High-Dimensional, Two Sample Testing

김지수 (Jisu KIM)

인공지능을 위한 이론과 모델링, 2023 가을학기

The lecture note is a minor modification of the lecture notes from Prof. Larry Wasserman's "Statistical Machine Learning".

1 Introduction

We observe two iid sample

$$X_1, \ldots, X_n \sim P, \qquad Y_1, \ldots, Y_m \sim Q$$

where $X_i, Y_i \in \mathbb{R}^d$. We want to test

$$H_0 : P = Q \quad \text{versus} \quad H_1 : P \neq Q.$$

Throughout, we will assume that $n/(n+m) \to \pi \in (0,1)$ as the sample size increases.

In low dimensions, there are many tests with good power. For example, we could use the test statistic

$$T = \sup_t |\hat{F}_n(t) - \hat{G}_n(t)|$$

where $\hat{F}_n$ and $\hat{G}_n$ are the empirical cdf's. To find the α-level critical value we can use asymptotic theory or permutation testing. But there are other approaches for the high-dimensional case.

Why are we interested in two-sample testing? We might be interested in testing whether two groups are the same for scientific reasons (treatment versus control, for example). Two sample testing can also be used to screen features for classification.

2 Metrics

One way to define a test is to first define a metric between distributions. For example

$$d(P,Q) = \sup_{g \in \mathcal{G}} \Big| \int g dP - \int g dQ \Big|$$

for some class of functions $\mathcal{G}$. Here are some examples. If $\mathcal{G} = \{g : ||g||_\infty \leq 1\}$ then $d(P,Q)$ is the total variation distance. If $\mathcal{G}$ is the set of g such that

$$\sup_{x \neq y} \frac{|g(y) - g(x)|}{||x - y||} \leq 1$$

then $d(P,Q)$ is the earth-mover distance (or Wasserstein distance). This is equivalent to $\inf_R \mathbb{E}_R ||X - Y||$ where the infimum is over all joint distributions R for (X,Y) with marginals P and Q. If $\mathcal{G} = \{I_{(-\infty,t]} : \ t \in \mathbb{R}^d\}$ then $d(P,Q)$ is the Kolmogorov-Smirnov distance. See [5] for more examples.

In general, estimating $d(P,Q)$ is difficult. But if we take $\mathcal{G}$ to be a RKHS defined by a kernel K, it can be shown that

$$\theta = d^2(P,Q) = \int\int K(x,y) dP(x) dP(y) + \int\int K(x,y) dQ(x) dQ(y) - 2 \int\int K(x,y) dP(x) dQ(y).$$

The plus-in estimator of $d^2(P,Q)$ is

$$T = \frac{2}{n(n-1)} \sum_{i<j} K(X_i, X_j) + \frac{2}{m(m-1)} \sum_{i<j} K(Y_i, Y_j) - \frac{2}{nm} \sum_{i,j} K(X_i, Y_j).$$

A related distance is the energy distance (Szekeley 1989, 2002) defined by

$$d^2(P,Q) = 2\mathbb{E}[||X-Y||] - \mathbb{E}[||X-X'||] - \mathbb{E}[||Y-Y'||].$$

The advantage of the energy distance is that there is no tuning parameter. (The RKHS distance actually requires a bandwidth.) The sample estimate is

$$\frac{2}{n_1 n_2}\sum_i\sum_j ||X_i - Y_j|| - \frac{1}{n_1^2}\sum_i\sum_j ||X_i - X_j|| - \frac{1}{n_2^2}\sum_i\sum_j ||Y_i - Y_j||.$$

How do we know when to reject H_0? One approach is to find the limiting distribution of T under H_0. This turns out to be, for the RKHS distance,

$$T \rightsquigarrow 2\sum_{j=1}^{\infty}\lambda_j(Z_j^2 - 1)$$

where the Z_j's are N(0,1) and the λ_j's are the eigenvalues defined by

$$\int L(x,y)\psi_j(x)dP(x) = \lambda_j\psi_j(y)$$

where $L(x,y) = K(x,y) - \mathbb{E}[K(x,X)] - \mathbb{E}[K(X,x)] + \mathbb{E}[K(X,Y)]$. This distribution is called a *Gaussian chaos.* This distribution has infinitely many nuisance parameters which makes it un-usable. Instead, we use the permutation distribution to choose the critical value.

