36-705: Intermediate Statistics **Fall 2019**

Lecture 36: December 6

Lecturer: Siva Balakrishnan

This and the next lecture are completely optional (and will not be on your final exam).

In the last lecture we spent most of our time discussing several popular f-divergences (KL, χ^2, TV and Hellinger), and some reasons why they are special:

1. (Lower Bounds) We discussed that if any of these distances are sufficiently small, then the distributions are indistinguishable.

2. (Upper Bounds) We discussed that at least for the TV and Hellinger there are strong converses. We showed that if the $H^2(P, Q) \gg 1/n$ then we can reliably distinguish P and Q when we have n samples from them.

3. We discussed that Hellinger and KL in particular are nice in the usual statistical learning setup when we have n samples, i.e. there is a simple relationship between the distance between P and Q, and the distance between n samples from P and n samples from Q.

4. We discussed minimum distance estimators (a generalization of maximum likelihood) and how these estimators can be robust to model-misspecification in different distances. In particular, we discussed that if we have (outlier) contaminated data then rather than use the MLE we might choose to try to use something that attempts to find the closest distribution to the sampling distribution in TV.

36.1 More Connections Between Testing and Estimation – Fano's Inequality

We focused primarily on the role of f-divergences in testing but they are equally fundamental in providing lower bounds for estimation. In estimation we obtain samples $X_1, \ldots, X_n \sim P_\theta$ where $P_\theta \in \mathcal{P}_\Theta$, and our goal is to estimate θ (say with small ℓ_2 error).

In an intuitive sense, estimation is a lot like like a multiple hypothesis testing problem of the following form – I give you samples from one of M distributions $\{P_{\theta_1}, \ldots, P_{\theta_M}\}$ and I ask you to figure out which one it was. Our estimator Ψ in this context simply takes in n samples and returns an index $\{1, \ldots, M\}$.

We could imagine the setting where we sample an index u uniformly from $\{1, \ldots, M\}$ and generate n samples from P_{θ_u}. We define the following notion of error:

$$\text{err} = P(\Psi(X_1, \ldots, X_n) \neq u).$$

Fano's inequality (and others like it) relate how hard this testing problem is (i.e. they lower bound err) to a function of some distance between the distributions $\{P_{\theta_1}, \ldots, P_{\theta_M}\}$. I will state some simplified version of Fano's inequality that will give us some interesting estimation minimax lower bounds, and as usual be pretty sloppy with constants.

Suppose that $M \geq 3$, and for some small constant $c_1 > 0$:

$$\frac{1}{M^2} \sum_{i=1}^{M} \sum_{j=1}^{M} \text{KL}(P_{\theta_i} \| P_{\theta_j}) \leq c_1 \log M,$$

then for some other $c_2 > 0$, err $> c_2$. In words, Fano's inequality says if the average pairwise KL divergence is small then multiple testing problem is difficult.

So how does this relate to estimation? Suppose we additionally ensure that for all pairs (i, j) we have that $\|\theta_i - \theta_j\|_2^2 \geq (2\delta)^2$, then it is easy to verify that the minimax estimation error:

$$\inf_{\widehat{\theta}} \sup_{\theta \in \Theta} \mathbb{E}\|\widehat{\theta} - \theta\|_2^2 \geq \inf_{\widehat{\theta}} \sup_{\theta \in \{\theta_1, \ldots, \theta_M\}} \mathbb{E}\|\widehat{\theta} - \theta\|_2^2 \geq \inf_{\widehat{\theta}} \frac{1}{M} \sum_{i=1}^{M} \mathbb{E}_{X_1, \ldots, X_n \sim P_{\theta_i}} \|\widehat{\theta} - \theta_i\|_2^2 \geq c_2 \delta^2,$$

using Markov's inequality.

So now we can try to understand the Fano inequality game. To produce a tight lower bound, we need to find many hypothesis (large M) such that their average KL divergence is sufficiently small, but their corresponding parameters are well-separated. Intuitively, this should make sense, if we can find parameters that are well-separated but the underlying distributions are very close, then the estimation problem should be difficult.

