CSE 847: Machine Learning Assignment #1

Instructor: Prof. Jun Wu
Out: Aug. 30, 2024; Due: Sep. 25, 2024

Submit electronically 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]

Assume that there are N i.i.d samples $x_1, \dots, x_N \in \mathbb{R}$ drawn from the same Gaussian distribution $x_i \sim N(\mu, \sigma^2), i = 1, 2, \cdots, N$.

1. (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.

Hint: The bias of an estimator of the parameter σ^2 is defined to be the difference between the expected value of the estimator and σ^2.

Solution.

$$\begin{aligned}
E[\hat{\sigma}^2_{MLE}] &= E[\frac{1}{N}\sum_{i=1}^{N}(x_i - \hat{\mu}_{MLE})^2] \\
&= \frac{1}{N}\sum_{i=1}^{N} E[(x_i-\mu)^2 + 2(x_i-\mu)(\mu-\hat{\mu}_{MLE}) + (\mu-\hat{\mu}_{MLE})^2] \\
&= \frac{1}{N}\sum_{i=1}^{N} E[(x_i-\mu)^2] - \frac{1}{N}\cdot N \cdot E[(\mu-\hat{\mu}_{MLE})^2] \\
&= \frac{1}{N}\sum_{i=1}^{N} \sigma^2 - E[(\mu-\hat{\mu}_{MLE})^2] \\
&= \sigma^2 - E[(\mu-\hat{\mu}_{MLE})^2] \\
&= \sigma^2 - E[(\mu - \frac{1}{N}\sum_{i=1}^{N} x_i)^2] \\
&= \sigma^2 - Var(\mu) \\
&= \sigma^2 - \frac{1}{N^2}Var(\sum_{i=1}^{N} x_i) \\
&= \frac{N-1}{N}\sigma^2
\end{aligned} \tag{1}$$

2. If the prior distribution for mean follows $\mu \sim N(\theta, \lambda)$, what is the MAP estimator $\hat{\mu}_{MAP}$ of μ?

Solution.

$$\hat{\mu}_{MAP} = \frac{\lambda \sum_{i=1}^{N} x_i + \sigma^2\theta}{\lambda N + \sigma^2}$$

2 Parameter Estimation [15 points]

For this question, assume that there are N integers $k_1, \ldots, k_N \in \mathbb{Z}$, which are i.i.d samples drawn from the same underlying distribution. Assume that the underlying distribution is Poisson distribution with PMF

$$P(k|\lambda) = \frac{\lambda^k e^{-\lambda}}{k!}$$

1. [10 points] Please provide the MLE estimator of λ.

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

2. [5 points] Let X be a discrete random variable with the Poisson distribution, What is the expectation $E[X]$?
   Hint: $k! = k \times (k-1) \times (k-2) \times \cdots \times 2 \times 1$ **and** $\sum_{k\geq 0} \frac{\lambda^k}{k!} = e^{\lambda}$

   Solution.

$$E[X] = \sum_{k\geq 0} k \cdot \frac{\lambda^k e^{-\lambda}}{k!} = \lambda e^{-\lambda} \cdot \sum_{k\geq 1} \frac{\lambda^{k-1}}{(k-1)!} = \lambda e^{-\lambda} \cdot \sum_{k\geq 0} \frac{\lambda^k}{k!} = \lambda e^{-\lambda} \cdot e^{\lambda} = \lambda$$

3 Naïve Bayes Classifier [20 points]

Given the training data set shown in Table 1, we train a Naïve Bayes classifier with it. Each row refers to an apple, where the categorical features (size, color and shape) and the class label (whether one apple is good) 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 your answers.

   Solution.

   There are $2 + 2 + 2 + 1 = 7$ parameters. We need a parameter per $P(X_i = x_i|Y = y)$:

   (1). $P(X_1 = x_1|Y = y)$,$x_1 \in \{Small, Large\}$,$y \in \{Yes, No\}$: $2 \times 1 = 2$;

   (2).$P(X_2 = x_2|Y = y)$,$x_2 \in \{Read, Green\}$,$y \in \{Yes, No\}$: $2 \times 1 = 2$;

   (3).$P(X_3 = x_3|Y = y)$,$x_3 \in \{Circle, Irregular\}$,$y \in \{yes, no\}$: $2 \times 1 = 2$;

   and finally a parameter for $p(Y = yes)$.

