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

Lecture 35: December 4

Lecturer: Siva Balakrishnan

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

Today we will discuss distances and metrics between distributions that are useful in statistics. I will be loose in distinguishing metrics, distances and divergences (i.e. some of these are not symmetric, and some do not satisfy triangle inequality). We will discuss them in two contexts:

1. There are metrics that are analytically useful in a variety of statistical problems, i.e. they have intimate connections with estimation and testing.

2. There are metrics that are useful in data analysis, i.e. given data we want to measure some notion of distance between (the distribution of) subsets of the data and use this in some way.

There is of course overlap between these contexts and so there are distances that are useful in both, but they are motivated by slightly different considerations.

35.1 The fundamental statistical distances

There are four notions of distance that have an elevated status in statistical theory. I will define them a bit crudely (trying to avoid some basic measure theory that is necessary) to try to focus on the main ideas. Throughout, we will let P, Q be two probability measures on some sample space, and we denote their densities with respect to some common underlying dominating measure be p and q. To further simplify things, I will assume that this common measure is the Lebesgue measure.

1. Total Variation: The TV distance between two distributions is:

$$\mathrm{TV}(P, Q) = \sup_A |P(A) - Q(A)| = \sup_A \left| \int (p(x) - q(x))dx \right|,$$

   where A is just any measurable subset of the sample space, i.e. the TV distance is measuring the maximal difference between the probability of an event under P versus under Q.

The TV distance is equivalent to the ℓ_1 distance between the densities, i.e. one can show that:

$$\mathrm{TV}(P,Q) = \frac{1}{2}\int |p(x) - q(x)|dx.$$

One can also write the TV distance as:

$$\mathrm{TV}(P,Q) = \sup_{\|f\|_\infty \le 1} |\mathbb{E}_P[f] - \mathbb{E}_Q[f]|\,.$$

We will return to this form later on in the lecture.

2. The χ^2 divergence: The χ^2 divergence is defined for distributions P and Q such that Q dominates P, i.e. if $Q(A) = 0$ for some set A then it has to be the case that $P(A)$ is also 0. For such distributions:

$$\chi^2(P,Q) = \int_{\{x:q(x)>0\}} \frac{p^2(x)}{q(x)}dx - 1.$$

Alternatively, one can write this as:

$$\chi^2(P,Q) = \int_{\{x:q(x)>0\}} \frac{(p(x)-q(x))^2}{q(x)}dx.$$

3. Kullback-Leibler divergence: Again we suppose that Q dominates P. The KL divergence between two distributions:

$$\mathrm{KL}(P,Q) = \int \log\frac{p(x)}{q(x)}p(x)dx.$$

4. Hellinger distance: The Hellinger distance between two distributions is,

$$H(P,Q) = \left[\int (\sqrt{p(x)} - \sqrt{q(x)})^2 dx\right]^{1/2},$$

i.e. the Hellinger distance is the ℓ_2 norm between $\sqrt{p}$ and $\sqrt{q}$. It might seem at first a bit weird to consider the ℓ_2 norm between $\sqrt{p}$ and $\sqrt{q}$ rather than p and q, but this turns out to be the right thing to do for statistical applications. The use of Hellinger is various statistical contexts was popularized by Lucien Le Cam, who advocated strongly (and convincingly) for thinking of square-root of the density as the central object of interest.

The Hellinger distance is also closely related to what is called the *affinity* (or Bhattacharya coefficient):

$$\rho(P,Q) = \int \sqrt{p(x)q(x)}dx.$$

In particular, note the equality:

$$H^2(P,Q) = 2\left(1 - \rho(P,Q)\right).$$

All of these fundamental statistical distances are special cases of what are known as f-divergences. The field of information theory has devoted considerable effort to studying families of distances (α, β, ϕ, f-divergences) and so on, and this has led to a fruitful interface between statistics and information theory. An f-divergence is defined for a *convex* function f with $f(1) = 0$:

$$D_f(P, Q) = \int q(x) f\left(\frac{p(x)}{q(x)}\right) dx.$$

You can look up (for instance on Wikipedia) which functions lead to each of the divergences we defined above.

