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]

   Using Bayes theorem we can write

   $$p(y|x) = \frac{p(x|y)p(y)}{p(x)} \tag{5}$$

where

$$p(x) = p(x|y=0)p(y=0) + p(x|y=1)p(y=1) \tag{6}$$

therefore we have

$$\begin{aligned} p(y=0|x) &= \frac{p(x|y=0)\pi_0}{p(x|y=0)\pi_0 + p(x|y=1)\pi_1} \\ p(y=1|x) &= \frac{p(x|y=0)\pi_1}{p(x|y=0)\pi_0 + p(x|y=1)\pi_1} \end{aligned} \tag{7}$$

factoring out a term $p(x|y=0)\pi_0$ and defining

$$f = \frac{\pi_1 p(x|y=1)}{\pi_0 p(x|y=0)} \tag{8}$$

we can write this expression as

$$\begin{aligned} p(y=0|x) &= \frac{1}{1+f} \\ p(y=1|x) &= \frac{f}{1+f} \end{aligned} \tag{9}$$

of using the fact that $y \in \{0,1\}$

$$p(y|x) = \frac{f^y}{1+f} \tag{10}$$

We now use the fact that $p(x|y)$ is Gaussian and that σ is the same for both classes to write

$$f = \frac{\pi_1}{\pi_0} e^{-\frac{1}{2}\frac{(x-\mu_1)^2}{\sigma^2} + \frac{1}{2}\frac{(x-\mu_0)^2}{\sigma^2}} \tag{11}$$

expanding the squares, and using the fact that the terms quadratic on x cancel out we obtain

$$f = e^{\frac{\mu_1-\mu_0}{\sigma^2}x - \frac{1}{2}\frac{\mu_1^2-\mu_0^2}{\sigma^2} + \log\frac{\pi_1}{\pi_0}} = e^{wx+b} \tag{12}$$

where we have made the identifications

$$\begin{aligned} w &= \frac{\mu_1-\mu_0}{\sigma^2} \\ b &= \log\frac{\pi_1}{\pi_0} - \frac{1}{2}\frac{\mu_1^2-\mu_0^2}{\sigma^2} \end{aligned} \tag{13}$$

.

Substituting expression (12) into equation (10) we finally obtain

$$p(y|x) = \frac{e^{y(wx+b)}}{1+e^{wx+b}} \tag{14}$$

with the loadings w and bias b defined by (13).

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

$$\hat{y}(x) = \begin{cases} 1 & \text{if} \quad x \geq \frac{\mu_0+\mu_1}{2} - \frac{\sigma^2}{\mu_1-\mu_0} \log \frac{\pi_1}{\pi_0} \\ 0 & \text{otherwise} \end{cases} \tag{15}$$

The optimal accuracy classifier will be such that

$$\hat{y}(x) = \begin{cases} 1 & p(y|x) \geq \frac{1}{2} \\ 0 & \text{otherwise} \end{cases} \tag{16}$$

When $\mu_1 > \mu_0$ whe have that $w > 0$ and the conditional probability $p(y|x)$ is an increasing function of x.
The condition $p(y|x) > \frac{1}{2}$ thus becomes equivalent to

$$\hat{y}(x) = \begin{cases} 1 & x \geq x_0 \\ 0 & \text{otherwise} \end{cases} \tag{17}$$

for the point x_0 given by

$$p(y|x_0) = \frac{1}{2}. \tag{18}$$

Given the functional form (14) equation (18) is equivalent to

$$wx_0 + b = 0 \tag{19}$$

solving for x_0 we find

$$x_0 = -\frac{b}{w} \tag{20}$$

substituting the expressions for w and b we find the stated result.

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

$$\hat{y}(x) = \begin{cases} 1 & \text{if} \quad x \leq \frac{\mu_0+\mu_1}{2} - \frac{\sigma^2}{\mu_1-\mu_0} \log \frac{\pi_1}{\pi_0} \\ 0 & \text{otherwise} \end{cases} \tag{21}$$

When $\mu_1 < \mu_0$ whe have that $w < 0$ and $p(y|x)$ is a decreasing function of x.
In that situation $p(y|x) > p(y|x_0)$ if $x < x_0$ so we have an optimal classifier

$$\hat{y}(x) = \begin{cases} 1 & x \leq x_0 \\ 0 & \text{otherwise} \end{cases} \tag{22}$$

where x_0 is defined as before as $p(y|x_0) = \frac{1}{2}$.

