Lecture Notes 7
36-705

In today's lecture we will begin to study what are known as uniform laws or uniform tail bounds. Roughly, these are LLNs or tail bounds that apply to a collection of random variables taken together. Results of the type we will develop in the next few lectures form the theoretical basis for the study of statistical estimators, and are core topics in statistics and machine learning. In statistics this area of study is known as *empirical process theory.* We'll start by studying these from a relatively classical viewpoint, discussing what are called Glivenko-Cantelli theorems, and then focus on providing motivation.

1 Uniform convergence of the CDF

A classical question that was already on the mind of probabilists in the early 1930s was:

How can one estimate the CDF of a univariate random variable given a random sample?

Recall that the cdf is defined by $F(x) = P(X \leq x)$. Suppose we observe $X_1, \ldots, X_n \sim F$. The most common estimator is the empirical cdf defined by

$$\widehat{F}_n(x) = \frac{1}{n}\sum_{i=1}^n \mathbb{I}(X_i \leq x).$$

This makes sense since $F(x)$ is the probability that a random X is less than or equal to x and $\widehat{F}_n(x)$ is the observed proportion of of observations less than or equal to x. At each value x, we essentiall have a coin-flipping experiment. But we are simultaneously conducting infinitely many such experiments. Note that $\widehat{F}_n$ is a non-decreasing step-function.

You might have noticed that unlike in a classical statistical estimation problem we are not estimating a simple parameter, rather we are estimating an entire function.

So let us back up a little bit. Suppose I fixed a value x and we decided to try to estimate $F_X(x)$. We could use the empirical CDF at x, but this time it is a rather simple problem. Observe that,

$$\mathbb{E}[\widehat{F}_n(x)] = \frac{1}{n}\sum_{i=1}^n \mathbb{E}[\mathbb{I}(X_i \leq x)] = \mathbb{P}(X \leq x) = F(x).$$

The indicators are bounded random variables so we could just use Hoeffding's bound to conclude that,

$$\mathbb{P}(|\widehat{F}_n(x) - F_X(x)| \geq \epsilon) \leq 2\exp(-2n\epsilon^2).$$

This shows that for every point x, we can use simple tail bounds to say that the empirical CDF is close to the true CDF. Does this imply that the whole function $\widehat{F}_n$ is close to F? It does not. This is the difference between pointwise and uniform convergence.

To see that there is a difference, consider the function $g(x) = 0$ for all x. Let

$$g_n(x) = \begin{cases} 1 & \text{if } x = n \\ 0 & \text{otherwise.} \end{cases}$$

For any fixed x we have that $|g_n(x) - g(x)| \to 0$ as $n \to \infty$. But $\sup_x |g_n(x) - g(x)| = 1$ which does not go to 0.

Back to the cdf, we would like to understand the behaviour of

$$\Delta_n = \sup_{x \in \mathbb{R}} |\widehat{F}_n(x) - F_X(x)|.$$

Reasoning about Δ_n requires us to reason about the CDF everywhere, hence the name *uniform* bounds or *uniform* LLNs. The Glivenko-Cantelli theorem says that for *any distribution*, Δ_n converges to 0 in probability.

Theorem 1 Glivenko-Cantelli Theorem. *Let $X_1, \ldots, X_n \sim F$ and define*

$$\widehat{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}(X_i \leq x).$$

Then

$$\sup_x |\widehat{F}_n(x) - F(x)| \xrightarrow{P} 0.$$

Notes:

1. The Glivenko-Cantelli theorem is like a WLLN but it is a uniform WLLN that ensures that the WLLN is true simultaeously at every point $x \in \mathbb{R}$.

2. There is a corresponding strong theorem that guarantees convergence almost surely.

3. One should pay particular attention to the fact that we can estimate the CDF of a random variable with *no assumptions*. This is contrast to estimating the density of a random variable which typically requires strong smoothness assumptions (we will re-visit this much later in the course).

2 Empirical Processes

For any set A, define the empirical probability as

$$\mathbb{P}_n(A) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}(X_i \in A).$$

The quantity Δ_n defined earlier can be written as,

$$\Delta_n = \sup_{A \in \mathcal{A}} |\mathbb{P}_n(A) - \mathbb{P}(A)|,$$

where $\mathcal{A}$ is a collection of sets,

$$\mathcal{A} = \{A(x) : A(x) = (-\infty, x]\},$$

since in this case, $\mathbb{P}(A(x)) = F_X(x)$.