35.2 Hypothesis Testing Lower Bounds

We covered this part earlier in the course as well. When we started this piece of the lecture we discussed that the above divergences are central in statistics. Let us try to substantiate this statement.

A basic hypothesis testing problem is the following: suppose that we have two distributions P_0 and P_1, and we consider the following experiment: I toss a fair coin and if it comes up heads I give you a sample from P_0 and if it comes up tails I give you a sample from P_1. Let $T = 0$ if the coin comes up heads and $T = 1$ otherwise. You only observe the sample X, and need to tell me which distribution it came from.

This is exactly like our usual simple versus simple hypothesis testing problem, except I pick each hypothesis with probability $1/2$ (formally this is what would be called a Bayesian hypothesis testing problem).

Now, suppose you have a test $\Psi : X \mapsto \{0, 1\}$, then I can define its error rate as:

$$\mathbb{P}(\Psi(X) \neq T) = \frac{1}{2}\left[\mathbb{P}_0(\Psi(X) \neq 0) + \mathbb{P}_1(\Psi(X) \neq 1)\right].$$

Roughly, we might believe that if P_0 and P_1 are close in an appropriate distance measure then the error rate of the best possible test should be high and otherwise the error rate should be low. This is known an Le Cam's Lemma.

Lemma 35.1 *For any two distributions P_0, P_1, we have that,*

$$\inf_{\Psi} \mathbb{P}(\Psi(X) \neq T) = \frac{1}{2}\left(1 - TV(P_0, P_1)\right).$$

Before we prove this result we should take some time to appreciate it. What Le Cam's Lemma tells us is that if two distributions are close in TV then *no test* can distinguish

them. In some sense TV is the right notion of distance for statistical applications. In fact, in theoretical CS, the TV distance is sometimes referred to as the *statistical distance* (I do not endorse this terminology). It is also the case that the likelihood ratio test achieves this bound exactly (should not surprise you since it is a simple-vs-simple hypothesis test).

Proof: For any test Ψ we can denote its acceptance region A, i.e. if $X \in A$ then $\Psi(X) = 0$. Then,

$$\begin{aligned}\frac{1}{2}\left[\mathbb{P}_0(\Psi(X) \neq 0) + \mathbb{P}_1(\Psi(X) \neq 1)\right] &= \frac{1}{2}\left[\mathbb{P}_0(X \notin A) + \mathbb{P}_1(X \in A)\right] \\ &= \frac{1}{2}\left[1 - (\mathbb{P}_0(X \in A) - \mathbb{P}_1(X \in A))\right].\end{aligned}$$

So to find the best test we simply minimize the RHS or equivalently:

$$\begin{aligned}\inf_{\Psi} \frac{1}{2}\left[\mathbb{P}_0(\Psi(X) \neq 0) + \mathbb{P}_1(\Psi(X) \neq 1)\right] &= \frac{1}{2}\left[1 - \sup_{A}(\mathbb{P}_0(X \in A) - \mathbb{P}_1(X \in A))\right] \\ &= \frac{1}{2}\left(1 - \mathrm{TV}(P_0, P_1)\right).\end{aligned}$$

Close analogues of Le Cam's Lemma hold for all of the other divergences above, i.e. roughly, if any of the χ^2, Hellinger or KL divergences are small then we cannot reliably distinguish between the two distributions. If you want a formal statement see Theorem 2.2 in Tsybakov's book.

35.3 Tensorization

Given the above fact, that all of the distances we defined so far are in some sense "fundamental" in hypothesis testing, a natural question is why do we need all these different distances?

The answer is a bit technical, but roughly, when we want to compute a lower bound (i.e. understand the fundamental statistical difficulty of our problem) some divergences might be easier to compute than others. For instance, it is often the case that for mixture distributions the χ^2 is easy to compute, while for many parametric models the KL divergence is natural (in part because it is closely related to the Fisher distance as you have seen in some earlier HW). Knowing which divergence to use when is a bit of an art but having many tools in your toolbox is always useful.