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

In that case x contains no information about the class y because

$$p(x|y) = \mathcal{N}(x; \mu, \sigma^2) \tag{23}$$

does not depend on class.

The optimal accuracy classifier would assign every sample to the class with larger probability π_y so $\hat{y}(x)$ will be a constant.

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

$$\pi_{y,a} = p(y, a) \tag{24}$$

where the normalization condition is

$$\sum_{y,a} \pi_{y,a} = 1 \tag{25}$$

- 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)

$$p(x|y, a) = \mathcal{N}(x; \mu_{y,a}, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\frac{(x-\mu_{y,a})^2}{\sigma^2}} \tag{26}$$

- 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.

using the result for exercise (1) and the fact that $\mu_{1,0} > \mu_{0,0}$ we have that

$$\hat{y}_0(x) = \begin{cases} 1 & \text{if} \quad x \geq x_0 \\ 0 & \text{otherwise} \end{cases} \tag{27}$$

where

$$x_0 = \frac{\mu_{0,0} + \mu_{1,0}}{2} - \frac{\sigma^2}{\mu_{1,0} - \mu_{0,0}} \log \frac{\pi_{1,0}}{\pi_{0,0}} \tag{28}$$

substituting the values from table (1) we have

$$x_0 = \frac{-0.5 + 0.5}{2} - \frac{0.25}{0.5 - (-0.5)} \log \frac{0.35}{0.45} \approx 0.06 \tag{29}$$

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

using again the results for exercise (1) and the fact that $\mu_{1,1} < \mu_{0,1}$ we have that

$$\hat{y}_1(x) = \begin{cases} 1 & \text{if} \quad x \leq x_1 \\ 0 & \text{otherwise} \end{cases} \tag{30}$$

where

$$x_1 = \frac{\mu_{0,1} + \mu_{1,1}}{2} - \frac{\sigma^2}{\mu_{1,1} - \mu_{0,1}} \log \frac{\pi_{1,1}}{\pi_{0,0}} \tag{31}$$

substituting the values from table (1) we have

$$x_1 = \frac{1 - 1}{2} - \frac{0.25}{-1 - (1)} \log \frac{0.15}{0.05} = 0.14 \tag{32}$$

(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

$$N(z) = \int_{-\infty}^{z} \mathrm{d}u \, \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}} \tag{33}$$

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

The true positive rate is the probability that we predict $\hat{y} = 1$ to samples of the true class $y = 1$

$$\mathrm{tpr} = p(\hat{y} = 1 | y = 1) \tag{34}$$

Given the class label y, and the group a $\hat{y}_a$ is a funtion of x, and x is a Gaussian distribution, we can, thus write

$$\mathrm{tpr}_a = \int \mathrm{d}x \mathcal{N}(x; \mu_{1,a}, \sigma^2) \hat{y}_a(x) \tag{35}$$

and then we have, for $a = 0$

$$\mathrm{tpr}_0 = \int_{x_0}^{\infty} \mathrm{d}x \mathcal{N}(x; \mu_{1,0}, \sigma^2), \tag{36}$$

where we have used that $\hat{y}_0(x)$ is 1 if $x > x_0$ and zero otherwise. Changing variables into

$$z = \frac{x - \mu_{1,0}}{\sigma} \tag{37}$$

we have

$$\mathrm{tpr}_0 = \int_{z_0}^{\infty} \mathrm{d}z \mathcal{N}(z; 0, 1) = 1 - N(z_0) \tag{38}$$

where

$$z_0 = \frac{x_0 - \mu_{1,0}}{\sigma} = \frac{0.06 - (0.5)}{0.5} = -0.88 \tag{39}$$

