Gary Chen
804361027

Homework 1

EE M146
April 10, 2019

1 Problem 1

Assume that there are two urns. The first urn contains 4 red balls, 3 blue balls, and 3 white balls. The second urn contains 2 red balls, 4 blue balls, and 4 white balls. You randomly select an urn and take two balls from the urn. The probability that you pick the first urn is 40%. What is the probability that

a) **the two balls are red?**

We may define the total probability of two events occurring after an urn is chosen as the union of all the probabilities over all complementary urn selections:

$$\begin{aligned} P(A,B) &= \sum_i P(A \cap B \cap u_i) \\ &= \sum_i P(u_i)P(A|u_i)P(B|(u_i \cap A)) \end{aligned} \tag{1}$$

given that we have two urns:

$$\begin{aligned} P(u_1) &= 0.4 \\ P(u_2) &= 0.6 \end{aligned}$$

Let $A = B = r$

$$\begin{aligned} P(r|u_1) &= 0.4 \\ P(r|r \cap u_1) &= \frac{1}{3} \\ P(r|u_2) &= 0.2 \\ P(r|r \cap u_2) &= \frac{1}{9} \end{aligned}$$

By using equation 1:

$$\boxed{P(r,r) = 0.0667}$$

b) **the second ball is blue?**

Let $B = b$ where $A = b, b^c$ are the two possible initial draws. This manifests itself as another union operation where there is no intersection as the two options are complementary. We can change equation 1 accordingly:

$$P(A,b) = \sum_i P(u_i)[P(b|u_i)P(b|(u_i \cap b)) + P(b^c|u_i)P(b|(u_i \cap b^c))] \tag{2}$$

Where the probabilities of interest are:

$$\begin{aligned} P(b|u_{1,2}) &= 0.3, 0.4 \\ P(b|(u_{1,2} \cap b)) &= \frac{2}{9}, \frac{1}{3} \\ P(b^c|u_{1,2}) &= 0.7, 0.6 \\ P(b|(u_{1,2} \cap b^c)) &= \frac{1}{3}, \frac{4}{9} \end{aligned}$$

By using equation 2:

$$\boxed{P(A, b) = 0.36}$$

c) **the second ball is blue given that the first ball is red?**

We use Baye's theorem:

$$P(b|r) = \frac{P(b)P(r|b)}{P(r)} \tag{3}$$

$$= \frac{\sum_i P(u_i)P(b|u_i)P(r|b|u_i)}{\sum_i P(r|u_i)P(u_i)} \tag{4}$$

Where the probabilities of interest are:

$$P(r|b|u_{1,2}) = \frac{4}{9}, \frac{2}{9}$$

By using equation 4:

$$\boxed{P(b|r) = 0.38}$$

2 Problem 2

A fair coin is tossed four times. Let Z be the number of coins that turn up heads.

The number of heads Z that one would get for $n = 4$ flips, where each flip has a probability of getting heads $p = 0.5$ follows a binomial distribution.

$$P(Z, n, p) = \binom{n}{Z} p^Z (1 = p)^{n-Z}$$

Where the binomial coefficient is defined as $\binom{n}{Z} = \frac{n!}{Z!(n-Z)!}$. But using this is probably too easy. Fundamentally, each event is a Bernoulli trial with expectation value $E[X] = 0.5$ and variance $var(X) = 0.25$. Z is the total number of heads, $Z = \sum_i^4 X_i$

a) **What is the expected value of Z?**

$$\begin{aligned} E[Z] &= E\left[\sum_i^4 X_i\right] \\ &= \sum_i^4 E[X] \\ &= 4E[X] \\ &= \boxed{2} \end{aligned}$$

b) **What is the variance of Z?**

$$\begin{aligned} var[Z] &= var\left[\sum_i X_i\right] \\ &= \sum_i^4 var[X] \\ &= 4var[X] \\ &= \boxed{1} \end{aligned}$$

3 Problem 3

Generally, students who study regularly have a probability of 85% to get an A, 10% to get a B, and 5% to get a C or lower. Unfortunately, some students only study right before a test. These students have a probability of 50% to get an A, 15% to get a B, and 35% to get a C or lower. Thankfully, the students who study regularly make up 90% of the student body.