It can be shown that

$$T - d^2(P,Q) = O_P\left(\frac{1}{\sqrt{N}}\right)$$

where $N = n \wedge m$. Thus, it appears that the quality of T does not depend on the dimension! This is false. What matters here is the power. As we shall see below, the minimax power, that is the smallest detectable difference, is

$$\left(\frac{1}{N}\right)^{\frac{2\beta}{4\beta+d}}$$

where β is the smoothness. This was proved by Arias-Castro, Pelletier and Saligrama (2016) based on techniques developed by Ingster (1987). We'll discuss this more below.

The problem is that the kernel is hiding a lot. To see this, note that T is essentially the same as

$$\int (\hat{p}_h(x) - \hat{q}_h(x))^2$$

where $\hat{p}_h$ and $\hat{q}_h$ are kernel density estimators. This test was proposed by Anderson, Hall and Titterington (1994). But remember, the kernel has a tuning parameter. If it is Gaussian, there is a bandwidth. The statement $T - d(P,Q) = O_P(1/\sqrt{N})$ assumes we do not change the bandwidth. But to have good power, we need to let the bandwidth go to zero and we no longer have the fast rate. The power of the RKHS test in general, nonparametric settings is not well studied.

Now suppose we want a confidence interval for $\theta = d^2(P,Q)$. Unfortunately, there is no known practical method if we use the above estimator. However, we can use the idea in [2] to get a simple (but statistically inefficient) method. Instead of using a U-statistic, we break the sample into blocks of size two. For simplicity, assume that $n_1 = n_2 = n$. Define

$$\hat{\theta} = \frac{2}{n}\sum_j h\Big((X_{2j-1}, Y_{2j-1}), (X_{2j}, Y_{2j})\Big) \equiv \frac{1}{m}\sum_j R_j$$

where $m = n/2$ and

$$h((x_i,y_i),(x_j,y_j)) = K(X_i,X_j) + K(Y_i,Y_j) - K(X_i,Y_j) - K(X_j,Y_i).$$

It follows from the CLT and Slutzky's theorem that $\sqrt{m}(\hat{\theta} - \theta)/s \rightsquigarrow N(0,1)$ where s^2 is the sample variance of $R_1, \ldots, R_m$. Hence, an asymptotic $1-\alpha$ confidence interval is $\hat{\theta} \pm sz_{\alpha/2}/\sqrt{m}$.

3 Graph Based Tests

Another class of tests is based on geometric graphs. Let $Z_1, \ldots, Z_N$ be the combined sample where $N = n + m$. Let $L_i = 1$ if Z_i is from group 1 and $L_i = 2$ if Z_i is from group 2.

Let N_i be the k-nearest neighbors of Z_i. Define

$$T = \frac{1}{nk} \sum_{i=1}^{n} \sum_{r=1}^{k} B_j(r)$$

where $B_j(r) = 1$ if the r^{th} nearest neighbor has the same label as Z_i. This corresponds to forming a k nearest neighbor graph and asking how many of the k nearest neighbors are from the same group as the node. The probability of getting the same label under H_0 is $\mu = \pi^2 + (1-\pi)^2$.

It can be shown that, under H_0,

$$\frac{\sqrt{nk}(T-\mu)}{\sigma} \rightsquigarrow N(0,1).$$

The proof is difficult because the test statistic is summing quantities that are not dependent. The variance σ^2 is known but is very, very complicated. See Schilling (1986a, 1986b). In practice, we can use the permutation distribution to get the critical value. Under H_1, the mean of T converges to

$$\theta = 1 - 2\pi(1-\pi) \int \frac{p(x)q(x)}{\pi p(x) + (1-\pi)q(x)} dx$$

which is a distance between p and q. In my experience, this test works well even with $k = 1$.

In high-dimensions we need to correct the test to account for some strange effects [3]. If P concentrates its data on a ring R and Q concentrates its data on a larger ring S that surrounds R, then every point in Q can be closer to a point from P.

