Lista - FECD - B

1. Consider the usual vector space $\mathbb{R}^2$. Explain why each subset W below is or is not a ~~set~~ vector sub-space of $\mathbb{R}^2$.

a) $W = \{(x,y) ;\ x^2+y^2 \leq 1\}$ (disk centered at $(0,0)$)

b) $W = \{(x,y);\ x \geq 0$ and $y \geq 0\}$ (first quadrant)

c) $W = \{(x,y);\ y = 3+2x\}$ (a certain straigth line)

d) $W = \{(x,y);\ y = 2x\}$ (another straigth line)

e) $W = \{(x,y);\ y = 3\}$ (yet another line)

f) $W = \{(x,y);\ y = 3+2x$ <u>or</u> $y = 5-3x\}$ (two straigth lines)

g) $W = \{(x,y);\ x \in \mathbb{Z}$ and $y \in \mathbb{R}\}$ (x-coordinate are integers)

h) $W = \{(x,y);\ (x,y) = (0,0)\}$ (a single point)

i) $W = \{(x,y);\ y = x^2\}$ (a parabola)

j) $W = \{(x,y) ;\ x+y = \pi\}$

k) $W = \{(x,y);\ \frac{y}{x} = 2$ when $x \neq 0\} \cup \{(x,y);\ x=0$ and $y=0\}$

2) The normal equation which leads to the linear regression solution $\hat{\beta}$ is

$$(X'X)\cdot\beta = X'Y$$

- Each element in the matrix can be seen as the inner product of two vectors present in the matrix X. Explain which vectors are these for a generic element of $X'X$.
- Is $X'X$ a symmetric matrix?
- Likewise, each element in the vector $X'Y$ is the result of an inner product of two vectors. Explain what are these vectors for each coordinate of $X'Y$.

---

3) Suppose we are fitting a simple straigth line to points $(x_i, y_i) \in \mathbb{R}^2$. The regression line $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ has coefficients $\hat{\beta} = \begin{pmatrix}\hat{\beta}_0\\ \hat{\beta}_1\end{pmatrix}$ that are the solution of the normal equation $(X'X)\cdot\beta = X'Y$. Assume that we center the feature x and the response y such that we use the points $(x_i - \bar{x}, y_i - \bar{y})$ where $\bar{x} = \sum_i x_i/n$ and $\bar{y} = \sum y_i/n$ are the means of

x and y, respectively. The fitted regression line ③ for these centered-data will be solution of the new normal equation:

$$\left((X^*)' X^*\right) \cdot \beta^* = (X^*)' \cdot Y^*$$

where $X^* = \begin{pmatrix} 1 & x_1 - \bar{x} \\ 1 & x_2 - \bar{x} \\ 1 & x_3 - \bar{x} \\ \vdots & \vdots \\ 1 & x_n - \bar{x} \end{pmatrix}$ and $Y^* = \begin{pmatrix} y_1 - \bar{y} \\ y_2 - \bar{y} \\ \vdots \\ y_n - \bar{y} \end{pmatrix}$

Equivalently, the normal equation can be written as $\frac{1}{n}\left((X^*)' X^*\right) \cdot \beta^* = \frac{1}{n} (X^*)' \cdot Y^*$

by dividing both sides by n.

- What is the matrix $\frac{1}{n} (X^*)' X^*$? Interpret its elements.
- What about the elements in $\frac{1}{n} (X^*)' \cdot Y^*$?
- Obtain $\left((X^*)' X^*\right)^{-1}$ and the solution

$$\hat{\beta}^* = \left((X^*)' \cdot X^*\right)^{-1} \cdot \left((X^*)' \cdot Y^*\right) = \left(\frac{1}{n}(X^*)' X^*\right)^{-1} \left(\frac{1}{n}(X^*)' \cdot Y^*\right)$$

$\hookrightarrow n$ will be cancelled

4) Generalize the previous problem for a multiple regression problem. Assume that we have 3 features X_1, X_2, and X_3 and we center each of them around their respective means $\bar{x}_1$, $\bar{x}_2$, and $\bar{x}_3$. The feature matrix X is then given by

$$X = \begin{pmatrix} 1 & x_{11}-\bar{x}_1 & x_{12}-\bar{x}_2 & x_{13}-\bar{x}_3 \\ 1 & x_{21}-\bar{x}_1 & x_{22}-\bar{x}_2 & x_{23}-\bar{x}_3 \\ 1 & x_{31}-\bar{x}_1 & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1}-\bar{x}_1 & x_{n2}-\bar{x}_2 & x_{n3}-\bar{x}_3 \end{pmatrix}$$

Likewise, center the response vector:

$$Y = \begin{pmatrix} y_1-\bar{y} \\ y_2-\bar{y} \\ \vdots \\ y_n-\bar{y} \end{pmatrix}$$

The normal equation is given by

$$\frac{1}{n}(X'X)\cdot\beta = \frac{1}{n}X'Y$$

Interpret the elements in the matrix $\frac{1}{n}(X'X)$ and in $\frac{1}{n}X'Y$

5) How the linear regression solution change when we center the features and the response?

To answer this, consider the simple ~~set~~ situation where we have a single feature and points $(x_i, y_i) \in \mathbb{R}^2$, $i = 1, \dots, n$.

Not centering the data, we have $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1)^t$ as the solution of

$$\hat{\beta} = \arg\min_{\beta_0, \beta_1} \sum_i \left(y_i - (\beta_0 + \beta_1 x_i)\right)^2$$

the centered-data linear regression coefficient $\hat{\beta}^* = (\hat{\beta}_0^*, \hat{\beta}_1^*)^t$ is the solution of

$$\hat{\beta}^* = \arg\min_{\beta_0^*, \beta_1^*} \sum_i \left((y_i - \bar{y}) - (\beta_0^* + \beta_1^*(x_i - \bar{x}))\right)^2$$

Find the relationship between $\hat{\beta}$ and $\hat{\beta}^*$.

---

6) Suppose that the features are orthogonal to each other. That is, the columns of the matrix X are orthogonal to each other. What is the normal equation in this case? How easy is it to solve the normal equation?

---

7) The linear regression solution is

$$\hat{\beta} = (X'X)^{-1} X'Y$$

and the predicted values for the n observations is the vector $\hat{Y} = X\hat{\beta} = X(X'X)^{-1}X'Y = H.Y$. The matrix $H = X(X'X)^{-1}X'$ is called the hat-matrix because "it puts the hat on top of Y" (Tukey). Show that:

i) $H^t = H$ (symmetric)

ii) $H^2 = H$ (idempotent)

iii) the residual sum of squares RSS is: $RSS = \sum_i (y_i - \hat{y}_i)^2 = \|\underset{\sim}{Y} - \underset{\sim}{\hat{Y}}\|^2$. Show that RSS is a quadratic form: $RSS = Y'.(I - H).Y$