CSE 575: Statistical Machine Learning Assignment #1

Instructor: Prof. Jingrui He
Out: Jan 19, 2018; Due: Feb 8, 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 Bayes Classifier [15 points]

Suppose that in your coin flip experiment, you observed a set of α_H heads and α_T tails. Let θ denote the probability of observing heads, whose prior distribution follows $Beta(\beta_H, \beta_T)$, where β_H and β_T are two positive parameters. Prove that the posterior distribution $P(\theta|D)$ (D denotes the observed coin flips) follows $Beta(\beta_H + \alpha_H, \beta_T + \alpha_T)$. What is the mean of $P(\theta|D)$? What is the MAP estimator $\hat{\theta}_{MAP}$ of θ?

2 Parameter Estimation [15 points]

For this question, assume that $x_1, \ldots, x_N \in \mathbb{R}$ are i.i.d samples drawn from the same underlying distribution. Assume that the underlying distribution is Gaussian $N(\mu, \sigma^2)$.

1. (5 points) Let $\hat{\mu}_{MLE}$ denote the MLE estimator of μ. Please prove that $\hat{\mu}_{MLE}$ is unbiased. **Hint: The bias of an estimator of the parameter μ is defined to be the difference between the expected value of the estimator and μ.**

2. (10 points) If the true value of μ is unknown, then the MLE estimator of σ^2 is as follows.

$$\hat{\sigma}^2_{MLE} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \hat{\mu}_{MLE})^2$$

   Please prove that $\hat{\sigma}^2_{MLE}$ is biased.

3 Naïve Bayes Classifier [20 points]

Given the training data set shown in Figure 1, we train a Naïve Bayes classifier with it. Each row refers to a person, where the categorical features (age, income etc.) and the class label (whether he/she buys a computer) are shown.

1. (5 points) How many independent parameters would be there for the Naïve Bayes classifier trained with this data? What are they? Justify the your answers.

2. (10 points) Using standard MLE, what are the estimated values for these parameters?

3. (5 points) Given a new person with features $x = (youth, medium, yes, fair)$, what is $P(y = yes|x)$? Would the Naïve Bayes classifier predict $y = yes$ or $y = no$ for this person?

RID	age	income	student	credit_rating	Class: buys_computer
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle_aged	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle_aged	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	medium	yes	fair	yes
11	youth	medium	yes	excellent	yes
12	middle_aged	medium	no	excellent	yes
13	middle_aged	high	yes	fair	yes
14	senior	medium	no	excellent	no

Figure 1: Training Data for Naïve Bayes Classifier

4 Logistic Regression [20 points]

Suppose we have two positive examples $x_1 = (1, 0)$ and $x_2 = (0, -1)$ and two negative examples $x_3 = (0, 1)$ and $x_4 = (-1, 0)$. Apply standard gradient ascent method to train a logistic regression classifier (without any regularization terms). Initialize the weight vector with two different values and set $w_0^0 = 0$ (e.g. $w_0 = (0, 0, 0)', w_0 = (0, 1, 0)'$). Would the final weight vector (w^*) be the same for the two different initial values? What are the values? Please explain your answer. You may assume the learning rate to be a positive real constant η.

5 Naïve Bayes Classifier and Logistic Regression [30 points]

1. (5 points) **Gaussian Naïve Bayes and Logistic Regression.** Suppose a logistic regression model and a Gaussian Naïve Bayes classifier are trained for a binary classification task $f : X \to Y$ where X is real-valued features $X =< X_1, ..., X_d > \in \mathbb{R}^d$, $Y = \{0, 1\}$ is the binary label. After training, we get the weight vector $w =< w_0, w_1, ..., w_d >$ for the logistic regression model.

   Recall that in Gaussian Naïve Bayes, each feature X_i $(i = 1, ..., d)$ is assumed to be conditional independent given the label Y so that $P(X_i|Y = k) = \mathcal{N}(\mu_{ik}, \sigma_{ik})$ $(k = 0, 1; i = 1, ..., d)$. We assume that the marginal distribution of class labels $P(Y)$ follows Bernoulli$(\theta, 1 - \theta)$ $(P(Y = 1) = \theta, P(Y = 0) = 1 - \theta)$.

   - How many independent parameters are there in this Gaussian Naïve Bayes classifier? What are them?
   - Can we translate w into the parameters of an equivalent Gaussian Naïve Bayes classifier without any extra assumption? If that is the case, justify your answer. Otherwise, please specify what extra assumption(s) you need to complete the translation and explain why.

2. (25 points) **Implementation of Gaussian Naïve Bayes and Logistic Regression.** Compare the two approaches on the bank note authentication dataset, which can be downloaded from

http://archive.ics.uci.edu/ml/datasets/banknote+authentication. Complete description of the dataset can be also found on this webpage. In short, for each row the first four columns are the feature values and the last column is the class label (0 or 1). You will observe the learning curves similar to those Dr. He mentioned in class. Implement a Gaussian Naïve Bayes classifier (recall the conditional independent assumption mentioned before) and a logistic regression classifier. Please write your own code from scratch and **do NOT use existing functions or packages which can provide you the Naïve Bayes Classifier/Logistic Regression class or fit/predict function (e.g. sklearn)**. But you can use some basic linear algebra/probability functions (e.g. numpy.sqrt(), numpy.random.normal()). For the Naïve Bayes classifier, assume that $P(x_i|y) \sim \mathcal{N}(\mu_{i,k}, \sigma_{i,k})$, where x_i is a feature in the bank note data, and y is the class label. Use three-fold cross-validation to split the data and train/test your models.

- (5 points) For each algorithm: briefly describe how you implement it by giving the pseudocode. The pseudocode must include equations for estimating the model parameters and for classifying a new example. Remember, **this should not be a printout of your code, but a high-level outline description**. Include the pseudocode in your pdf file (or .doc/.docx file). Submit the actual code as a single zip file named yourFirstName-yourLastName.zip **IN ADDITION TO** the pdf file (or .doc/.docx file).
- (10 points) Plot a learning curve: the accuracy vs. the size of the training set. Plot 6 points for the curve, using [.01 .02 .05 .1 .625 1] **RANDOM** fractions of you training set and testing on the **full** test set each time. Average your results over 5 runs using each random fraction (e.g. 0.05) of the training set. Plot both the Naïve Bayes and logistic regression learning curves on the same figure. For logistic regression, do not use any regularization term.
- (10 points) Show the power of generative model: Use your trained Naïve Bayes classifier (with the complete training set) to generate 400 examples from class $y = 1$. Report the mean and variance of the generated examples and the corresponding training data (for each fold, over 1 run). and compare with those in your training set (examples in training set with $y = 1$). Try to explain what you observed in this comparison.