Machine Learning for Business Decisions - Lecture 2 Notes

Somani, Neel - Bao, Jason

1. Basic Concepts

- **Vector**: point in some vector space with some rules
  - Can represent points in some n-dimensional space
  - Can represent data points
- **Matrix**:
  - Can hold data / vectors
  - Can represent special functions (e.g. linear transformations)
- **Norm**:
  - Used to find the "size"of a vector (or matrix)
  - Must be greater than or equal to zero
  - e.g. $\mathcal{L}2$ Norm: $\|x\|_2 = \sqrt{\sum_{i=0}^{n} x_i^2}$

2. Linear Least Squares

We have a system of equations:

$$2w_1 + 3w_2 = 5$$
$$w_1 - w_2 = 0$$

This system of equations can represent:

Point	Label
(2, 3)	Label 5
(1, -1)	Label 0

In the equation $Xw = b$ where X is a matrix and w and b are column vectors, you can view the components of w as "weights"for each column of X. You might also view each row of X (and b) as a constraint on the solution.

In real life, your data might be noisy, making the matrix inconsistent. You might want to find w (a linear combination of the columns of X) that minimizes the error function:

$$loss = \|Xw - y\|_2^2 \tag{1}$$

(this is known as the $\mathcal{L}2$ norm of $Xw - y$ squared).

The answer to this problem is known as the Ordinary Least Squares (OLS) Problem that we will touch on later.

2.1. Non-linear Functions

You can use linear least squares to fit nonlinear functions, too. Suppose you have two sets of data, X and Y, and you think that the relationship is:

$$y(x) = w_1 * x^2 + w_2 * sin(x) + w_3 * x + w_4 \tag{2}$$

The goal is to find the the value for w that minimizes the loss function (1).

You can substitute the values of your samples X into (2) to create a system of linear equations. For example, if $x_1 = 10$, then:

$$y_1 = w_1 * 10^2 + w_2 * sin(10) + w_3 * 10 + w_4$$
$$y2 = \dots$$

which creates a system of equations that are linear with respect to w.

3. Feature Engineering

What if you think that the relationship between x and y is a circle? You might want to choose "features" (functions that take in your data point x as input) that allow the data to be linearly separable (since we cannot separate the data with a line if it forms a circle). We will go into this more later but tools like neural nets allow us to not even have to handpick the features ourselves!

4. Ordinary Least Squares

Let's solve the ordinary least squares problem from earlier, for the single variable case.

$$loss = \sqrt{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2} = \sqrt{\sum_{i=1}^{n}(x_i * w - y_i)^2}$$

We want to minimize the loss with respect to w.

$$\frac{d}{dw} = \frac{1}{2}(\sum_{i=1}^{n}(x_i * w - y_i)^2)^{\frac{-1}{2}} * \sum_{i=1}^{n}(2x_i^2 * w - 2x_i * y_i)$$

Setting this to 0 we can solve for w to get the solution:

$$w = \frac{\sum_{i=1}^{n}(x_i * y_i)}{\sum_{i=1}^{n}(x_i * x_i)}$$

While it's not within the scope of this course to derive the following equation, here is the closed form solution for the multi-variable ordinary least squares:

$$w^* = (X^T X)^{-1} X^T y$$

5. Business Applications

Computer vision: There aren't too many business applications at first sight. This is used in the postal service.
Now we have ways to quantify the risk associated with, e.g., a stock investment.