as before

$$\mathrm{tpr}_0 \approx 1 - 0.19 = 0.81 \tag{40}$$

For the protected group $a = 1$, using that $\hat{y}_1(x)$ is 1 if $x < x_1$ and zero otherwise, we have

$$\mathrm{tpr}_1 = \int_{-\infty}^{x_1} \mathrm{d}x \mathcal{N}(x; \mu_{1,1}, \sigma^2) = N(z_1) \tag{41}$$

normalizing again in terms of z we find

$$z_1 = \frac{x_1 - \mu_{1,1}}{\sigma} = \frac{0.14 - (-1)}{0.5} = 2.28 \tag{42}$$

and looking up on an integral table we find

$$\mathrm{tpr}_1 \approx 0.99 \tag{43}$$

(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).

The false positive rate is the probability of assigning $\hat{y} = 1$ to samples of the fall class $y = 0$

$$\text{fpr} = p(\hat{y} = 1|y = 0) \tag{44}$$

Given the class label y, and the group a $\hat{y}_a$ is,again, a funtion of x, and x is a Gaussian distribution, we can, thus write

$$\text{fpr}_a = \int \mathrm{d}x \mathcal{N}(x; \mu_{0,a}, \sigma^2)\hat{y}_a(x) \tag{45}$$

and then we have, for $a = 0$

$$\text{fpr}_0 = \int_{x_0}^{\infty} \mathrm{d}x \mathcal{N}(x; \mu_{0,0}, \sigma^2) \tag{46}$$

changing variables into

$$z' = \frac{x - \mu_{0,0}}{\sigma} \tag{47}$$

we have

$$\text{fpr}_0 = \int_{z_0'}^{\infty} \mathrm{d}z \mathcal{N}(z; 0, 1) = 1 - N(z_0') \tag{48}$$

where

$$z_0' = \frac{x_0 - \mu_{0,0}}{\sigma} = \frac{0.06 - (-0.5)}{0.5} = 1.12 \tag{49}$$

consulting a table for the cumulative normal distribution we can find that

$$\text{fpr}_0 \approx 1 - 0.87 = 0.13 \tag{50}$$

For the protected group $a = 1$ we then have

$$\text{fpr}_1 = \int_{-\infty}^{x_1} \mathrm{d}x \mathcal{N}(x; \mu_{0,1}, \sigma^2) = N(z_1') \tag{51}$$

where

$$z_1' = \frac{x_1 - \mu_{0,1}}{\sigma} = \frac{0.14 - (+1)}{0.5} = -1.72 \tag{52}$$

and, therefore

$$\text{fpr}_1 = N(-1.72) \approx 0.04 \tag{53}$$

(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 $a = 1$.

The fraction of the population belonging to the protected group $a = 1$ is

$$p_1 = \pi_{0,1} + \pi_{1,1} = 0.05 + 0.15 = 0.2 \tag{54}$$

The blended true positive rate, will be

$$(1 - p_1)\text{tpr}_0 + p_1\text{tpr}_1 = 0.8 * 0.81 + 0.2 * 99 \approx 85\% \tag{55}$$

and the false positive rate will be

$$(1 - p_1)\text{fpr}_0 + p_1\text{fpr}_1 = 0.8 * 0.13 + 0.2 * 99 \approx 11\% \tag{56}$$

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

The accuracy for group a is given by

$$\text{acc}_a = \frac{\text{tpr}_a * \pi_{1,a} + (1 - \text{fpr}_a)\pi_{0,a}}{\pi_{0,a} + \pi_{1,a}} \tag{57}$$

substituting the rates we computed before we have

$$\text{acc}_0 = \frac{0.81 \cdot 0.35 + (1 - 0.13)0.45}{0.8} \approx 0.84 \tag{58}$$

$$\text{acc}_1 = \frac{0.99 \cdot 0.15 + (1 - 0.04)0.05}{0.2} \approx 0.98 \tag{59}$$

(g) The accuracy of the blended classifier

The accuracy of the blended classifier will be

$$(1 - p_1)\text{acc}_0 + p_1\text{acc}_1 = 0.8 * 0.84 + 0.2 * 98 \approx 87\% \tag{60}$$