Table 1: Training Data for Naïve Bayes Classifier

RID	Size	Color	Shape	Class: good_apple
1	Small	Green	Irregular	No
2	Large	Red	Irregular	Yes
3	Large	Red	Circle	Yes
4	Large	Green	Circle	No
5	Large	Green	Irregular	No
6	Small	Red	Circle	Yes
7	Large	Green	Irregular	No
8	Small	Red	Irregular	No
9	Small	Green	Circle	No
10	Large	Red	Circle	Yes

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

   Solution.

   First, we estimate $p(Y = Yes) = \frac{4}{10} = \frac{2}{5}$.

   Then, we iterate through each dimension of X:

   (1). $P(X_1 = Small|Y = Yes) = \frac{1}{4}, P(X_1 = Small|Y = No) = \frac{3}{6} = \frac{1}{2}$.

   (2). $P(X_2 = Green|Y = Yes) = \frac{0}{4} = 0, P(X_2 = Green|Y = No) = \frac{5}{6}$.

   (3).$P(X_3 = Circle|Y = Yes) = \frac{3}{4}$ $P(X_3 = Circle|Y = No) = \frac{2}{6} = \frac{1}{3}$

3. (5 points) Given a new apple with features $x = (Small, Red, Circle)$, what is $P(y = No|x)$? Would the Naïve Bayes classifier predict $y = Yes$ or $y = No$ for this apple?

   Solution.

   $P(y = No|X = (Small, Red, Circle))$
   $= \frac{P(X=(Small,Red,Circle)|y=No)P(y=No)}{P(X=(Small,Red,Circle)|y=No)P(y=No)+P(X=(Small,Red,Circle)|y=Yes)P(y=Yes)}$
   $= \frac{2}{11}$

   It will predict this apple to be $y = Yes$.

4 Logistic Regression [20 points]

Suppose we have three positive examples $x_1 = (1, 0, 0)$, $x_2 = (0, 0, 1)$ and $x_3 = (0, 1, 0)$ and three negative examples $x_4 = (-1, 0, 0)$, $x_5 = (0, -1, 0)$ and $x_6 = (0, 0, -1)$. Apply the standard gradient ascent method to train a logistic regression classifier (without regularization terms). Initialize the weight vector with two different values and set $w_0^0 = 0$ (e.g. $w_0 = (0, 0, 0, 0)', w_0 = (0, 0, 1, 0)'$). Would the final weight vector (w^*) be the same for the two different initial values? What are the values? Please explain your answer in detail. You may assume the learning rate to be a positive real constant η.

Solution.

With a positive learning rate η, w_t^0 will be 0 while w_t^1, w_t^2 and w_t^3 will increase monotonically, so we conclude that $w^* = (0, +\infty, +\infty, +\infty)$ for different initialization.

We use $Y = 1$ for positive samples, $Y = 0$ for negative ones. Suppose we have the weight $w_t = (0, a, b, c)'$ at t-th iteration, then
$w_t^0 = w_t^0 + \eta \sum_j [y^j - P(Y^j = 1|x^j, w_t)]$
$= 0 + \eta[(1 - \frac{1}{1+\exp{(-a)}}) + (1 - \frac{1}{1+\exp{(-c)}}) + (1 - \frac{1}{1+\exp{(-b)}}) + (0 - \frac{1}{1+\exp{(a)}}) + (0 - \frac{1}{1+\exp{(b)}}) + (0 - \frac{1}{1+\exp{(c)}})] = 0$
$w_t^1 = a + \eta \frac{2}{1+\exp(a)}$
$w_t^2 = b + \eta \frac{2}{1+\exp(b)}$
$w_t^3 = c + \eta \frac{2}{1+\exp(c)}$
Therefore, w_t^1, w_t^2 and w_t^3 will increase monotonically.

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 $f : X \rightarrow Y$ 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 they?

     Solution.

     There are $4 \times d + 1$ parameters: 1 parameters for each $\mathcal{N}(\mu_{ik}, \sigma_{ik})$, $i = 1, ..., d$ and $k \in \{0, 1\}$ and 1 parameter for θ.

   - 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.

     Solution.

