Introduction to Machine Learning

Problems: Logistic Regression

2. Suppose that a logistic regression model for a binary class label $y = 0, 1$ is given by

$$P(y = 1|\mathbf{x}) = \frac{1}{1 + e^{-z}}, \quad z = \beta_0 + \beta_1 x_1 + \beta_2 x_2,$$

where $\boldsymbol{\beta} = [1, 2, 3]^\mathsf{T}$. Describe the following sets:

(a) The set of $\mathbf{x}$ such that $P(y = 1|\mathbf{x}) > P(y = 0|\mathbf{x})$.

(b) The set of $\mathbf{x}$ such that $P(y = 1|\mathbf{x}) > 0.8$.

(c) The set of x_1 such that $P(y = 1|\mathbf{x}) > 0.8$ and $x_2 = 0.5$.

Solution:

(a). From the question, we know:

$p(y=1|x) = \frac{1}{1+e^{-z}}$ and it's a binary class label $y=0, 1$

$\therefore p(y=0|x) = 1 - \frac{1}{1+e^{-z}}$

According to $p(y=1|x) > p(y=0|x)$, we have:

$$\frac{1}{1+e^{-z}} > 1 - \frac{1}{1+e^{-z}}$$

$\therefore e^{-z} < 1$

$\therefore z > 0$

$\therefore z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 > 0$

(b). $p(y=1|x) > 0.8$

$\therefore \frac{1}{1+e^{-z}} > 0.8$

$1 > 0.8 + 0.8e^{-z}$

$e^{-z} < \frac{1}{4}$

$-z < \ln(\frac{1}{4})$

$\therefore z > \ln 4$

(c). According to the result in (b), we have:

$\beta_0 + \beta_1 x_1 + \beta_2 x_2 > \ln 4$ and $x_2 = 0.5$

$\therefore \beta_0 + \beta_1 x_1 + \frac{1}{2}\beta_2 > \ln 4$

$1 + 2x_1 + \frac{1}{2} \cdot 3 > \ln 4$

$\therefore x_1 > \frac{2\ln 4 - 5}{4}$

3. A data scientist is hired by a political candidate to predict who will donate money. The data scientist decides to use two predictors for each possible donor:

   - x_1 = the income of the person(in thousands of dollars), and
   - x_2 = the number of websites with similar political views as the candidate the person follow on Facebook.

   To train the model, the scientist tries to solicit donations from a randomly selected subset of people and records who donates or not. She obtains the following data:

Income (thousands $), x_{i1}	30	50	70	80	100
Num websites, x_{i2}	0	1	1	2	1
Donate (1=yes or 0=no), y_i	0	1	0	1	1

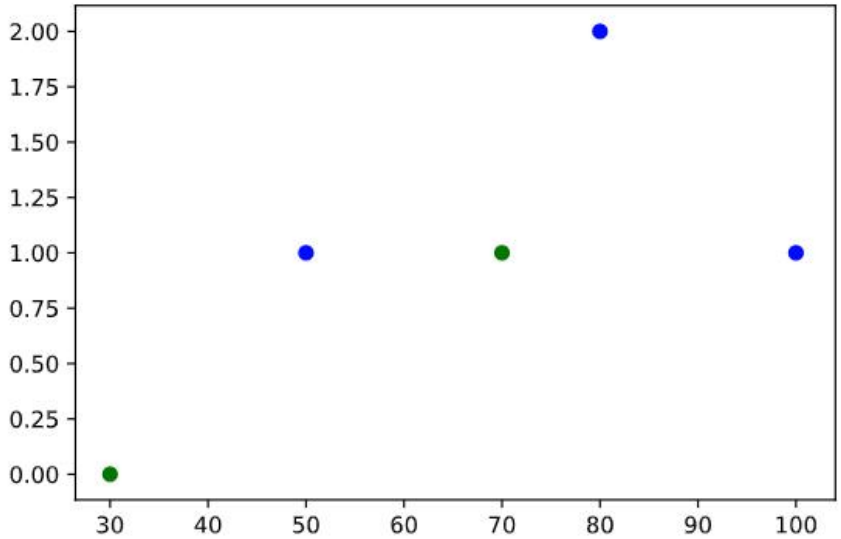
(a)

Solution

```
import numpy as np
import matplotlib
import matplotlib.pyplot as plt
#matplotlib inline
%config InlineBackend.figure_format = 'svg'

x1 = np.array([30, 50, 70, 80, 100])
x2 = np.array([0, 1, 1, 2, 1])
y = np.array([0, 1, 0, 1, 1])
n = len(y)

for i in range(n):
    if y[i] == 0:
        plt.scatter(x1[i], x2[i], c='g')
    if y[i] == 1:
        plt.scatter(x1[i], x2[i], c='b')
```



(b)(c)(d)

Solution:

(b)
A simple classifier is to use $x_2 = 0.5$.

So $z_i = x_2 - 0.5 = w^T x + b$

$\therefore w = 1, \quad b = -0.5.$

(c).
From question, we have:

$$p(y_i = 1 \mid x_i) = \frac{1}{1+e^{-z_i}}$$

$$\therefore p(y_i = 0 \mid x_i) = 1 - \frac{1}{1+e^{-z_i}} = \frac{1}{e^{z_i}+1}$$

$$\therefore p(y_i \mid x_i) = \begin{cases} \frac{1}{1+e^{-z_i}}, & y = 1 \\ \frac{1}{1+e^{z_i}}, & y = 0. \end{cases}$$

$z_i = x_2 - 0.5$

So

x_{i1}	30	50	70	80	100
x_{i2}	0	1	1	2	1
y_i	0	1	0	1	1
z_i	-0.5	0.5	0.5	1.5	0.5

$\therefore$ The smallest sample is $i = 4$

(d).
since $\alpha > 0$, $\hat{y}$ will not change. But the likelihood will change under the demand of whether $z_i > 0$ or $z_i < 0$, ~~because~~ since:

~~$z_i' = \alpha w$~~

~~w'~~ $z_i' = (w')^T x_i + b' = \alpha[w^T x_i + b] = \alpha z_i.$

4. Suppose we collect data for a group of students in a machine learning class with variables X_1 = hours studied, X_2 = undergrad GPA, and Y = receive an A. We fit a logistic regression and produce estimated coefficient, $\beta_0 = -6$, $\beta_1 = 0.05$, $\beta_2 = 1$.

(a) Estimate the probability that a student who studies for 40 h and has an undergrad GPA of 3.5 gets an A in the class.

(b) How many hours would the student in part (a) need to study to have a 50 % chance of getting an A in the class?

(a)

From the question, we know:

$$z = \beta_0 + \beta_1 x_1 + \beta_2 x_2, \quad \beta = [-6, 0.05, 1]$$

$$\therefore z = -6 + 0.05 x_1 + x_2$$

when $x_1 = 40$, $x_2 = 3.5$

$$z = -6 + 0.05 \times 40 + 3.5 = -0.5$$

$$\therefore p(y=1|x) = \frac{1}{1+e^{-z}} = \frac{1}{1+e^{0.5}} \approx 0.375$$

(b). From the question, we have:

$$p(y_i|x_i) = \frac{1}{1+e^{-z}} = \frac{1}{2}$$

$$\therefore z = 0$$

$$\therefore z = -6 + 0.05 x_1 + x_2 = 0$$

$$-6 + 0.05 x_1 + 3.5 = 0.$$

$$\therefore x_1 = 50$$

The student in part (a) needs to study 50 hours to have a 50% chance of getting an A in the class.