One natural thing that will arise in statistical applications is that unlike the above setting of Le Cam's Lemma we will observe n i.i.d. samples $X_1, \ldots, X_n \sim P$, rather than just one sample. Everything we have said so far works in exactly the same way, except we need to calculate the distance between the product measures, i.e. $d(P^n, Q^n)$, which is just the

distance between the distributions:

$$p(X_1, \ldots, X_n) = \prod_{i=1}^{n} p(X_i).$$

For the TV distance this turns out to be quite difficult to do directly. However, one of the most useful properties of the Hellinger and KL distance is that they *tensorize*, i.e. they behave nicely with respect to product distributions. In particular, we have the following useful relationships:

$$\begin{aligned}
&\mathrm{KL}(P^n, Q^n) = n\mathrm{KL}(P, Q) \\
&H^2(P^n, Q^n) = 2\left(1 - \left(1 - \frac{H^2(P,Q)}{2}\right)^n\right) = 2\left(1 - \rho(P,Q)^n\right),
\end{aligned}$$

where ρ is the affinity defined earlier. The key point is that when we see n i.i.d samples it is easy to compute the KL and Hellinger.

35.4 Hypothesis Testing Upper Bounds

One can ask if there are analogous upper bounds, i.e. for instance if the distance between P_0 and P_1 gets larger, are there quantitatively better tests for distinguishing them?

For Hellinger and TV the answer turns out to be yes (and for χ^2 and KL the answer is yes under some assumptions). Formally, given n samples from either P_0 or P_1 you can construct tests that distinguish between P_0 and P_1 such that for some constant $c_1, c_2 > 0$:

$$\inf_{\Psi} \frac{1}{2}\left[\mathbb{P}_0(\Psi(X) \neq 0) + \mathbb{P}_1(\Psi(X) \neq 1)\right] \leq c_1 \exp(-c_2 n \mathrm{TV}(P_0, P_1)),$$

and similarly

$$\inf_{\Psi} \frac{1}{2}\left[\mathbb{P}_0(\Psi(X) \neq 0) + \mathbb{P}_1(\Psi(X) \neq 1)\right] \leq c_1 \exp(-c_2 n H^2(P_0, P_1)).$$

These are sometimes called large-deviation inequalities (and in most cases precise constants are known).

It turns out that even for distinguishing two hypotheses that are separated in the Hellinger distance, the likelihood ratio test is optimal (and achieves the above result). The proof is short and elegant so we will cover it.

Proof: We recall the elementary bound:

$$\log x \leq x - 1, \quad \text{for all } x \geq 0.$$

Which in turn tells us that:

$$\log \rho(P_0, P_1) \leq -\frac{H^2(P_0, P_1)}{2}.$$

So now let us analyze the LRT which rejects the null if:

$$\prod_{i=1}^{n} \frac{P_0(X_i)}{P_1(X_i)} \leq 1.$$

Let us study its Type I error (its Type II error bound follows essentially the same logic). We note that:

$$\begin{aligned} P_0\left(\prod_{i=1}^{n} \frac{P_0(X_i)}{P_1(X_i)} \leq 1\right) &= P_0\left(\prod_{i=1}^{n} \frac{P_1(X_i)}{P_0(X_i)} \geq 1\right) \\ &= P_0\left(\prod_{i=1}^{n} \sqrt{\frac{P_1(X_i)}{P_0(X_i)}} \geq 1\right) \\ &\leq \mathbb{E}_{P_0} \prod_{i=1}^{n} \sqrt{\frac{P_1(X_i)}{P_0(X_i)}}, \end{aligned}$$

using Markov's inequality. Now, using independence we see that:

$$\begin{aligned} P_0\left(\prod_{i=1}^{n} \frac{P_0(X_i)}{P_1(X_i)} \leq 1\right) &\leq \left[\mathbb{E}_{P_0} \sqrt{\frac{P_1(X)}{P_0(X)}}\right]^n \\ &= \exp(n \log \rho(P_0, P_1)) \\ &\leq \exp(-nH^2(P_0, P_1)/2). \end{aligned}$$

Putting this together with an identical bound under the alternate we obtain,

$$\inf_{\Psi} \frac{1}{2} \left[\mathbb{P}_0(\Psi(X) \neq 0) + \mathbb{P}_1(\Psi(X) \neq 1)\right] \leq \exp(-nH^2(P_0, P_1)/2).$$

The result is quite nice – it says that the LRT can distinguish two distributions reliably provided their Hellinger distance is large compared to $1/\sqrt{n}$. Furthermore, the bound has an exponential form so you might imagine that it will interact nicely with a union bound (in a multiple testing setup where we want to distinguish between several distributions).

