Assignment 2

Homework assignments will be done individually: each student must hand in their own answers. Use of partial or entire solutions obtained from others or online is strictly prohibited. Electronic submission on Canvas is mandatory.

1. **Linear Discriminant Analysis** (20 points) Please download the Iris data set from the UCI Machine Learning repository and implement Linear Discriminant Analysis for each pair of the classes and report your results. Note that there are three (3) class labels in this data set. Write down each step of your solution. **Do not use any package/tool**.

2. **Generative methods vs Discriminative methods** (50 points) Please download the breast cancer data set from UCI Machine Learning repository. **Do not use any package/tool for implementing the algorithms; You can use packages for matrix/vector operations and data processing**.

   1. (10 pts) Show that the derivative of the error function in Logistic Regression with respect to $\mathbf{w}$ is:

   $$\nabla_{\mathbf{w}} E(\mathbf{w}) = \sum_{n=1}^{N} (f(\mathbf{x}_n) - y_n)\mathbf{x}_n$$

   2. (20 pts) Implement a logistic regression classifier with maximum likelihood (ML) estimator using Stochastic gradient descent and Mini-Batch gradient descent algorithms. Divide the data into training and testing. Choose a proper learning rate. Use cross-validation on the training data to choose the best model and report the recall, precision, and accuracy on malignant class prediction (class label malignant is positive) on the test data using the best model. Write down each step of your solution.
   3. (20 pts) Implement a probabilistic generative model (the one in our lecture) for this problem. Use cross-validation on the training data and report the recall, precision, and accuracy on malignant class prediction (class label malignant is positive) on the test data using the best model. Write down each step of your solution.

3. **Naive Bayes** (20 points) From Project Gutenberg, we downloaded two files: The Adventures of Sherlock Holmes by Arthur Conan Doyle (pg1661.txt) and The Complete Works of Jane Austen(pg31100.txt). Please develop a multinomial Naive Bayes Classifier that will learn to classify the authors from a snippet of text into: Conan Doyle or Jane Austen. A multinomial Naive Bayes uses a feature vector $\mathbf{x} = \{x_1, ..., x_D\}$ as a histogram and model the posterior probability as:

$$p(C_k|\mathbf{x}) \propto p(C_k) \prod_{i=1}^{D} p(x_i|C_k) \tag{1}$$

where $p(x_i|C_k)$ can be estimated by the number of times word i was observed in class C_k plus a smoothing factor divided by the total number of words in C_k

In the testing phase, given a new example $\mathbf{x}_t$, you can output the class assignment for this example by comparing $\log p(C_1|\mathbf{x}_t)$ and $\log p(C_2|\mathbf{x}_t)$. If $\log p(C_2|\mathbf{x}_t) > \log p(C_1|\mathbf{x}_t)$, assign C_2 to this example.

You need to divide the data into training and testing. Make sure the testing data has equal number of samples from Conan Doyle and Jane Austen. Report accuracy on test data using your Naive Bayes classifier.**Do not use any package/tool**.

4. **Linear classification** (10 points) Please prove that 1) the multinomial naive Bayes classifier in log-space essentially translates to a linear classifier. 2) Logistic regression is a linear classifier.

Please follow the below instructions when you submit the assignment.

1. You are allowed to use packages for preprocessing data, and cross-validation
2. You shall submit a zip file named Assignment2_LastName_FirstName.zip which contains:
   - a pdf file contains all your solutions for the written part
   - python files (jupyter notebook or .py files)