One could generalize the CDF question from the previous section further to ask more generally about other interesting collections of sets $\mathcal{A}$, i.e. we are interested in collections of sets $\mathcal{A}$, for which we have uniform convergence, i.e.

$$\Delta_n(\mathcal{A}) = \sup_{A \in \mathcal{A}} |\mathbb{P}_n(A) - \mathbb{P}(A)|,$$

converges to 0 (in probability, say). This line of inquiry forms the basis for what is called *Vapnik-Cervonenkis* theory who were amongst the first to ask this general question.

Even more generally, one can replace the indicators with general (integrable) functions, i.e. let $\mathcal{F}$ be a class of integrable, real-valued functions, and suppose we have an i.i.d. sample $X_1, \ldots, X_n \sim P$, then we could be interested in,

$$\Delta_n(\mathcal{F}) = \sup_{f \in \mathcal{F}} \left| \frac{1}{n}\sum_{i=1}^{n} f(X_i) - \mathbb{E}[f] \right|.$$

This quantity is known as an *empirical process* and empirical process theory is the area of statistics that asks questions about the convergence in probability, almost surely or in distribution for the quantity $\Delta(\mathcal{F})$ for interesting classes of functions $\mathcal{F}$. If $\mathcal{F} = \{I_A(\cdot) : A \in \mathcal{A}\}$ for some class of sets $\mathcal{A}$, then we get back $\Delta_n(\mathcal{A})$.

If $\Delta_n(\mathcal{F}) \xrightarrow{p} 0$ then we say that $\mathcal{F}$ is a *Glivenko-Cantelli* class. The class of functions: We refer to classes for which $\Delta(\mathcal{F}) \xrightarrow{p} 0$, as a *Glivenko-Cantelli* class. The class of functions:

$$\mathcal{F} = \{\mathbb{I}(-\infty, x],\ x \in \mathbb{R}\},$$

which defines the uniform convergence of the CDF is an example of a Glivenko-Cantelli class.

3 Failure of Glivenko-Cantelli

In general, very complex classes of functions or sets will fail to be Glivenko-Cantelli and one of the goals of the next few lectures is to find ways to measure the complexity of a class of functions. For example, suppose we draw $X_1, \ldots, X_n \sim P$ where P is some continuous distribution over $[0, 1]$. Suppose further that $\mathcal{A}$ is all subsets of $[0, 1]$ with finitely many elements.

Since the distribution is continuous we have that, $\mathbb{P}(A) = 0$ for each $A \in \mathcal{A}$, however for the finite set $\{X_1, \ldots, X_n\}$ we have that $\mathbb{P}_n(A) = 1$, i.e.

$$\Delta(\mathcal{A}) = \sup_{A \in \mathcal{A}} |\mathbb{P}_n(A) - \mathbb{P}(A)| = 1,$$

no matter how large n is. So the collection of sets $\mathcal{A}$ is not Glivenko-Cantelli. Roughly, the collection of sets is “too large”.

4 Estimation of Statistical Functionals

We discussed estimating the CDF of a random variable. In this section we provide several examples of problems where we use estimates of the CDF. Furthermore, as we will see, we can develop a unified understanding of such estimators using the Glivenko-Cantelli theorem.

Often we want to estimate some quantity which can be written as a simple functional of the CDF, and a natural estimate just replaces the true CDF with the empirical CDF (such estimators are known as plug-in estimators). As an aside, a functional is just a function of a function. A statistical functional is a functional of the CDF. Here are some examples:

1. **Expectation Functionals:** For a given function g, we can view the usual empirical estimator of its expectation as a plug-in estimate where we replace the population CDF by the empirical CDF,

$$\widehat{\mathbb{E}}[g(X)] = \frac{1}{n} \sum_{i=1}^{n} g(X_i) = \int_x g(x) d\widehat{F}_n(x).$$

2. **Quantile Functionals:** For an $\alpha \in [0, 1]$, the α-th quantile of a distribution is given as:

$$Q_\alpha(F) = \inf\{t \in \mathbb{R} | F(t) \geq \alpha\}.$$

Taking $\alpha = 0.5$ gives the median. A natural plug-in estimator of $Q_\alpha(F)$ is to simply take $Q_\alpha(\widehat{F}_n)$.

3. **Goodness-of-fit Functionals:** We will re-visit this topic in more detail when we talk about hypothesis testing but often in data analysis we want to test the hypothesis that data we have are i.i.d. from some known distribution F_0. The rough idea is we form a statistic to test this hypothesis which (hopefully) takes large values when the distribution is not F_0 and takes small values otherwise. Typical tests of this form include the Kolmogorov-Smirnov test, where we compute the plug-in quantity:

   $$\widehat{T}_{KS} = \sup_{x \in \mathbb{R}} |\widehat{F}_n(x) - F_0(x)|,$$

   which is natural because if the true distribution is F_0 we know by the Glivenko-Cantelli theorem that T_{KS} is small. Similarly, one can use the Cramer-von Mises test which uses the plug-in statistic,

   $$\widehat{T}_{CvM} = \int_x (\widehat{F}_n(x) - F_0(x))^2 dF_0(x).$$

There are many other statistical functionals for which the usual estimators can be thought of as plug-in estimators. For example: the variance, correlation, and higher moments can all be expressed in this fashion.