The conditional probabilities of the grades earned given if the student studies regularly (S), or right before the test (N) can be shown to be:

$$P(S) = 0.90 \qquad P(N) = 0.10$$
$$P(A|S) = 0.85 \qquad P(A|N) = 0.50$$
$$P(B|S) = 0.10 \qquad P(B|N) = 0.15$$
$$P(\leq C|S) = 0.05 \qquad P(\leq C|N) = 0.35$$

Where the total probability of finding any grade is:

$$P(G) = P(G|S)P(S) + P(G|N)P(N)$$

a) **Given that a student received an A on the test, what is the probability that they studied?**

$$\begin{aligned} P(S|A) &= \frac{P(A|S)P(S)}{P(A)} \\ &= \frac{P(A|S)P(S)}{P(A|S)P(S) + P(A|N)P(N)} \\ &= \boxed{0.94} \end{aligned}$$

b) **Given that a student received a B or lower, what is the probability that they didn't study?**

We consider the union of mutually exclusive events B and $\leq C$

$$\begin{aligned} P(N|B \cup \leq C) &= \frac{P(B \cup \leq C|N)P(N)}{P(B \cup \leq C)} \\ &= \frac{P(N)[P(B|N) + P(\leq C|N)]}{P(S)[P(B|S) + P(\leq C|S)] + P(N)[P(B|N) + P(\leq C|N)]} \\ &= \boxed{0.27} \end{aligned}$$

4 Problem 4

Let X and Y be discrete random variables. Let $E[X]$ and $var[X]$ be the expected value and variance, respectively, of a random variable X.

a) **Show that $E[X + Y] = E[X] + E[Y]$.**

$$\begin{aligned}
E[X+Y] &= \sum_x \sum_y (x+y)p(x,y) \\
&= \sum_x \sum_y xp(x,y) + \sum_x \sum_y yp(x,y) \\
&= \sum_x x \sum_y p(x,y) + \sum_y y \sum_x p(x,y) \\
&= \sum_x xp(x,y) + \sum_y yp(x,y) \\
&= \boxed{E[X]+E[Y]}
\end{aligned}$$

b) **If X and Y are independent, show that $var[X+Y] = var[X] + var[Y]$.**

$$\begin{aligned}
var[X+Y] &= E[(X+Y)^2] - E[X+Y]^2 \\
&= E[X^2 + 2XY + Y^2] - (E[X]+E[Y])^2 \\
&= E[X^2] + 2E[XY] + E[Y^2] - E[X]^2 - 2E[X]E[Y] - E[Y]^2
\end{aligned}$$

For independent variables X and Y, $E[XY] = E[X]E[Y]$

$$\begin{aligned}
&= E[X^2] - E[X]^2 + E[Y^2] - E[Y]^2 \\
&= \boxed{var[X] + var[Y]}
\end{aligned}$$

If X and Y are were not independent, there would be a non-zero covariance term $cov(X,Y) = E[XY] - E[X]E[Y]$.

5 Problem 5

Suppose that you are waiting at a bus stop. The waiting time until a bus arrives is T where T is an exponentially distributed random variable with parameter λ i.e. $P(T \leq t) = 1 - e^{-\lambda t}, \forall t \geq 0$.

a) **Given that you have already waited r seconds, what is the probability that the bus will not arrive within d more seconds?**

Let the amount of time already waited $t = r$ and the further wait time $t = r + d$

$$\begin{aligned} P(T > r + d|T > r) &= \frac{P(T > r + d)P(T > r|T > r + d)}{P(T > r)} \\ &= \frac{P(T > r + d)}{P(T > r)} \\ &= \frac{1 - P(T \leq r + d)}{1 - P(T \leq r)} \\ &= \frac{e^{-\lambda(r+d)}}{e^{-\lambda r}} \\ &= \boxed{e^{-\lambda d}} \end{aligned}$$

b) **What is the average waiting time for the bus i.e. the expected value of T? Hint: Recall that one way to solve $\int udv$ is by integration by parts.**

The expected value of a continuous random variable is $\int_0^\infty xp(x)dx$ where $p(x)$ is the probability density function. We first determine $p(t)$:

$$\begin{aligned} p(t) &= \frac{d}{dt}P(T \leq t) \\ &= \lambda e^{-\lambda t} \end{aligned}$$

Where then the expectation value is:

$$\begin{aligned} E[T] &= \int_0^\infty tp(t)dt \\ &= \int_0^\infty t\lambda e^{-\lambda t} \end{aligned}$$

Applying integration by parts, we yield $u = t$, $v = -e^{-\lambda t}$

$$\begin{aligned} &= -te^{-\lambda t}\Big|_0^\infty + \int_0^\infty e^{\lambda t}dt \\ &= 0 - \frac{1}{\lambda e^{-\lambda t}}\Big|_0^\infty \\ &= \boxed{\frac{1}{\lambda}} \end{aligned}$$