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

$$\mu_y = \mathbb{E}[x|y] \tag{61}$$

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

We need to estimate the marginal probability per Y class, the mean per class Y, and a single variance shared by both classes

$$\begin{aligned}
\pi_Y =& \pi_{Y,0} + \pi_{Y,1} \\
\mu_Y =& \frac{\pi_{Y,0}\mu_{Y,0} + \pi_{Y,1}\mu_{Y,1}}{\pi_Y} \\
\tilde{\sigma}^2 =& \sum_{a=0,1} \pi_{Y,a} \left\{ \sigma^2_{Y,a} + (\mu_{Y,a} - \mu_Y)^2 \right\}.
\end{aligned} \tag{62}$$

As we will see we will not really need to compute $\tilde{\sigma}$ to build the classifier.

Substituting value from table 1 we have

$$\pi_0 = 0.45 + 0.05 = 0.5 \tag{63}$$

$$\pi_1 = 0.35 + 0.15 = 0.5 \tag{64}$$

for the the means

$$\begin{aligned}
\mu_0 =& \frac{0.45(-0.5) + 0.05(1)}{0.45 + 0.05} = -0.35 \\
\mu_1 =& \frac{0.35(0.5) + 0.15(-1)}{0.35\% + 0.15} = 0.05
\end{aligned} \tag{65}$$

and variance is given by

$$\begin{aligned}
\tilde{\sigma}^2 =& 0.45 \left\{ 0.25 + (-0.5 - (-0.35))^2 \right\} \quad +0.05 \left\{ 0.25 + (1 - (-0.35))^2 \right\} \\
&+ 0.35 \left\{ 0.25 + (0.5 - 0.05)^2 \right\} \qquad +0.15 \left\{ 0.25 + (-1 - (0.05))^2 \right\} \\
&\approx 0.59
\end{aligned} \tag{66}$$

(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.

We need to build a classifier

$$\hat{y}_{\mathrm{F}}(x) = \begin{cases} 1 & \text{if} \quad p(y = 1|x) \geq \frac{1}{2} \\ 0 & \text{if} \quad p(y = 1|x) < \frac{1}{2} \end{cases} \tag{67}$$

And give the QDA assumption, we will assume that x is Gaussian conditional on the class label y

$$p(x|y) = \mathcal{N}(\mu_y, \sigma_y^2) \tag{68}$$

so we need to estimate the marginal probability per Y class, the mean per class Y, and a single variance shared by both classes

$$\begin{aligned}\pi_Y &= \pi_{Y,0} + \pi_{Y,1} \\ \mu_Y &= \frac{\pi_{Y,0}\mu_{Y,0} + \pi_{Y,1}\mu_{Y,1}}{\pi_Y} \\ \sigma^2 &= \sum_{a=0,1} \pi_{Y,a}\left\{\sigma^2_{Y,a} + (\mu_{Y,a} - \mu_Y)^2\right\}\end{aligned} \tag{69}$$

For the probability π_Y of y marginalized on a and x we have

$$\pi_0 = 0.45 + 0.05 = 0.5 \tag{70}$$

$$\pi_1 = 0.35 + 0.15 = 0.5 \tag{71}$$

Substitution the values in Table (1) we find the means

$$\begin{aligned}\mu_0 &= \frac{0.45(-0.5) + 0.05(1)}{0.45 + 0.05} = -0.35 \\ \mu_1 &= \frac{0.35(0.5) + 0.15(-1)}{0.35\% + 0.15} = 0.05\end{aligned} \tag{72}$$

The variance is given by

$$\begin{aligned}\sigma^2 = {} & 0.45\left\{0.25 + (-0.5 - (-0.35))^2\right\} \quad +0.05\left\{0.25 + (1 - (-0.35))^2\right\} \\ & + 0.35\left\{0.25 + (0.5 - 0.05)^2\right\} \qquad +0.15\left\{0.25 + (-1 - (0.05))^2\right\} \\ & \approx 0.59\end{aligned} \tag{73}$$

