Mathematics of Big Data I

Lecture 1: Overview of Big Data Analytics

September 5th, 2016

1 Introduction – The Big Picture of Big Data

We will explore the key ideas of machine learning today to frame the subject. We will answer

- Where does big data come from?
- What are different ways of describing big data?
- Data can be structured, semi-structured, or constructured. How can we understand it?
- What are the challenges with data (e.g. "dirty" data)?

We will look at different methods of data processing, including ETL, data structuring, and dimensionality reduction. Once we process data using these methods we will obtain a **data representation**. These representations can include

- Discrete
- Real-valued
- Vector-valued
- Matrix-valued
- Manifold-valued
- Mixed, etc.

We want to use data technology to get to "data to decision" or D2D technologies. All of these concepts fall under large subject called **machine learning**. Within machine learning, we can **supervised** and **supervised learning**.

The basics steps of D2D technologies include

- Data visualization
- Mathematical modeling
- Computational methods
- Validation & verification.

The key steps to doing this involves the same basic procedure:

1. Define a metric on your data space in order to define similarities (data points close in your metric space are "similar").

2. Use strategies for effective optimization and computation (we need to be able to discern which methods are faster and well-suited to the data we are processing).

There are different **modeling approaches** we can take to new data. These include **statistical calculus, geometric/analytic methods**, and **probability**. Each one has it's own merit, but require different skill sets. For instance, using the statistical calculus we could linear regression; geometric methods we will be able to see the answer intuitively; probabilistic methods are well-suited for machines since computers can quantify large amounts of data and many big data problems can be made into discrete probabilistic problems.

Concepts we need to explore in all of this is to look at problems and see whether we can compute things globally or locally (like in the case of manifolds), and whether problems can be thought of either discretely or smoothly. We can utilizing the power of computers using the following methods:

- Hadoop – useful for large data sets
- HDF – useful for fault resilience
- MapReduce – useful for a "divide and conquer" strategy
- Spark – useful for fast computation in memory, and
- Zookeeper – useful for orchestration.

Just as important as utilizing our computational tool set, we want to utilize mathematics to this end as well. We will unify mathematics from

- Linear algebra
- Statistics and Probability
- Multivariable calculus
- Geometry and discrete mathematics.

All of these concepts give us a big picture to all of the tools and concepts needed in data analytics. In today's lecture we will explore the two questions of the origin of big data, and how we can represent it. Later we will look at linear regression as an example of data analytics.

1.1 Where Does Big Data Come From?

Big data comes from many sources. Primarily we can get data from organizations such as retail stores. Data also comes from machines. For instance, an unmanned aerial vehicle (UAV), GPS devices, cell phones, computers, etc. Lastly, people also generate data. For instance, through social media, online shopping preferences, etc. Data is not new, but in our modern era the scale of data has increased drastically. The way people now use data is far different than it was in earlier generations.

1.2 What are the Types of Data?

There are broadly speaking three types of data:

1. **Structured Data**: This type of data is often generated by organizations (for instance email data, shopping records).
2. **Semi-Structured Data**: This comes from machines typically.
3. **Unstructured Data**: This is often human-generated.

1.3 What Exactly is "Big" Data

We might think that "big" refers to volume, but it actually refers to five "V"-words:

- Volume (size)
- Velocity (speed)
- Variety (types)
- Veracity (quality)
- Valence (relationships; multidimensional data often has dependencies on each other).

We can think of these 5 Vs as being the main points needed in order to master D2D technologies. In Data to Decision we want to program machines to take actions based on data and learned knowledge.

