Assignment 2

CS 5824: Advanced Machine Learning
Spring 2022

Due Date: Mar 10th, 2022

- Feel free to talk to other members of the class when doing the homework. We are more concerned that you learn how to solve the problem than that you demonstrate that you solved it entirely on your own. You should, however, write down your solution only yourself. Please try to keep the solution brief and clear.
- Please use the discussion section on Canvas (`https://canvas.vt.edu/courses/145285/discussion_topics`) first if you have questions about the homework. Also feel free to send us e-mails and come to office hours.
- The homework is due at 11:59 PM on the due date. We will be using Canvas (`https://canvas.vt.edu/courses/145285`) for collecting homework assignments. Please do NOT hand in a scan of your handwritten solution, only the typed solution (e.g., Microsoft Word, Latex, etc) will be accepted for grading. Contact the TAs if you are having technical difficulties in submitting the assignment. We do NOT accept late homework!
- The homework submission should include two files: (1) a PDF using the name convention `yourFirstName-yourLastName.pdf`, and (2) a Python file using the name convention `yourFirstName-yourLastName.py`.
- For each question, you will NOT get full credit if you only give out a final result. Necessary calculation steps are required. If the result is not an integer, round your result to 3 decimal places.

Problem 1. Naïve Bayes Classifier (**20** points total)
Given the training data set shown in Table 1, we train a Naïve Bayes classifier with it. Each row refers to an apple, where the categorical features (size, color, and shape) and the class label (whether one apple is good) are shown.

(a) (5 points) How many independent parameters would be there for the Naïve Bayes classifier trained with this data? What are they? Justify your answer.

(b) (10 points) Using standard MLE, what are the estimated values for these parameters?

(c) (5 points) Given a new apple with features $x = (Small, Red, Circle)$, what is $P(y = No|x)$? Would the Naïve Bayes classifier predict $y = Yes$ or $y = No$ for this apple?

Table 1: Training Data for Naïve Bayes Classifier

RID	Size	Color	Shape	Class: good_apple
1	Small	Green	Irregular	No
2	Large	Red	Irregular	Yes
3	Large	Red	Circle	Yes
4	Large	Green	Circle	No
5	Large	Green	Irregular	No
6	Small	Red	Circle	Yes
7	Large	Green	Irregular	No
8	Small	Red	Irregular	No
9	Small	Green	Circle	No
10	Large	Red	Circle	Yes

Problem 2. Support Vector Machines (**15** points total)
Given a linearly separable data set where each sample is either from class $y_i = 1$ or class $y_i = -1$. Then a linear SVM can always find the separating hyperplane with the maximum margin. Assuming we know the two nearest training samples from different classes are $\mathbf{x}_j, y_j = 1$ and $\mathbf{x}_k, y_k = -1$, and the parameters for the linear SVM are $\mathbf{w}$ and b. The linear SVM is trying to maximize the margin as:

$$\arg\max_{\mathbf{w},b} d = d^+ - d^- = \frac{\mathbf{w}^T\mathbf{x}_j}{||\mathbf{w}||_2} - \frac{\mathbf{w}^T\mathbf{x}_k}{||\mathbf{w}||_2} \tag{1}$$

where d^+ (d^-) denotes the distance of $\mathbf{x}_j$ ($\mathbf{x}_k$) to the separating hyperplane.

(a) (5 points) Please formulate an optimization problem (including objective function, constraints) of the "hard-margin" linear SVM (i.e., SVM without slack variables).

(b) (5 points) In practice, it is often the case that data points cannot be well-separated via a "hard margin". Can you provide a solution to solve this problem?

(c) (5 points) Is it possible to apply SVM to multi-class classification? Justify your answer.

Problem 3. Classification Evaluation (**20** points total)
Suppose we have a confusion matrix shown in the following table 2:

Table 2: Confusion Matrix.

Actual \ Predicted	0	1	Total
0	174	7	181
1	23	18	41
Total	197	25	222

Here there are two categories involved, 0 (False) and 1 (True).

