ECE523: Engineering Applications of Machine Learning and Data Analytics
Due 01/29/2020 @ 11:59PM (D2L)

I acknowledge that this homework is solely my effort. I have done this work by myself. I have not consulted with others about in any way. I have not received outside aid (outside of my own brain). I understand that violation of these rules contradicts the class policy on academic integrity.

Name: Zhaoyang Bu

Signature: Zhaoyang Bu

Date: 1/31/2020

Instructions: Partial credit is given for answers that are partially correct. No credit is given for answers that are wrong or illegible. All work must be supported and code must be submitted for credit.

Theory: ____________

Practice: ____________

Total: ____________

Part A: Theory (20pts)

(3pts) Maximum Posterior vs Probability of Chance

Show/explain that $\mathbb{P}(\omega_{\max}|\mathbf{x}) \geq \frac{1}{c}$ when we are using the Bayes decision rule, where c is the number of classes. Derive an expression for $\mathbb{P}(\text{error})$. Let $\omega_{\max}$ be the state of nature for which $P(\omega_{\max}|\mathbf{x}) \geq P(\omega_i|\mathbf{x})$ for $i =, 1 \dots, c$. Show that $\mathbb{P}(\text{error}) \leq (c-1)/c$ when we use the Bayes rule to make a decision. Hint, use the results from the previous questions.

(3pts) Bayes Decision Rule Classifier

Let the elements of a vector $\mathbf{x} = [x_1, \dots, x_d]^{\mathsf{T}}$ be binary valued. Let $\mathbb{P}(\omega_j)$ be the prior probability of the class ω_j ($j \in [c]$), and let

$$p_{ij} = \mathbb{P}(x_i = 1|\omega_j)$$

with all elements in $\mathbf{x}$ being independent. If $\mathbb{P}(\omega_1) = \mathbb{P}(\omega_2) = \frac{1}{2}$, and $p_{i1} = p > \frac{1}{2}$ and $p_{i2} = 1 - p$, show that the minimum error decision rule is

$$\text{Choose } \omega_1 \text{ if } \sum_{i=1}^{d} x_i > \frac{d}{2}$$

Hint: Think back to ECE503 and types of random variables then start out with

$$\text{Choose } \omega_1 \text{ if } P(\omega_1)P(\mathbf{x}|\omega_1) > P(\omega_2)P(\mathbf{x}|\omega_2)$$

(3pts) The Ditzler Household Growing Up

My parents have two kids now grown into adults. Obviously there is me, Greg. I was born on a Wednesday. What is the probability that I have a brother? You can assume that $\mathbb{P}(\text{boy}) = \mathbb{P}(\text{girl}) = \frac{1}{2}$.

(10pts) Linear Classifier with a Margin

Show that, regardless of the dimensionality of the feature vectors, a data set that has just two data points, one from each class, is sufficient to determine the location of the maximum-margin hyperplane. Hint #1: Consider a data set of two data points, $\mathbf{x}_1 \in \mathcal{C}_1$ ($y_1 = +1$) and $\mathbf{x}_2 \in \mathcal{C}_2$ ($y_2 = -1$) and set up the minimization problem (for computing the hyperplane) with appropriate constraints on $\mathbf{w}^{\mathsf{T}}\mathbf{x}_1 + b$ and $\mathbf{w}^{\mathsf{T}}\mathbf{x}_2 + b$ and solve it. Hint #2: This can be formed as a constrained optimization problem.

$$\arg\min_{\mathbf{w} \in \mathbb{R}^p} \|\mathbf{w}\|_2^2$$
$$\text{Subject to: (some constraint)}$$

What is $\mathbf{w}$? b? Hint: What are the constraints? How did we solve the constrained optimization problem in Fisher's linear discriminate?