2 Analytic Approaches – Illustrating Big Data Analysis with Linear Regression

There are three basic approaches to data analysis we discussed in the beginning:

2.1 Statistical Calculus Approach

Our objective in data analysis is in reducing or minimizing the total error in our method. However, the errors e_i some could be positive and some could be negative. We can't simply sum the errors since we run the risk of having zero error or negative error. How can we fix this problem? We can make error as something nonnegative. We can consider an objective or cost function, for instance:

$$J(m,b) \;=\; \sum_i e_i^2 \;=\; \sum_i (y_i - mx_i - b)^2.$$

This is an example of a quadratic. Can we possibly use $\sum_i |e_i|$? The answer is yes! This is what's referred to as the L^1-norm, which can have many uses in data analysis.

To illustrate this concept of a cost function, we will consider linear regression of sampled data points $(x_1, y_1), \ldots, (x_n, y_n)$. The goal here is to find the line of best fit, or rather, find the best pairing (m, b) for slope and y-intercept that minimizes the error in fitting these data points. We can find the minimum of $J(m, b)$ with respect to these variables by finding the values of m and b where $\frac{\partial J}{\partial m} = 0$ and $\frac{\partial J}{\partial b} = 0$. The partial derivatives here are given by

$$\begin{aligned} \frac{\partial J}{\partial m} &= \sum_i 2(y_i - mx_i - b)(-x_i) \\ \frac{\partial J}{\partial b} &= \sum_i 2(y_i - mx_i - b)(-1). \end{aligned}$$

Since we are setting these both equal to zero, we can simply the coefficient 2 in front of both of these equations. To solve for m and b in these equations we have that

$$\begin{aligned} m\sum_i x_i^2 + b\sum_i x_i &= \sum_i x_i y_i \\ m\sum_i x_i + b\underbrace{\sum_i 1}_{=n} &= \sum_i y_i. \end{aligned}$$

One method we can use is from linear algebra called *Cramer's rule*. Essentially the method we are using above is by taking the gradient of the objective function J and setting it equal to zero. In general we can consider functions in the following situations:

1. **Single Variable**: These are the ordinary functions we see in single variable calculus which are $f : \mathbb{R} \to \mathbb{R}$.

2. **Many-to-One**: We can have scalar-valued functions of many variables $f : \mathbb{R}^n \to \mathbb{R}$.

3. **One-to-Many**: Functions of the form $f : \mathbb{R} \to \mathbb{R}^n$ define curves in higher-dimensional Euclidean spaces or manifolds.

4. **Many-to-Many**: Functions of the form $f : \mathbb{R}^n \to \mathbb{R}^m$ define n-dimensional vector fields embedded into $\mathbb{R}^m$. Typically in this case we will use an upper case function of the form $\vec{F} : \mathbb{R}^n \to \mathbb{R}^m$ to denote vector fields.

The way we take derivatives of each of these types of functions is different since we have to take derivatives with respect to different numbers of variables. We will look at these more and more as time goes on.

Returning to our example of linear regression, we would like to express our equations in a more compact form, by writing our equations in terms of a vector equation. Instead of writing $J(m, b)$ in terms of a sum over it's components we can instead write

$$\vec{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \qquad \vec{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}$$

and we can therefore write

$$J(m, b) = (\vec{y} - m\vec{x} - b) \cdot (\vec{y} - m\vec{x} - b)$$

where the $\cdot$ indicates the dot product. We now want to take the gradient ∇J with respect to the variables m and b. Noting the product rule of functions $f, g : \mathbb{R}^n \to \mathbb{R}$ we obtain:

$$\nabla(fg) = (\nabla f)g + f(\nabla g)$$

so therefore we compute

$$0 = \nabla J = 2\nabla(\vec{y} - m\vec{x} - b\mathbf{1}) \cdot (\vec{y} - m\vec{x} - b\mathbf{1})$$

and this implies the following equations:

$$\begin{aligned} (-\vec{x}) \cdot (\vec{y} - m\vec{x} - b\mathbf{1}) &= 0 \\ (-\mathbf{1}) \cdot (\vec{y} - m\vec{x} - b\mathbf{1}) &= 0. \end{aligned}$$