     For the GBN:
     $P(Y = 1|X) = \frac{P(Y=1)P(X|Y=1)}{P(Y=0)P(X|Y=0)+P(Y=1)P(X|Y=1)} = \frac{1}{1+\frac{P(Y=0)P(X|Y=0)}{P(Y=1)P(X|Y=1)}}$
     $= \frac{1}{1+\exp{(\ln \frac{P(Y=0)P(X|Y=0)}{P(Y=1)P(X|Y=1)})}}$
     $= \frac{1}{1+\exp(\ln \frac{1-\theta}{\theta} + \sum_{i=1}^{d} \ln \frac{P(X_i|Y=0)}{P(X_i|Y=1)})}$
     To match the form of logistic regression, we have to let $\ln \frac{P(X_i|Y=0)}{P(X_i|Y=1)}$ be linear to x_i. So it has $\sigma_{i0} = \sigma_{i1}$, which is the extra assumption we need to complete this translation.

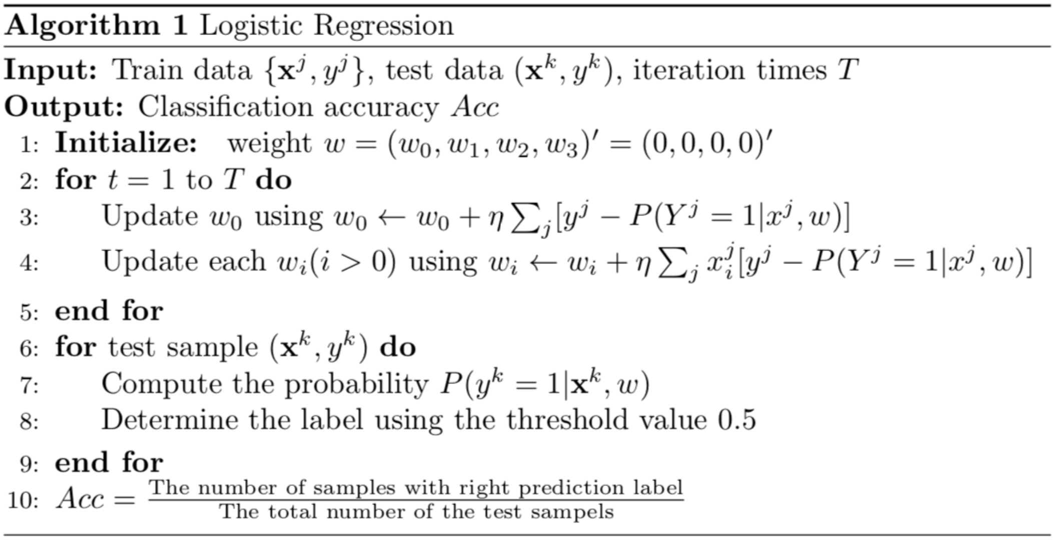
2. (25 points) **Implementation of Gaussian Naïve Bayes and Logistic Regression.** Compare the two approaches on the bank note authentication dataset, which can be downloaded from http://archive.ics.uci.edu/ml/datasets/banknote+authentication. Complete description of the

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). 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).

  Solution.

  The pseudocode of Logistic Regression is described as follows:



```
**Algorithm 1** Logistic Regression
**Input:** Train data $\{\mathbf{x}^j, y^j\}$, test data $(\mathbf{x}^k, y^k)$, iteration times $T$
**Output:** Classification accuracy $Acc$
**Initialize:**   weight $w = (w_0, w_1, w_2, w_3)' = (0, 0, 0, 0)'$
**for** $t = 1$ to $T$ **do**
    Update $w_0$ using $w_0 \leftarrow w_0 + \eta \sum_j [y^j - P(Y^j = 1|x^j, w)]$
    Update each $w_i (i > 0)$ using $w_i \leftarrow w_i + \eta \sum_j x_i^j [y^j - P(Y^j = 1|x^j, w)]$
**end for**
**for** test sample $(\mathbf{x}^k, y^k)$ **do**
    Compute the probability $P(y^k = 1|\mathbf{x}^k, w)$
    Determine the label using the threshold value 0.5
**end for**
$Acc = \frac{\text{The number of samples with right prediction label}}{\text{The total number of the test sampels}}$
```

