STAT3006 Assignment 3—Classification

Due Date: 15th October 2021
Weighting: 30%

Instructions

- The assignment consists of **three (3) problems, each problem is worth 10 marks, and each mark is equally weighted.**
- The mathematical elements of the assignment can be completed by hand, **in LaTeX (preferably)**, or in Word (or other typesetting software). The mathematical derivations and manipulations should be accompanied by clear explanations in English regarding necessary information required to interpret the mathematical exposition.
- Computation problems can be answered using your programming language of choice, although R is generally recommended, or Python if you are uncomfortable with R. As with the mathematical exposition, you may choose to typeset your answers to the problems in whatever authoring or word processing software that you wish. You should also maintain a copy of any codes that you have produced.
- Computer generated plots and hand drawn graphs should be included together with the text where problems are answered.
- The assignment will require four (4) files containing data, that you can can download from the Assignment 3 section on Blackboard. These files are: `p2_1ts.csv`, `p2_1cl.csv`, `p2_2ts.csv`, `p3_1x.csv`, `p3_1y.csv`, and `data_bank_authentification.txt`.
- Submission files should include the following (which ever applies to you):
  - Scans of handwritten mathematical exposition.
  - Typeset mathematical exposition, outputted as a `pdf` file.
  - Typeset answers to computational problems, outputted as a `pdf` file.
  - Program code/scripts that you wish to submit, outputted as a `txt` file.

- **All submission files should be labeled with your name and student number and archived together in a `zip` file and submitted at the TurnItIn link on Blackboard.** We suggest naming using the convention:

  `FirstName_LastName_STAT3006A3_[Problem_XX/Problem_XX_Part_YY].[FileExtension]`.

- As per `my.uq.edu.au/information-and-services/manage-my-program/student-integrityand-conduct/academic-integrity-and-student-conduct`, what you submit should be your own work. Even where working from sources, you should endeavour to write in your own words. You should use consistent notation throughout your assignment and define whatever is required.

Problem 1 [10 Marks]

Let $X \in \mathbb{X} = [0,1]$ and $Y \in \{0,1\}$. Further, suppose that

$$\pi_y = \mathrm{P}\left(Y = y\right) = 1/2$$

for both $y \in \{0,1\}$, and that the conditional distributions of $[X|Y = y]$ are characterized by the probability density functions (PDFs):

$$f\left(x|Y=0\right) = 2 - 2x$$

and

$$f\left(x|Y=1\right) = 2x.$$

Part a [2 Marks]

Consider the Bayes' classifier for $Y \in \{0,1\}$ is

$$r^*\left(x\right) = \begin{cases} 1 & \text{if } \tau_1\left(x\right) > 1/2, \\ 0 & \text{otherwise,} \end{cases}$$

where

$$\tau_1\left(x\right) = \mathrm{P}\left(Y = 1|X = x\right).$$

Derive the explicit form of $\tau_1\left(x\right)$ in the current scenario and plot $\tau_1\left(x\right)$ as a function of x.

Part b [2 Marks]

Define the classification loss function for a generic classifier $r : \mathbb{X} \to \{0,1\}$ as

$$\ell\left(x, y, r\left(x\right)\right) = \left[\!\left[r\left(x\right) \neq y \right]\!\right],$$

where $\ell : \mathbb{X} \times \{0,1\} \times \{0,1\}$, and consider the associated risk

$$L\left(r\right) = \mathrm{E}\left(\left[\!\left[r\left(X\right) \neq Y \right]\!\right]\right).$$

It is known that the Bayes' classifier is optimal in that it minimizes the classification risk, that is

$$L\left(r^*\right) \leq L\left(r\right).$$

In the binary classification case,

$$L\left(r^*\right) = \mathrm{E}\left(\min\left\{\tau_1\left(X\right), 1 - \tau_1\left(X\right)\right\}\right) = \frac{1}{2} - \frac{1}{2}\mathrm{E}\left(\left|2\tau_1\left(X\right) - 1\right|\right).$$

Calculate $L\left(r^*\right)$ for the current scenario.

