CSE 575: Statistical Machine Learning Assignment #1

Instructor: Prof. Jingrui He
Out: Aug. 23rd, 2016; Due: Sep. 20th, 2016

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 [20 points]

Prove that the Bayes classifier is the optimal, i.e., the expected risk of a Bayes classifier is minimal among all the possible classifiers. You only need to show this for binary classifiers.

Solution: For any given example x, the risk of the Bayes classifier is $r_{bayes} = min(q_1(x), q_2(x))$, which is smaller than or equal to the risk of any other classifier, where $q_i(x)$ $(i = 1, 2) = p(y = i|x)$.

Therefore, the expected risk of the Bayes classifier must be the smallest among all possible (binary) classifiers.

2 Parameter Estimation [20 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) What is the MLE estimator of μ?

   Solution. $\hat{\mu}_{MLE} = \frac{\sum_{i=1}^{N} x_i}{N}$

2. (5 points) Is your MLE estimator of μ a random variable? **Explain.**

   Solution. Yes. $\hat{\mu}_{MLE}$ is a function of $x_1, \ldots, x_N$. Each of them is a random variable. So $\hat{\mu}_{MLE}$ is also a random variable.

3. (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.**

   Solution. $E(\hat{\mu}_{MLE}) = E(\frac{\sum_{i=1}^{N} x_i}{N}) = \frac{1}{N}\sum_{i=1}^{N} E(x_i) = \mu$. So $\hat{\mu}_{MLE}$ is unbiased.

4. (5 points) If the true value of μ is known, then the MLE estimator of σ^2 is as follows.

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

   Please prove that $\hat{\sigma}^2_{MLE}$ is unbiased. Notice that this estimator is different from the one we introduced in class due to the fact that we already know the true value of μ.

   Solution. $E(\hat{\sigma}^2_{MLE}) = E(\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2) = \frac{1}{N}\sum_{i=1}^{N} E(x_i - \mu)^2 = \mu$. So $\hat{\sigma}^2_{MLE}$ is unbiased.

3 Naive Bayes Classifier [20 points]

Given the training data set in Figure 2, 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.

Sky	Temp	Humid	Wind	Water	Forecst	EnjoySpt
Sunny	Warm	Normal	Strong	Warm	Same	Yes
Sunny	Warm	High	Strong	Warm	Same	Yes
Rainy	Cold	High	Strong	Warm	Change	No
Sunny	Warm	High	Strong	Cool	Change	Yes

Figure 1: Training Data Set for Naive Bayes Classifiers

1. (5 points) How many independent parameters are there in your Naive Bayes classifier? What are they? Justifiy your answer.

   Solution: (1) $P(y = 1)$ (i.e., enjoy sports), (2) $P(x_1 = sunny|y = i)$ $(i = 1, 0)$, (3) $P(x_2 = warm|y = i)$ $(i = 1, 0)$, (4) $P(x_3 = normal|y = i)$ $(i = 1, 0)$, (5) $P(x_4 = strong|y = i)$ $(i = 1, 0)$, (6) $P(x_5 = warm|y = i)$ $(i = 1, 0)$, (7) $P(x_7 = same|y = i)$ $(i = 1, 0)$. 13 independent parameters in total.

2. (10 points) What are your estimations for these parameters? (say using standard MLE).

   Solution: (1) $P(y = 1) = 3/4$ (i.e., enjoy sports), (2) $P(x_1 = sunny|y = 1) = 1$ and $P(x_1 = sunny|y = 0) = 0$, (3) $P(x_2 = warm|y = 1) = 1$ and $P(x_2 = warm|y = 0) = 0$, (4) $P(x_3 = normal|y = 1) = 1/3$ and $P(x_3 = normal|y = 0) = 0$, (5) $P(x_4 = strong|y = 1) = 1$ and $P(x_4 = strong|y = 0) = 1$, (6) $P(x_5 = warm|y = 1) = 2/3$ and $P(x_5 = warm|y = 0) = 1$, (7) $P(x_7 = same|y = 1) = 2/3$ and $P(x_7 = same|y = 0) = 0$.

