ECE 445 (Fall 2020) – Notebook Exercise #4 (60 points)

Last updated: November 9, 2020

Rationale and learning expectations: This exercise reinforces fundamental classification concepts within machine learning, which range from the ideal Bayes classifier to k-nearest neighbor classifier, as well as introduces students to various performance metrics that are often used to evaluate the performance of classification methods. The data being used in this exercise has been synthetically generated from two distinct two-dimensional Gaussian distributions, which allows us to compare different classification methods with respect to the ideal Bayes decision boundary. At the conclusion of this activity, students attempting this exercise are expected to understand the key aspects of classification methods derived under the principle of minimization of the 0–1 loss function.

General Instruction: All parts of this exercise must be done within a Notebook, with text answers (and other discussion) provided as markdown / LaTeX cells. Please refer to the solution template for Exercise #4 as a template for your own submission of this exercise. In addition, make sure that the version of the notebook submitted by you has fully executed cells (i.e., submit it after a complete run of all the cells in the notebooks).

Restrictions: You are free to use `numpy`, `pandas`, `scipy.stats`, `matplotlib`, `mpl_toolkits.mplot3d`, `seaborn`, and `IPython.display` packages within your code. Unless explicitly permitted by the instructor, you are not allowed to use any other packages or modules.

Notebook Preamble: In the case of this exercise, I suggest importing the different packages/modules in the preamble as follows (but you are allowed to use any other names of your liking):

```
import numpy as np
import pandas as pd
from scipy import stats as sps
from matplotlib import pyplot as plt
from mpl_toolkits.mplot3d import Axes3D
from IPython.display import display, Latex
```

1 Machine Learning for Diagnosis of 'Senioritis'

Medical researchers at Rutgers University have isolated two chemicals, named `ChemA` and `ChemB`, within the blood serum of college students that can potentially help them diagnose the debilitating *Senioritis* disease. The researchers are wondering whether machine learning can be used for automated diagnosis of Senioritis and, to this end, they have recruited 400 college students into a medical study. The data collected from these 400 students, 200 of which are suffering from Senioritis, correspond to concentrations of the two chemicals (`ChemA` and `ChemB`) within their blood serums. These data are further equally and randomly divided into *training data* and *test data*, with each training and test dataset having exactly 100 students without Senioritis and 100 students with Senioritis. These training and test data are then stored in `csv` files, respectively named as `SenioritisTrainingData.csv` and `SenioritisTestData.csv`, which are being provided to you to help the medical researchers finalize a machine learning algorithm for diagnostic purposes.

1.1. (4 points) Provide two scatter plots, one for training data and one for test data, in which samples of healthy students are colored green and samples of students suffering from Senioritis are colored red. The horizontal axes of these scatter plots should correspond to the concentration of `ChemA` and their vertical axes should correspond to the concentration of `ChemB` for each sample. The plots should have their axes appropriately labeled and should have legends that help distinguish between samples having different labels.

1.2. Use the training data to train the following machine learning classification methods that *must* be implemented from scratch, i.e., you are not allowed to use libraries such as `scikit-learn` for these classification methods.

   (a) (5 points) *Linear discriminant analysis* (LDA)

 waheed.bajwa@rutgers.edu

(b) (5 points) *Quadratic discriminant analysis* (QDA)

(c) (5 points) *Gaussian naïve Bayes* (GNB) classifier

(d) (2 points) *k-nearest neighbor* (k-NN) classifier

1.3. Use the test data to evaluate the performance of each of the four classification methods by reporting the following performance metrics for *each* classifier, where you should use the Euclidean distance metric and $k = 3$ for the k-NN classifier.

(a) (8 points) Empirical probability of misclassification (*aka*, (empirical) probability of error)

(b) (2 points) *True positives* (TP), defined as the number of students with Senioritis that are correctly classified as having Senioritis, and *true positive rate* (TPR), defined as the fraction of students with Senioritis that are correctly classified as having Senioritis

(c) (2 points) *False positives* (FP), defined as the number of students without Senioritis that are incorrectly classified as having Senioritis, and *false positive rate* (FPR), defined as the fraction of students without Senioritis that are incorrectly classified as having Senioritis

(d) (2 points) *True negatives* (TN), defined as the number of students without Senioritis that are correctly classified to be without Senioritis, and *true negative rate* (TNR), defined as the fraction of students without Senioritis that are correctly classified to be without Senioritis

(e) (2 points) *False negative rate* (FNR), defined as the number of students with Senioritis that are incorrectly classified to be without Senioritis, and *false negative rate* (FNR), defined as the fraction of students with Senioritis that are incorrectly classified to be without Senioritis

Note: Several other performance metrics / metric names are used in the machine learning literature to characterize the performance of classification methods. Some of these include:

- *Precision*, which is equal to TP/(TP + FP)
- *Sensitivity* and *recall*, both of which simply mean TPR
- *Specificity* and *selectivity*, both of which simply mean TNR
- *Probability of false alarm*, which simply means FPR
- *Probability of missed detection*, which simply means FNR

1.4. (3 points) Based on your inspection of the performance metrics on the test data, which classifier would you recommend for automated diagnostics of Senioritis? Justify your answer.

1.5. Display the following classification decision boundaries, along with appropriate legends, by overlaying them on top of the scatter plot of training data.

(a) (2 points) Decision boundary corresponding to the trained LDA classifier

(b) (2 points) Decision boundary corresponding to the trained QDA classifier

(c) (2 points) Decision boundary corresponding to the trained GNB classifier

(d) (2 points) Decision boundary corresponding to the k-NN classifier ($k = 3$)

(e) (2 points) Decision boundary corresponding to the ideal Bayes classifier, based on the assumption that the concentrations of the two chemicals, `ChemA` and `ChemB`, when stacked into a two-dimensional vector $\mathbf{x} = \begin{bmatrix}\texttt{ChemA} & \texttt{ChemB}\end{bmatrix}^{\mathrm{T}}$, follow a two-dimensional Gaussian distribution. That is,

$$\begin{aligned}\mathbf{x}|Y = \{\textsf{No Senioritis}\} &\sim \mathcal{N}(\boldsymbol{\mu}_0, \mathbf{C}_0), \text{ and}\\ \mathbf{x}|Y = \{\textsf{Senioritis}\} &\sim \mathcal{N}(\boldsymbol{\mu}_1, \mathbf{C}_1),\end{aligned}$$

where $\boldsymbol{\mu}_0 = \begin{bmatrix}-2 & 2\end{bmatrix}^{\mathrm{T}}$, $\boldsymbol{\mu}_1 = \begin{bmatrix}2 & 0\end{bmatrix}^{\mathrm{T}}$, $\mathbf{C}_0 = \begin{bmatrix}3 & 0\\ 0 & 1\end{bmatrix}$, and $\mathbf{C}_1 = \begin{bmatrix}2 & -1\\ -1 & 2\end{bmatrix}$.

1.6. Consider the k-NN classifier with the Euclidean distance metric and evaluate the probability of error on both training data and test data, i.e., training error and test error, as a function of k for $k = 1, 2, \ldots, 10$.

(a) (6 points) Provide plots of training error and test error, as a function of k, overlayed on top of each other in a single figure. Appropriately label the axes and add legends to the figure.

(b) (4 points) Based on your inspection of the training and test errors, what value of k would you recommend for the k-NN classifier that should go into production for diagnostics of Senioritis? Justify your answer.

 waheed.bajwa@rutgers.edu