The threshold will follow formula (21) with

$$x_0 = \frac{\mu_1 + \mu_0}{2} - \frac{\sigma^2}{\mu_1 - \mu_2} \log \frac{\pi_1}{\pi_0} \tag{74}$$

were the dependence of σ^2 drops out because the prevalence of both classes is equal $\pi_1 = \pi_0$ and we find

$$x_0 = \frac{0.05 + -0.035}{2} = -0.15 \tag{75}$$

Because $\mu_1 > \mu_0$ the optimal classifier is

$$\hat{y}_{\mathrm{F}}(x) = \begin{cases} 1 & \text{if} \quad x \geq x_0 \end{cases} \tag{76}$$

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

The expression for the true positive rate on each population is

$$\mathrm{tpr}_a \tag{77}$$

where the classifier now does not depend on the population.
Substituting values we have

$$\mathrm{tpr}_0 = \int_{x_0}^{\infty} \mathrm{d}x \mathcal{N}(x; \mu_{1,0}, \sigma^2) \tag{78}$$

$$\mathrm{tpr}_1 = \int_{x_0}^{\infty} \mathrm{d}x \mathcal{N}(x; \mu_{1,1}, \sigma^2) \tag{79}$$

$$\tag{80}$$

which we can express in terms of the re-scaled variables

$$\tilde{z}_0 = \frac{x_0 - \mu_{1,0}}{\sigma} = -1.3 \tag{81}$$

$$\tilde{z}_1 = \frac{x_0 - \mu_{1,1}}{\sigma} = 1.7 \tag{82}$$

$$\tag{83}$$

as

$$\mathrm{tpr}_0 = 1 - N(\tilde{z}_0) \approx 90\% \tag{84}$$

$$\mathrm{tpr}_1 = 1 - N(\tilde{z}_0) \approx 4\% \tag{85}$$

$$\tag{86}$$

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

$$\mathrm{fpr}_a \tag{87}$$

where the classifier now does not depend on the population.
Substituting values we have

$$\mathrm{fpr}_0 = \int_{x_0}^{\infty} \mathrm{d}x \mathcal{N}(x; \mu_{0,0}, \sigma^2) \tag{88}$$

$$\mathrm{fpr}_1 = \int_{x_0}^{\infty} \mathrm{d}x \mathcal{N}(x; \mu_{0,1}, \sigma^2) \tag{89}$$

$$\tag{90}$$

which we can express in terms of the re-scaled variables

$$\tilde{z}'_0 = \frac{x_0 - \mu_{0,0}}{\sigma} = 0.7 \tag{91}$$

$$\tilde{z}'_1 = \frac{x_0 - \mu_{0,1}}{\sigma} = -2.3 \tag{92}$$

$$\tag{93}$$

as

$$\text{fpr}_0 = 1 - N(\tilde{z}_0) \approx 24\% \tag{94}$$
$$\text{fpr}_1 = 1 - N(\tilde{z}_0) \approx 99\% \tag{95}$$
$$\tag{96}$$

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

$$\text{acc}_a = \frac{\text{tpr}_a * \pi_{1,a} + (1 - \text{fpr}_a)\pi_{0,a}}{\pi_{0,a} + \pi_{1,a}} \tag{97}$$

substituting the rates we computed before we have

$$\text{acc}_0 = \frac{0.9 \cdot 0.35 + (1 - 0.24)0.45}{0.8} \approx 0.82 \tag{98}$$
$$\text{acc}_1 = \frac{0.04 \cdot 0.15 + (1 - 0.99)0.05}{0.2} \approx 0.03 \tag{99}$$

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

$$\text{`acc} = p_1 acc_1 + (1 - p_1) acc_0 = 0.2 * 0.03 + 0.8 * 0.82 \approx 66\% \tag{100}$$

