CSE 575: Statistical Machine Learning Assignment #1

Instructor: Prof. Hanghang Tong
Out: Aug. 21th, 2018; Due: Sep. 13th, 2018

Submit electronically, using the submission link on Blackboard for Assignment #1, a file named `yourFirstName-yourLastName.pdf` *containing your solution to this assignment (a .doc or .docx file is also acceptable, but .pdf is preferred).*

1 Probability [10 points]

1. [2 points] Independent and disjoint.
   - [1 points] If A and B are **INDEPENDENT**, and $P(A) > 0$, $P(B) > 0$, what is the value of $P(A|B)$?
   - [1 points] If A and B are **DISJOINT** events, and $P(A) > 0$, $P(B) > 0$, what is the value of $P(A|B)$?

2. [4 points] Suppose X is a random variable with the following PDF:

$$f(x) = \begin{cases} c(3 + x^3), & \text{for } 0 \leq x \leq 2 \\ 0, & \text{otherwise} \end{cases}$$

   - [2 points] What is the value of c?
   - [2 points] What is the value of $P(X \leq 2)$?

   HINT: what is the definition of PDF?

3. [4 points] Suppose you are getting tested for a heart condition. The probability that the test comes out with accurate results is 0.98 if the patient has the heart condition. However, the probability that the test is falsely positive is 0.0004. Suppose that on average 1 in 1000 people have this heart condition.
   - [2 points] Can you partition the sample space into **2** events? What are they?
   - [2 points] What is the probability that you actually have the disease if the test comes back positive?

   HINT: this is a Bayes Rule problem.

2 Maximum Likelihood Estimation [10 points]

1. [5 points] Given a set of i.i.d samples $x_1, x_2, \cdots, x_n \leq \theta$ following the uniform distribution Uniform$(0, \theta)$. Find the maximum likelihood estimation of the parameter θ.

2. [5 points] Suppose that X is a discrete random variable with the following probability mass function: $P(X = 0) = 3\theta/4; P(X = 1) = \theta/4; P(X = 2) = 2(1 - \theta)/3; P(X = 3) = (1 - \theta)/3$ where $\theta \in [0, 1]$. We have the following 10 i.i.d samples taken from such a distribution: (3,0,2,1,3,2,1,0,2,1). What is the maximum likelihood estimation of the parameter θ?

3 Bayes Classifier [20 points]

1. **Continuous Bayes Classifier**. We want to build a Bayes Classifier for a binary classification task ($y = 1$ or $y = 2$) with a 1-dimensional input feature (x). We know the following quantities: (1) $P(y = 1) = 0.8$; (2) $P(x|y = 1) = 0.5$ for $0 \leq x \leq 2$ and $P(x|y = 1) = 0$ otherwise; and (3) $P(x|y = 2) = 0.5$ for $0 \leq x \leq 4$ and $P(x|y = 2) = 0$ otherwise.

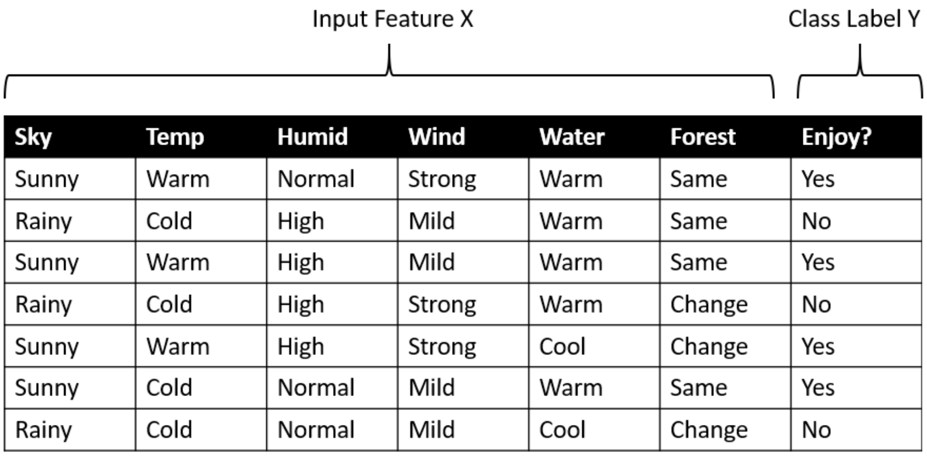
   - [2 points] What is the prior of the class label $y = 2$?
   - [2 points] What is $P(y = 1|x)$?
   - [2 points] For $x = 1$, what is class label your classifier will assign? What is the risk of this decision?
   - [2 points] What is the decision boundary of your Bayes classifier?
   - [2 points] What is the Bayes error of your Bayes classifier?

