Normal Equation

Notation and Definition of the Problem

- Here we use Machine Learning terminology.
- Let n and m be non-zero natural numbers.
- The general case has n features, i.e., n independent variables $x_1, \ldots, x_n$.
- As usual, we define $x_0 \equiv 1$.
- These quantities are written as components of the $(n+1)$-dimensional column vector $\mathbf{x}$:

$$\mathbf{x} = \begin{bmatrix} x_0 \\ x_1 \\ \vdots \\ x_n \end{bmatrix}.$$

- The corresponding dependent variable is denoted by y.
- A training example is represented by the ordered pair $(\mathbf{x}, y)$.
- In general, we have m training examples $\left(\mathbf{x}^{(1)}, y^{(1)}\right), \ldots, \left(\mathbf{x}^{(m)}, y^{(m)}\right)$.
- Next, we use the column vectors $\mathbf{x}^{(i)}$ $(i = 1, \ldots, m)$ to define the m by $n+1$ design matrix X.
- By definition, the i-th row of X is the row vector $\left(\mathbf{x}^{(i)}\right)^T$:

$$X \equiv \begin{bmatrix} \left(\mathbf{x}^{(1)}\right)^T \\ \vdots \\ \left(\mathbf{x}^{(m)}\right)^T \end{bmatrix} = \begin{bmatrix} x_0^{(1)} & x_1^{(1)} & \cdots & x_n^{(1)} \\ \vdots & \vdots & \vdots & \vdots \\ x_0^{(m)} & x_1^{(m)} & \cdots & x_n^{(m)} \end{bmatrix}.$$

- We write the values of the variable y (observed values) as components of the m-dimensional column vector $\mathbf{y}$:

$$\mathbf{y} = \begin{bmatrix} y^{(1)} \\ \vdots \\ y^{(m)} \end{bmatrix}.$$

- Assumption: The relation between the dependent variable and the independent ones is linear.
- In other words, the observed value is a linear function of the features.
- This function has $n+1$ coefficients $\theta_0, \theta_1, \ldots, \theta_n$.
- They are written as components of the $(n+1)$-dimensional column vector $\boldsymbol{\theta}$:

$$\boldsymbol{\theta} = \begin{bmatrix} \theta_0 \\ \theta_1 \\ \vdots \\ \theta_n \end{bmatrix}.$$

- Then the mathematical form of our hypothesis is

$$y = h_{\boldsymbol{\theta}}(\mathbf{x}) = \theta_0 x_0 + \theta_1 x_1 + \ldots + \theta_n x_n = \boldsymbol{\theta}^T \mathbf{x}.$$

- An alternative equation for y is

$$y = \sum_{j=0}^{n} \theta_j x_j.$$

- If this assumption is correct, the observed values can be expressed as

$$y^{(i)} = \boldsymbol{\theta}^T \mathbf{x}^{(i)} = \sum_{j=0}^{n} \theta_j x_j^{(i)}, \quad i = 1, \ldots, m.$$

- We continue by writing the last sum in terms of the design matrix X.
- To do so, we consider the i-th component of the column vector $X\theta$:

$$(X\theta)^{(i)} = \sum_{j=0}^{n} X_{ij}\theta_j = \sum_{j=0}^{n} \theta_j x_j^{(i)} = y^{(i)}.$$

- Therefore, the matrix version of our hypothesis is

$$\mathbf{y} = X\theta.$$

- However, in many situations, this assumption is not entirely correct.
- Nevertheless, it can be used as an approximation.
- In other words, the observed values can be approximately described by a linear function of the features.
- In these cases, there are errors (also called residuals) $\epsilon^{(i)}$:

$$\epsilon^{(i)} = y^{(i)} - (X\theta)^{(i)}.$$

- This equation can be put in matrix form if we introduce the m-dimensional column vector $\boldsymbol{\epsilon}$:

$$\boldsymbol{\epsilon} = \begin{bmatrix} \epsilon^{(1)} \\ \vdots \\ \epsilon^{(m)} \end{bmatrix}.$$

- This definition allows us to write the formula for the residuals as

$$\boldsymbol{\epsilon} = \mathbf{y} - X\theta.$$

- Finally, we can present the statement of the problem we shall solve:
  Suppose we have a set of training examples that are approximately described by a linear function h_θ.
  Determine the function h_θ corresponding to the best approximation.
- This is the so-called linear regression problem.
- To continue, we have to state it more precisely.
- Our goal is to find a linear function, which is characterized by its coefficients θ.
- Then solving the problem means finding a specific vector θ.
- This is the vector that minimizes the residuals.