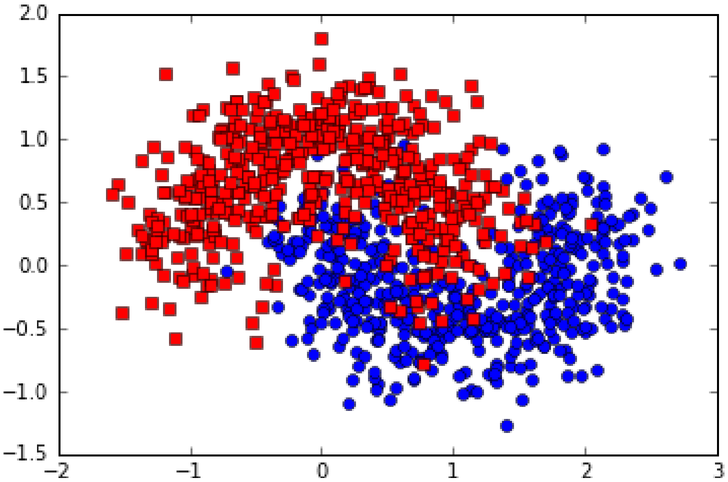

Figure 1: Example of the half moon data set.

(1pt) Decision Making with Bayes

The Bayes decision rule describes the approach we take to choosing a class ω for a data point $\mathbf{x}$. This can be achieved modeling $P(\omega|\mathbf{x})$ or $P(\mathbf{x}|\omega)P(\omega)/P(\mathbf{x})$. Compare and contrast these two approaches to modeling and discuss the advantages and disadvantages. For the latter model, why might knowing $P(\mathbf{x})$ be useful?

Part B: Practice (20pts)

You are free to use functions already implemented in Python with the exception of the sampling problem. I recommend using Python's Scikit-learn (`http://scikit-learn.org/stable/`) as is implements most of the methods we will be discussing in this course... as well as problems in this homework!

(10pts) Half Moon Data Generator and Linear Classifier

Write a script to generate the "half moon" data set shown in Figure 1. Implement a linear classifier (e.g., logistic regression or $\text{sign}(\mathbf{w}^\mathsf{T}\mathbf{x})$) to discriminate between the two classes. Show the decision boundary between the two classes. For example, one approach could be to plot the posterior over a 2D grid where the data lie. Note that you must use a linear classifier. I have posted example code on Github.

(5pts) Naïve Bayes Spam Filter

A Spam data set has been uploaded to the ECE523 Github page (use `data/spambase_train.csv`). Using whatever library you wish, implement a naïve Bayes classifier and report the 5-fold cross validation error.

(5pts) Comparing Classifiers

A text file, `hw1-scores.txt`, containing classifier errors measurements has been uploaded to D2L. Each of the columns represents a classifier and each row a data set that was evaluated. Are all of the classifiers performing equally? Is there one or more classifiers that is performing better than the others? Your response should be backed by statistics. Suggested reading:

- Janez Demšar, "Statistical Comparisons of Classifiers over Multiple Data Sets," Journal of Machine Learning Research, vol. 7, 1–20.

Read the abstract to get an idea about the theme of comparisons. Sections 3.1.3 and 3.2.2 can be used to answer the question posed here.

(5pts) Sampling from a Distribution

Let the set $\mathcal{N} \in [1, \dots, n, \dots, N]$ be a set of integers and $\mathbf{p}$ be a probability distribution $\mathbf{p} = [p_1, \dots, p_n, \dots, p_N]$ such that p_k is the probability of observing $k \in \mathcal{N}$. Note that since $\mathbf{p}$ is a distribution then $1^\mathsf{T}\mathbf{p} = 1$ and $0 \leq p_k \leq 1\ \forall n$ Write a function `sample`$(M, \mathbf{p})$ that returns M indices sampled from the distribution $\mathbf{p}$. Provide evidence that your function is working as desired. Note that all sampling is assumed to be i.i.d. You must include a couple of paragraphs and documented code that discusses how you were able to accomplish this task.