6 Problem 6

In this exercise, you will implement the perceptron algorithm in MATLAB. You will be provided with 3 datasets: data1.csv, data2.csv, and data3.csv. Each dataset will have three columns. The first two columns are the attributes of the datapoint and the third column is the label for each datapoint. The attributes have been normalized so that $\|x\| \leq 1$. Each label is either 1 or -1.

a) **Plot all datasets. Which datasets are linearly separable?**

(Figure 1) Data1 is clearly separable, data2 may be separable within a small margin, but data3 is clearly mixed and not linearly separable.

b) **Implement the perceptron algorithm as shown in chapter 4 of *A Course in Machine Learning*. To allow for the same results, initialize the hyperplane parameters as 0, iterate through data points in the order provided. Set the maximum iteration number to 1000. For each dataset, provide the hyperplane parameters that are learned by the perceptron algorithm (w and b) and report the total number of updates performed (u). In addition, for each data set, provide a plot that shows both the data and the decision boundary, i.e., the line defined by $w^T x + b = 0$. Based on the total number of updates performed, comment on the convergence of perceptron algorithm for each data set.**

(Figure 2) Data sets:

1) Required 2 updates to find a solution for $\vec{w}$, converging rapidly as expected.
2) Required 4 updates to find a solution for $\vec{w}$, converging rapidly as expected. The convergence was not as fast as that of data set 1, showing that the margin for the separable line is more narrow.
3) Took 4953 updates, but never fully converged.

c) **Now, you will compare the rate of convergence for the linearly separable datasets. Recall that the margin $\gamma_{w,b}$ is the distance between the hyperplane defined by w, b and the nearest point of a set. The margin γ of a set is the largest $\gamma_{w,b}$ for all hyperplanes w, b that separate the set. As shown in lecture, the number of updates needed to converge is upper bounded by $\frac{1}{\gamma^2}$. Unfortunately, we currently do not have the tools to find γ (will be discussed when the course reaches SVMs). Fortunately, we can use the hyperplane (defined by w and b) found by the perceptron algorithm to get an lower bound on the margin since by definition $\gamma_{w,b} \leq \gamma$ which implies that $\frac{1}{\gamma^2} \leq \frac{1}{\gamma_{w,b}^2}$**

For each linearly separable dataset, calculate the margin $\gamma_{w,b}$ using the learned parameters and compare the upper bound $\left(\frac{1}{\gamma_{w,b}^2}\right)$ to the number of updates that you actually had.

$$d_i = \frac{|\vec{w} \cdot \vec{x}_i + b|}{\|\vec{w}\|}$$
$$\gamma_{w,b} = min(d_i)$$
$$k = \frac{1}{\gamma_{w,b}}$$

For the linearly separable data sets 1 and 2, we find that $\gamma_{w,b,1} = 0.19$ and $\gamma_{w,b,2} = 0.01$. With these values, the upper bounds on the updates (k) are found to be $k_1 = 27$ and $k_2 = 7252$. The number of updates that our algorithm performed were far below these values.

