DSE 210: Final Exam

Professor: A. Enis Çetin

Teaching Assistant: Shivani Agrawal

Joshua Wilson **A53228518**

Problem 1

Given :
Urn A contains 2 white balls and 4 red balls;
Urn B contains 8 white balls and 5 red balls;
Urn C contains 1 white ball and 3 red balls.
$p(Urn\ A) = 0.5$, and $p(Urn\ B) = p(Urn\ C)$

(a) $p(Urn\ A \mid red) = \dfrac{p(red \mid Urn\ A) \times p(Urn\ A)}{p(red)} = \dfrac{\frac{4}{6} \times \frac{1}{2}}{p(red)}$, and

$$\begin{aligned} p(red) &= p(red \mid Urn\ A) \times p(Urn\ A) + p(red \mid Urn\ B) \times p(Urn\ B) + p(red \mid Urn\ C) \times p(Urn\ C) \\ &= \left(\frac{4}{6}\right)\left(\frac{1}{2}\right) + \left(\frac{5}{13}\right)\left(\frac{1}{4}\right) + \left(\frac{3}{4}\right)\left(\frac{1}{4}\right) \\ &= \left(\frac{4}{12}\right) + \left(\frac{5}{52}\right) + \left(\frac{3}{16}\right) = \frac{208 + 60 + 117}{624} = \frac{385}{624} \approx 0.617, \text{ so} \end{aligned}$$

$$p(Urn\ A \mid red) = \frac{\frac{4}{6} \times \frac{1}{2}}{\frac{385}{624}} = \frac{\frac{4}{12}}{\frac{385}{624}} = \boxed{\frac{208}{385} \approx 0.54}$$

(b) $p(white) = 1 - p(red) = 1 - \dfrac{385}{624} = \boxed{\dfrac{239}{624} \approx 0.38}$

Problem 2

Let set $X = \{x_1, x_2, \ldots, x_n\}$.

(a) We can create a subset of X by choosing any k elements of X, such that there are $\binom{n}{k}$ possible subsets of size k. There are therefore $\sum_{k=0}^{n} \binom{n}{k}$ total subsets of X.

By applying the binomial theorem, we have

$$\sum_{k=0}^{n} \binom{n}{k} = \sum_{k=0}^{n} \binom{n}{k} 1^{n-k} 1^k = (1+1)^n = 2^n$$

Alternatively, every element $x_i \in X$ will either be included or excluded from a subset of X, and we can create a subset of X by either including or excluding each x_i, and the choice of including or excluding each x_i is independent.

Since we have 2 choices for n elements, there are 2^n possible subsets of X that we can create.

(b) Since there are 2^n total subsets of X as shown in part (a), and one is the empty set $\varnothing$, there are $2^n - 1$ nonempty subsets of X. We will construct all nonempty subsets of X as follows:

Step 1: Take element x_1, and create the union of x_1 with every possible subset of the n-1 elements in $\{x_2, x_3, \ldots, x_n\}$. As shown in part (a), the number of subsets of $\{x_2, x_3, \ldots, x_n\}$ is 2^{n-1}. Because we are adding x_1 to each of these subsets, none will be empty, and they will all include the element x_1.

Step 2: Take element x_2, and create the union of x_2 with every possible subset of the n-2 elements in $\{x_3, x_4, \ldots, x_n\}$. We will have 2^{n-2} nonempty subsets of X, each of which includes element x_2, and none of which include element x_1. Thus, we have created 2^{n-2} new subsets of X.

Problem 2 continued on next page...

$\vdots$

Step k: Take element x_k, and create the union of x_k with every possible subset of the n-k elements in $\{x_{k+1}, x_{k+2}, \ldots, x_n\}$. We will have 2^{n-k} nonempty subsets of X, each of which includes element x_k, and none of which include elements in $\{x_1, x_2, \ldots, x_{k-1}\}$. Thus, we have created 2^{n-k} new subsets of X.

$\vdots$

Step n: Take element x_n, and create the union of x_n with the empty set $\varnothing$. We will have a single nonempty subset of X (i.e. $\{x_n\}$), which includes no elements in $\{x_1, x_2, \ldots, x_{n-1}\}$. Thus, we have created $1 = 2^0$ new subset of X.

At each step k above, we created 2^{n-k} unique, nonempty subsets of X. The total number of nonempty subsets of X is therefore $\sum_{k=1}^{n} 2^{n-k}$, so $2^n - 1 = \sum_{k=1}^{n} 2^{n-k}$.

