Machine Learning
COMP 5630/ COMP 6630/ COMP 6630 - D01

Instructor: Dr. Shubhra ("Santu") Karmaker
TA 1: Dongji Feng
TA 2: Souvika Sarkar
Department of Computer Science and Software Engineering
Auburn University
Fall, 2022

September 3, 2022

Assignment #2
Naive Bayes Classifiers

Submission Instructions

This assignment is due Thursday, September 8, 2022, at 11:59pm. Please submit your solutions via Canvas (https://auburn.instructure.com/). You should submit your assignment as a typeset PDF. Please do not include scanned or photographed equations as they are difficult for us to grade.

Late Submission Policy

The late submission policy for assignments will be as follows unless otherwise specified:

1. 75% credit within 0-48 hours after the submission deadline.
2. 50% credit within 48-96 hours after the submission deadline.
3. 0% credit after 96 hours after the submission deadline.

Tasks

1 Independent Events and Bayes Theorem [20 pts]

1. [**5 Points**] For events A, B prove:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|\neg A)P(\neg A)}$$

($\neg A$ denote the event that A does not occur.)

2. Let X, Y, and Z be random variables taking values in {0, 1}. The following table lists the probability of each possible assignment of 0 and 1 to the variables X, Y, and Z:

	Z = 0		Z = 1	
	X = 0	X = 1	X = 0	X = 1
Y = 0	0.1	0.05	0.1	0.1
Y = 1	0.2	0.1	0.175	0.175

   (a) [**5 Points**] Is X independent of Y ? Why or why not?

   (b) [**5 Points**] Is X conditionally independent of Y given Z? Why or why not?

   (c) [**5 Points**] Calculate $P(X \neq Y|Z = 0)$.

2 Maximum Likelihood Estimation [40 pts]

This problem explores maximum likelihood estimation (MLE), which is a technique for estimating an unknown parameter of a probability distribution based on observed samples. Suppose we observe the values of n i.i.d.[1] random variables X_1, . . . , X_n drawn from a single Bernoulli distribution with parameter . In other words, for each X_i, we know that:

$$P(X_i = 1) = \theta \qquad \text{and} \qquad P(X_i = 0) = 1 - \theta$$

Our goal is to estimate the value of θ from these observed values of X_1 through X_n.

For any hypothetical value $\hat{\theta}$, we can compute the probability of observing the outcome X_1, . . . , X_n if the true parameter value θ were equal to $\hat{\theta}$. This probability of the observed data is often called the data likelihood, and the function $L(\hat{\theta}) = P(X_1, ..., X_n|\hat{\theta})$ that maps each $\hat{\theta}$ to the corresponding likelihood is called the likelihood function. A natural way to estimate the unknown parameter θ is to choose the $\hat{\theta}$ that maximizes the likelihood function. Formally,

$$\hat{\theta}^{MLE} = \underset{\hat{\theta}}{\operatorname{argmax}}\, L(\hat{\theta})$$

Often it is more convenient to work with the log likelihood function $l‘(\hat{\theta}) = logL(\hat{\theta})$. Since the log function is increasing, we also have:

$$\hat{\theta}^{MLE} = \underset{\hat{\theta}}{\operatorname{argmax}}\, l(\hat{\theta})$$

1. [**8 Points**] Write a formula for the log likelihood function, $l(\hat{\theta})$. Your function should depend on the random variables X_1, . . . , X_n, the hypothetical parameter $\hat{\theta}$, and should be simplified as far as possible (i.e., don't just write the definition of the log likelihood function). Does the log likelihood function depend on the order of the random variables?

2. [**8 Points**] Consider the following sequence of 10 samples: $X = (0, 1, 0, 1, 1, 0, 0, 1, 1, 1)$.

   Compute the maximum likelihood estimate for the 10 samples. Show all of your work (hint: recall that if x^* maximizes $f(x)$, then $f'(x^*) = 0$).

[1] iid means Independent, Identically Distributed

3. [**8 Points**] Now we will consider a related distribution. Suppose we observe the values of m iid random variables Y_1,....,Y_m drawn from a single Binomial distribution $B(n, \theta)$. A Binomial distribution models the number of 1's from a sequence of n independent Bernoulli variables with parameter. In other words,

$$P(Y_i = k) = \binom{n}{k} \theta^k (1 - \theta)^{n-k} = \frac{n!}{k!(n-k)!} \cdot \theta^k (1 - \theta)^{n-k}$$

   Write a formula for the log likelihood function, $l(\hat{\theta})$. Your function should depend on the random variables Y_1, . . . , Y_m and the hypothetical parameter $\hat{\theta}$.

4. [**8 Points**] Consider two Binomial random variables Y_1 and Y_2 with the same parameters, $n = 5$ and θ. The Bernoulli variables for Y_1 and Y_2 resulted in $(0, 1, 0, 1, 1)$ and $(0, 0, 1, 1, 1)$, respectively. Therefore, $Y_1 = 3$ and $Y_2 = 3$. Compute the maximum likelihood estimate for the 2 samples. Show your work.

5. [**8 Points**] How do your answers for parts 1 and 3 compare? What about parts 2 and 4? If you got the same or different answers, why was that the case?

3 Implementing Naive Bayes [40 pts]

We will now learn how to use Naive Bayes and Logistic Regression to solve a real world problem: text categorization. Text categorization (also referred as text classification) is the task of assigning documents to one or more topics. For our homework, we will use a benchmark dataset that is frequently used in text categorization problems. This dataset, Reuters-21578, consists of documents that were appeared in Reuters newswire in 1987. Each document was then manually categorized into a topic among over 100 topics. In this homework we are only interested in earn and acquisition (acq) topics, so we will be using a shortened version of the dataset (documents assigned to topics other than "earn" or "acq" are not in the dataset provided for the homework). As features, we will use the frequency (counts) of each word occurred in the document. This model is known as bag of words model and it is frequently used in text categorization. You can download Assignment 2 data from the Canvas. In this folder you will find:

- **train.csv:** Training data. Each row represents a document, each column separated by commas represents features (word counts). There are 4527 documents and 5180 words.
- **train labels.txt:** labels for the training data
- **test.csv:** Test data, 1806 documents and 5180 words
- **test labels.txt:** labels for the test data
- **word indices:** words corresponding to the feature indices.

Implement Naive Bayes. To avoid 0 probabilities, choose a Beta distribution with equal valued parameters as a prior when estimating Naive Bayes parameters using MAP. You may need to implement with log probabilities to avoid underflow.

Train your classifier on the training set that is given. For each of the classifier, report training accuracy, testing accuracy and the amount of time spent training the classifier.

Disclaimers: This assignment re-uses some materials from the publicly available website: CMU Introduction to Machine Learning Course, 10-315, Spring 2019. I personally thank Prof. Maria-Florina Balcan for sharing her teaching materials publicly. This assignment is exculively used for instructional purposes.