ECE M146 Homework 1 Solution
Introduction to Machine Learning
Instructor: Lara Dolecek
TA: Zehui (Alex) Chen, Ruiyi (John) Wu

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?

   (b) the second ball is blue?

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

   Solution:

   (a) $0.4 * \frac{\binom{4}{2}}{\binom{10}{2}} + 0.6 * \frac{1}{\binom{10}{2}}$

   (b) $0.4 * \frac{7\times3+3\times2}{10\times9} + 0.6 * \frac{6\times4+4\times3}{10\times9}$

   (c)
   $$\frac{0.4 \times \frac{4}{10} \times \frac{3}{9} + 0.6 \times \frac{2}{10} \times \frac{4}{9}}{0.4 \times \frac{4}{10} + 0.6 \times \frac{2}{10}}$$

2. Suppose 6 identical dice each with faces numbered 1 through 6 are tossed at the same time. What is the probability of the event "the result of the outcome is such that three different numbers each appear twice?"
   Solution: The problem can be done by the following step:
   Pick 2 out of 6 dice, and let them be one number. Then pick 2 out of the remaining 4 dice, and let them to be a different number. Finally, let the remaining 2 dice be a third number.There are totally $\binom{6}{3}$ ways to pick 3 different numbers out of 6 numbers. So the probability is
   $$\binom{6}{2}\left(\frac{1}{6}\right)^2\binom{4}{2}\left(\frac{1}{6}\right)^2\binom{4}{2}\left(\frac{1}{6}\right)^2\binom{2}{2}\left(\frac{1}{6}\right)^2\binom{6}{3} = \frac{25}{648}.$$

3. In a bolt factory machines A, B, C manufacture, respectively 25, 35 and 40 per cent of the total. Of their product 5, 4, and 2 per cent are defective bolts. A bolt is drawn at random from the produce and is found defective. What are the probabilities that it was manufactured by machines A, B and C?
   Solution:
   Let D denote the event that a bolt randomly drawn from the produce is defective and A, B, C denote the events that it was manufactured by machines A, B and C

respectively. We are interested in the probabilities $P(A|D)$, $P(B|D)$, $P(C|D)$. We have,

$$\begin{aligned} P(D) &= P(D|A)P(A) + P(D|B)P(B) + P(D|C)P(C) \\ &= 0.05 \cdot 0.25 + 0.04 \cdot 0.35 + 0.02 \cdot 0.4 \\ &= 0.0345 \end{aligned} \tag{1}$$

By the Bayesian rule, we get

$$P(A|D) = \frac{P(D|A)P(A)}{P(D)} = \frac{0.05 \cdot 0.25}{0.0345} = 0.3623,$$

$$P(B|D) = \frac{P(D|B)P(B)}{P(D)} = \frac{0.04 \cdot 0.35}{0.0345} = 0.4058,$$

and

$$P(C|D) = \frac{P(D|C)P(C)}{P(D)} = \frac{0.02 \cdot 0.40}{0.0345} = 0.2319.$$

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

   (a) Show that $\mathbb{E}[X+Y] = \mathbb{E}[X] + \mathbb{E}[Y]$.

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

   Solution:

   (a)

$$\begin{aligned} \mathbb{E}[X+Y] &= \sum_x \sum_y (x+y)P(x,y) \\ &= \sum_x \sum_y xP(x,y) + \sum_x \sum_y yP(x,y) \\ &= \sum_x x \sum_y P(x,y) + \sum_y y \sum_x P(x,y) \\ &= \sum_x xP(x) + \sum_y yP(y) \\ &= \mathbb{E}[X] + \mathbb{E}[Y] \end{aligned}$$

   (b)

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

$$\begin{aligned}\mathbb{E}[XY] &= \sum_x \sum_y xyP(x,y) \\ &= \sum_x \sum_y xyP(x)P(y) \\ &= \sum_x xP(x) \sum_y yP(y) \\ &= \mathbb{E}[X]\mathbb{E}[Y]\end{aligned}$$

where the 2nd line comes from the independence assumption.

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

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?

   (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.