Problem 3

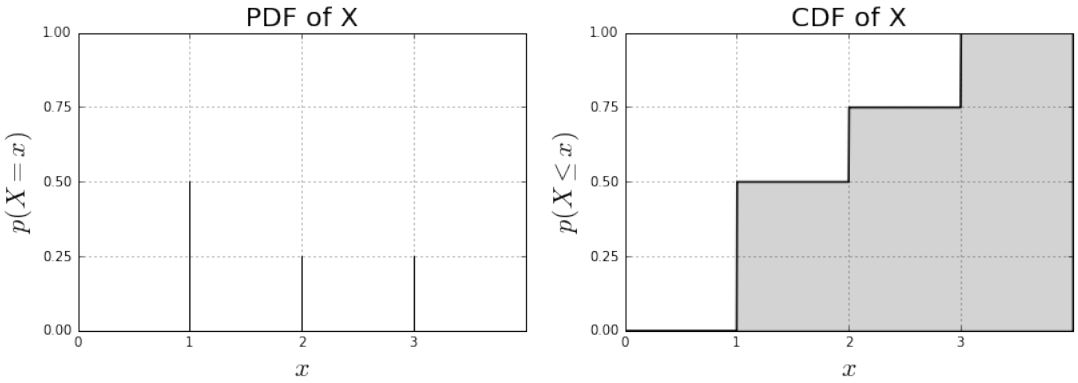
Given :
$p(X = 1) = 0.5$
$p(X = 2) = 0.25$
$p(X = 3) = 0.25$

(a) $\mathbb{E}[X] = \sum_x x \times p(X = x) = 1 \times 0.5 + 2 \times 0.25 + 3 \times 0.25 = 0.5 + 0.5 + 0.75 = \boxed{1.75}$

(b)

$var[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2$, and
$\mathbb{E}[X^2] = 1^2 \times 0.5 + 2^2 \times 0.25 + 3^2 \times 0.25 = 0.5 + 4 \times 0.25 + 9 \times 0.25 = 0.5 + 1 + 2.25 = 3.75$, and
$\mathbb{E}[X]^2 = 1.75^2 = 3.0625$, so
$var[X] = 3.75 - 3.0625 = \boxed{0.6875}$

(c)

