Machine Learning: Homework Assignment 5

E4525 Spring 2019,
IEOR, Columbia University

Due: March 29t, 2019

1. **Binary LDA**

   Given the following assumptions

   - input data $x \in \mathbb{R}$ is one dimensional.
   - target label $y \in \{0, 1\}$ is binary.
   - the marginal probabilities of y (independent of x) are given by

   $$\pi_y = p(y) \tag{1}$$

   - Conditional on the label y the distribution of x is Gaussian, with a variance σ^2 that does not depend on the target label y

   $$p(x|y) = \mathcal{N}(x; \mu_y, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\frac{(x-\mu_y)^2}{\sigma^2}} \tag{2}$$

   Show that

   (a) The probability of the labels y conditional on the input data x is given by

   $$p(y|x) = \frac{e^{y(wx+b)}}{1 + e^{wx+b}} \tag{3}$$

   where

   $$\begin{aligned} w &= \frac{\mu_1 - \mu_0}{\sigma^2} \\ b &= \log\frac{\pi_1}{\pi_0} - \frac{1}{2}\frac{\mu_1^2 - \mu_0^2}{\sigma^2} \end{aligned} \tag{4}$$

   [HINT: This is a special case of the multi-dimensional LDA boundary geometry problem in class lecture notes]

   (b) when $\mu_1 > \mu_0$ the classifier with the optimal accuracy is defined by

   $$\hat{y}(x) = \begin{cases} 1 & \text{if} \quad x \geq \frac{\mu_0+\mu_1}{2} - \frac{\sigma^2}{\mu_1-\mu_0}\log\frac{\pi_1}{\pi_0} \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

(c) show that when $\mu_1 < \mu_0$ the classifier with the optimal accuracy is defined by

$$\hat{y}(x) = \begin{cases} 1 & \text{if} \quad x \leq \frac{\mu_0+\mu_1}{2} - \frac{\sigma^2}{\mu_1-\mu_0} \log \frac{\pi_1}{\pi_0} \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

(d) What happens when $\mu_1 = \mu_0$?

2. **Multi Group Classifier**

We have a population that is separated into two groups labeled by a variable $A \in \{0, 1\}$. The group $a = 1$ is a minority group $p(a = 1) \ll p(a = 0)$ that we will consider the "protected group".

In this problem we would like to predict the target binary attribute y making use of the information provided by the protected attribute A, and a continuous variable x.

We make the following assumptions

- input data $x \in \mathbb{R}$ is one dimensional.
- target label $y \in \{0, 1\}$ is binary.
- the marginal probabilities of the pairs (y, a) (the probabilty that a point belongs to group a and has class label y, independent of the value of x) are given by the 2×2 matrix

$$\pi_{y,a} = p(y, a) \tag{7}$$

where the normalization condition is

$$\sum_{y,a} \pi_{y,a} = 1 \tag{8}$$

- conditional on the class label y and the group a the distribution of x is Gaussian with a variance σ^2 that does not depend of the pair (y, a)

$$p(x|y, a) = \mathcal{N}(x; \mu_{y,a}, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\frac{(x-\mu_{y,a})^2}{\sigma^2}} \tag{9}$$

- The variance for each (y, a) combination is $\sigma^2 = 0.25$. The marginal probabilities $\pi_{y,a}$ and conditional means $\mu_{y,a}$ are given by the following table

y	a	$\pi_{y,a}$	$\mu_{y,a}$
0	0	45%	-0.5
0	1	5%	1
1	0	35%	0.5
1	1	15%	-1

Figure 1: Distribution values as a function of class y and group a

Using the results of problem (1) compute the following

(a) The classifier function $\hat{y}_0(x)$ with optimal accuracy for data examples of the majority group $a = 0$. Compute the classification threshold x_0 numerical value explicitly.

(b) The classifier function $\hat{y}_1(x)$ with optimal accuracy for data examples of the minority group $a = 1$. Be explicit.

(c) The true positive rate tpr_a for for the populations with $a = 0$, and $a = 1$ (compute two separate explicit numerical values, one for each group), assuming we use classifier with optimal accuracy.

