

# **Data Mining CS573: Midterm**

March 11, 2016

**Yu-Chen Chang**

## Contents

<b>Problem 1</b>	<b>3</b>
<b>Problem 2</b>	<b>5</b>

## Problem 1

a.

i. Naive Bayes Classifier:

From the naive bayes classifier, we have the formula:

$P(C|X) = \frac{P(X|C)P(C)}{P(X)} \propto P(X|C)P(C)$  (Bayes rule), where  $C$  is the class random variable and  $X$  is the attribute random vector.

In NBC, there is an assumption that attributes are conditionally independent given the class, Therefore, we have the naive Bayes classifier:

$P(C|X) \propto P(X|C)P(C) \propto \prod_{i=1}^m P(X_i|C)P(C)$ , where  $m$  is the number of attributes and  $X_i$  is the  $i$ -th attribute random variable.

Because we don't know the distribution of  $P(X_i|C)$  and  $P(C)$ , Therefore, we need likelihood function to determine unknown parameters based on known outcomes. Assume the data  $D$  are independently sampled from the same distribution. Let  $D = [x_1, \dots, x_n]$ , where  $n$  is the number of samples:

$$L(\theta|D) = \prod_{i=1}^n P(x_i, c_i|\theta) \quad (\text{general likelihood}) \quad (1)$$

$$\propto \prod_{i=1}^n P(x_i|c_i, \theta)P(c_i|\theta) \quad (\text{Bayes rule}) \quad (2)$$

$$\propto \prod_{i=1}^n \prod_{j=1}^m P(x_{ij}|c_i, \theta)P(c_i|\theta) \quad (\text{Naive assumption}) \quad (3)$$

We apply Maximum Likelihood estimation to learn the best parameters by finding the value  $\theta$  that maximizes likelihood:

$$\hat{\theta}_{MLE} = \underset{\theta}{argmax} L(\theta) \quad (4)$$

For Multinomials, Let  $A \in \{1, \dots, k\}$  be a discrete random variable with  $k$  values, where  $P(A = j) = \theta_j$ . Then  $P(A)$  is a multinomial distribution:

$$P(A|\theta) = \prod_{j=1}^k \theta_j^{I(A=j)}, \text{ where } I(A=j) \text{ is an indicator function.} \quad (5)$$

The likelihood for a data set  $D$  is:

$$P(D|\theta) = \prod_{i=1}^n \prod_{j=1}^k \theta_j^{I(A=j)} = \prod_j \theta_j^{n_j} \quad (6)$$

Therefore, by using Lagrange multipliers, the maximum likelihood estimates for each parameter are:

$$\hat{\theta}_j = \frac{n_j}{n} \quad (7)$$

which means that in multinomial case, MLE can be determined analytically by counting.

For continuous inputs  $X_i$ , the common way to represent the distributions  $P(X_i|Y)$  to assume that

for each possible discrete value  $y_k$  of  $Y$ , the distribution of each continuous  $X_i$  is Gaussian, and is defined by a mean and standard deviation specific to  $X_i$  and  $y_k$ .

$$\mu_{ik} = E[X_i|Y = y_k] \quad (8)$$

$$\sigma_{ik}^2 = E[(X_i - \mu_{ik})^2|Y = y_k] \quad (9)$$

Again, by MLE, we get:

$$\hat{\mu}_{ik} = \frac{1}{\sum_j \delta(Y^j = y_k)} \sum_j X_i^j \delta(Y^j = y_k) \quad (10)$$

$$\hat{\sigma}_{ik}^2 = \frac{1}{\sum_j \delta(Y^j = y_k)} \sum_j (X_i^j - \hat{\mu}_{ik})^2 \delta(Y^j = y_k) \quad (11)$$

Then we can estimate continuous attributes using Gaussian distribution with  $\hat{\mu}_{ik}$  and  $\hat{\sigma}_{ik}^2$ .

In this question, we are given 11 attributes.

1. Record Number
2. Amount Requested
3. Interest Rate Percentage
4. Loan Length in Months
5. Loan Title
6. Loan Purpose
7. Monthly Payment
8. Total Amount Funded
9. Debt-To-Income Ratio Percentage
10. FICO Range
11. Status

The Record Number is used as id and will not be considered as an attribute and Status is the classification goal that we are interested in and used as the class random variable. Therefore, the potential attributes are from the 2 to 10 entry, which forms our attribute random vector.

In the step of classifying out-of-sample items, we will use the above shown formula to calculate the  $P(C|X)$  and compare  $P(C = c_1|X)$  with  $P(C = c_2|X)$  to see whether the out-of-sample with its attributes given in  $X$  should belong to  $c_1$  or  $c_2$  class.

- ii. From the MLE, we have the formula

$$\hat{\theta}_j = \frac{n_j}{n} \quad (12)$$

The prior is estimated from the dataset by counting the number of each class among the entire dataset. However, if the real value prior is far from the estimated one, it will have significant impacts on the correctness of the prediction. For example, if we have a dataset with half of people with cancer and other half are healthy while in really life the probability that a person has a cancer is nearly 0.01%, then in this situation the prior will be estimated wrong (50%), which should be 0.01% for cancer class and 99.99% for healthy class, and cause large false positive in this prediction. Therefore, we can see that the wrong prior in NBC will cause either false positive or false negative to increase depending on the difference between real prior and the estimated one. That's the reason why prior in NBC is important.

## Problem 2

a.

b.