Cost Function

- To proceed, we define the so-called cost function $J(\theta)$:

$$J(\theta) \equiv \frac{1}{2m} |\boldsymbol{\epsilon}|^2 = \frac{1}{2m} |\mathbf{y} - X\theta|^2.$$

- By minimizing the residuals, we obtain the minimum value of $|\boldsymbol{\epsilon}|$.
- In turn, this gives us the minimum of the cost function.
- Hence, we can solve our problem by minimizing $J(\theta)$.

Solution to the Problem: Derivation of the Normal Equation

- To minimize $J(\theta)$, we have to determine the vector θ for which the following equation is satisfied:

$$\nabla J(\theta) = 0.$$

- Before evaluating the gradient of the cost function, we write an alternative formula for $J(\theta)$:

$$J(\theta) = \frac{1}{2m} \sum_{i=1}^{m} \left(\epsilon^{(i)}\right)^2 = \frac{1}{2m} \sum_{i=1}^{m} \left[y^{(i)} - (X\theta)^{(i)}\right]^2.$$

- Next, we differentiate the last expression with respect to θ_k ($k = 0, 1, \ldots, n$):

$$\frac{\partial}{\partial \theta_k} [J(\theta)] = -\frac{1}{m} \sum_{i=1}^{m} \left[y^{(i)} - (X\theta)^{(i)}\right] \frac{\partial}{\partial \theta_k} \left[(X\theta)^{(i)}\right].$$

- The derivative on the right-hand side of the above equation is given by

$$\frac{\partial}{\partial\theta_k}\left[(X\theta)^{(i)}\right]=\frac{\partial}{\partial\theta_k}\left(\sum_{j=0}^{n}\theta_j x_j^{(i)}\right)=\sum_{j=0}^{n}\delta_{jk}x_j^{(i)}=x_k^{(i)}.$$

- This result allows us to write the k-th component of $\nabla J(\theta)$ as follows:

$$[\nabla J(\theta)]_k=\frac{1}{m}\sum_{i=1}^{m}\left[(X\theta)^{(i)}-y^{(i)}\right]x_k^{(i)}=\frac{1}{m}\sum_{i=1}^{m}\left(X^T\right)_{ki}\left[(X\theta)^{(i)}-y^{(i)}\right]=\frac{1}{m}\left(X^TX\theta-X^T\mathbf{y}\right)_k.$$

- Therefore, the gradient of the cost function is

$$\nabla J(\theta)=\frac{1}{m}\left(X^TX\theta-X^T\mathbf{y}\right).$$

- To obtain the normal equation, we take the expression on the right-hand side and set it equal to zero:

$$X^TX\theta-X^T\mathbf{y}=0 \quad\Rightarrow\quad X^TX\theta=X^T\mathbf{y}.$$

- Later we explain the reason for the name "normal equation".
- We are not finished, since our goal is to derive a formula for the coefficients θ.
- To do so, we assume the following: all the rows of the design matrix are linearly independent.
- This is the same as assuming that the vectors $\mathbf{x}^{(i)}$ are linearly independent.
- In this case, one can prove that the matrix X^TX is invertible.
- Then we can multiply both sides of the normal equation by the inverse matrix $\left(X^TX\right)^{-1}$ to obtain

$$\theta=\left(X^TX\right)^{-1}X^T\mathbf{y}.$$

- This is the solution to the normal equation, i.e., the solution to the linear regression problem.

Why "Normal" Equation?

