MATH 156S HOMEWORK 4

Due Apr. 28

Exercise 1. In this exercise, we solve a simple optimization problem related to the multinomial naive Bayes classifier. Fix a finite set $\mathscr{X} = \{x_1, \ldots, x_M\}$ and real numbers $c_1, \cdots, c_M \geq 0$. Consider the following optimization problem

$$\hat{\mathbf{w}} = \underset{\mathbf{w}=[w_1,\ldots,w_M]^T \in \mathbb{R}^M}{\arg\max} \left(L(\mathbf{w}) := \sum_{i=1}^{M} c_i \log w_i \right) \quad \text{subject to: } \mathbf{w} \text{ is a PMF on } \mathscr{X}. \tag{1}$$

We will show that the solution is given by

$$\hat{\mathbf{w}} = \left[\frac{c_1}{\sum_{i=1}^M c_i}, \ldots, \frac{c_M}{\sum_{i=1}^M c_i} \right]^T. \tag{2}$$

(i)) The constraint set of (1) is the 'simplex' $\mathscr{C} = \{\mathbf{w} \in \mathbb{R}^M \mid 0 \leq w_1, \cdots, w_M \leq 1, \sum_{i=1}^M w_i = 1\}$. Denote the larger constraint set $\overline{\mathscr{C}} = \{\mathbf{w} \in \mathbb{R}^M \mid \sum_{i=1}^M w_i = 1\}$. Let $\lambda \in \mathbb{R}$ be a Lagrange multiplier for $\overline{\mathscr{C}}$. Then the Lagrangian is

$$g(\lambda, \mathbf{w}) = L(\mathbf{w}) - \lambda \left(\sum_{i=1}^{M} w_i - 1 \right). \tag{3}$$

Show that λ needs to satisfy

$$\frac{\partial g(\lambda, \mathbf{w})}{\partial \mathbf{w}} = \left[\frac{c_1}{w_1}, \ldots, \frac{c_M}{w_M} \right]^T - \lambda \, [1, \ldots, 1]^T = 0. \tag{4}$$

Conclude that (2) is the global maximum $L(\mathbf{w})$ over $\overline{\mathscr{C}}$.

(ii) From **(i)**, conclude that (2) is the global maximum $L(\mathbf{w})$ over $\mathscr{C}$.

Exercise 2. Consider the following training data: The goal is to predict whether new subjects with given

age	income	student	credit_rating	buys_computer
≤ 30	high	no	fair	no
≤ 30	high	no	excellent	no
$[31,40]$	high	no	fair	yes
> 40	medium	no	fair	yes
> 40	low	no	fair	yes
> 40	low	yes	excellent	no
$[31,40]$	low	yes	excellent	yes
≤ 30	medium	yes	fair	yes
≤ 30	low	no	fair	yes
> 40	medium	yes	fair	yes
≤ 30	medium	yes	excellent	yes
$[31,40]$	medium	yes	excellent	yes
$[31,40]$	high	no	fair	yes
> 40	medium	no	excellent	no

features (age, income, student, credit rating) would buy a computer, using naive Bayes classifier.

(i) Compute the maximum likelihood prior distribution on the label $\{0,1\}$, where 0 =‘does not buy computer’ and 1 =‘buys computer’.

(ii) We will model the feature space as the following 10-dimensional product space of binary values $\{0,1\}$:[1]

$$\boldsymbol{\phi}(\mathbf{x}) = \left[\mathbf{1}(\text{age} \leq 30), \mathbf{1}(\text{age} \in [31,40]), \mathbf{1}(\text{age} > 40), \ldots, \mathbf{1}(\text{credit} = \text{ecellent})\right] \in \{0,1\}^{10}. \tag{5}$$

Compute the maximum likelihood class-conditional probabilities of each feature.

(iii) Suppose we have the following testing examples:

$$\mathbf{x}_1 = \text{“ age} \leq 30\text{, medium income, student, fair credit rating”} \tag{6}$$

$$\mathbf{x}_2 = \text{“ age} \in [31,40]\text{, low income, not student, fair credit rating”} \tag{7}$$

$$\mathbf{x}_3 = \text{“ age} > 40\text{, high income, not student, excellent credit rating”} \tag{8}$$

Compute the predictive probabilities (posterior distribution of class labels) of each testing examples. Make the prediction. For each example, can you tell which factor affected the most for the classification result?

(iv) Write a python script (in Jupyter notebook) that implements your computation automatically. Test your code on a similar but larger (≥ 100 training examples and ≥ 20 testing examples) dataset of your choice.

REFERENCES

[ZL12] Songfeng Zheng and Weixiang Liu, *Functional gradient ascent for probit regression*, Pattern recognition **45** (2012), no. 12, 4428–4437.

[1]The resulting naive Bayes classifier is a special instance of the multinomial naive Bayes called *Bernoulli naive Bayes*.