ECE M146 — Homework 1 Solution
Introduction to Machine Learning — Monday, March 29, 2021
Instructor: Lara Dolecek — Due: Monday, April 5, 2021
TA: Zehui (Alex) Chen — chen1046@ucla.edu

Please upload your homework to Gradescope by April 5, 4:00 pm.
Please submit a single PDF directly on Gradescope
You may type your homework or scan your handwritten version. Make sure all the work is discernible.

1. Let

$$\boldsymbol{x} = \begin{bmatrix} 2 \\ 2 \\ 3 \end{bmatrix}, \boldsymbol{y} = \begin{bmatrix} -1 \\ 0 \\ 2 \end{bmatrix}.$$

   Also let θ be the angle between $\boldsymbol{x}$ and $\boldsymbol{y}$. Calculate the following expressions:

   (a) $\boldsymbol{x^T y}$.

   (b) $\|\boldsymbol{x}\|_1$ and $\|\boldsymbol{y}\|_1$.

   (c) $\|\boldsymbol{x}\|_2$ and $\|\boldsymbol{y}\|_2$.

   (d) $cos(\theta)$.

   Solution:

   (a) $\boldsymbol{x^T y} = -2 + 6 = 4$.

   (b) $\|\boldsymbol{x}\|_1 = 2 + 2 + 3 = 7$ and $\|\boldsymbol{y}\|_1 = 1 + 2 = 3$.

   (c) $\|\boldsymbol{x}\|_2 = \sqrt{2^2 + 2^2 + 3^2} = \sqrt{17}$ and $\|\boldsymbol{y}\|_2 = \sqrt{1 + 2^2} = \sqrt{5}$.

   (d) $cos(\theta)=\frac{\boldsymbol{x^T y}}{\|\boldsymbol{x}\|_2 \|\boldsymbol{y}\|_2} = \frac{4}{\sqrt{95}} = \frac{4\sqrt{95}}{95}$.

2. In a bolt factory machines A, B, C manufacture, respectively 25, 35 and 40 per cent of the total products. Of their product 5, 4, and 2 per cent are defective bolts. A bolt is drawn at random from the products 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 products 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 Bayes 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

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

3. 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.

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

4. (a) Let $x_1, x_2, \cdots, x_n$ be identically distributed random variables. A random vector, $\boldsymbol{x}$, is defined as $\boldsymbol{x} = [x_1, x_2, \cdots, x_n]^T$. What is $\mathbb{E}[\boldsymbol{Ax} + \boldsymbol{b}]$ in terms of $\mathbb{E}[\boldsymbol{x}]$, given that $\boldsymbol{A}$ and $\boldsymbol{b}$ are deterministic?

Solution: Because expectation is a linear operator and vector transformations are linear operations:

$$\mathbb{E}[\boldsymbol{Ax} + \boldsymbol{b}] = \boldsymbol{A}\mathbb{E}[\boldsymbol{x}] + \boldsymbol{b}.$$

(b) Let

$$\boldsymbol{cov}(\boldsymbol{x}) = \mathbb{E}[(\boldsymbol{x} - \mathbb{E}[\boldsymbol{x}])(\boldsymbol{x} - \mathbb{E}[\boldsymbol{x}])^T].$$

What is $\boldsymbol{cov}(\boldsymbol{Ax} + \boldsymbol{b})$ in terms of $\boldsymbol{cov}(\boldsymbol{x})$, given that $\boldsymbol{A}$ and $\boldsymbol{b}$ are deterministic?

Solution:

$$\begin{aligned}
\boldsymbol{cov}(\boldsymbol{Ax} + \boldsymbol{b}) &= \mathbb{E}\left[\left(\boldsymbol{Ax} + \boldsymbol{b} - \boldsymbol{A}\mathbb{E}[\boldsymbol{x}] - \boldsymbol{b}\right)\left(\boldsymbol{Ax} + \boldsymbol{b} - \boldsymbol{A}\mathbb{E}[\boldsymbol{x}] - \boldsymbol{b}\right)^T\right] \\
&= \mathbb{E}\left[\left(\boldsymbol{Ax} - \boldsymbol{A}\mathbb{E}[\boldsymbol{x}]\right)\left(\boldsymbol{Ax} - \boldsymbol{A}\mathbb{E}[\boldsymbol{x}]\right)^T\right] \\
&= \mathbb{E}\left[\boldsymbol{A}(\boldsymbol{x} - \mathbb{E}[\boldsymbol{x}])(\boldsymbol{x} - \mathbb{E}[\boldsymbol{x}])^T\boldsymbol{A}^T\right] \\
&= \boldsymbol{A}\mathbb{E}\left((\boldsymbol{x} - \mathbb{E}[\boldsymbol{x}])(\boldsymbol{x} - \mathbb{E}[\boldsymbol{x}])^T\right)\boldsymbol{A}^T \\
&= \boldsymbol{Acov}(\boldsymbol{x})\boldsymbol{A}^T.
\end{aligned}$$

(c) Let $\boldsymbol{x}$ be a random vector that follows a multivariate Gaussian distribution defined by its mean $\mathbb{E}[\boldsymbol{x}]$ and covariance matrix $\boldsymbol{cov}(\boldsymbol{x}) = \mathbb{E}((\boldsymbol{x}-\mathbb{E}[\boldsymbol{x}])(\boldsymbol{x}-\mathbb{E}[\boldsymbol{x}])^T)$. What is the distribution of $\boldsymbol{y} = \boldsymbol{Ax} + \boldsymbol{b}$, given that $\boldsymbol{A}$ and $\boldsymbol{b}$ are deterministic?