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 $\theta(x)$

$$p(y|x) = \theta(x)^y \left(1 - \theta(x)\right)^{1-y} \tag{101}$$

- The parameter θ depends on x in the following form

$$\theta(x) = \frac{e^{\eta(x)}}{1 + e^{\eta(x)}} \tag{102}$$

- where η is a linear function of x

$$\eta(x) = w^T x + b = \sum_d w_d x_d + b \tag{103}$$

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

$$\hat{E}(w, b; \{y_i, x_{i,d}\}) = \frac{1}{N} \sum_{i=1}^{N} l_i(\eta_i) \tag{104}$$

where

$$l_i(\eta_i) = \log\left(1 + e^{\eta_i}\right) - y_i \eta_i \tag{105}$$

and

$$\eta_i = \eta(x_i) = w^T x_i + b \tag{106}$$

Taking logs of $p(y|x)$ and averaging over all observations we find

$$\hat{E} = -\frac{1}{N} \sum_i y_i \log \theta(x_i) + (1 - y_i) \log(1 - \theta(x_i)) \tag{107}$$

using that

$$\begin{aligned} \log \theta &= \eta - \log(1 + e^{\eta}) \\ \log(1 - \theta) &= -\log(1 + e^{\eta}) \end{aligned} \tag{108}$$

and substituting in the expression for $\hat{E}$ we find

$$\hat{E} = \frac{1}{N} \sum_i \log(1 + e^{\eta(x_i)}) - y_i \eta(x_i) = \frac{1}{N} \sum_i l_i(\eta_i) \tag{109}$$

where

$$l_i(\eta) = \log(1 + e^{\eta}) - y_i \eta \tag{110}$$

and

$$\eta_i = w^T x + b = \sum_d w_d x_{i,d} + b \tag{111}$$

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

$$\begin{aligned} \frac{\partial \hat{E}}{\partial b} &= \frac{1}{N} \sum_i \left(\theta(x_i) - y_i\right) \\ \frac{\partial \hat{E}}{\partial w_d} &= \frac{1}{N} \sum_i \left(\theta(x_i) - y_i\right) x_{i,d} \end{aligned} \tag{112}$$

As we have shown

$$\hat{E} = \frac{1}{N} \sum_i l_i(\eta_i) \tag{113}$$

Using the chain rule we have

$$\begin{aligned}\frac{\partial l_i(\eta_i)}{\partial b} &= \frac{\partial l_i}{\partial \eta_i}\frac{\partial \eta_i}{\partial b}\\ \frac{\partial l_i(\eta_i)}{\partial w_d} &= \frac{\partial l_i}{\partial \eta_i}\frac{\partial \eta_i}{\partial w_d}\end{aligned} \tag{114}$$

using that

$$\eta_i = w^T x_i + b \tag{115}$$

we have that

$$\begin{aligned}\frac{\partial \eta_i}{\partial b} &= 1\\ \frac{\partial \eta_i}{\partial w_d} &= x_{i,d}\end{aligned} \tag{116}$$

Therefore

$$\begin{aligned}\frac{\partial l_i(\eta_i)}{\partial b} &= \frac{\partial l_i(\eta_i)}{\partial \eta_i}\\ \frac{\partial l_i(\eta_i)}{\partial w_d} &= \frac{\partial l_i(\eta_i)}{\partial \eta_i}x_{i,d}\end{aligned} \tag{117}$$

taking the derivative of

$$l_i(\eta_i) = \log(1 + e^{\eta_i}) - y_i\eta_i \tag{118}$$

we find

$$\frac{\partial l_i(\eta_i)}{\partial \eta_i} = \frac{e^{\eta_i}}{1 + e^{\eta_i}} - y_i = \theta(\eta(x_i)) - y_i \tag{119}$$

and, finally

$$\begin{aligned}\frac{\partial \hat{E}}{\partial b} &= \frac{1}{N}\sum_i (\theta(x_i) - y_i)\\ \frac{\partial \hat{E}}{\partial w_d} &= \frac{1}{N}\sum_i (\theta(x_i) - y_i)\, x_{i,d}\end{aligned} \tag{120}$$

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

$$\begin{aligned}\sum_i y_i &= \sum_{i=1}^N \theta(x_i; \hat{w}, \hat{b})\\ \sum_i y_i x_{i,d} &= \sum_{i=1}^N \theta(x_i; \hat{w}, \hat{b}) x_{i,d}\end{aligned} \tag{121}$$

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

It follows immediately after setting zero the expressions for the gradient on the previous problem.