**
  This question uses the variables `dis` (the weighted mean of distances to five Boston employment centers) and `nox` (nitrogen oxides concentration in parts per 10 million) from the `Boston` data. We will treat `dis` as the predictor and `nox` as the response. Include your code for each subquestion.

  1. **(1 pt)** Use the `poly()` function to fit a cubic polynomial regression to predict `nox` using `dis`. Report the fitted regression summary, and plot the resulting data and polynomial fits.
  2. **(2 pts)** Plot the polynomial fits for a range of different polynomial degrees, from {1, 3, 5, 7, 10}, and report the associated residual sum of squares.
  3. **(2 pts)** Perform 10-fold cross-validation to select the optimal degree for the polynomial, and explain your results.
  4. **(2 pts)** Use the `bs()` function to fit a regression spline to predict `nox` using `dis`. Report the output for the fit using four degrees of freedom. Specify how you choose the knots and plot the resulting fit.
  5. **(2 pts)** Now fit a regression spline for a range of degrees of freedom, from {4, 6, 8, 10}, and plot the resulting fits and report the resulting RSS. Describe the results obtained.
  6. **(2 pts)** Perform 10-fold cross-validation to select the best degrees of freedom for a regression spline on this data. Describe your results.