Part c [2 Marks]

Assume now that $\pi_1 \in [0,1]$ is now unknown. **Derive an expression for $L\left(r^*\right)$ that depends on π_1.**

Part d [2 Marks]

Assume again that $\pi_1 \in [0,1]$ is unknown. **Argue that we can write**

$$L\left(r^*\right) = \int_{\mathbb{X}} \min\left\{\left(1 - \pi_1\right) f\left(x | Y = 0\right), \pi_1 f\left(x | Y = 1\right)\right\} \mathrm{d}x.$$

Then, assuming that $\pi_0 = \pi_1 = 1/2$, **argue that we can further write**

$$L\left(r^*\right) = \frac{1}{2} - \frac{1}{4}\int_{\mathbb{X}} \left|f\left(x | Y = 1\right) - f\left(x | Y = 0\right)\right| \mathrm{d}x.$$

Part e [2 Marks]

Consider now that $\pi_1 \in [0,1]$ is unknown, as are $f\left(x|Y=0\right)$ and $f\left(x|Y=1\right)$. That is, we only know that $f\left(\cdot|Y=y\right) : \mathbb{X} \to \mathbb{R}$ is a density function on $\mathbb{X} = [0,1]$, for each $y \in \{0,1\}$, in sense that $f\left(x|Y=y\right) \geq 0$ for all $x \in \mathbb{X}$ and that $\int_{\mathbb{X}} f\left(x|Y=y\right) \mathrm{d}x = 1$.

Using the expressions from **Part d**, **deduce the minimum and maximum values of $L(r^*)$ and provide conditions on π_1, $f(\cdot|Y=0)$ and $f(\cdot|Y=1)$ that yield these values.**

Problem 2 [10 Marks]

Suppose that we observe an independent and identically distributed sample of $n = 300$ random pairs $(\boldsymbol{X}_i, Y_i)$, for $i \in [n]$, where $\boldsymbol{X}_i = (X_{i1}, \dots, X_{id})$ is a mean-zero time series of length $d = 100$ and $Y_i \in \{1, 2, 3\}$ is a class label. Here, X_{it} is the observation of time series $i \in [n]$ at time $t \in [d]$ and we may say that $\boldsymbol{X}_i \in \mathbb{X} = \mathbb{R}^d$.

We assume that the label Y_i, for $i \in [n]$, is such that each class occurs in the general population with unknown probability

$$\pi_y = \mathrm{P}(Y_i = y),$$

for each $y \in \{1, 2, 3\}$, where $\sum_{y=1}^{3} \pi_y = 1$. Further, we know that X_{it} is first-order autoregressive, in the sense that the distribution of $[\boldsymbol{X}_i | Y = y]$ can be characterized by the fact the conditional probability densities

$$f(x_{i1}|Y = y) = \phi\left(x_{i1}; 0, \sigma_y^2\right)$$

and for each $t \geq 2$,

$$f(x_{ir}|X_{i1} = x_{i1}, X_{i2} = x_{i2}, \dots, X_{i,r-1} = x_{i,r-1}, Y_i = y) = \phi\left(x_{ir}; \beta_y x_{i,r-1}, \sigma_y^2\right),$$

where $\boldsymbol{x}_i = (x_{i1}, \dots, x_{id})$ is a realization of $\boldsymbol{X}_i$, and for each $y \in \{1, 2, 3\}$, $\sigma_y^2 \in (0, \infty)$ and $\beta_y \in [-1, 1]$. Here,

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

is the univariate normal probability density function with mean $\mu \in \mathbb{R}$ and variance $\sigma^2 \in (0, \infty)$.

