# Math 390.4
# Second Theoretical Lecture

Joseph Peltroche

2/1/2018

Following the conceptual example from the previous lecture, recall

$$y \in \{0, 1\} = Y(\text{the output space})$$

and $y = t(z_1, z_2 z_3)$, where $Y$ is the output space. This is known as a *true system*. By definition it is not a model, but it has also been called "the true model". For such a model deciding credit worthiness, the following causal inputs were declared

- $z_1$: having sufficient funds
- $z_2$: unforeseen emergencies
- $z_3$: criminal intent

The features $z_1, z_2, z_3$ are reasonable to consider, yet impossible to actually access and obtain any predictions. The problem with this model is that the inputs are inaccessible and unobservable. There is no way to obtain values for these inputs, so, the next best thing is to attempt to define a collection of information that influences the above features.

Consider the following features $x_1, x_2$,and $x_3$:

- $x_1$: A measurement of the average salary over a certain time frame (say over 5 years)
- $x_2$: Statuses of any previous loan repayment $\in \{0, 1\}$

- $x_3$: Criminal record $\in$ {no crime, infraction, misdemeanor, felony}

where $x_n$ is related to $z_n$ for $n \in \{1, 2, 3\}$. Notice these variables are not true inputs but are the best next thing, influential to our true inputs and accesible. Data on $x_n$ can help create predictions that are close to the truth. It is also worth noting that in reality just a subset of this data is available, restricting the accessiblity of the inputs and, consequently, the relability of the model. Social media has become a means of extracting information to better obtain features that can predict reality.

Consider a man named Bob. Bob's information is given by the "observation", $\vec{x}$,

$$\vec{x} = [x_1, x_2 x_3] \in X$$

otherwise known as the record. The inputs $x_1, x_2$, and $x_3$ are also known as features, variables, attributes, regressors, or covariates. The dimensions of $\vec{x}$, $\dim(\vec{x})$, are known as the predictive features, $d$ or $p$. Here, the features themselves distinguish from each other:

- $x_1 \in \mathbb{R}$ a continuous variable
- $x_1 \in \{0, 1\}$ a binary/dummy variable
- $x_1 \notin \mathbb{R}$ a "categorical variable" (has unique possible values)

To proceed in creating a model, we must mathematically represent $x_3$:

(a) We can proceed by creating an order for the categorical variable, e.g. $x_3 \in \{0, 1, 2, 3\}$ where each number can represent an eye color. The fault with using an ordinal cateogrical predictor is that it creates an undesired hierarchy for the objects in consideration. Going back to the previous example, this approach would also not distinguish the difference in judgement between a person with multiple felonies and a person with one felony.

(b) We can turn $x_3$ into multiple binary variables $x_{3a}, x_{3b}, x_{3c}, x_{3d}$, otherwise known as "Dummification". Now

- $x_{3a}$ is a binary for no crime
- $x_{3a}$ is a binary for infraction
- $x_{3a}$ is a binary for misdemeanor
- $x_{3a}$ is a binary for felonies

This will cause the predictive features to increase $p \to 6$. However, now we have a inputs that can influence the true causal inputs.

In an effort to explain $y$, we must create a model $f$ that never outputs anything equal to $y$, the reality, but results in a close approximation.

$$y \approx f(x_1, x_2, x_3)$$

or

$$y = f(x_1, x_2, x_3) + \delta$$

where $\delta$ is known as the error due to ignorance, $\delta = t(\vec{z}) - f(\vec{x})$. Obtaining $f$ is crucial for our system, so to get $f$ it must be understood that there is no analytical solution. Instead, we use an "empirical solution" i.e. use the available data, otherwise known as "learning from data" or "supervised learning". This will require 3 ingredients:

1. $\mathbf{D}$, the training data

2. $\mathbf{H}$, the set of all candidate functions for $f$

3. $\mathbf{A}$, the algorithim which produces $g = \mathbf{A}(\mathbf{D}, \mathbf{H})$