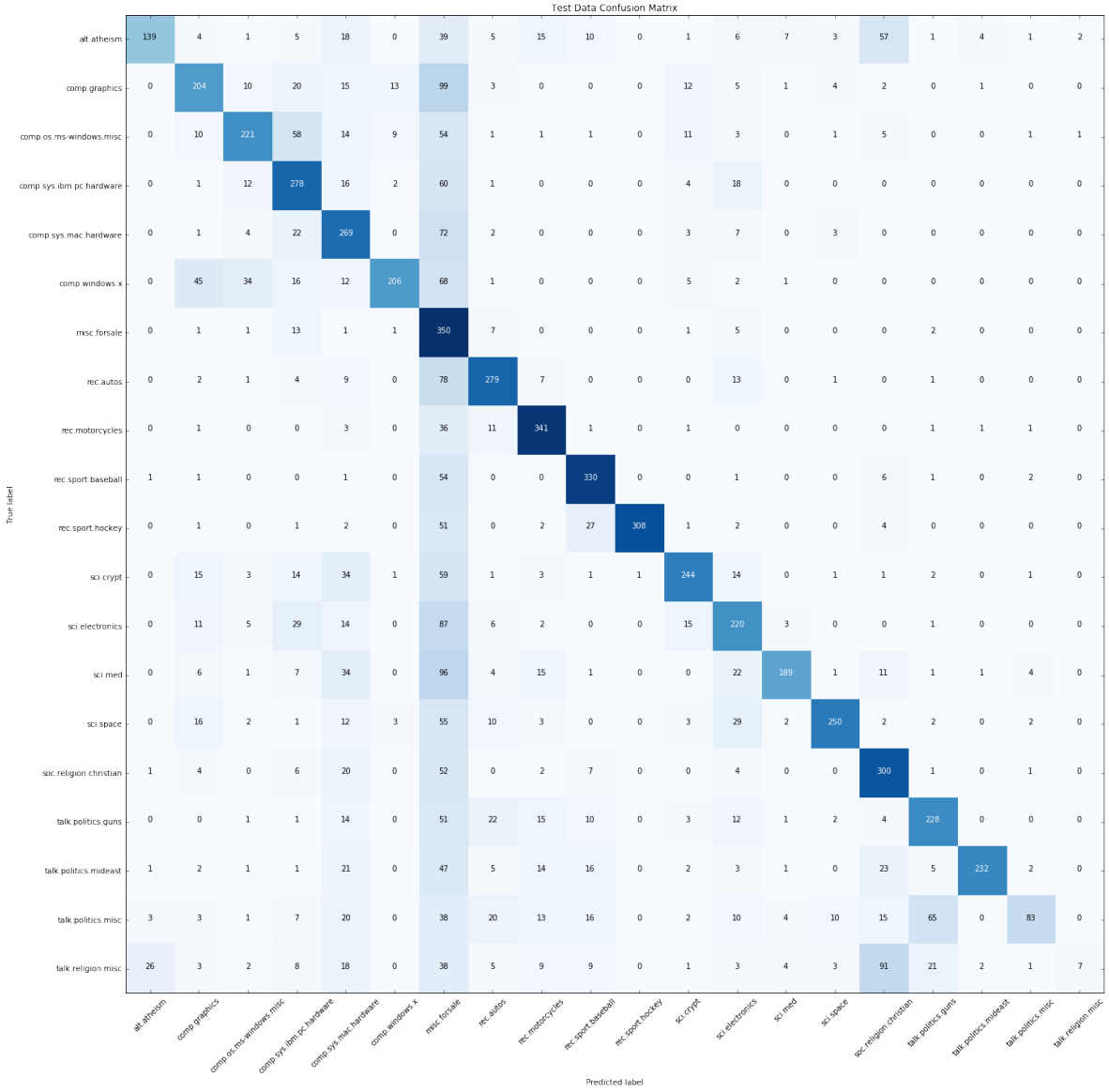


Problem 4

See DSE210_Final_Q4.ipynb notebook at https://github.com/mas-dse/jsw037/tree/master/DSE210.

(a) Error rate on the test dataset is 0.3767.

(b) Confusion Matrix:



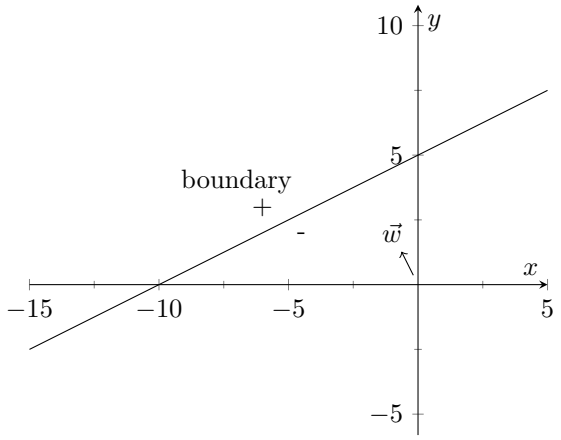
Problem 5

Given : $\vec{w} = \begin{bmatrix} -1 \\ 2 \end{bmatrix}, \theta = 10$. Find the decision boundary $\vec{w} \bullet \vec{x} \geq \theta$ in $\mathbb{R}^2$.

Let $\vec{x} = \begin{bmatrix} x \\ y \end{bmatrix}$, then $\vec{w} \bullet \vec{x} = \begin{bmatrix} -1 \\ 2 \end{bmatrix} \bullet \begin{bmatrix} x \\ y \end{bmatrix} = -1x + 2y \geq 10 \implies y \geq 5 + \frac{1}{2}x$.

If $x = 0$, then $y \geq 5$, and if $y = 0$, then $x \leq -10$.