Part a [2 Marks]

Let $(\boldsymbol{X}, Y)$ arise from the same population distribution as $(\boldsymbol{X}_1, Y_1)$. Using the information above, **derive expressions for the a posteriori probabilities**

$$\tau_y(\boldsymbol{x}; \boldsymbol{\theta}) = \mathrm{P}(Y = y | \boldsymbol{X} = \boldsymbol{x}),$$

for each $y \in \{1, 2, 3\}$, as functions of the parameter vector

$$\boldsymbol{\theta} = \left(\pi_1, \pi_2, \pi_3, \beta_1, \beta_2, \beta_3, \sigma_1^2, \sigma_2^2, \sigma_3^2\right).$$

Further, **use the forms of the a posteriori probabilities to produce an explicit form of the Bayes classifier** (i.e., a form that is written in terms of the parameters $\boldsymbol{\theta}$).

Part b [1 Marks]

Using the information above, **construct the likelihood function**

$$L\left(\boldsymbol{\theta}; \mathbf{Z}_n\right) = \prod_{i=1}^{n} f\left(\boldsymbol{z}_i; \boldsymbol{\theta}\right)$$

based on the random sample $\mathbf{Z}_n = \left(\boldsymbol{Z}_1, \dots, \boldsymbol{Z}_n\right)$, where $\boldsymbol{Z}_i = \left(\boldsymbol{X}_i, Y_i\right)$ (for $i \in [n]$), and **write the log-likelihood function** $\log L\left(\boldsymbol{\theta}; \mathbf{Z}_n\right)$**.** Here, $f\left(\boldsymbol{z}_i; \boldsymbol{\theta}\right)$ is the joint density of $\boldsymbol{Z}_i$, deduced from the problem description, and where $\boldsymbol{\theta}$ is defined in **Part a**.

Part c [2 Marks]

Using the form of the log-likelihood function from the problem above, derive closed-form expressions of the maximum likelihood estimator

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta} \in \left\{(\pi_1, \pi_2, \pi_3) : \pi_y \geq 0, \sum_{y=1}^{3} \pi_y = 1\right\} \times [-1,1]^3 \times (0,\infty)^3}{\arg\max} \log L\left(\boldsymbol{\theta}; \mathbf{Z}_n\right).$$

Part d [1 Marks]

The data set `p2_1ts.csv`[1] contains a realization $\mathbf{x}_n = \left(\boldsymbol{x}_1, \dots, \boldsymbol{x}_n\right)$ of the $n = 300$ time series $\mathbf{X}_n = \left(\boldsymbol{X}_1, \dots, \boldsymbol{X}_n\right)$, and the data set `p2_1cl.csv` contains a realization $\mathbf{y}_n = \left(y_1, \dots, y_n\right)$ of the associated $n = 300$ class labels $\mathbf{Y}_n = \left(Y_1, \dots, Y_n\right)$. Using the notion the mth order auto-covariances of a time series $\boldsymbol{X} = \left(X_1, \dots, X_d\right)$:

$$\rho_m = \mathrm{E}\left\{\left[X_t - \mathrm{E}\left(X_t\right)\right]\left[X_{t+m} - \mathrm{E}\left(X_{t+m}\right)\right]\right\}$$

for $m \geq 0$, and appropriate sample estimators, **attempt to visualize these data in a manner that demonstrates the differences between the three class specific distributions.**

Part e [2 Marks]

For the data sets from **Part d**, using the maximum likelihood estimator from **Part c**, **derive the expressions of the estimate** $\tau_y\left(\boldsymbol{x}; \hat{\boldsymbol{\theta}}\right)$ **of** $\tau_y\left(\boldsymbol{x}; \boldsymbol{\theta}\right)$**, for each** $y \in \{1, 2, 3\}$. Furthermore, **provide an explicit form of the estimated Bayes' classifier** (i.e., a classifier $r\left(\boldsymbol{x}; \hat{\boldsymbol{\theta}}\right)$, dependent

[1] Each row of the CSV file is a time series and each column is a time point.

on $\hat{\boldsymbol{\theta}}$). Finally, **use the estimated Bayes' classifier to compute the so-called in-sample empirical risk**:

$$\bar{L}_n\left(r\left(\cdot;\hat{\boldsymbol{\theta}}\right)\right) = \frac{1}{n}\sum_{i=1}^{n}\left[\!\left[r\left(\boldsymbol{X}_i;\hat{\boldsymbol{\theta}}\right) \neq Y_i\right]\!\right],$$

where the averaging is over the same sample $\mathbf{Z}_n$ that is used to compute $\hat{\boldsymbol{\theta}}$.

