Submitted by: Ranji Raj

Assignment 2.3

For our data we can derive *categorical* and *numerical* features. Describe 2 probabilistic distributions/models and their application for a categorical and a numerical support.

By Gaussian distribution:

Given by: $$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$$

Where μ = sample mean, σ = sample standard deviation, σ^2 = sample variance

Consider a dataset having ***1 numerical*** column & ***3 categorical*** columns where *Defaulted Borrower* is the target column.

And want to classify a new record:
X=(Home Owner=N, Marital Status=Married, Annual Income=90) = ?

#	Home Owner	Marital Status	Annual Income	Defaulted Borrower
0	Y	Single	125	N
1	N	Married	100	N
2	N	Single	70	N
3	Y	Married	120	N
4	N	Divorced	95	Y
5	N	Married	60	N
6	Y	Divorced	220	N
7	N	Single	85	Y
8	N	Married	75	N
9	N	Single	90	Y

Submitted by: Ranji Raj

Counts for Home Owner (Distribution table)

Features	DB=Y	DB=N
Home Owner		
Y	0	3
N	3	4

Counts for Marital Status (Distribution table)

Features	DB=Y	DB=N
Marital Status		
Single	2	2
Married	0	4
Divorced	1	1

Mean & Std. deviation for Annual Income

$$\mu^{AnnualIncome}{}_{DB=Y} = \frac{95 + 85 + 90}{3} = 90$$

$$\mu^{AnnualIncome}{}_{DB=N} = \frac{125 + 100 + 70 + 120 + 60 + 220 + 75}{3} = 110$$

$$\sigma^{AnnualIncome}{}_{DB=Y} = \sqrt{\frac{\sum_{i \in \{1,..n\}} \left(x_i - 90\right)^2}{2}} = 5$$

$$\sigma^{AnnualIncome}{}_{DB=N} = \sqrt{\frac{\sum_{i \in \{1,..n\}} \left(x_i - 110\right)^2}{6}} = 54.5$$

Probabilities for Annual Income = 90

P(Annual Income=90 | DB=Y) = $\frac{1}{5\sqrt{2\pi}} \cdot e^{-\frac{(90-90)^2}{2 \cdot 5^2}} = 0.08$

P(Annual Income=90 | DB=N) = $\frac{1}{54.5\sqrt{2\pi}} \cdot e^{-\frac{(90-110)^2}{2 \cdot 54.5^2}} = 0.007$

Now we classify our record by using **Joint Probability** as:

$$P(DB=Y|X) = P(Home\ Owner=N|Y) \cdot P(Marital\ Status=Married|Y) \cdot P(Annual\ Income=90|Y) \cdot P(Y)$$

$$\frac{3}{3} \cdot \frac{0}{3} \cdot 0.08 \cdot \frac{3}{10} = 0$$

$$P(DB=N|X) = P(Home\ Owner=N|N) \cdot P(Marital\ Status=Married|N) \cdot P(Annual\ Income=90|N) \cdot P(N)$$

$$\frac{4}{7} \cdot \frac{4}{7} \cdot 0.007 \cdot \frac{7}{10} = 0.0016$$

Since P(DB=N | X) > P(DB=Y | X), the record is classified as **DB = N.**

By Logistic regression:

It is a *probabilistic model* that classifies a given record to either 0 or 1.

The function by which it classifies a record to 0 or 1 is by means of a *sigmoid* function which is given as, $\frac{1}{1+e^{-y}}$

In order to be able to do the classification using logistic regression for our problem we need to transform our dataset as follows:

1. Assigning **N** as **1**, **Y** as **0**
2. Converting categorical column '**Marital Status**' into '**IS_SINGLE**', '**IS_MARRIED**', '**IS_DIVORCED**'

So our new dataset looks like:

#	Home Owner	IS_SINGLE	IS_MARRIED	IS_DIVORCED	Annual Income	Defaulted Borrower
0	0	1	0	0	125	1
1	1	0	1	0	100	1
2	1	1	0	0	70	1
3	0	0	1	0	120	1
4	1	0	0	1	95	0
5	1	0	1	0	60	1
6	0	0	0	1	220	1
7	1	1	0	0	85	0
8	1	0	1	1	75	1
9	1	1	0	0	90	0

Let us train our model on some of the instances:

We want to predict the class of 'Defaulted Borrower' (target).

From the basic equation of line, $y = w_0 + w_1 \cdot x_1 + w_2 \cdot x_2 + w_3 \cdot x_3 + w_4 \cdot x_4$

Where, w0 = 'Home Owner'
w1 = '**IS_SINGLE**'
w2 = '**IS_MARRIED**'
w3 = '**IS_DIVORCED**'
w4 = 'Annual Income'

Classify on #0: $y_0 = 0 + 1 + 0 + 0 + 125 = 126 \Rightarrow \dfrac{1}{1 + e^{-126}} = 1$ ⇒ Correct classification
(The explanatory variables x_i are taken as value 1 when 'Home Owner'=0)

Classify on #9: $y_9 = 1 + 1 \cdot 0 + 0 + 0 + 90 \cdot 0 = 1 \Rightarrow \dfrac{1}{1 + e^{-1}} = 0.73$ ⇒ Correct classification
(The explanatory variables x_i are taken as value 0 when 'Home Owner'=1)

Now train on our test sample **X=(Home Owner=N, Marital Status=Married, Annual Income=90) =?**

Classify on the test: $y_{test} = 1 + 0 + 1 \cdot 0 + 0 + 90 \cdot 0 = 1 \Rightarrow \frac{1}{1 + e^{-1}} = 0.73 \Rightarrow$

Correct classification (i.e. **DB=N**) the same as the case with Gaussian distribution.