An Application: So why do we need Fano's inequality? Suppose we consider establishing a lower bound for estimating the mean of a Normal distribution. We have already seen that the minimax error is at least $\sigma^2 d/n$ (but this required a complicated Bayes argument). On the other using a simple versus simple testing problem (with two separated Normals) we can easily show via Le Cam's lemma that the error is at least σ^2/n. To get the right dimension dependence we will need to use the full power of Fano's inequality.

Let us see how this works. We need to know a few facts: one is that there is a packing of the radius 4δ sphere of size $c2^d$ (for some $c > 0$), such that:

$$\|\theta_i\|_2 \leq 4\delta, \text{ and } \|\theta_i - \theta_j\| \geq 2\delta$$

i.e. there are roughly 2^d vectors in the 4δ-sphere that are well-separated (i.e. are separated by at least 2δ).

Additionally, we can calculate the KL divergence between n-samples from $N(\theta_i, I_d)$ and $N(\theta_j, I_d)$,

$$\mathrm{KL}(P_{\theta_i}, P_{\theta_j}) = \frac{n}{2\sigma^2}\|\theta_i - \theta_j\|_2^2 \leq \frac{cn\delta^2}{\sigma^2}.$$

So (ignoring constants) $n\delta^2 \ll \sigma^2 d$, then our minimax estimation error is at least δ^2. So this gives us that the minimax estimation error is at least $c\sigma^2 d/n$.

Again, it's worth remarking – this argument is very simple (and robust in the sense that it generalizes more easily to other settings) relative to the Bayes argument (which really required a lot of precise analytical computations). On the other hand the bound is off by constants.

36.2 Distances for data analysis

Although the distances we have discussed so far are fundamental for many analytic purposes, we do not often use them in data analysis. This is because they can be difficult to estimate in practice. Ideally, we want to roughly be able to trade-off how "expressive" the distance is versus how easy it is to estimate (and the distances so far are on one end of the spectrum).

To be a bit more concrete lets think about a typical setting (closely related to two-sample testing that we have discussed earlier): we observe samples $X_1, \ldots, X_n \sim P$ and $Y_1, \ldots, Y_n \sim Q$, and we want to know how different the two distributions are, i.e. we want to *estimate* some divergence between P and Q given samples.

This idea has caught on again recently because of GANs, where roughly we want to generate samples that come from a distribution that is close to the distribution that generated the training samples, and often we do this by estimating a distance between the training and generated distributions and then trying to make it small.

Another popular task is *independence testing* or *measuring dependence* where we are given samples $(X_1, Y_1), \ldots, (X_n, Y_n) \sim P_{XY}$ and we want to estimate/test how far apart P_{XY} and $P_X P_Y$ are (i.e. we want to know how far apart the joint and product of marginals are).

36.3 Integral Probability Metrics

If two distributions P, Q are identical then it should be clear that for any (measurable) function f, it must be the case that:

$$\mathbb{E}_{X\sim P}[f(X)] = \mathbb{E}_{Y\sim Q}[f(Y)].$$

One might wonder if the reverse implication is true, i.e. is it that if $P \neq Q$ then there must be some *witness* function f such that:

$$\mathbb{E}_{X\sim P}[f(X)] \neq \mathbb{E}_{Y\sim Q}[f(Y)].$$

It turns out that this statement is indeed true. In particular, we have the following classical lemma.

Lemma 36.1 *Two distributions P, Q are identical if and only if for every continuous function $f \in C(\mathcal{X})$*

$$\mathbb{E}_{X\sim P}[f(X)] = \mathbb{E}_{Y\sim Q}[f(Y)].$$

This result suggests that to measure the distance between two distributions we could use a so-called integral probability metric (IPM):

$$d_{\mathcal{F}}(P, Q) = \sup_{f\in\mathcal{F}} |\mathbb{E}_{X\sim P}[f(X)] - \mathbb{E}_{Y\sim Q}[f(Y)]|,$$

where $\mathcal{F}$ is a class of functions. We note that the TV distance is thus just an IPM with $\mathcal{F} = \{f : \|f\|_\infty \leq 1\}$. This class of functions (as well as the class of all continuous functions) is too large to be useful statistically, i.e. it is the case that these IPMs are not easy to estimate from data, so we instead use function classes that have *smooth* functions.

