Assignment 2
Classification

STAT3612: Statistical machine learning (Spring 2025)

due: **March 16, 2025, Sunday, 11:59 PM**

Goal

The first part of this assignment is to make you familiar with basic concepts and formula derivation in classification models via basic computations. The second part encourages you to grasp the coding skills in basic classification applications. **Note that you should complete this assignment using Python 3.8+**.

Submission

Please submit the following files in Moodle for grading:

- PDF files containing your answers to all questions, including written parts as well as coding parts. You are recommended to convert the Jupyter notebook file into PDF format.
- The completed Jupyter notebook file `assign2.ipynb`.

Part 1: Conceptual Questions and Formula Derivation

Q1 What does 'naive' mean in Naive Bayes? (Single-choice) **[TOTAL: 5 points]**

A. The full Bayes' Theorem is not used. The 'naive' in naive bayes specifies that a simplified version of Bayes' Theorem is used.

B. The Bayes' Theorem makes estimating the probabilities easier. The 'naive' in the name of classifier comes from this ease of probability calculation.

C. The model assumes that the input features are statistically independent of one another. The 'naive' in the name of classifier comes from this naive assumption.

Q2 Normally in which phase are model parameters adjusted? (Single-choice) **[TOTAL: 5 points]**

A. Testing phase.

B. Training phase.

C. Data preparation phase.

D. Model parameters are always constant throughout the modeling process.

Q3 Suppose we have collected data from a group of students in a Statistical machine learning class with variables X_1 = hours studied, X_2 = grade point average, and Y = a binary output indicates if the student received grade 90 ($Y = 1$) or not ($Y = -1$). We learn a logistic regression model:

$$p(Y = 1 \mid X) = \frac{e^{\theta_0+\theta_1 X_1+\theta_2 X_2}}{1 + e^{\theta_0+\theta_1 X_1+\theta_2 X_2}}$$

with parameters $\hat{\theta}_0 = -6, \hat{\theta}_1 = 0.05, \hat{\theta}_2 = 1$. **[TOTAL: 15 points]**

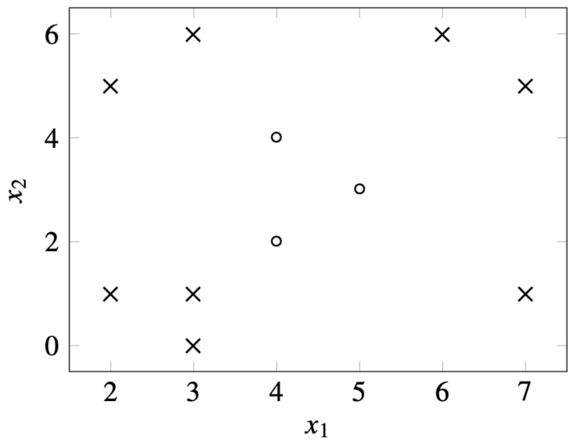

Figure 1: Dataset

(a) According to the logistic regression model, estimate the probability that a student who studies for 40h and has the grade point average of 3.5 gets a 90 grade in the Statistical machine learning class. **[5 points]**

(b) According to the logistic regression model, how many hours would the student in part (a) need to study to have 50% chance of getting a 90 grade in the class? **[10 points]**

Q4 Considering the cross entropy loss for softmax regression and a simple expression can be acquired. We use the linear model $\eta(\hat{\mathbf{x}}) = \hat{\mathbf{W}}\hat{\mathbf{x}}$ and the loss function can be represented as:

$$L(\hat{\mathbf{W}}) = -\frac{1}{n}\sum_{i=1}^{n} \mathbf{y_i} \log \hat{\mathbf{y}}_\mathbf{i} = -\frac{1}{\mathrm{n}}\sum_{\mathrm{i=1}}^{\mathrm{n}} \log \frac{\mathrm{e}^{\eta_{y_\mathrm{i}}(\hat{\mathbf{x}}_\mathbf{i})}}{\mathrm{e}^{\eta_1(\hat{\mathbf{x}}_\mathbf{i})} + \ldots + \mathrm{e}^{\eta_\mathrm{K}(\hat{\mathbf{x}}_\mathbf{i})}}$$

We assume there are K classes in the classification problem.
Prove the following equation

$$\nabla_{\hat{\mathbf{W}}} L(\hat{\mathbf{W}}) = \frac{1}{n}[\mathrm{softmax}(\hat{\mathbf{W}}\hat{\mathbf{X}}) - \mathbf{Y}]\hat{\mathbf{X}}^\mathrm{T}$$

Note that $\hat{\mathbf{W}} \in \mathbb{R}^{K\times(p+1)}$, $\hat{\mathbf{X}} = [\hat{\mathbf{x}}_1, \ldots, \hat{\mathbf{x}}_\mathbf{n}] \in \mathbb{R}^{(p+1)\times n}$, and $\mathbf{Y} \in \mathbb{R}^{K\times n}$. **[TOTAL: 5 points]**

Hint: You may try to first calculate the derivative of each item in the summation.