- It is important to explain why the equation we have derived is called "normal".
- Consider a real matrix M.
- By definition, this matrix is normal if it commutes with its transpose:

$$\left[M,M^T\right]=MM^T-M^TM=0.$$

- The normal equation has this name, because it involves the matrix X^TX, which is normal.
- Let us quickly prove this fact.
- We begin by computing the transpose of X^TX:

$$\left(X^TX\right)^T=X^T\left(X^T\right)^T=X^TX.$$

- Next, we use this result to evaluate the commutator $\left[X^TX,\left(X^TX\right)^T\right]$:

$$\left[X^TX,\left(X^TX\right)^T\right]=X^TX\left(X^TX\right)^T-\left(X^TX\right)^TX^TX=X^TXX^TX-X^TXX^TX=0.$$

- Hence, X^TX is a normal matrix.
- There is also a geometrical reason for the name "normal equation".
- To interpret this equation geometrically, first we rewrite it:

$$X^TX\theta=X^T\mathbf{y} \quad\Rightarrow\quad X^T(\mathbf{y}-X\theta)=0 \quad\Rightarrow\quad X^T\boldsymbol{\epsilon}=0.$$

- By evaluating the transpose of both sides of the last relation, we obtain

$$\boldsymbol{\epsilon}^TX=0.$$

- To continue, consider any vector $X\mathbf{v}$ belonging to the column space of the design matrix.
- Due to the above formula, the inner product of $X\mathbf{v}$ and the residual vector $\boldsymbol{\epsilon}$ equals zero:

$$\boldsymbol{\epsilon}^T X\mathbf{v} = 0.$$

- Therefore, $\boldsymbol{\epsilon}$ is orthogonal (normal) to the column space of X.
- These are two ways of justifying the name "normal equation".

Particular Case: One Independent Variable

- An important particular case is the one with a single independent variable.
- In other words, this is the case with a single feature, i.e., $n = 1$.
- Next, we consider this case and derive the corresponding formula for the coefficients $\boldsymbol{\theta}$.
- When $n = 1$, the design matrix can be written as

$$X = \begin{bmatrix} x_0^{(1)} & x_1^{(1)} \\ \vdots & \vdots \\ x_0^{(m)} & x_1^{(m)} \end{bmatrix}.$$

- Since X is a m by 2 matrix, its transpose is 2 by m:

$$X^T = \begin{bmatrix} x_0^{(1)} & \cdots & x_0^{(m)} \\ x_1^{(1)} & \cdots & x_1^{(m)} \end{bmatrix}.$$

- Then the product $X^T X$ is a 2 by 2 matrix whose elements are given by

$$\left(X^T X\right)_{ab} = \sum_{i=1}^{m} \left(X^T\right)_{ai} X_{ib} = \sum_{i=1}^{m} x_a^{(i)} x_b^{(i)} \quad (a, b = 0, 1).$$

- The last expression allows us to write $X^T X$ as follows:

$$X^T X = \begin{bmatrix} m & \sum_{i=1}^{m} x_1^{(i)} \\ \sum_{i=1}^{m} x_1^{(i)} & \sum_{i=1}^{m} \left(x_1^{(i)}\right)^2 \end{bmatrix}.$$

- This equation can be put in a simpler form if we introduce the averages

$$\overline{x} \equiv \frac{1}{m} \sum_{i=1}^{m} x_1^{(i)},$$
$$\overline{x^2} \equiv \frac{1}{m} \sum_{i=1}^{m} \left(x_1^{(i)}\right)^2.$$

- By using these definitions, we can rewrite $X^T X$ as

$$X^T X = m \begin{bmatrix} 1 & \overline{x} \\ \overline{x} & \overline{x^2} \end{bmatrix}.$$

- To continue, we have to find the inverse of this matrix.
- Consider the following invertible 2 by 2 matrix:

$$A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}.$$

- The formula for the corresponding inverse matrix is

$$A^{-1} = \frac{1}{\det(A)} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}.$$

- We shall use this result to determine $\left(X^T X\right)^{-1}$.

- To do so, first we compute the determinant of $X^T X$:

$$\det\left(X^T X\right) = m^2\left(\overline{x^2} - \overline{x}^2\right) = m^2 \sigma_x^2,$$

where $\sigma_x^2 = \overline{x^2} - \overline{x}^2$ is the variance of the independent variable x.

