Statistical Machine Learning for Data Science
Assignment 2

Due Date: Sunday, January 21, 2024 - By 11:59pm

Exercise 1: Nearest Neighbors as a Kernel Method

Now let's consider a binary response Y with the specific coding of $Y \in \{-1, +1\}$. Let the kNearest Neighbors learning machine be the method under consideration. Recognizing that the $\mathrm{w}_i(\mathbf{x})$ is the inverse of the distance (dissimilarity) between $\mathbf{x}$ and $\mathbf{x}_i$, we adopt the reformulation of the weight $\mathrm{w}_i(\mathbf{x})$ as a measure of the similarity between $\mathbf{x}$ and $\mathbf{x}_i$, and using a bivariate function denoted by $\mathcal{K} : \mathcal{X} \times \mathcal{X} \longrightarrow \mathbb{R}_+$, we now write for the negative exponential weighting scheme

$$\mathcal{K}(\mathbf{x}, \mathbf{x}_i) = \mathrm{w}_i(\mathbf{x}) = \frac{e^{-\gamma d(\mathbf{x}, \mathbf{x}_i)}}{\displaystyle\sum_l e^{-\gamma d(\mathbf{x}, \mathbf{x}_l)}}$$

where γ is real positive number representing the bandwidth. Let $\alpha_i = \mathbb{1}(\mathbf{x}_i \in \mathcal{V}_k(\mathbf{x}))$ be the neighborhood indicator. Note that $\mathcal{V}_k(\mathbf{x}) \subseteq \mathcal{D} \subset \mathcal{X}$.

1. Explain clearly why and how $\widehat{\mathrm{f}}_{\texttt{kNN}}(\mathbf{x})$ does indeed compute the predicted label of $\mathbf{x}$

$$\widehat{\mathrm{f}}_{\texttt{kNN}}(\mathbf{x}) = \texttt{sign}\left(\sum_{i=1}^{n} \mathrm{y}_i \alpha_i \mathcal{K}(\mathbf{x}, \mathbf{x}_i)\right)$$

2. Explain succinctly what $\widehat{\pi(\mathbf{x})}$ is estimating in this case

$$\widehat{\pi(\mathbf{x})} = \sum_{i=1}^{n} \mathbb{1}(\mathrm{y}_i = 1)\alpha_i \mathcal{K}(\mathbf{x}, \mathbf{x}_i)$$

3. Explain clearly the relationship between $\widehat{\mathrm{f}}_{\texttt{kNN}}(\mathbf{x})$ and

$$\widehat{\mathrm{g}}_{\texttt{kNN}}(\mathbf{x}) = 2\,\mathbb{1}\left(\widehat{\pi(\mathbf{x})} > \frac{1}{2}\right) - 1$$

4. Consider the function space $\mathscr{H}$ and the task of finding $f \in \mathscr{H}$ with $f : \mathscr{X} \longrightarrow \mathscr{Y}$.

$$\mathscr{H} := \left\{ f \text{ s.t } \forall \mathbf{x} \in \mathscr{X},\ \ f(\mathbf{x}) = \mathbb{1}\left(\widehat{\pi(\mathbf{x})} > \frac{1}{2}\right) \right\}$$

   What then is $\mathscr{Y}$ in this case?

5. Specify any other function (hypothesis) space defined in a similar way that we explored just recently during the dissection of the seven wheels of statistical machine learning.

Exercise 2: When can we compute the Bayes' Risk?

Note: This exercise is provided **solely** as a way to help you comprehend the concepts of decision boundary and aspects of the Bayes' learning machine and the corresponding Bayes' risk in th context of binary classification in the plane (2-dimensional Euclidean space)

Consider the lecture notes lec-sml@aims-basics-seven-wheels-of-sml-1.pdf. Throughout this exercise, you will find slides 103 and slide 104 very useful.