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

   Solution: $P(x|y = 1)P(y = 1) = P(x_1 = sunny|y = 1)P(x_2 = warm|y = 1)P(x_3 = high|y = 1)P(x_4 = strong|y = 1)P(x_5 = cool|y = 1)P(x_6 = change|y = 1)P(y = 1) = 1 \times 1 \times 2/3 \times 1 \times 1 \times 1/3 \times 1/3 \times 3/4 = 1/18$. $P(x|y = 0)P(y = 0) = 0$. Therefore, $P(y = 1|x) = \frac{P(x|y=1)P(y=1)}{P(x|y=1)P(y=1)+P(x|y=0)P(y=0)} = 1$. The assigned label will be $y = 1$.

4 Logistic Regression [20 points]

Suppose we have two positive examples $x_1 = (1, 1)$ and $x_2 = (1, -1)$; and two negative examples $x_3 = (-1, 1)$ and $x_4 = (-1, -1)$. We use the standard gradient ascent method (without any additional regularization terms) to train a logistic regression classifier. What is the final weight vector w? Justifiy your answer. You can assume that the weight vector starts at the origin, i.e., $w_0 = (0, 0, 0)'$. How would you explain the final weight vector w you get?

Solution: The final weight vector $w = (0, \infty, 0)'$. We can prove this by induction.

Suppose at t^{th} iteration, the current weight vector $w_t = (0, c, 0)$, where $c \geq 0$, we will show that after the update, $w_{t+1} = (0, c + d, 0)$, where $d > 0$.

We can verify that $P(y_1 = 1|x_1) = P(y_2 = 1|x_2) = \frac{e^c}{1+e^c} = a$, and $P(y_3 = 1|x_3) = P(y_4 = 1|x_4) = \frac{1}{1+e^c} = b$; and $a + b = 1$.

Thus, by gradient ascent, we have $w_{t+1} = w_t + \eta \sum_{i=1}^{4} x_i(y_i - P(y_i = 1|x_i, w_t)) = w_t + \eta(2(1 - a - b), 2(1 + b - a), 0)' = (0, c + 2\eta(1 + b - a), 0$, where $\eta > 0$ is the learning rate; and $0 < a, b < 1$. Therefore $d = 2\eta(1 + b - a) > 0$.

The intuition is that this final weight vector maximizes the likelihood of the training set.

5 Naïve Bayes Classifier and Logistic Regression [20 points]

1. (5 points) **Gaussian Naïve Bayes and Logistic Regression.** Suppose we want to train Gaussian Naïve Bayes to learn a boolean/binary classifier: $f : X \to Y$, where X is a vector of n dimensional real-valued features: $X =< X_1, ..., X_n >$; and Y is boolean class label (i.e., $Y = 1$ or $Y = 0$). Recall that in Gaussian Naïve Bayes, we assume all X_i $(i = 1, ..., n)$ are conditionally independent given the class label Y, i.e., $P(X_i|Y = k) \sim \mathcal{N}(\mu_{ik}, \sigma_i)$ $(k = 0, 1; i = 1, ..., n)$. We also assume that $P(Y)$ follows Bernoulli$(\theta, 1 - \theta)$ (i.e., $P(Y = 1) = \theta$).

   – How many independent model parameters are there in this Gaussian Naïve Bayes classifier?

   – Prove that the Gaussian Naïve Bayes assumption imply that $P(Y|X)$ follow the form of $P(Y = 1|X =< X_1, ..., X_n >= \frac{1}{1+exp(w_0, \sum_{i=1}^{n} w_i X_i)}$. In particular, you need to express w_i $(i = 0, ..., n)$ by the model parameters (i.e., $\theta, \mu_{ik}, \sigma_i$ $(k = 0, 1; i = 1, ..., n)$).
   Solution:

$$
\begin{aligned}
P(Y = 1|X) &= \frac{P(Y = 1)P(X|Y = 1)}{P(Y = 1)P(X|Y = 1) + P(Y = 0)P(X|Y = 0)} \quad \text{(Bayes Rule)} \\
&= \frac{1}{1 + \frac{P(Y=0)P(X|Y=0)}{P(Y=1)P(X|Y=1)}} \\
&= \frac{1}{1 + exp(ln(\frac{P(Y=0)P(X|Y=0)}{P(Y=1)P(X|Y=1)}))} \\
&= \frac{1}{1 + exp(ln(\frac{1-\theta}{\theta}) + \sum_i \; ln(\frac{P(X_i|Y=0)}{P(X_i|Y=1)}))} \quad \text{(Naïve Bayes Assumption)}
\end{aligned}
$$