The decision boundary is depicted below:



Problem 6

Given :
Urn A $\sim N(0,1)$, $\quad p(A) = \frac{2}{3}$, $\quad$ Urn B $\sim N(5,2)$, $\quad p(B) = \frac{1}{3}$

(a) $p(A \mid x = 2.5) = \dfrac{p(x = 2.5 \mid A) \times p(A)}{p(x = 2.5)} = \dfrac{p(x = 2.5 \mid A) \times p(A)}{p(x = 2.5 \mid A) \times p(A) + p(x = 2.5 \mid B) \times p(B)}$,

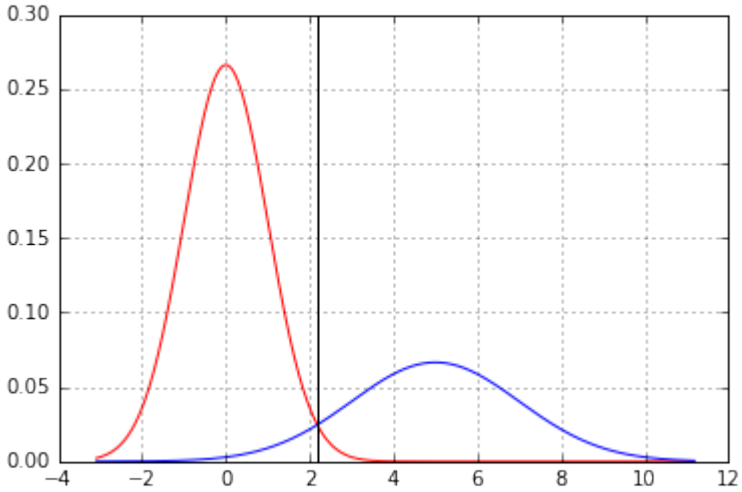
where $p(x = 2.5 \mid A) = \dfrac{1}{\sqrt{2\pi\sigma_A^2}} \exp\left(- \dfrac{(2.5 - \mu_A)^2}{2\sigma_A^2} \right) = \dfrac{1}{\sqrt{2\pi(1)^2}} \exp\left(- \dfrac{(2.5 - (0))^2}{2(1)^2} \right) \approx 0.01753,$

and $p(x = 2.5 \mid B) = \dfrac{1}{\sqrt{2\pi\sigma_B^2}} \exp\left(- \dfrac{(2.5 - \mu_B)^2}{2\sigma_B^2} \right) = \dfrac{1}{\sqrt{2\pi(2)^2}} \exp\left(- \dfrac{(2.5 - (5))^2}{2(2)^2} \right) \approx 0.09132,$

so $p(A \mid x = 2.5) \approx \dfrac{(0.01753) \times \frac{2}{3}}{(0.01753) \times \frac{2}{3} + (0.09132) \times \frac{1}{3}} \approx \dfrac{0.01169}{0.04213} = \boxed{0.2774}$

(b) Our prediction is based on the higher of $p(A) \times p(x \mid A)$ and $p(B) \times p(x \mid B)$. The decision boundary is the point at which $p(A) \times p(x \mid A) = p(B) \times p(x \mid B)$. This value is approximately $\boxed{2.181}$.

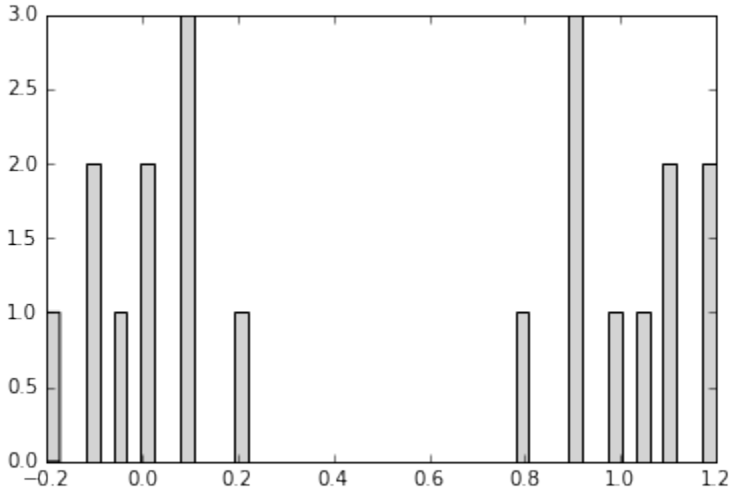
$p(A) \times p(x \mid A)$ and $p(B) \times p(x \mid B)$ are plotted below in red and blue, respectively, along with the vertical decision boundary line.



See DSE210_Final_Scratch.ipynb notebook at https://github.com/mas-dse/jsw037/tree/master/DSE210.

Problem 7

A histogram of X = {-0.1, -0.2, 0.1, 0.2, 0, 0.1, -0.1, 0, -0.05, 0.1, 1.05, 1.1, 0.9, 0.8, 0.9, 1, 1.2, 1.1, 1.2, 0.9} is below.