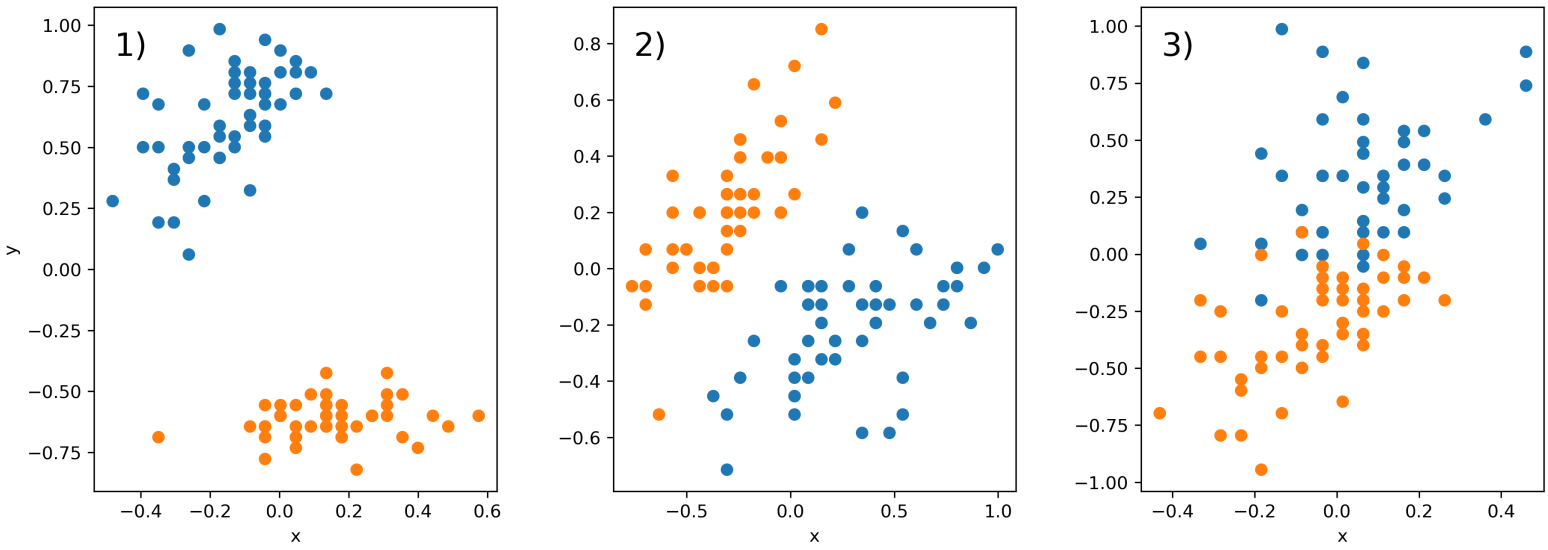

Figure 1: Corresponding plots of data sets 1-3 where (+1) labeled data in orange, and (-1) in blue. Sets 1 and 2 are linearly separable, while 3 is not.

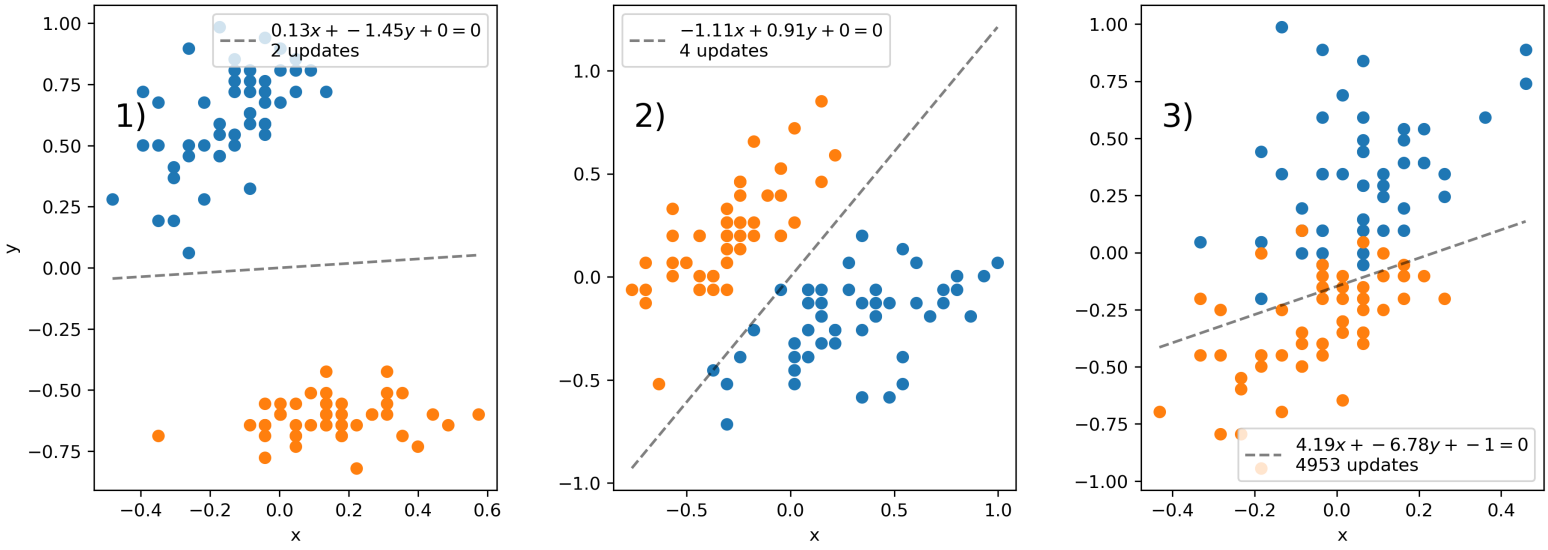

Figure 2: Corresponding plots of data sets 1-3 where (+1) labeled data in orange, and (-1) in blue. Sets 1 and 2 are shown to be linearly separable with a dividing hyperplane (dashed grey) with its equation and number of updates.