2. How We Do Modeling and Decisions

GULP Winter 2015

January 20, 2015

1 A Statistical Model

We we say we are dealing with a statistical model, we mimic the randomness in the real world by some model whose structure is clearly defined, and we use the data to nail down the parameters - which we think define the model. How do we generate this process into the formal mathematical language? We have to do some rough but more formal definitions. They may differ from advanced textbooks a little bit.

Population - The target of your model. This is the pool of all the possible samples. If you happened to know the population, you could see from above and do all the estimations accurately. Unfortunately this is never true. The examples include: people's height; living time after the treatment; the possible waiting time of one server; etc. We often denote the population by using random variables, say X.

Distribution - This is the main assumption equipped by the population. We **assume** that the population is born with some specific distribution on it. The waiting time is exponential; the height of human is normal; the number of cars passing one crossroad is Poisson, and so on. This is never going to be true, but we always do the assumptions as accurate as possible. We say $X \sim F$, meaning that X is following some distribution F.

Parameters - This is the reason why we are using the distributions to mimic the real world. The distribution is **defined** when the parameters are settled down. If you have λ for Poisson(λ), you can tell every single possible probability under the distribution. If you have the mean and the variance of a normal distribution, everything will be well defined. The target of learning the distribution is therefore simplified to leaning the parameters. We denote this as $X \sim F_\theta$. Unfortunately, the parameter is also **unknown**.

Sample - This is the so called data, which is a realization of the population in your hand. They can be identically independently distributed (i.i.d.), which is the most common and most fancy case. Or they can be awfully samples, which needs some special treatment. Usually they are labeled as $X_i, i = 1, 2, \ldots, n$. Notice that the sample is random variables.

2 Estimators

When we have the data in hand, we would like to use the data to estimate the parameters. However we are not always interested in all the original parameters. We may only be inter-

ested in some function of the parameters, denoted as $g(\theta)$. The weapon in our hands to nail it down, is called statistics.

Statistic - This is a (measurable) function of the data, and the **data only**, which means that it can not have any argument of the unknown parameters. Examples include sample mean, sample variance, indicator $\mathbb{1}(X_i \leq c)$ for some known c, empirical distribution, and so on. We denote the statistic as $\delta(\mathbf{X})$, where $\mathbf{X}= \{X_i, i = 1, 2, \ldots, n\}$.

When we are using $\delta(\mathbf{X})$ to estimate $g(\theta)$, we say that $\delta(\mathbf{X})$ is an **estimator** of $g(\theta)$. The next thing is that, how to measure the behavior of an estimator? We have the following tools to do this.

Loss function - This is defined as $L(g(\theta), \delta(\mathbf{X}))$. This function can take a lot of forms, but some common loose rules will apply:

$$L \geq 0$$

$$L(g(\theta), g(\theta)) = 0, \forall \theta$$

and L is monotone non-decreasing with the increase of the distance between $g(\theta)$ and $\delta(\mathbf{X})$. These regulations are natural.

Since the loss function involves the data, this is also a random variable. To accurately describe the 'loss', we need the following tool.

Risk function - This function is defined as $R_\theta(\delta, g) = \mathbb{E}_\theta\, L(g(\theta), \delta(\mathbf{X}))$. Notice that risk function may differ with different θ values. The argument inside is the function form δ and the function interested g.

Example(Regression). The data is defined as

$$\mathbf{Y} = \mathbf{X}\beta + \mathbf{e}.$$

The loss function is taking the form as

$$L(\beta, \hat{\beta}) = (\mathbf{Y} - \mathbf{X}\hat{\beta})^{\mathbf{2}}.$$

Minimizing using derivatives will yield

$$\mathbf{X'Y} = \mathbf{X'X}\hat{\beta},$$

which is

$$\hat{\beta} = (\mathbf{X'X})^{-\mathbf{1}}\mathbf{X'Y}.$$