- Then the inverse matrix $\left(X^T X\right)^{-1}$ can be written as

$$\left(X^T X\right)^{-1} = \frac{1}{m\sigma_x^2}\begin{bmatrix} \overline{x^2} & -\overline{x} \\ -\overline{x} & 1 \end{bmatrix}.$$

- We proceed by calculating the product $X^T \mathbf{y}$, which is a 2-dimensional column vector.
- The components of this vector are given by

$$\left(X^T \mathbf{y}\right)_a = \sum_{i=1}^{m}\left(X^T\right)_{ai} y^{(i)} = \sum_{i=1}^{m} x_a^{(i)} y^{(i)} \quad (a = 0, 1).$$

- This result allows us to write the following formula for $X^T \mathbf{y}$:

$$X^T \mathbf{y} = \begin{bmatrix} \sum_{i=1}^{m} y^{(i)} \\ \sum_{i=1}^{m} x_1^{(i)} y^{(i)} \end{bmatrix} = \begin{bmatrix} m\overline{y} \\ \sum_{i=1}^{m} x_1^{(i)} y^{(i)} \end{bmatrix},$$

where $\overline{y}$ denotes the average of the dependent variable y.

- To simplify the last expression, we define the covariance of x and y:

$$\sigma_{x,y} \equiv \frac{1}{m}\sum_{i=1}^{m}\left(x_1^{(i)} - \overline{x}\right)\left(y^{(i)} - \overline{y}\right).$$

- It is useful to derive an alternative equation for this quantity:

$$\sigma_{x,y} = \frac{1}{m}\sum_{i=1}^{m}\left(x_1^{(i)} y^{(i)} - x_1^{(i)}\overline{y} - \overline{x} y^{(i)} + \overline{x}\,\overline{y}\right) = \frac{1}{m}\sum_{i=1}^{m} x_1^{(i)} y^{(i)} - \overline{x}\,\overline{y} = \overline{xy} - \overline{x}\,\overline{y},$$

where we have defined

$$\overline{xy} \equiv \frac{1}{m}\sum_{i=1}^{m} x_1^{(i)} y^{(i)}.$$

- With the aid of the last equation for $\sigma_{x,y}$, we obtain our final result for $X^T \mathbf{y}$:

$$X^T \mathbf{y} = m\begin{bmatrix} \overline{y} \\ \overline{xy} \end{bmatrix} = m\begin{bmatrix} \overline{y} \\ \sigma_{x,y} + \overline{x}\,\overline{y} \end{bmatrix}.$$

- Finally, we multiply our expressions for $\left(X^T X\right)^{-1}$ and $X^T \mathbf{y}$:

$$\left(X^T X\right)^{-1} X^T \mathbf{y} = \frac{1}{\sigma_x^2}\begin{bmatrix} \overline{x^2} & -\overline{x} \\ -\overline{x} & 1 \end{bmatrix}\begin{bmatrix} \overline{y} \\ \sigma_{x,y} + \overline{x}\,\overline{y} \end{bmatrix} = \frac{1}{\sigma_x^2}\begin{bmatrix} \sigma_x^2\overline{y} - \overline{x}\sigma_{x,y} \\ \sigma_{x,y} \end{bmatrix}.$$

- Recall that this product is equal to θ.
- In the case $n = 1$, this vector is 2-dimensional.
- Therefore, we can write

$$\begin{bmatrix} \theta_0 \\ \theta_1 \end{bmatrix} = \frac{1}{\sigma_x^2}\begin{bmatrix} \sigma_x^2\overline{y} - \overline{x}\sigma_{x,y} \\ \sigma_{x,y} \end{bmatrix}.$$

- Then we conclude that the slope of the desired linear function is

$$\theta_1 = \frac{\sigma_{x,y}}{\sigma_x^2}.$$

- The y-intercept of this function is given by

$$\theta_0 = \frac{\sigma_x^2\overline{y} - \overline{x}\sigma_{x,y}}{\sigma_x^2} = \overline{y} - \overline{x}\theta_1.$$

- These are the well-known $n = 1$ formulas for the linear regression coefficients.