Part f [2 Marks]

The data set p2_2ts.csv[2] contains realization $\mathbf{x}'_n = (\boldsymbol{x}'_1, \dots, \boldsymbol{x}'_n)$ of $n' = 20$ partially observed time series $\boldsymbol{X}'_i = (X_{i1}, \dots X_{i50})$, where $\boldsymbol{X}'_i$ contains the first 50 time points of a fully observe time series $\boldsymbol{X}''_i = (X_{i1}, \dots, X_{i100})$, for each $i \in [n']$. Under the assumption that $\boldsymbol{X}''_i$ has the same distribution as $\boldsymbol{X}_1$, as described at start of the problem, **argue that you can use the maximum likelihood estimates from Part e to produce a Bayes' classifier for the partially observed time series $\boldsymbol{X}'_i$, and produce classifications for each of the $n' = 20$ times series.**

Problem 3 [10 Marks]

Let $\mathbf{Z}_n = (\boldsymbol{Z}_1, \dots, \boldsymbol{Z}_n)$ be an independent and identically distributed sample of n pairs $\boldsymbol{Z}_i = (\boldsymbol{X}_i, Y_i)$ of features $\boldsymbol{X}_i \in \mathbb{X} = \mathbb{R}^d$ and labels $\mathbb{Y} = \{-1, 1\}$, where $i \in [n]$. Further, let

$$\rho(\boldsymbol{x};\boldsymbol{\theta}) = \alpha + \boldsymbol{\beta}^\top \boldsymbol{x}$$

be a linear classification rule and let

$$r_\rho(\boldsymbol{x};\boldsymbol{\theta}) = \operatorname{sign}(\rho(\boldsymbol{x};\boldsymbol{\theta}))$$

be the classifier based on $\rho(\cdot;\boldsymbol{\theta}) : \mathbb{X} \to \mathbb{R}$. Here $\boldsymbol{\theta} = \left(\alpha, \boldsymbol{\beta}^\top\right)^\top \in \mathbb{R}^{d+1}$ is a parameter vector and

$$\operatorname{sign}(r) = \begin{cases} -1 & \text{if } r \leq 0, \\ 1 & \text{otherwise.} \end{cases}$$

Consider the least-squares loss function

$$\ell(\boldsymbol{x}, y, \rho(\boldsymbol{x})) = [1 - y\rho(\boldsymbol{x})]^2$$

and define the estimator

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}\in\mathbb{R}^{d+1}}{\arg\min}\ \bar{L}_n(\rho(\cdot;\boldsymbol{\theta})) + \lambda\|\boldsymbol{\beta}\|_2^2,$$

[2] Again, each row of the CSV file is a time series and each column is a time point.

where $\lambda > 0$ is a fixed penalty constant and

$$\bar{L}_n\left(\rho\left(\cdot;\boldsymbol{\theta}\right)\right) = \frac{1}{n}\sum_{i=1}^{n}\ell\left(\boldsymbol{X}_i, Y_i, \rho\left(\boldsymbol{X}_i\right)\right),$$

is the empirical risk. We say that the classifier

$$r_\rho\left(\boldsymbol{x};\hat{\boldsymbol{\theta}}\right) = \operatorname{sign}\left(\rho\left(\boldsymbol{x};\hat{\boldsymbol{\theta}}\right)\right)$$

is the so-called linear least-squares support vector machine.