   Solution:

   (a)

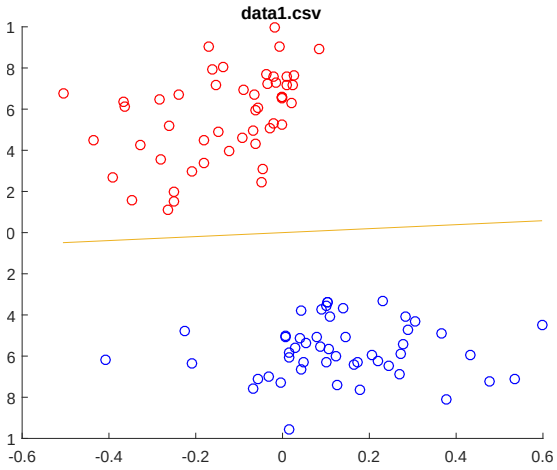
$$\begin{aligned}P(T > r + d | T > r) &= \frac{P(T > r + d, T > r)}{P(T > r)} \\ &= \frac{P(T > r + d)}{P(T > r)} \\ &= \frac{e^{-\lambda*(r+d)}}{e^{-\lambda*r}} \\ &= e^{-\lambda*d}\end{aligned}$$

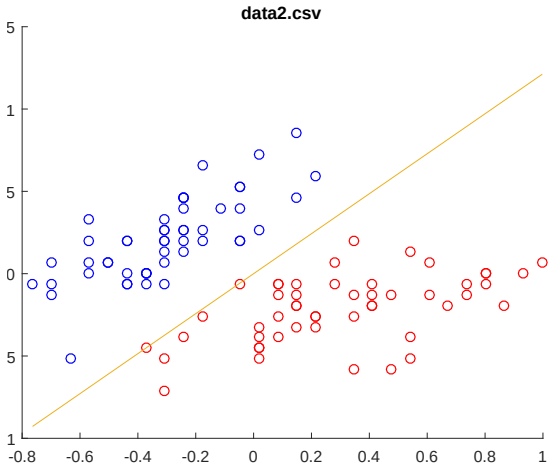
   (b)

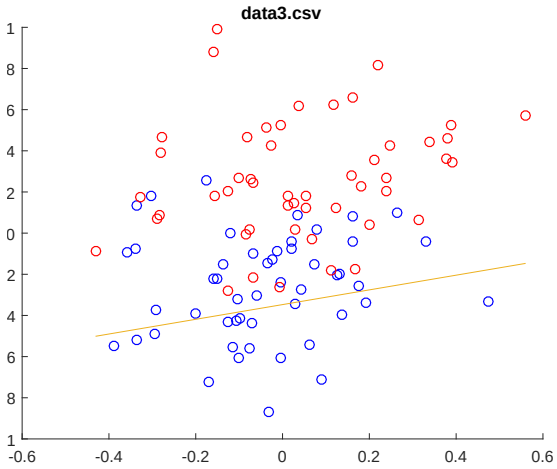
$$\begin{aligned}\mathbb{E}[T] &= \int_0^\infty \lambda\, t e^{-\lambda*t} dt \\ &= -t * e^{-\lambda*t}|_0^\infty + \int_0^\infty e^{-\lambda*t} dt \\ &= 0 + \frac{1}{\lambda}(-e^{-\lambda*t}|_0^\infty) \\ &= \frac{1}{\lambda}\end{aligned}$$

6. In this exercise, you will implement the perceptron algorithm. 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 the datasets. Which datasets are linearly separable?
   Solutions:







The linearly separable datasets are sets 1 and 2.

(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.

Solutions: The above plots already have the hyperplanes plotted. To plot the separating hyperplane, which is a line in 2D, you only need to find x_1, x_2 that satisfy the equation $w_1 x_1 + w_2 x_2 + b = 0$. For *data1*, we have $w_1 = 0.1422, w_2 = -1.4732$, $b = 0$ and $u = 2$. For *data2*, we have $w_1 = -1.1092, w_2 = 0.9134$, $b = 0$ and $u = 4$. For *data3*, we have $w_1 = 1.0291, w_2 = -2.8797$, $b = -1$ and $u = 4501$. For *data3*, the algorithm does not converge since the data is not linearly separable and we can see it take a large number of updates (4501) until the max number of iteration.

(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 ($\frac{1}{\gamma_{w,b}^2}$) to the number of updates that you actually had.

Solutions: For data1.csv, the empirical margin is 0.1369 which means that the upper bound on iterations is 53.3820. Our algorithm converged in 2 updates.

For data2.csv, the empirical margin is 2.6664×10^{-4} which means that the upper bound on iterations is $1.4065 * 10^7$. Our algorithm converged in 4 updates.

In both cases, the actual margin was significantly larger than what we obtained empirically which is intuitive since the algorithm (for linearly separable data) terminates as soon as a separating hyperplane is identified. This separating hyperplane will typically have a small margin.