2. **Discrete Bayes Classifier**. We want to build a Bayes Classifier for a binary classification task ($y = 1$ or $y = 2$) with two binary features (x_1 and x_2). We know the following quantities: (1) $P(y = 1) = 0.6$; (2) $P(x_1 = 0, x_2 = 0|y = 1) = 0.2$, $P(x_1 = 0, x_2 = 1|y = 1) = 0.4$, $P(x_1 = 1, x_2 = 0|y = 1) = 0.1$, and $P(x_1 = 1, x_2 = 1|y = 1) = 0.3$; and (3) $P(x_1 = 0, x_2 = 0|y = 2) = 0.1$, $P(x_1 = 0, x_2 = 1|y = 2) = 0.2$, $P(x_1 = 1, x_2 = 0|y = 2) = 0.3$, and $P(x_1 = 1, x_2 = 1|y = 2) = 0.4$;

   - [2 points] What is the prior of the class label $y = 2$?
   - [2 points] What is $P(y = 1|x_1, x_2)$?
   - [2 points] For an example with the following features $x_1 = 1, x_2 = 0$, what is class label your classifier will assign? What is the risk of this decision?
   - [2 points] What is the decision boundary of your Bayes classifier?
   - [2 points] What is the Bayes error of your Bayes classifier?

4 Naive Bayes Classifier [10 points]

Given the training data set in Figure 1, we want to train a binary classifier, with (1) the last column being the class label (i.e., whether or not to enjoy the sport); and (2) each column of X being a binary feature.


Input Feature X
Class Label Y

Sky	Temp	Humid	Wind	Water	Forest	Enjoy?
Sunny	Warm	Normal	Strong	Warm	Same	Yes
Rainy	Cold	High	Mild	Warm	Same	No
Sunny	Warm	High	Mild	Warm	Same	Yes
Rainy	Cold	High	Strong	Warm	Change	No
Sunny	Warm	High	Strong	Cool	Change	Yes
Sunny	Cold	Normal	Mild	Warm	Same	Yes
Rainy	Cold	Normal	Mild	Cool	Change	No

Figure 1: Training Data Set for Naive Bayes Classifiers.

1. [2 points] How many independent parameters are there in your Naive Bayes classifier? What are they? **ONLY** list the independent parameters. Justify your answer.

2. [3 points] What are your estimations for these parameters? (say using standard MLE).

3. [5 points] Now, given a new (test) example $x = (sunny, cold, high, strong, cool, same)$, what is $P(y = 1|x)$? Which class label will the naive Bayes classifer assign to this example? Justify your answer.

5 Exploring Data [20 points]

Download the SkillCraft data set from UCI Machine Learning Repository (Click Here). The data has 3,395 instances and 20 attributes. The detailed attribute information can be referred to on the website. In this problem, you need to write MATLAB code for some basic feature analysis:

- [6 points] Plot the frequency histogram on 'GapBetweenPACs' and 'ActionsInPAC' attributes, respectively. What is your observation regarding the figures?

- [2 points] Plot the scatter figure for the following pairs of features: (1) UniqueHotkeys vs. AssignToHotkeys, and (2) MinimapAttacks vs. MinimapRightClicks, respectively.

- [12 points] For all the attributes from the 6-th attribute to the 20-th attribute, compute the Pearson Correlation Coefficient (PCC) between each two different attributes and save the results as a matrix in a .mat file or in a .txt file (each line of which is separated by spaces). What are the maximum and minimum PCC values and what are corresponding attribute pairs? Plot the scatter figures for these attribute pairs. What is the main difference between these two scatter figures?

6 Naive Bayes for Review Sentiment Classification [30 points]

Download the movie review data from Blackboard. There are two folders (i.e., training data and test data), each of which includes a folder of positive reviews and a folder of negative reviews. All the reviews in the positive review folders are labeled as 1, whereas the reviews in the negative review folders are labeled as 0. In this problem, the goal is to implement and train a Naive Bayes classifier **from scratch** to predict whether a review in the test data is positive or negative.

Use the bag-of-words model and the number of occurrence of words in each review as its feature. Besides, you assume the positions of words do not matter and each attribute value is independently generated. For Logistic Regression, please **DO NOT** add regularization terms. In case a word in the test data has not been seen in the training data, please use Laplace (add-1) smoothing for Naive Bayes, i.e.,

$$p(w_i|y) = \frac{\text{count}(w_i, y) + 1}{\text{count}(y) + |V|}.$$

where $|V|$ represents the number of elements in the vocabulary set.

- [10 points] Remove the stop words and use the text files to generate the data matrix. Each row of the matrix represents a word (i.e., feature), and each column represents a review (i.e., instance). The value of a certain matrix entry is the count of the word in the corresponding review. For example, a triplet representation of an entry (r, c, v) means the number of occurrence of word-r in the review-c is v. Make sure you generate the data matrix as a **sparse** matrix. Save and submit the data matrix as either a .mat file or a .txt file (each line of which is a nonzero triplet, i.e., $v \neq 0$).

- [20 points] Implement the Naive Bayes classifier without using any existing packages and functions. Plot the learning curve: the classification accuracy vs. the size of the training data. Generate 6 points on the curve, using $[0.1, 0.3, 0.5, 0.7, 0.8, 0.9]$ random fraction of your training set and testing on the full test set each time. Average your results over 5 runs when using each random fraction of the training set.

Submit your source code and results in a .zip file.