Part a [2 Marks]

Using the information from the problem description, for any fixed $\lambda > 0$, **provide a closed-form expression for the estimator $\hat{\boldsymbol{\theta}}$**.

Part b [2 Marks]

A realization of a random sample $n = 1000$ observations $\boldsymbol{Z}_i = (\boldsymbol{X}_i, Y_i)$ for $i \in [n]$ is contained in the files `p3_1x.csv` and `p3_1y.csv`. Here the feature data $\mathbf{X}_n = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)$ are contained in `p3_1x.csv`[3] and the label data $\mathbf{Y}_n = (Y_1, \ldots, Y_n)$ are contained in `p3_1y.csv`[4].

For $\lambda = 1$, using the estimator from **Part a**, **provide an explicit form of the linear least-squares support vector machine classifier based on the provided data and plot the decision boundary. Explore whether different values of $\lambda > 0$ change the decision boundary and propose some strategy to choose the value using $\mathbf{Z}_n$.**

Part c [2 Marks]

A realization $\mathbf{z}_n = (\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n)$ of a random sample $n = 1372$ observations $\boldsymbol{Z}_i = (\boldsymbol{X}_i, Y_i)$, for $i \in [n]$, is contained in the file `data_bank_authentification.txt`. The data set consists of features extracted from genuine and forged banknote-like documents that were digitized into gray-scale images.

Features of the image are then extract to form the feature vector (i.e. $\boldsymbol{x}_i$) of dimension $d = 4$, which are stored in the first four columns of the data set. The features are the variance (`variance`), skewness (`skewness`) and kurtosis (`kurtosis`) of a wavelet transformation of the image, and the entropy (`entropy`) of the image. All of the features can be considered real-valued. The final column of the data set contains the class label, where a label of zero indicates a genuine banknote and a label of 1 indicates a forgery[5].

[3] Each row of the CSV file is a feature vector of dimension 2.

[4] Note that the label data are not in the appropriate form for use within the large-margin framework.

[5] You will have to transform the label data to the appropriate form for use within the large-margin framework.

Implement a linear least-squares support vector machine classifier and provide an explicit form the decision boundary. You should use the convention that each realization can be written as

$$\begin{aligned}\boldsymbol{x}_i &= (\texttt{variance}_i, \texttt{skewness}_i, \texttt{kurtosis}_i, \texttt{entropy}_i) \\ &= (x_{i1}, x_{i2}, x_{i3}, x_{i4})\end{aligned}$$

and that $y_i = -1$ indicates a genuine banknote.

Part d [1 Marks]

Let $r\left(\boldsymbol{x}; \hat{\boldsymbol{\theta}}\right)$ denote the classifier from **Part c**, and **compute an estimate of the in-sample empirical classification risk**

$$\bar{L}_n\left(r\left(\cdot; \hat{\boldsymbol{\theta}}\right)\right) = \frac{1}{n}\sum_{i=1}^{n} \left[\!\left[r\left(\boldsymbol{X}_i; \hat{\boldsymbol{\theta}}\right) \neq Y_i \right]\!\right].$$

Then, **visualize the data in a manner that displays the realizations that are misclassified by $r\left(\cdot; \hat{\boldsymbol{\theta}}\right)$, in the sense that $\boldsymbol{x}_i$ is misclassified if $r\left(\boldsymbol{x}_i; \hat{\boldsymbol{\theta}}\right) \neq y_i$.**

Part e [1 Marks]

Upon inspection of the plot (or plots) from **Part d**, **discuss why a linear classifier is insufficient for the task of distinguishing between banknotes using the available data. Suggest some modifications to the least-squares support vector machine construction from the problem description that would alleviate any perceived inadequacies of the linear classifier.**

Part f [2 Marks]

Implement your suggested modifications from Part e and compare the performance of your suggested classifier via visualization of the data and estimation of the in-sample empirical classification risk.