This can subsequently be written as

$$\begin{aligned} \vec{x}^T\vec{y} - m\vec{x}^T\vec{x} - b\vec{x}^T\mathbf{1} &= 0 \\ \mathbf{1}^T\vec{y} - m\mathbf{1}^T\vec{x} - nb &= 0 \end{aligned}$$

where we note that $\mathbf{1}^T\mathbf{1} = 1 + \ldots + 1 = n$ where we sum over all of the components of $\mathbf{1}$. There is in fact a more compact way of doing this. If we instead make the substitutions $\vec{x}_0 = \mathbf{1}$ and $\vec{x}_1 = \vec{x}$ we can instead write the compact equation

$$\begin{bmatrix} \vec{x}_0^T \\ \vec{x}_1^T \end{bmatrix} \left(\vec{y} - \begin{bmatrix} \vec{x}_0 & \vec{x}_1 \end{bmatrix} \begin{bmatrix} b \\ m \end{bmatrix} \right) = 0.$$

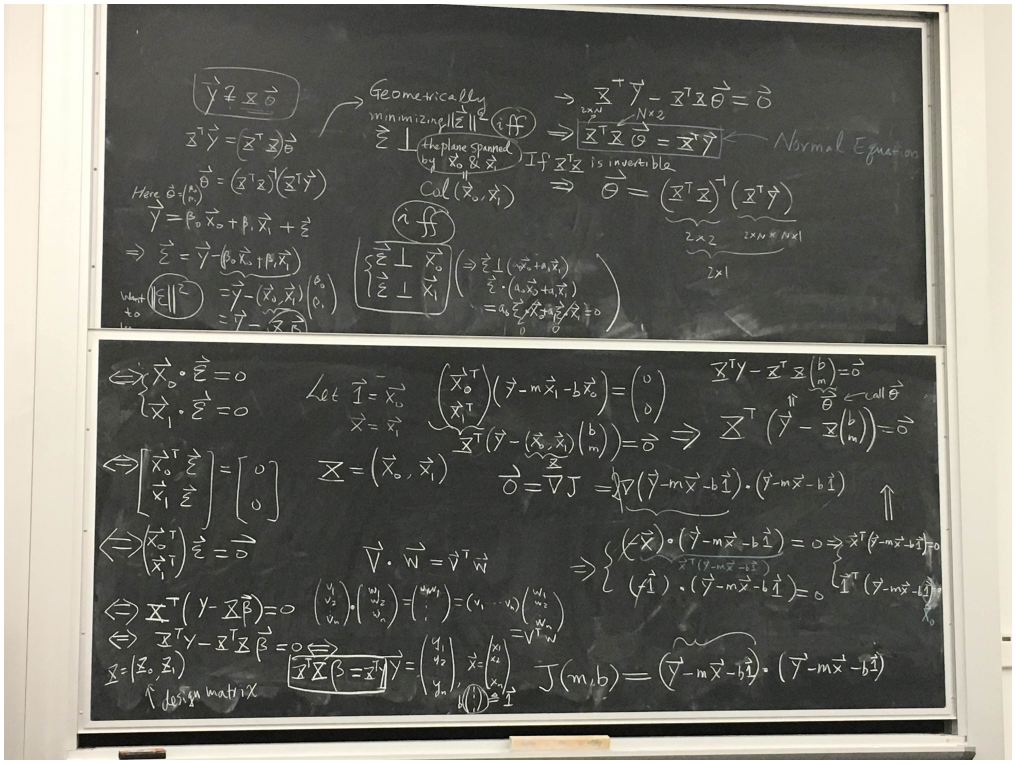
Making the substitutions

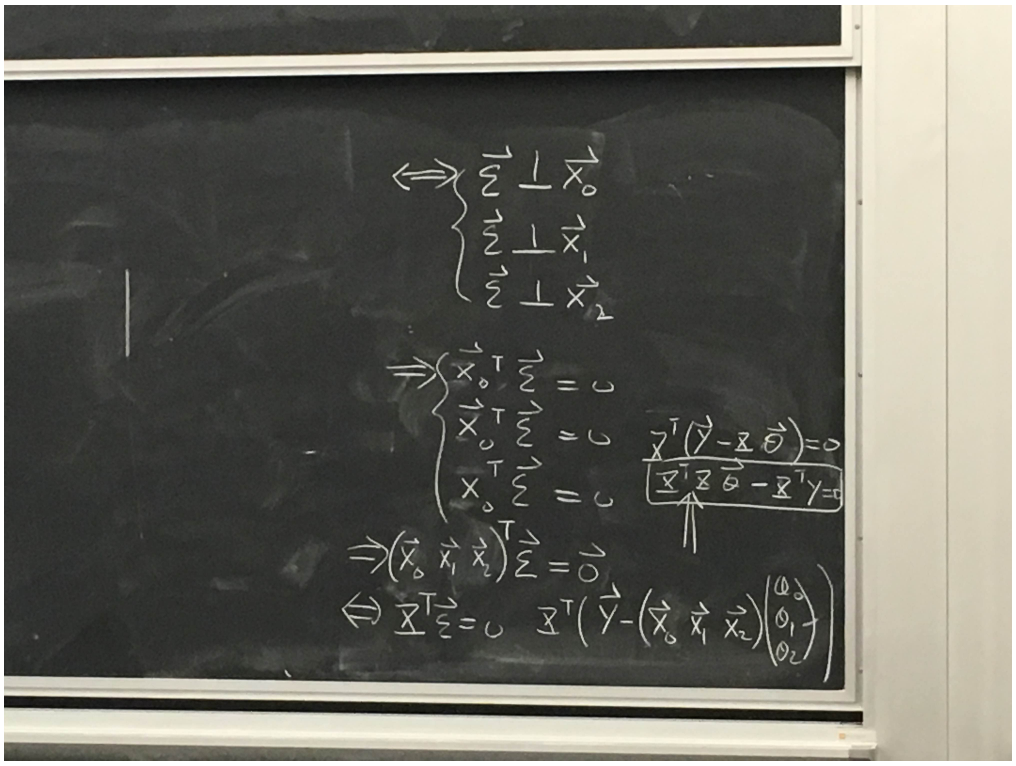
$$X = \begin{bmatrix} \vec{x}_0 & \vec{x}_1 \end{bmatrix} \qquad \vec{\theta} = \begin{bmatrix} b \\ m \end{bmatrix}$$

we find that

$$X^T(\vec{y} - X\vec{\theta}) \;=\; 0$$

and therefore $X^T X\vec{\theta} = X^T\vec{y}$. In the case that $X^T X$ is invertible we can solve this explicitly as $\vec{\theta} = (X^T X)^{-1}X^T\vec{y}$.





