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**.

   First we read in the dataset and replace the class labels with numbers to make it easier. Then we compute the mean vectors that contains the mean for the sepal length, sepal width, petal length, and petal width for each class. Next, we compute the scatter matrices which contains the within-class and between-class for each pair of classes. Next, we solve the generalized eigenvalue problem for each pair of classes. We use this to select linear discriminants for the new feature subspace. With these eigenvalues, we sort the eigenvectors by decreasing eigenvalues. We select the 2 most informative eigenpairs for each pair of classes thereby reducing the initial 4-dimensional feature space into a 2-dimensional feature subspace. We transform the samples onto the new subspace for each pair. The primary use of LDA is dimensional reducation for pre-processing for other applications. By reducing the dimensions, it makes the program easier to solve.

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:

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

   The loss function is:

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

   Therefore the derivative of the error function is:

   $$\nabla_{\mathbf{w}} E(W) =$$

   $$= -\sum_{n=1}^{N} \frac{\sigma(\mathbf{w}^T\mathbf{x}_n + \omega_0)(1 - \sigma(\mathbf{w}^T\mathbf{x}_n + \omega_0))\mathbf{x}_n y_n}{\sigma(\mathbf{w}^T\mathbf{x}_n + \omega_0)} - \frac{\sigma(\mathbf{w}^T\mathbf{x}_n + \omega_0)(1 - \sigma(\mathbf{w}^T\mathbf{x}_n + \omega_0))\mathbf{x}_n(1 - y_n)}{1 - \sigma(\mathbf{w}^T\mathbf{x}_n + \omega_0)}$$

$$= -\sum_{n=1}^{N}(1-\sigma(\mathbf{w}^T\mathbf{x}_n+\omega_0))\mathbf{x}_n y_n - \sigma(\mathbf{w}^T\mathbf{x}_n+\omega_0)\mathbf{x}_n(1-y_n)$$

$$= -\sum_{n=1}^{N}(\mathbf{x}_n y_n - \mathbf{x}_n y_n \sigma(\mathbf{w}^T\mathbf{x}_n+\omega_0) - \sigma(\mathbf{w}^T\mathbf{x}_n+\omega_0)\mathbf{x}_n + \sigma(\mathbf{w}^T\mathbf{x}_n+\omega_0)\mathbf{x}_n y_n)$$

$$= -\sum_{n=1}^{N}(\mathbf{x}_n y_n - \mathbf{x}_n \sigma(\mathbf{w}^T\mathbf{x}_n+\omega_0)) = \sum_{n=1}^{N}(f(\mathbf{x}_n)-y_n)\mathbf{x}_n$$

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.

   First, I got the dataset and divided it into X, which was the predictors, and y, which was whether it was malignant or benign. Then using MinMaxScaler from sklearn, I scaled the features, X. Then, split the data into train and test sets. I then implement Logistic Regression using Stochastic gradient descent and using Mini-Batch gradient descent. I then created a classification report for both models to compare them. Stochastic gradient descent came out on top by hvaing the higher accuracy along with precision, recall, and F1 score.

Classification Report for SGD logistic Regression:

	precision	recall	f1-score	support
0	0.93	0.99	0.96	67
1	0.99	0.96	0.97	121
accuracy			0.97	188
macro avg	0.96	0.97	0.97	188
weighted avg	0.97	0.97	0.97	188

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.

   First, I got the dataset and divided it into X, which was the predictors, and y, which was whether itwas malignant or benign. Then using MinMaxScaler from sklearn, I scaled the features, X. Then, split the data into train and test sets. Implementation of probabilistic generative model showed that the accuracy compared to logistic regression was less. This leads us to believe that this model is not as good. This is understandable since this model only gets one round of training compared to the epoch number of rounds logistic regression gets.

Classification Report for Probabilistic Generative Model:

	precision	recall	f1-score	support
0	0.95	0.88	0.91	67
1	0.94	0.98	0.96	121
accuracy			0.94	188
macro avg	0.94	0.93	0.94	188
weighted avg	0.94	0.94	0.94	188

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.

The multinomial naive Bayes classifier in log-space essentially translates to a linear classifier:
You can write any naive bayes classifier as,

$$p(c = 1|\mathbf{x}) = \frac{p(\mathbf{x}|c = 1)p(c = 1)}{p(\mathbf{x}|c = 1)p(c = 1) + p(\mathbf{x}|c = 0)p(c = 0)} = \frac{1}{1 + \frac{p(\mathbf{x}|c=0)p(c=0)}{p(\mathbf{x}|c=1)p(c=1)}}$$

$$= \frac{1}{1 + \exp(-\log \frac{p(\mathbf{x}|c=1)p(c=1)}{p(\mathbf{x}|c=0)p(c=0)})} = \sigma\bigg(\sum_i \log \frac{p(x_i|c = 1)}{p(x_i|c = 0)} + \log \frac{p(c = 1)}{p(c = 0)}\bigg)$$

where σ is logistic function. if $p(x_i|c)$ is from exponential family, we can write it as:

$$p(x_i|c) = h_i(x_i) \exp(\mathbf{u_{ic}^T} \phi_\mathbf{i}(\mathbf{x_i}) - \mathbf{A_i}((\mathbf{u_{ic}}))),$$

and hence:

$$p(c = 1|\mathbf{x}) = \sigma\bigg(\sum_i (\mathbf{w}_i^T \phi_i(x_i) + b)\bigg)$$

where:

$$\mathbf{w}_i = \mathbf{u}_{i1} - \mathbf{u}_{i0},$$

$$b = \log \frac{p(c = 1)}{p(c = 0)} - \sum_i (A_i(\mathbf{u}_{i1}) - A_i(\mathbf{u}_{i0}))$$

This is similar to logistic regression. The feature space is defined by the ϕ_i. For more than two classes, we similarly get multinomial logistic (or) softmax regression. If $p(x_i|c)$ is gaussian, then $\phi_i(x_i) = (x_i, x_i^2)$ and we should have:

$$w_{i1} = \frac{\mu_i}{\sigma_1^2} - \frac{\mu_0}{\sigma_0^2},$$

$$w_{i2} = \frac{2}{\sigma_0^2} - \frac{2}{\sigma_1^2},$$

$$b_i = \log \sigma_0 - \log \sigma_1,$$

assuming $p(c = 1) = p(c = 0) = \frac{1}{2}$.

Logistic regression is a linear classifier:
Logistic regression is linear regression in the sense that the prediction can be written as,

$$\hat{p} = \frac{1}{1 + e^{-\hat{\mu}}}, \text{where} \hat{\mu} = \hat{\theta} \cdot x.$$

Thus, the prediction can be written in terms of $\hat{\mu}$, where is a linear function of x. (More precisely, the predicted log-odds is a linear function of x).

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)