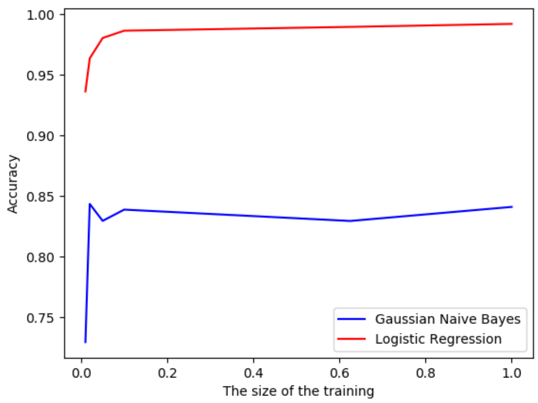
  The pseudocode of Gaussian Naïve Bayes is described as follows:

Algorithm 2 Gaussian Naïve Bayes

```
Input: Train data $\{\mathbf{x}^j, y^j\}$, test data $(\mathbf{x}^k, y^k)$
Output: Classification accuracy $Acc$
Detect positive instance set $X_p$ with label $y = 1$
Detect negative instance set $X_n$ with label $y = 0$
Compute the mean $\mu_p = (\mu_{11}, \mu_{21}, \mu_{31}, \mu_{41})$ and $\sigma_p = (\sigma_{11}, \sigma_{21}, \sigma_{31}, \sigma_{41})$
Compute the mean $\mu_n = (\mu_{10}, \mu_{20}, \mu_{30}, \mu_{40})$ and $\sigma_n = (\sigma_{10}, \sigma_{20}, \sigma_{30}, \sigma_{40})$
Compute the parameter $\theta = \frac{\#(\mathbf{x},y)|_{y=1}}{\#(\mathbf{x},y)}$
for test sample $(\mathbf{x}^k, y^k)$ do
    Compute $P(\mathbf{x}^k|y^k = 1) = \prod_i \frac{1}{\sigma_{i1}\sqrt{2\pi}} e^{\frac{-(\mathbf{x}_i^k - \mu_{i1})}{2\sigma_{i1}^2}}$
    Compute $P(\mathbf{x}^k|y^k = 0) = \prod_i \frac{1}{\sigma_{i0}\sqrt{2\pi}} e^{\frac{-(\mathbf{x}_i^k - \mu_{i0})}{2\sigma_{i0}^2}}$
    Compute $P(y^k = 1|\mathbf{x}^k) = \frac{P(y^k=1)P(\mathbf{x}^k|y^k=1)}{P(y^k=1)P(\mathbf{x}^k|y^k=1)+P(y^k=0)P(\mathbf{x}^k|y^k=0)}$
    Determine the label using the threshold value 0.5
end for
$Acc = \frac{\text{The number of samples with right prediction label}}{\text{The total number of the test sampels}}$
```

- (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.

  Solution.

- (10 points) Show the power of the 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.

Solution.

The first fold:

The generate samples have the mean and variance: [-1.94219658, -0.56993282, 2.14199124, -1.26416] and [3.68469771, 27.76032313, 28.20610056, 4.33278749]

The original samples have the mean and variance: [-1.91025813 -0.73091174 1.96220477 -1.25155333] and [3.52261124 27.58649393 26.63236394 4.21267962]

The second fold:

The generate samples have the mean and variance: [-1.90620877 -0.80405518 2.13434387 -1.32494127] and [3.64208516 30.48297452 28.15374916 4.25340743]

The original samples have the mean and variance: [-1.83055077 -0.99092589 2.11860643 -1.27812395] and [3.61541267 29.83309153 28.01796137 4.20636802]

The third fold:

The generate samples have the mean and variance: [-1.85434711 -1.47418077 2.08448615 -1.15848495] and [3.45757556 29.72243267 28.35352783 4.33542487]

The original samples have the mean and variance: [-1.86152052 -1.27117161 2.37194477 -1.21087113] and [3.46057388 30.01817177 28.24404385 4.42571264]

The code is attached in other submitted file.

It can be seen from these results that the mean and variance are not identical absolutely, but they are very close because these synthetic samples are created based on the Gaussian distribution using the original samples with $y = 1$. The synthetic samples are limited. That is, the mean and variance of these samples may not exactly be equal to the original mean and variance.