Homework Problem: Given the four points $(0,1),(2,3),(3,6),(4,8)$. Find $y = mx + b$ based on Cramer's rule. Use the normal formula to find the solution and compare it with the first part. Plot the data points and draw $y = mx + b$. Find another 100 points near the line $y = mx + b$, then find the least squares approximations again and plot both the data points and the new line.

Suppose we want to extend linear regression to multiple dimensions? For instance, what if instead of fitting points in $\mathbb{R}^2$ with a line, we fit points in $\mathbb{R}^3$ with a plane? We can generalize what we wrote earlier to parameterize the plane as

$$y = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \vec{\varepsilon}.$$

Given a sampling of points $(x_1^{(1)}, x_2^{(1)}, y^{(1)}), \ldots (x_1^{(n)}, x_2^{(n)}, y^{(n)})$ we can then write all of these in matrix form as

$$\underbrace{\begin{bmatrix} y^{(1)} \\ \vdots \\ y^{(n)} \end{bmatrix}}_{\vec{y}} = \theta_0 \underbrace{\begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}}_{\vec{x}_0} + \theta_1 \underbrace{\begin{bmatrix} x_1^{(1)} \\ \vdots \\ x_1^{(n)} \end{bmatrix}}_{\vec{x}_1} + \theta_2 \underbrace{\begin{bmatrix} x_2^{(1)} \\ \vdots \\ x_2^{(n)} \end{bmatrix}}_{\vec{x}_2} + \underbrace{\begin{bmatrix} \varepsilon^{(1)} \\ \vdots \\ \varepsilon^{(n)} \end{bmatrix}}_{\vec{\varepsilon}}.$$