Solution: The linear transformation of a multivariate Gaussian RV is also a multivariate Gaussian RV. A multivariate Gaussian distribution is completely defined by its mean and covariance matrix, so we only need to find $\mathbb{E}[\boldsymbol{y}]$ and $\mathbb{E}((\boldsymbol{y} - \mathbb{E}[\boldsymbol{y}])(\boldsymbol{y} - \mathbb{E}[\boldsymbol{y}])^T)$. From part (a) and part (b), we find that $\boldsymbol{y}$ follows a multivariate Gaussian distribution with mean $\boldsymbol{A}\mathbb{E}[\boldsymbol{x}] + \boldsymbol{b}$ and covariance matrix $\boldsymbol{Acov}(\boldsymbol{x})\boldsymbol{A}^T$.

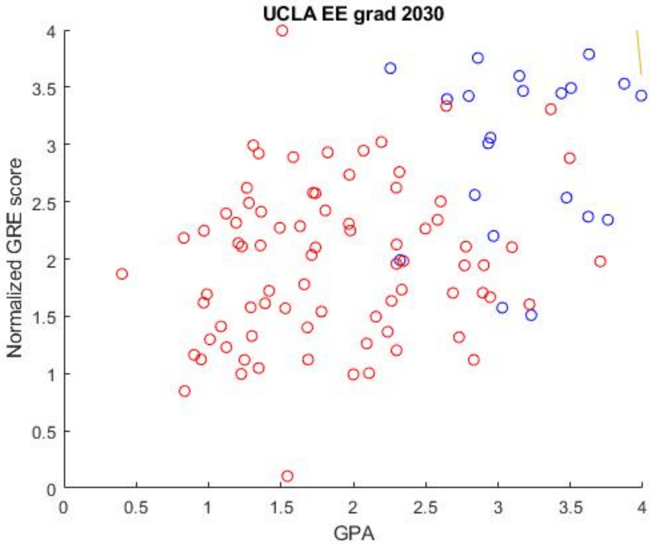
5. You will implement the perceptron algorithm in this problem. You are provided with 2 datasets: *UCLA_EE_grad_2030.csv* and *UCLA_EE_grad_2031.csv*. These datasets contain UCLA graduate student admission data in 2030 and 2031, respectively. Each dataset will have three columns. The first two columns contain two features for the applicants. The first column represents the GPA of applicants and the second column represents the normalized GRE score of applicants. The third column contains labels denoting whether the applicant is admitted into UCLA or not. Each label is either 1 or -1 where 1 denotes a student being admitted and -1 denotes a student not being admitted. Throughout this quarter, we will use these two datasets repetitively to text different learned algorithms.

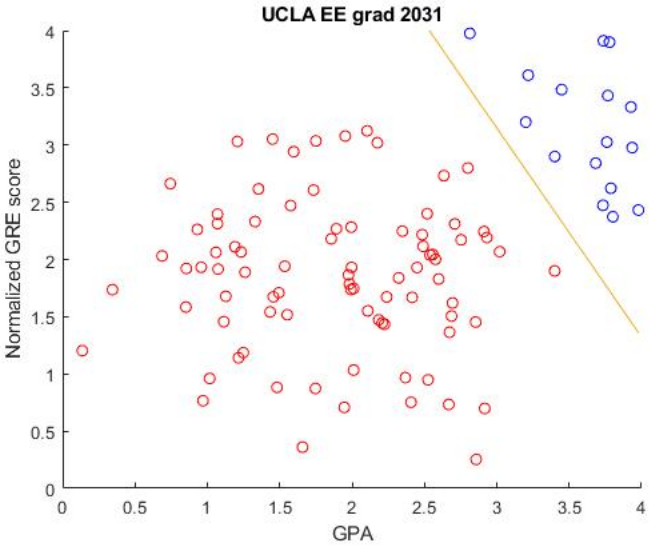
   (a) Plot all the datasets. Which datasets are linearly separable?

   Solutions: The linearly separable datasets is *UCLA_EE_grad_2031.csv*.

   (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 learned 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 the 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 will need to find x_1, x_2 that satisfy the equation $w_1x_1 + w_2x_2 + b = 0$. For *UCLA_EE_grad_2030.csv*, we have $w_1 = 14.3, w_2 = 1.09$, $b = -61$ and $u = 11299$. The algorithm does not converge since the data is not linearly separable and we can see it takes a large number of updates (11299) until the max number of iteration. For *UCLA_EE_grad_2031.csv*, the perception algorithm converges in $u = 158$ iterations with $w_1 = 5.9, w_2 = 3.2$, and $b = -28$.

(c) Recall that the empirical margin $\gamma_{w,b}$ is the distance between the hyperplane defined by $\{w, b\}$ and the nearest point of a set. Calculate $\gamma_{w,b}$ for the linearly separable dataset with your learned parameters.

Solution:

The margin for *UCLA_EE_grad_2031.csv* is 0.2002. To find this margin, iterate through all the data points and find the minimum distance between those points and the learned hyperplane. We find that the point that has the minimum distance is $(3.2, 3.2)$ and we can use the following formula to confirm that the margin is 0.2002:

$$d = \frac{|w^T x + b|}{\|w\|} = \frac{[5.9282, 3.2447][3.2, 3.2]^T - 28}{6.7581} \approx 0.2002.$$