HW#1

ECE523

Zhaoyang Bu

1. if $P(W_{max}|x) < \frac{1}{c}$

then $P(W_i|x) < \frac{1}{c}$ for $i=1,\cdots,c$

$\therefore \sum_{i=1}^{c} P(W_i|x) < \frac{1}{c} \cdot c$

$\sum_{i=1}^{c} P(W_i|x) < 1$

contradiction

$\therefore P(W_{max}|x) \geq \frac{1}{c}$

$P(err) = \int_x p(err, x) dx$

$= \int p(x) dx - \int p(W_{max}|x) p(x) dx$

$= 1 - P(W_{max}|x) \leq 1 - \frac{1}{c} = \frac{c-1}{c}$

$\therefore P(err) \leq \frac{c-1}{c}$

2. choose W_1 if $p(w_1) p(x|w_1) > p(w_2) p(x|W_2)$

$\because P(W_1) = P(W_2) = \frac{1}{2}$

$\therefore p(x|W_1) > p(x|W_2)$

$x = [x_1 \cdots x_d]^T$

$\Rightarrow \prod_{i=1}^{d} p(x_i|W_1) > \prod_{i=1}^{d} P(x_i|W_2)$

$\because P_{i1} = P,\ P_{i2} = 1-P$

$\therefore$ we have $P_{i1} = P(x_i = 1|W_1) = P$

$P_{i1}' = P(x_i = 0|W_1) = 1-P$

$P_{i2} = P(x_i = 1|W_2) = 1-P$

$P_{i2}' = P(x_i = 0|W_2) = P$

let's say we have an integer k, which $k \in [1, d]$

then $P^k (1-P)^{d-k} > (1-P)^k P^{d-k}$

$P^{k-d+k} > (1-P)^{k-d+k}$

$P^{2k-d} > (1-P)^{2k-d}$

$\because P > \frac{1}{2}.\ \therefore 1-P < \frac{1}{2},\ P > 1-P$

$\therefore 2k - d > 0$

$2k > d$

$k > \frac{d}{2}$

$\therefore$ min error dicision rule is choose W_1 if $\sum_{i=1}^{d} x_i > \frac{d}{2}$

3.

A: boy, born on Wed

B: Two kids are boy

$P(B|A) = \frac{P(A|B)\,P(B)}{P(A)}$

$P(A) = \frac{14+14-1}{14 \times 14} = \frac{27}{196}$

$P(B) = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}$

$P(A|B) = \frac{7+7-1}{7 \times 7} = \frac{13}{49}$

$\therefore P(B|A) = \frac{13}{49} \cdot \frac{1}{4} \div \frac{27}{196}$

$= \frac{13}{\cancel{49}} \cdot \frac{1}{\cancel{4}} \cdot \frac{\cancel{196}}{27}$

$= \frac{13}{27}$

4. $\begin{cases} x_1 \in C_1 (y_1 = +1) \\ x_2 \in C_2 (y_2 = -1) \end{cases} \Rightarrow \begin{matrix} w^T x + b = 1 = y_1 \\ w^T x_2 + b = -1 = y_2 \end{matrix}$

$L(w) = \frac{1}{2}\|w\|_2^2 + \lambda_1 (w^T x_1 + b - 1) + \lambda_2 (w^T x_2 + b + 1)$

$\begin{cases} \frac{dL}{dw} = w + \lambda_1 x_1 + \lambda_2 x_2 = 0 \\ \frac{dL}{db} = \lambda_1 + \lambda_2 = 0 \end{cases}$

$\Rightarrow \lambda_1 = -\lambda_2,$

$w = -\lambda_1 x_1 - \lambda_2 x_2$

$= -\lambda_1 x_1 + \lambda_1 x_2$

$= \lambda_1 (x_2 - x_1)$

$w^T x_1 + b + w^T x_2 + b = 1 - 1 = 0$

$\Rightarrow b = -\frac{1}{2}(w^T x_1 + w^T x_2)$

5. Disadvantage:

We could not get $p(w|x)$ directly, we need to get $p(x|w)$, $p(w)$ and $p(x)$ first. However, there may exist some error in $p(x|w)$, $p(w)$ and $p(x)$, so the error of $p(w|x)$ will become lager.

Advantage:

If we want to get $p(w|x)$ use bayes rule is a easier way to get the probability. we do not need to do alot of statistic stuff, it can be very fast.

Latter model:

Because $P(X)$ may help us to do prediction.