The cluster centers are located at $\boxed{0.005 \text{ and } 1.015}$.

See DSE210_Final_Scratch.ipynb notebook at https://github.com/mas-dse/jsw037/tree/master/DSE210.

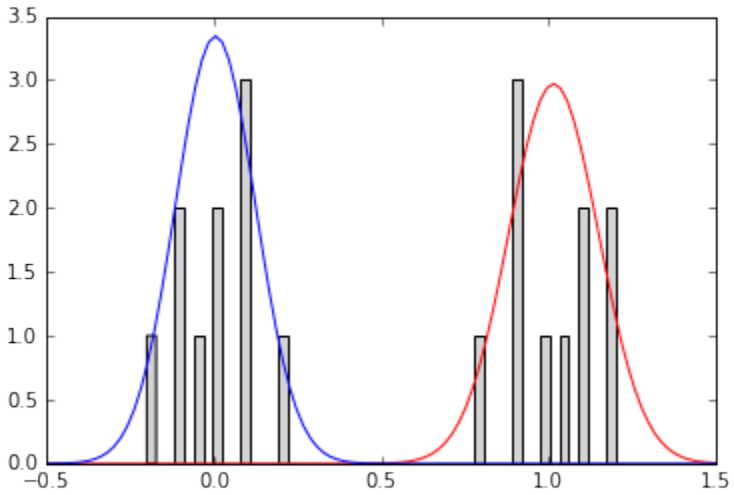
Problem 8

The best estimate for the number of mixtures is two:

$N(\mu_1 = 0.005, \sigma_1^2 = 0.014225)$ and $N(\mu_2 = 1.015, \sigma_2^2 = 0.018025)$.

The best estimate was determined by trying from 1 to 10 mixtures and choosing the mixture with the lowest Akaike information criterion and Bayesian information criterion.

A plot of the original data and the Gaussians is below:



See DSE210_Final_Scratch.ipynb notebook at https://github.com/mas-dse/jsw037/tree/master/DSE210.

Problem 9

Given :
$n = 1000$, sample average 307, sample standard deviation 30,

The estimate of the nationwide average is $\hat{\mu} = 307$.

The standard deviation of this estimate is equal to the sample standard deviation divided by $\sqrt{n}$, so $\hat{\sigma} = \dfrac{30}{\sqrt{1000}} \approx \boxed{0.9487}$.

Problem 10

Given :
A random sample of 48 men and 55 women produced the following results:

	Men	Women
Never Married	43.8%	16.4%
Married	41.7%	70.9%
Widowed, divorced, separated	14.6%	12.7%
Total	100%	100%

We will test the hypothesis that there is no differences in the distribution of results for men and women by performing a χ^2 test.

Define the null hypothesis as H_0 : The results for men and women come from the same distribution.

To do so, we estimate the underlying distribution and expected frequencies for each of the two samples by aggregating the results:

	Men %	Women %	Men Obs	Women Obs	Total Obs	Exp %
Never Married	43.8%	16.4%	21	8	30	29.1%
Married	41.7%	70.9%	20	39	59	57.3%
Widowed, divorced, separated	14.6%	12.7%	7	7	14	13.6%
Total	100%	100%	48	55	103	100%

Based on these expected percentages, we can calculate the following expected results for men and women, based on the number of each surveyed:

	Exp Men	Exp Women
Never Married	13.968	16.005
Married	27.504	31.515
Widowed, divorced, separated	6.528	7.48
Total	48	55

Now we can compute the χ^2 statistic for this data:

$$\chi^2 = \sum_{outcomes} \frac{\left((\text{observed frequency}) - (\text{expected frequency})\right)^2}{(\text{expected frequency})}$$

$$= \frac{(21-13.968)^2}{13.968} + \frac{(20-27.504)^2}{27.504} + \frac{(7-6.528)^2}{6.528} + \frac{(9-16.005)^2}{16.005} + \frac{(39-31.515)^2}{31.515} + \frac{(7-7.48)^2}{7.48}$$

$$\approx 3.54 + 2.05 + 0.03 + 3.07 + 1.78 + 0.03 \approx 10.5$$

$$\implies p\text{-value} \approx 0.005 \ \therefore \ \text{reject } H_0$$

The results we observe are very unlikely to have happened if the results for men and women come from the same distribution, so we reject the null hypothesis and conclude that the distributions really are different for men and women.