Math 390.4
Third Theoretical Lecture

Joseph Peltroche

2/5/2018

The training data, $\mathcal{D}$, can represented by the following:
$\mathcal{D} = \{< \vec{x}_1, y_1 >, < \vec{x}_2, y_2 >, \cdots, < \vec{x}_n, y_n >\}$, where $\vec{x}_n$ are the n-th person's characteristics and $y \in \{0, 1\}$ indicating whether the person has paid his loan.

A matrix X can be formulated such that

$$X = \begin{bmatrix} \vec{x}_1 \\ \vec{x}_2 \\ \cdots \\ \vec{x}_n \end{bmatrix} \in \mathcal{X}^n$$

where $\mathcal{X}$ is the n-th co-variate space. Simlilary for y:

$$\vec{y} = \begin{bmatrix} y_1 \\ y_2 \\ \cdots \\ y_n \end{bmatrix} \in \mathcal{Y}^n$$

where † is the n-th output space. Hence

$$\mathcal{D} =< X, \vec{y} >$$

In order to model the desired phenomena, we resort to the previous explanation:

$$y = \underbrace{f(\vec{x})}_{\text{approx. to y}} + \delta, \qquad \delta = t(\vec{z}) - f(\vec{x})$$

After obtaining the training data, the next step is to pick an estimate f. g is the best approximation in $\mathcal{H}$.

In the unlikely case that $f \in \mathcal{H}$ then our model has less errors to consider and is reduced to:

$$y = g(\vec{x}) + \underbrace{f(\vec{x}) - g(\vec{x})}_{\text{parameter estimation error}} + \underbrace{(t(\vec{z}) - f(\vec{x}))}_{\delta\text{: error due to ignorance}}$$

However, the likely case is that $f \notin \mathcal{H}$, in which case we consider an h^* that is the best approximation of $f \in \mathcal{H}$. This cause more sources of error, and our model becomes:

$$y = g(\vec{x}) + \underbrace{h^*(\vec{x}) - g(\vec{x})}_{\text{estimation error}} + \underbrace{(f(\vec{x}) - h^*(\vec{x}))}_{\text{misspecification error}} + \underbrace{(t(\vec{x}) - f(\vec{x}))}_{\text{error due to ignorance}}$$

The three sources of error can be minimized: estimation error can be minimized with a bigger population n, the misspecification error can minimized with a better algorithm $\mathcal{A}$, and better data can help decrease the error due to ignorance.