By writing the matrix

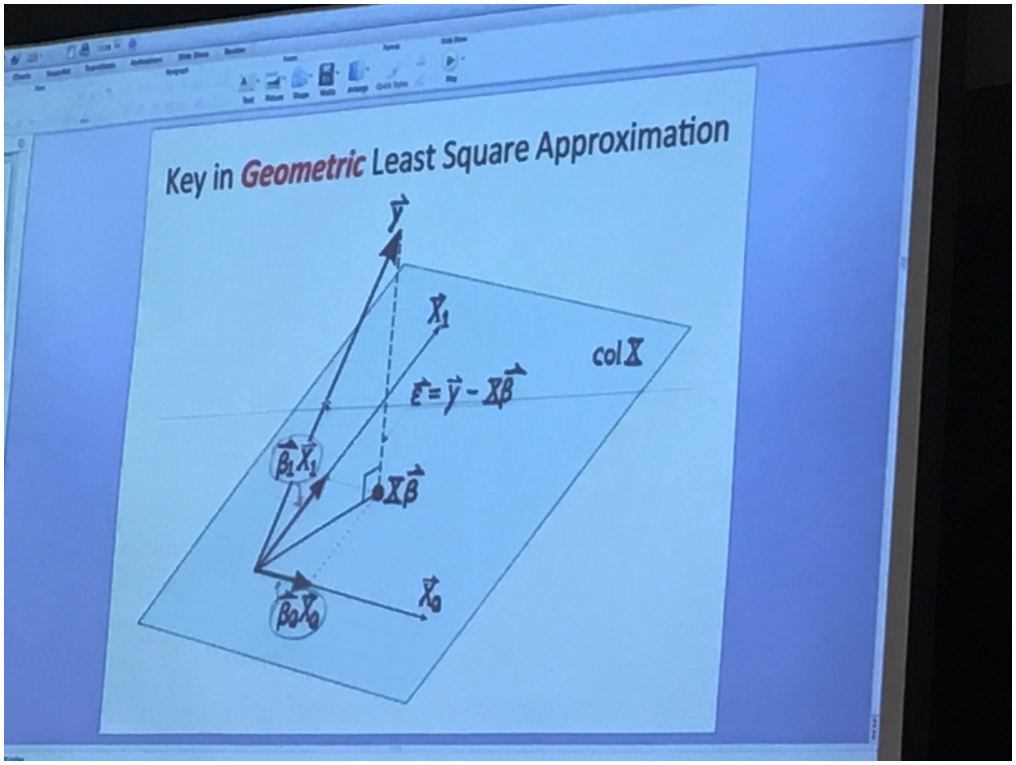
$$X = \begin{bmatrix} \vec{x}_0 & \vec{x}_1 & \vec{x}_2 \end{bmatrix} \qquad \vec{\theta} = \begin{bmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \end{bmatrix}$$

we obtain the matrix equation

$$\vec{y} = X\vec{\theta} + \vec{\varepsilon}.$$

We can solve this again by multiplying this equation by X^T and obtaining a similar equation. A key difference here is that we would like minimize $\|\vec{\varepsilon}\|^2$. How can we do this?