35.5 Inequalities

Given the fact that these four distances are fundamental and that they have potentially different settings where they are easy to use it is also useful to have inequalities that relate

these distances. There are many (I refer to them as Pinsker-type inequalities) and many textbooks will list them all.

The following inequalities reveal a sort of hierarchy between the distances:

$$\mathrm{TV}(P,Q) \le H(P,Q) \le \sqrt{\mathrm{KL}(P,Q)} \le \sqrt{\chi^2(P,Q)}.$$

This chain of inequalities should explain why it is the case that if any of these distances are too small we cannot distinguish the distributions. In particular, if any distance is too small then the TV must be small and we then use Le Cam's Lemma.

There are also reverse inequalities in some cases (but not all). For instance:

$$\frac{1}{2}H^2(P,Q) \le \mathrm{TV}(P,Q) \le H(P,Q),$$

so up to the square factor Hellinger and TV are closely related.

35.6 Distances from parametric families

We have encountered these before: they are usually only defined for parametric families:

1. Fisher information distance: for two distributions $P_{\theta_1}, P_{\theta_2}$ we have that,

$$d(P_{\theta_1}, P_{\theta_2}) = (\theta_1 - \theta_2)^T I(\theta_1)^{-1}(\theta_1 - \theta_2).$$

2. Mahalanobis distance: for two distributions $P_{\theta_1}, P_{\theta_2}$, with means μ_1, μ_2 and covariances Σ_1, Σ_2, the Mahalanobis distance would be:

$$d(P_{\theta_1}, P_{\theta_2}) = (\mu_1 - \mu_2)^T \Sigma_1^{-1}(\mu_1 - \mu_2).$$

   This is just the Fisher distance for the Gaussian family with known covariance.

35.7 Robustness to Model Misspecification

Another strong motivation for studying estimation or testing in various metrics stems from robustness considerations. We have already seen that Maximum Likelihood inherits a certain type of KL-robustness, i.e. if we observe samples $X_1, \ldots, X_n \sim P$ where possibly $P \notin \mathcal{P}_\theta$, then MLE still makes sense and asymptotically (under some regularity conditions) will find us a distribution $P_{\widehat{\theta}}$ such that,

$$\mathrm{KL}(P, P_{\widehat{\theta}}) \le \mathrm{KL}(P, P_\theta) \quad \forall \ \theta \in \Theta.$$

One way to interpret this statement is that the MLE is robust to model-misspecification in the KL distance, i.e. if P was close to $\mathcal{P}_\theta$ in the sense that for some $P_\theta \in \mathcal{P}_\theta$, $\mathrm{KL}(P, P_\theta) \leq \epsilon$ then the MLE would automatically find us a distribution which was asymptotically at most ϵ-far in KL from P.

Of course, this is not the only notion of model mis-specification that we might care about, and a general idea is to tailor the estimation procedure to the notion of model mis-specification that we expect.

These lead to so-called minimum distance estimators. These are estimators that attempt to find a distribution in $\mathcal{P}_\theta$ that is close to the samples or the true distribution P in the some distance – say the KL/TV/Hellinger etc. Minimum distance estimators are typically robust to mis-specification in their native distance, i.e. the TV minimum distance estimator will be robust to model-misspecification in TV.

Classically, outlier robustness was studied in something called the Huber ϵ-contamination model, where we observe samples:

$$X_1, \ldots, X_n \sim (1-\epsilon)P_\theta + \epsilon Q,$$

where Q is an arbitrary distribution, i.e. ϵ-fraction of the samples are arbitrarily corrupted (outliers). It is easy to see that,

$$\mathrm{TV}((1-\epsilon)P_\theta + \epsilon Q, P_\theta) \leq \epsilon,$$

so that Huber's model is very closely related to model mis-specification in TV. As a result, the TV minimum distance estimator is very robust to outliers. The main drawback is however a computational one: the TV minimum distance estimator is often difficult to compute.

35.8 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).

35.9 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 35.2 *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.

35.10 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.

35.11 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:

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

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.