Since $P(X_i|Y = k) \sim \mathcal{N}(\mu_{ik}, \sigma_i)$ $(k = 0, 1; i = 1, ..., n)$, we have $P(X_i = x|Y = k) = \frac{1}{\sigma_i\sqrt{2\pi}} e^{\frac{-(x-\mu_{ik})^2}{2\sigma_i^2}}$, which implies

$$
ln(\frac{P(X_i|Y = 0)}{P(X_i|Y = 1)}) = \sum_i \; (\frac{\mu_{i0} - \mu_{i1}}{\sigma_i^2} X_i + \frac{\mu_{i1}^2 - \mu_{i0}^2}{2\sigma_i^2})
$$

which completes the proof, with $w_0 = ln(\frac{1-\theta}{\theta}) + \sum_i \frac{\mu_{i1}^2 - \mu_{i0}^2}{2\sigma_i^2}$ and $w_i = \frac{\mu_{i0} - \mu_{i1}}{\sigma_i^2}$.

2. (15 points) Compare the two approaches on the Breast Cancer data set, which can be downloadded from Blackboard. Complete description of the data set can be found at `http://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29`. In the data file, please discard the first column of each row, which is the id number. The second to the 10^{th} columns are the features, and the last column is the class label (2 for benign, 4 for malignant). Please replace class label 2 with +1, and class label 4 with -1.

   In this problem you will obtain the learning curves similar to those from the lecture notes.

   Implement a Naive Bayes classifier and a logistic regression classifier with the assumption that each attribute value for a particular record is independently generated. Please write your own code and do NOT use existing functions or packages. For the Naive Bayes classifier, assume that $P(x_i|y)$, where x_i is a feature in the breast cancer data, and y is the label, is of the following multinomial distribution form:

$$\forall x_i \in \{v_1, v_2, \ldots, v_n\}, \;\; p(x_i = v_k|y = j) = \theta_{i,k}^j, \; s.t. \; \forall i, j : \sum_{k=1}^{n} \theta_{i,k}^j = 1$$

   where $0 \leq \theta_{i,k}^j \leq 1$. It may be easier to think of this as a normalized histogram or as a multi-value extension of the Bernoulli.

   Use the first $\frac{2}{3}$ of the examples as the training set and the remaining $\frac{1}{3}$ as the test set.

   For each algorithm:

   - (5 points) Briefly describe how you implement it by giving the pseudocode. The pseudocode must include equations for estimating the classification parameters and for classifying a new example. Remember, this should not be a printout of your code, but a high-level outline. 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 data. Generate 6 points on the curve, using [.01 .02 .03 .125 .625 1] **RANDOM** fractions of you training set and testing on the full test set each time. Average your results over 5 runs using 5 random fractions of the training set. Plot both the Naive Bayes and Logistic Regression learning curves on the same figure. For Naive Bayies, add 1 to each bin. For Logistic Regression, do not use the regularization term.

     Solution:

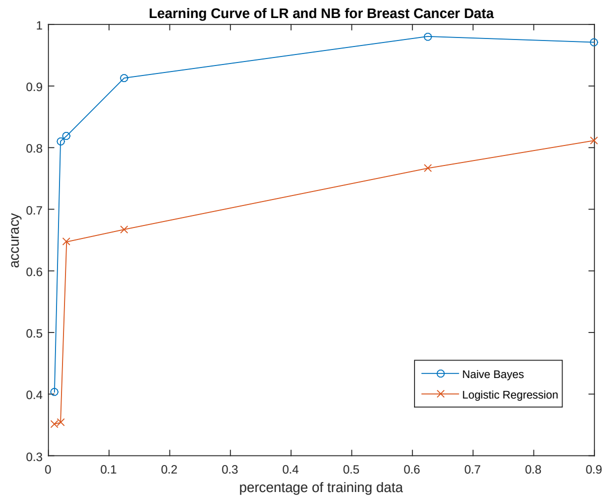

Figure 2: Learning Curve for Naive Bayes and Logistic Regression.