36.4 Wasserstein distance

A natural class of smooth functions is the class of 1-Lipschitz functions, i.e.

$$\mathcal{F}_L = \{f : f\text{continous }, |f(x) - f(y)| \leq \|x - y\|\}.$$

The corresponding IPM is known as the Wasserstein 1 distance:

$$W_1(P, Q) = \sup_{f\in\mathcal{F}_L} |\mathbb{E}_{X\sim P}[f(X)] - \mathbb{E}_{Y\sim Q}[f(Y)]|.$$

As with the TV there are many alternative ways of defining the Wasserstein distance. In particular, there are very nice interpretations of Wasserstein as a distance between "couplings" and as a so-called transportation distance.

The Wasserstein distance has the somewhat nice property of being well-defined between a discrete and continuous distribution, i.e. the two distributions you are comparing do not need to have the same support. This is one of the big reasons why they are popular in ML.

In particular, a completely reasonable estimate of the Wasserstein distance between two distributions, given samples from each of them is the Wasserstein distance between the corresponding empirical measures, i.e. we estimate:

$$\widehat{W}_1(P, Q) = W_1(\mathbb{P}_n, \mathbb{Q}_n),$$

where $\mathbb{P}_n$ (for instance) is the distribution that puts mass $1/n$ on each sample point.

There are many other nice ways to interpret the Wasserstein distance, and it has many other elegant properties. It is central in the field of optimal transport. A different expression for the Wasserstein distance involves coupling: i.e. a joint distribution J over X and Y, such that the marginal over X is P and over Y is Q. Then the W_1 distance (or more generally W_p distance) is:

$$W_p^p(P, Q) = \inf_J \mathbb{E}\|X - Y\|_2^p.$$

One can also replace the Euclidean distance by any metric on the space on which P and Q are defined. More generally, there is a way to view Wasserstein distances as measuring the cost of optimally moving mass of the distribution P to make it look like the distribution Q (hence, the term optimal transport).

The Wasserstein distance also arises frequently in image processing. In part, this is because the Wasserstein barycenter (a generalization of a mean) of a collection of distributions preserves the shape of the distribution. This is quite unlike the "usual" average of the distributions.

36.5 Maximum Mean Discrepancy

Another natural class of IPMs is where we restrict $\mathcal{F}$ to be a unit ball in a Reproducing Kernel Hilbert Space (RKHS). An RKHS is a function space associated with a kernel function k that satisfies some regularity conditions. We will not be too formal about this but you should think of an RKHS as another class of smooth functions (just like the 1-Lipschitz functions) that has some very convenient properties. A kernel is some measure of similarity, a commonly used one if the RBF kernel:

$$k(X, Y) = \exp\left[-\frac{\|X - Y\|_2^2}{2\sigma^2}\right]$$

The MMD is a typical IPM:

$$\text{MMD}(P, Q) = \sup_{f \in \mathcal{F}_k} |\mathbb{E}_{X \sim P}[f(X)] - \mathbb{E}_{Y \sim Q}[f(Y)]|,$$

but because the space is an RKHS with a kernel k, it turns out we can write this distance as:

$$\mathrm{MMD}^2(P,Q) = \mathbb{E}_{X,X'\sim P}k(X,X') + \mathbb{E}_{Y,Y'\sim Q}k(Y,Y') - 2\mathbb{E}_{X\sim P,Y\sim Q}k(X,Y).$$

Intuitively, we are contrasting how similar the samples from P look to each other and Q look to each other to how similar the samples of P are to Q. If P and Q are the same then all of these expectations should be the same and the MMD will be 0.

The key point that makes the MMD so popular is that it is completely trivial to estimate the MMD since it is a bunch of expected values for which we can use the empirical expectations. We have discussed U-statistics before, the MMD can be estimated by a simple U-statistic:

$$\widehat{MMD}^2(P,Q) = \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}k(X_i,X_j) + \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}k(Y_i,Y_j) - \frac{2}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}k(X_i,Y_j).$$

In particular, if the kernel is bounded then we can use a Hoeffding-style bound to conclude that we can estimate the MMD at $1/\sqrt{n}$ rates.

However, the usefulness of the MMD hinges not on how well we can estimate it but on how strong a notion of distance it is, i.e. for distributions that are quite different (say in the TV/Hellinger/...distances), is it the case that the MMD is large? This turns out to be quite a difficult question to answer.