(a) (15 points) Given this confusion matrix, calculate the following evaluation measures: (i) Accuracy (ii) Error rate (iii) Precision (iv) Recall (v) F1-measure

(b) (5 points) Suppose this confusion matrix is obtained via evaluating the corresponding model on the test set. What problem can you observe regarding the model/data from the evaluation results? Justify your answer.

Problem 4. Logistic Regression (**45** points total)
Let $\{X_1, X_2\}$ denote two feature dimensions (e.g., age and income), and $Y \in \{0, 1\}$ denote the class labels of the observations. In Figure 1, we have three positive observations ("+" for $Y = 1$) and one negative observation ("-" for $Y = 0$). Please answer the following questions.

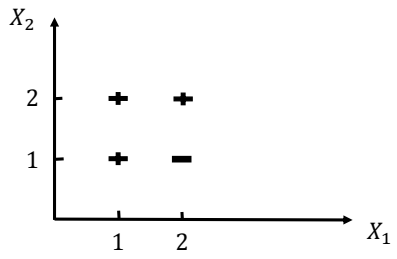

Figure 1: Four data points in a two-dimensional feature space.

(a) (10 points) Is there any logistic regression classifier using X_1 and X_2 that can perfectly classify the examples in Figure 1? What if we change the label of point $(1, 2)$ from "+" to "-"?

(b) (5 points) For predicting the samples in Figure 1, which model is better: Logistic Regression or Linear Regression. Please explain the reason.

(c) (5 points) Both logistic regression and linear SVM are linear classifiers. Explain the difference between these two kinds of models.

(d) (25 points) [**You MUST implement LR from scratch to get points**]
In this sub-problem, you are required to implement a logistic regression classifier to identify digit number 0. You will work on implementing LR using the MNIST dataset, which includes a large amount of handwritten digits. MNIST is very popular in the machine learning and data mining community for verifying learning models. It is accessible at this website `http://yann.lecun.com/exdb/mnist/`. By default, it is split into the training set and testing set, where the training set has 60k images of size 28×28 pixels, and the testing set has 10k images of the same image size. The total number of categories is 10 in the original MNIST dataset, i.e., the digits from 0 to 9. Please note that these images are stored in grayscale, so you will have to treat them as images with **one color channel**. For our case, since the goal is to classify digit number 0, it means

that only digit 0 is positive ($Y = 1$) while all other digits are negative ($Y = 0$). The posterior probabilities are shown as follows:

$$P(Y = 1|X = \mathbf{x}) = \frac{exp(\mathbf{w}^T\mathbf{x})}{1+exp(\mathbf{w}^T\mathbf{x})}$$
$$P(Y = 0|X = \mathbf{x}) = \frac{1}{1+exp(\mathbf{w}^T\mathbf{x})}$$

where $\mathbf{w}^T$ denotes the transpose of the $\mathbf{w}$ weight vector and $\mathbf{w}^T\mathbf{x}$ is the inner product of $\mathbf{w}$ and $\mathbf{x}$. Please check the detailed instructions and requirements below:

- Comment your codes properly to make them readable and understandable. Please note that you **MUST** implement the entire classifier from scratch. No off-the-shelf classification libraries are allowed, but you are allowed to use packages like Numpy and Scipy for the purpose of scientific computing and Pandas for the purpose of data formatting.
- For simplicity, you can flatten each image of size 28×28 pixels into a feature vector $\mathbf{x} \in \mathbb{R}^{784\times 1}$. For each value in $\mathbf{x}$, please divide it by 255 to make it in the range of $[0, 1]$.
- Please use the **gradient ascent algorithm** to maximize the corresponding log-likelihood. To simplify the model, we **IGNORE** the bias term. And please **DO NOT** add regularization.
- Please initialize $\mathbf{w}$ as $\mathbf{w} = \mathbf{0}$, i.e., a vector filled with 0.
- For hyperparameter settings, the number of iterations is 100 and the learning rate is $0.1/N$ where N is the number of images in the training set.
- For the results, please plot (1) the training & testing accuracy vs. the number of iterations and (2) log-likelihood vs. the number of iterations. Write your observations and explanations of the result from both figures. The figures should be placed together with your observations and explanations.
- For submission, you should submit (1) a `.pdf` file that includes everything except the codes (including the figures), and (2) a `.py` file that includes your codes.