Let $\mathbf{x} = (\mathrm{x}_1, \mathrm{x}_2)^\top \in \mathbb{R}^2$ be a bivariate predictor (input or explanatory variable), and consider the corresponding response variable $Y \in \{0, 1\}$. Now, consider the task of building a classifier $f : \mathbb{R}^2 \longrightarrow \{0, 1\}$. Let's assume that we are given the class conditional densities

$$p_X(\mathbf{x}|\mathrm{y} = 0) = \frac{1}{\sqrt{(2\pi)^2|\Sigma|}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_0)^\top \Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu}_0)},$$

and

$$p_X(\mathbf{x}|\mathrm{y} = 1) = \frac{1}{\sqrt{(2\pi)^2|\Sigma|}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_1)^\top \Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu}_1)}.$$

It is also revealed from by the data scientist running the case study, that $\Pr[Y = 1] = \psi$, where $\psi \in (0, 1)$. Also $\boldsymbol{\mu}_0 = (-2, -1)^\top$, $\boldsymbol{\mu}_1 = (0, 1)^\top$ and

$$\Sigma = \begin{bmatrix} 1 & -\frac{3}{4} \\ -\frac{3}{4} & 2 \end{bmatrix}.$$

1. Sketch the contours of the two classes in the plane. *You may use software to do this, from Wolfram, to ChatGPT or even* R *or* Python. *Be sure to place the underlying implicit decision boundary on your plot.*

2. Find and write down the simplified expression of the Bayes' classifier for this task, obtained by apply the natural logarithm appropriately inside the indicator function to end with an expression of the form $\mathbf{w}^\top \mathbf{x} + b > 0$.

$$f^\star(\mathbf{x}) = \mathbb{1}\left(\frac{\mathbb{P}[Y = 1|\mathbf{x}]}{\mathbb{P}[Y = 0|\mathbf{x}]} \geq 1\right)$$

3. Deduce the expression of the decision boundary for this task. *Slides 103 and/or 104 mentioned above might be helpful.*

4. Find and write down the theoretical expression of the Bayes' risk for this task.

5. Consider the setting where ψ is assumed to be 1/2. Now recall that $P_{XY}(\mathbf{x}, \mathrm{y})$ denotes the joint probability distribution of X and Y. Using all the above information, compute the value of

$$R^\star = \min_f \int_{\mathbb{R}^2 \times \{0,1\}} \mathbb{1}(\mathrm{y} \neq f(\mathbf{x})) dP_{XY}(\mathbf{x}, \mathrm{y}).$$

Exercise 3: Detecting and Recognizing Speaker Accent

The datasets `accent-raw-data-1.csv` and `accent-mfcc-data-1.csv` respectively contain audio tracks and transformed audio tracks of a total of 328 readings of English words. Most of the readings are done by US born speakers of English while the remaining ones are done by speakers born outside the US. The transformed version of the data was processed using an adaptation of Fourier like transforms known as MFCC.

1. First consider the dataset The datasets `accent-raw-data-1.csv`.
   1. Comment on the peculiarities of the dataset from a point of view dimensionality.
   2. Use the `ts.plot()` function to plot speakers $\{9, 45, 81, 99, 126, 234\}$. You must make this into a numeric

      ```
      xy <- read.csv('accent-raw-data-1.csv')
      x  <- as.matrix(xy[,-1])
      y  <- xy[,1]
      ```

   3. Comment comparatively on the features of the plotted soundtracks. For the fun of it, consider using the function `play()` from the `library(audio)` to hear the sound of each of the speakers. plotted.
   4. Comment on the use of Classification Trees as learning machines for classifying the speaking accent using this data.
   5. Comment on the use of kNearest Neighbors as learning machines for classifying the speaking accent using this data.
2. Consider now the dataset `accent-mfcc-data-1.csv` along with the binary classification task of US Born versus Non-US Born speakers. You are to compare the following methods of classification: (1) kNearest Neighbors (2) Trees (3) Support Vector Machines
   1. Generate separate confusion matrices for each of the three methods
   2. Plot comparative ROC curves for all the three methods
   3. Use a $60\% - 40\%$ Training-Test set split to generate comparative boxplots of the test error based on 100 replications.
   4. Comment on the predictive performances.
   5. Reconsider the confusion matrix of the best method and comment on the similarity between speaking accents.

Warning: On the Use of AI Large Language Models (LLMs)

You are absolutely allowed to use large language models (LLMs) like OpenAI ChatGPT, Google bard, Microsoft bing, Entropic claude or any other. However, you must be careful not to expect it to give you the right answer automatically. Those AI systems make catastrophic mistakes and generate disturbing awful errors. First of all, using those mindlessly is likely to waste your time as you seek to correct errors. Second of all, if your submission appears to be a mindless copy of output from an AI system, you will lose substantial of points and may incur a zero for that exercise! Using these AI tools as aids and assistants to personal exploration towards understanding and mastery is the only wise use. Selling your soul to them in the hope of dispensing of thinking and personal effort is both unwise and potential harmful.