2.2 Geometric/Analytic Approach



We can rewrite this equation as $\vec{\varepsilon} = \vec{y} - X\vec{\theta}$, we would like to have an intuitive way of minimizing $\vec{\varepsilon}$. We can see using an intuitive geometric approach that $\vec{\varepsilon}$ minimizes it's length when it is perpendicular to the column space of X. We can see that $\vec{\varepsilon} \perp \vec{x}_i$ if and only if $\vec{\varepsilon} \cdot \vec{x}_i = \vec{x}_i^T \vec{\varepsilon} = 0$.
Observe we can writ these conditions compactly as

$$\begin{bmatrix} \vec{x}_0^T \vec{\varepsilon} \\ \vec{x}_1^T \vec{\varepsilon} \\ \vec{x}_2^T \vec{\varepsilon} \end{bmatrix} = \begin{bmatrix} \vec{x}_0^T \\ \vec{x}_1^T \\ \vec{x}_2^T \end{bmatrix} \begin{bmatrix} \varepsilon_0 \\ \varepsilon_1 \\ \varepsilon_2 \end{bmatrix} = X^T \vec{\varepsilon} = 0.$$

We therefore are able to minimize $\|\vec{\varepsilon}\|^2$ if and only if $X^T \vec{\varepsilon} = 0$. Therefore this minimized if and only if $X^T \vec{y} = X^T X\vec{\theta} + X^T \vec{\varepsilon}$ if and only if $X^T \vec{y} = X^T X\vec{\theta}$. We therefore see that by perturbing our original equation by a noise vector $\vec{\varepsilon}$ our original solution was in fact the optimal solution. We can conclude that in the case that $X^T X$ is invertible $\vec{\theta} = (X^T X)^{-1} X^T \vec{y}$. We see here that the main intuition came from a geometric approach to data analysis.

2.3 Probabilistic Approach – Maximal Likelihood

How can we utilize probability in the case of linear regression? Observe our original equation again

$$y^{(i)} = \vec{\theta}^T \vec{x}^{(i)} + \varepsilon^{(i)}$$

where again the $\varepsilon^{(i)}$ is an error term that captures some unmodeled affect or random noise in the system. We can use probability by making some assumptions on $\varepsilon^{(i)}$. We can assume that the noise observations $\varepsilon^{(i)}$ are independently and identically distributed (IID) random variables. It is a good assumption that they are each distributed from a Gaussian distribution with mean 0 and variance σ^2. We can denote this relationship by $\varepsilon^{(i)} \sim N(0, \sigma^2)$. This is to say explicitly that

$$p\left(\varepsilon^{(i)}\right) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{(y^{(i)}-\theta^T\vec{x}^{(i)})^2}{2\sigma^2}\right].$$

By the original equation, we can therefore derive a conditional probability of $\vec{y}^{(i)}$ from the observations $\vec{x}^{(i)}$ and $\vec{\theta}$. We therefore have that

$$p(y^{(i)}|\vec{x}^{(i)};\vec{\theta}) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{(y^{(i)}-\theta^T\vec{x}^{(i)})^2}{2\sigma^2}\right]$$

therefore we can conclude from this that the conditional variable $y^{(i)}|\vec{x}^{(i)};\vec{\theta} \sim N(\vec{\theta}^T\vec{x}^{(i)},\sigma^2)$. These equations however were made for only one observation of y. With observations of this type for $i = 1,\ldots,n$ we are again able to collapse this into a matrix equation