In each of the above cases we are interested in estimating some functional $\gamma(F)$ and we use the plug-in estimator $\gamma(\widehat{F}_n)$. Analogous to the continuous mapping theorem, there is a Glivenko-Cantelli theorem that provides a WLLN for these estimators. We need to first define a notion of continuity. Suppose γ satisfies the property that for every $\epsilon > 0$, there is a $\delta > 0$ such that if,

$$\sup_x |\widehat{F}_n(x) - F(x)| \leq \delta,$$

then

$$|\gamma(F) - \gamma(\widehat{F}_n)| \leq \epsilon.$$

For such functionals γ, it is a simple consequence of the Glivenko-Cantelli theorem that $\gamma(\widehat{F}_n)$ converges in probability to $\gamma(F)$.

5 Risk Minimization

Perhaps the most compelling motivation for studying uniform convergence is to understand a procedure known as empirical risk minimization. Estimators of this type include maximum likelihood estimators, and many estimators we encounter in machine learning (SVMs, Boosting and so on). We will study this in detail in the next lecture.

Binary Classification: In the typical binary classification setting we observe a training set $\{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ that we assume are drawn i.i.d from some distribution P. Each $X_i \in \mathbb{R}^d$, $Y_i \in \{-1, +1\}$.

A classifier $f : \mathbb{R}^d \mapsto \{-1, +1\}$ is simply a function that takes an x outputs a label in $\{-1, +1\}$. The broad goal of classification is to try to find a function that has low error on future unseen data, i.e. we want a function that has low mis-classification error: $\mathbb{P}(f(X) \neq y)$.

For a given classifier f we can estimate its mis-classification error (risk) as:

$$\widehat{R}_n(f) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}(f(X_i) \neq y_i),$$

which is simply its error on the training set. If f is some fixed classifier we know by Hoeffding's bound (why?) that,

$$\mathbb{P}(|\widehat{R}_n(f) - \mathbb{P}(f(X) \neq y)| \geq t) \leq 2\exp(-2nt^2).$$

If we are trying to pick a good classifier from some set of classifiers $\mathcal{F}$, then a natural way to do this is to find the one that looks best on the training set, i.e. to choose

$$\widehat{f} = \arg\min_{f \in \mathcal{F}} \widehat{R}_n(f).$$

This procedure is known as *empirical risk minimization*. The terminology will be clearer later on in the course. For now though, we would like to understand this procedure better. How do we argue that in some cases this procedure will indeed select a good classifier? This question is intricately tied to uniform convergence.

Let f^* be the best classifier in $\mathcal{F}$. We would like to bound the excess risk of the classifier we chose, i.e.

$$\Delta = \mathbb{P}(\widehat{f}(X) \neq y) - \mathbb{P}(f^*(X) \neq y).$$

The typical way to do this is to consider the decomposition:

$$\Delta = \underbrace{\mathbb{P}(\widehat{f}(X) \neq y) - \widehat{R}_n(\widehat{f})}_{T_1} + \underbrace{\widehat{R}_n(\widehat{f}) - \widehat{R}_n(f^*)}_{T_2} + \underbrace{\widehat{R}_n(f^*) - \mathbb{P}(f^*(X) \neq y)}_{T_3}.$$

Since $\widehat{f}$ minimizes the empirical risk we know that $T_2 \leq 0$. We know that T_3 is small just by the Hoeffding argument from before, since f^* is a fixed classifier (i.e. does not depend on the training data).

The key point, one that you should really think carefully about is that we cannot use Hoeffding for the first term. The reason is that the classifier $\widehat{f}$ is data dependent so its empirical risk is not the sum of independent RVs.

Instead we have to rely on a *uniform* convergence bound, i.e. suppose we can show that with probability at least $1 - \delta/2$,

$$\sup_{f \in \mathcal{F}} \left[\mathbb{P}(f(X) \neq y) - \widehat{R}_n(f) \right] \leq \Theta,$$

then we can conclude that the excess risk with probability at least $1 - \delta$ satisfies

$$\Delta = \mathbb{P}(\widehat{f}(X) \neq y) - \mathbb{P}(f^*(X) \neq y) \leq \Theta + \sqrt{\frac{2 \log(2/\delta)}{n}},$$

so everything boils down to showing uniform convergence of the empirical risk to the true error over the collection of classifiers we are interested in.

We will discuss this is more detail later.