[HINT: Because all conditional distributions $p(x|y, a)$ are Gaussian, and the classifiers are step functions, the computation of tpr_a can be reduced to normal density integrals

$$N(z) = \int_{-\infty}^{z} \mathrm{d}u \, \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}} \tag{10}$$

that you can look up on a table, or compute using `scipy.stats.norm.cdf` python function.]

(d) The false positive rate fpr_a for the populations with $a = 0$, and for $a = 1$ (compute two separate explicit numerical values, one for each group).

(e) The blended true positive rate and false positive rates for the full population, averaging over $a = 1$ and $a = 1$ groups. Assume that we use classifier $\hat{y}_0(x)$ when $a = 0$, and $\hat{y}_1(x)$ when $a = 1$.

(f) The accuracy of of the classifiers $\hat{y}_0(x)$ and $\hat{y}_1(x)$ when used on samples of their respective groups.

(g) The accuracy of the blended classifier

3. **Fair (Anti-Classification) Classifier**

We have again the same population described in problem (2) with the marginal probabilities and group means listed in table (1)

We would now like to predict a target binary attribute **without using** the group attribute a of each sample

(a) Given the information in table (1 compute the marginal probabilities $\pi_y = p(y)$ and class mean

$$\mu_y = \mathbb{E}[x|y] \tag{11}$$

for target labels $y = 0$ and $y = 1$.

(b) Build the fair (in the anti-classification sense) LDA classifier with best possible accuracy that predicts Y as a function of X, ignoring the value of the group attribute A.

(c) The true positive rate for the populations with $a = 0$, and $a = 1$, assuming we use the fair classifier $\hat{y}_{\mathrm{F}}(x)$.

(d) The false positive rate for the populations with $a = 0$, and $a = 1$, assuming we use the fair classifier. Be explicit.

(e) The classification accuracy for populations with $a = 0$, and $a = 1$, assuming we use the fair classifier. Be explicit.

(f) The blended accuracy rate of the fair classifier.

4. **Binary Logistic Regression**

In Binary Logistic regression we make the following assumptions

- $x \in \mathbb{R}^D$
- y is a binary random variable taking values $\{0, 1\}$.
- Conditional on the value of x, y is distributed as a Bernoulli random variable with parameter $\theta(x)$

$$p(y|x) = \theta(x)^y \left(1 - \theta(x)\right)^{1-y} \tag{12}$$

- The parameter θ depends on x in the following form

$$\theta(x) = \frac{e^{\eta(x)}}{1 + e^{\eta(x)}} \tag{13}$$

- where η is a linear function of x

$$\eta(x) = w^T x + b = \sum_d w_d x_d + b \tag{14}$$

We are given a set of N observations $\{y_i, x_{i,d}\}$ for $i = 1, \ldots, N$ and $d = 1, \ldots, D$. Show that

(a) The average log likelihood loss is given by

$$\hat{E}(w, b; \{y_i, x_{i,d}\}) = \frac{1}{N} \sum_{i=1}^{N} l_i(\eta_i) \tag{15}$$

where

$$l_i(\eta_i) = \log\left(1 + e^{\eta_i}\right) - y_i \eta_i \tag{16}$$

and

$$\eta_i = \eta(x_i) = w^T x_i + b \tag{17}$$

(b) Show that the gradient of the loss function is

$$\frac{\partial \hat{E}}{\partial b} = \frac{1}{N} \sum_i \left(\theta(x_i) - y_i\right)$$

$$\frac{\partial \hat{E}}{\partial w_d} = \frac{1}{N} \sum_i \left(\theta(x_i) - y_i\right) x_{i,d} \tag{18}$$

(c) The maximum likelihood estimate of the parameters $\hat{w}_d$ and $\hat{b}$ satisfy the equations

$$\sum_i y_i = \sum_{i=1}^{N} \theta(x_i; \hat{w}, \hat{b})$$
$$\sum_i y_i x_{i,d} = \sum_{i=1}^{N} \theta(x_i; \hat{w}, \hat{b}) x_{i,d} \tag{19}$$

where we have written explicitly the dependence of θ on w and b.