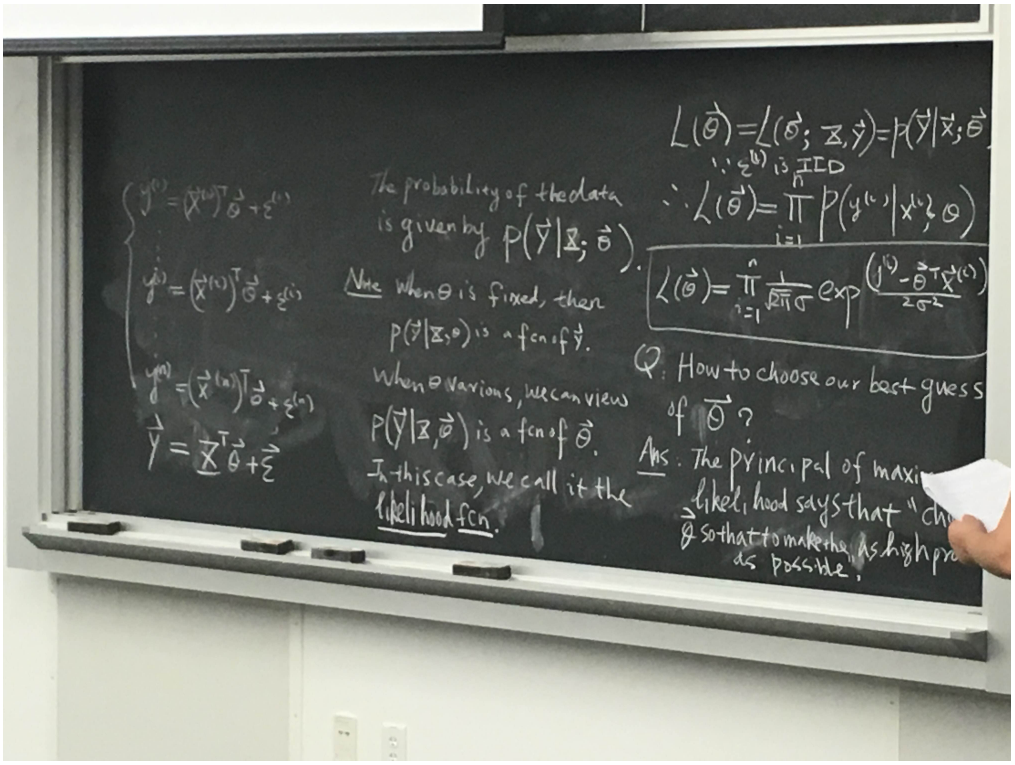
$$\vec{y} = X^T\vec{\theta} + \vec{\varepsilon}.$$

There are two major interpretations we can make here on the vector $\vec{\theta}$ that we are trying to fit: In one way we can view $\vec{\theta}$ is fixed as a function $\vec{y}$, in another way we can view $\vec{\theta}$ as varying hence $p(\vec{y}|X,\vec{\theta})$ is a function of $\vec{\theta}$. In this latter case we call it the **likelihood function**. We can let

$$L(\vec{\theta}) = L(\vec{\theta};X,\vec{y}) = p(\vec{y}|\vec{x};\vec{\theta})$$

and in the case that the $\varepsilon^{(i)}$ are IID, we then have by the rules of probability that

$$L(\vec{\theta}) = \prod_{i=1}^{n} p(y^{(i)} \mid \vec{x}^{(i)};\vec{\theta}) = \prod_{i=1}^{n}\frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{(y^{(i)}-\theta^T\vec{x}^{(i)})^2}{2\sigma^2}\right].$$



The question here now becomes how do we choose the optimal $\vec{\theta}$ to get the highest probability as possible? The **principle of maximal likelihood** states to choose the $\vec{\theta}$ to give this function the highest probability possible. In other words, we want to choose $\vec{\theta}$ to maximize $L(\vec{\theta})$. Analytically this is equivalent to maximizing $\log L(\vec{\theta}) = l(\vec{\theta})$, which from our assumption of IID Gaussian noise, we obtain

$$l(\vec{\theta}) \quad = \quad m \log \frac{1}{\sqrt{2\pi}\sigma} - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y^{(i)} - \vec{\theta}\vec{x}^{(i)})^2.$$

Note however that in order to maximize $l(\vec{\theta})$ we simply need to minimize the last term in the equation above. In other words, we want to minimize

$$\sum_{i=1}^{n} (y^{(i)} - \vec{\theta}\vec{x}^{(i)})^2 \quad \equiv \quad J(\vec{\theta}).$$

We observe here that we recover our original minimization problem with respect to the function $J(\vec{\theta})$.