Q5 Consider a neural network used for multi-class classification with K classes. The output layer uses a softmax activation function to produce the predicted probabilities $\hat{\mathbf{y}} = \{\hat{y}_1, \hat{y}_2, ..., \hat{y}_K\}$, and the true label is represented by a one-hot encoded vector $\mathbf{y} = \{y_1, y_2, ..., y_K\}$. The cross-entropy loss is defined as:

$$L(\hat{\mathbf{y}}, \mathbf{y}) = -\sum_{i=1}^{K} y_i \log(\hat{y}_i)$$

Which of the following statements about cross-entropy loss is true? **[TOTAL: 5 points]**

A. The cross-entropy loss is minimized when the predicted probability for the true class is exactly 0.5.

B. Cross-entropy loss always outputs a value between 0 and 1.

C. The cross-entropy loss is always non-negative and reaches 0 when the predicted probability for the true class is 1.

D. Cross-entropy loss penalizes correct predictions more than incorrect ones.

Q6 Suppose that we wish to predict whether a given stock will issue a dividend this year (Yes or No) based on X, last years percent profit. We examine a large number of companies and discover that the mean value of X for companies that issued a dividend was $\overline{X}_{\text{Yes}} = 10$, while the mean for those that didnt was $\overline{X}_{\text{No}} = 0$. In addition, the variance of X for these two sets of companies was $\hat{\sigma}^2 = 36$. Finally, 80% of companies issued dividends. Assuming that X follows a normal distribution, predict the probability that a company will issue a dividend this year given that its percentage profit was $X = 14$ last year. **[TOTAL 5 points]**

Hint: Recall that the density function for a normal random variable is:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

You will need to use Bayes theorem.

Q7 Consider the dataset in the Fig. 1. We will study tree-based methods.

[TOTAL 15 points]

(a) Construct a classification tree using recursive binary splitting and Gini index as impurity measure. Stop when all leaves contains a single class. Write necessary calculation procedure and you can present the tree similar to the one in Fig. 2. **[5 points]**

(b) Draw a graph showing the partitioning induced by the decision boundaries of the classification tree in Fig. 2. **Note: DO NOT draw the partitioning for the tree you got in Q7 (a).** **[5 points]**

(c) Besides tree-based methods, what other classification methods you will use to achieve zero training error (with some (hyper-)parameter setting) for the above dataset? List three other choices and briefly explain your answers. **[5 points]**

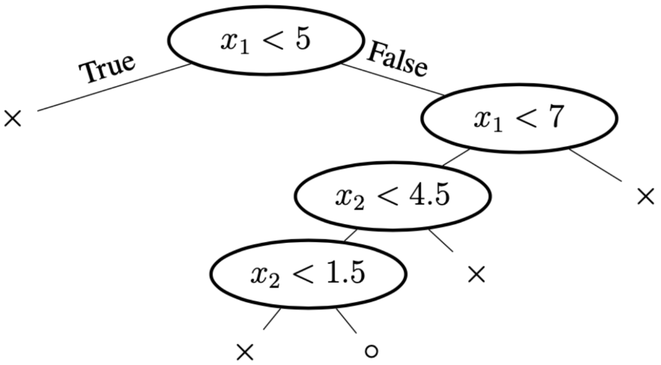

Figure 2: Decision Tree

Part 2: Programming

Q7 Linear Discriminant Analysis. Finish the code in `assign2.ipynb`. **[TOTAL 15 points]**

(a) Implement LDA to classify the MNIST images and print the testing data accuracy. **[5 points]**

(b) Remove all the images that do not corresponds to the digit 0 or 1 in the datasets (train/test). Then perform two-class image classification. Treat image (digit) 1 as positive and image (digit) 0 as negative. Compute Recall, Precision, and F1-Score. **[10 points]**

Hint: You can use the function in sklearn library. Refer to this link.

Q8 Regularized Softmax Regression. Finish the code in `assign2.ipynb`. **[TOTAL 30 points]**

(a) Perform Softmax Regression with L1 penalty for CIFAR-10 image classification. Adjust the weight of the L1 penalty and plot the figure showing accuracy vs. weight curve. Use the package in sklearn for this subquestion. **[15 points]**

(b) Perform Softmax Regression with penalty items for CIFAR-10 image classification. Write the expression of your penalty items and the parameter updating in your SGD step. After training, find the class that corresponds to the highest testing accuracy and visualize one image that belongs to that class. You should implement it from scratch for this subquestion. **[15 points]**

Hint: (a) Refer to this link. You should use 'saga' solver according to the warning box in the page. And you can use a small max_iter (e.g., 10) to cut down training time. **(b)** In each SGD step, we take the derivative of the loss function. You can use a certain kind of penalty items like $L_1 = \|\hat{\mathrm{W}}\|_1 = \sum_{ij} |\hat{w}_{ij}|$. Adding penalty does not means the results will be better but we want to have performance gain after some attempts. Reasonable performances should be obtained by adjusting the weight of the extra loss item.