Here is an example. Let's take $k = 1$ and $n = m$. Let $B_i = 1$ if its nearest neighbor is from the same group. The test statistic is $T = n^{-1} \sum_i B_i$. We are testing

$$H_0 : P(B_i) = \frac{1}{2} \quad \text{versus} \quad H_1 : P(B_i) > \frac{1}{2}.$$

Suppose that $X_1, X_2 \sim N(\mu_1, \sigma_1^2 I)$ and $Y_1, Y_2 \sim N(\mu_2, \sigma_2^2 I)$. Take $\mu_1 = (a, \ldots, a)$ and $\mu_2 = (b, \ldots, b)$. Now,

$$\frac{1}{d}||X_1 - X_2||^2 \xrightarrow{P} 2\sigma_1^2, \quad \frac{1}{d}||Y_1 - Y_2||^2 \xrightarrow{P} 2\sigma_2^2, \quad \frac{1}{d}||X_1 - Y_2||^2 \xrightarrow{P} \sigma_1^2 + \sigma_2^2 + (a-b)^2.$$

Let $a = 0$, $b = 0.2$, $\sigma_1^2 = 1$, $\sigma_2^2 = 1.2$. Then

$$2\sigma_1^2 < \sigma_1^2 + \sigma_2^2 + (a-b)^2 < 2\sigma_2^2.$$

Every observation from Q is closer to an observation from P.

The data will look like this:

	X_1	X_2	$\ldots$	X_n	Y_1	Y_2	$\ldots$	Y_n
B_i	1	1	$\ldots$	1	0	0	$\ldots$	0

We will not reject H_0 in this case since $(2n)^{-1} \sum_i B_i = 1/2$. The problem is that $P(B_i = 1 | L_i = 1) = 1$ and $P(B_i = 1 | L_i = 2) = 0$ but $P(B_i = 1) = 1/2$. However, if we do a two-sided test, separately within each group, we would reject. [3] suggest taking

$$U = (T_1 - \theta)^2 + (T_2 - \theta)^2$$

where $T_j = (nk)^{-1} \sum_{i:L_i=j} \sum_{Z_j \in N_i} I(L_i = L_j)$. However, this test can have low power in other cases. The best strategy is to use both tests i.e. $W = T \vee U$.

A similar test, called the *cross-match test*, was defined by by [4]. We take the pooled sample and partition the data into pairs $W_1 = (Z_1, Z_2), W_2 = (Z_3, Z_4), \ldots$. The partition is chosen to minimize $\sum_j ||Z_{2j} - Z_{2j-1}||^2$. Let

$$T = \sum_i A_i$$

where $A_i = 1$ if the i^{th} pair has differing labels (i.e. (0,1) or (1,0)) and $A_i = 0$ otherwise. We reject when T is small. The exact distribution of T under H_0 is known; it is hypergeometric. It can accurately be approximated with a $N(\mu, \sigma^2)$ where

$$\mu = \frac{mn}{(N-1)}, \quad \sigma^2 = \frac{2n(n-1)m(m-1)}{(N-3)(N-1)^2}.$$

This accurate, simple limiting distribution for T under the null is the main advantage of this test. However, seems to have less power than the NN test. Also, the distribution of T under H_1 is not known. We could have defined $T = \sum_i B_i$ where $B_i = 1 - A_i$ and and rejected when T is large. This is then the same as the k-NN test with $k = 1$ except that we allow no overlap between groups.

4 Smooth Tests

Neyman (1937) introduced a method for testing that takes advantage of smoothness. First, consider one dimensional data $Y_1, \ldots, Y_n \sim P$. Suppose we want to test

$$H_0 : P = \text{Uniform}(0,1) \quad H_1 : P \neq \text{Uniform}(0,1).$$

If we want to have power against smooth alternatives, Neyman proposed that we define

$$p_\theta(x) = c(\theta) \exp\left(\sum_{j=1}^{k} \theta_j \psi_j(x) \right)$$

where $\psi_1, \psi_2, \ldots,$ are orthonormal functions and

$$c(\theta) = \frac{1}{\int \exp\left(\sum_{j=1}^{k} \theta_j \psi_j(x) \right) dx}.$$

The null hypothesis corresponds to $\theta = (\theta_1, \ldots, \theta_k) = (0, \ldots, 0)$. One way to test H_0 is to use the likelihood ratio test $T = 2(\ell(\hat{\theta}) - \ell(0))$. Under H_0, $T \rightsquigarrow \chi^2_k$. But Neyman pointed out that there is a computationally easier test,

$$U = n \sum_j \overline{\psi}_j^2$$

where

$$\overline{\psi}_j = \frac{1}{n} \sum_i \psi_j(X_i).$$

This also has the property that, under H_0, $U \rightsquigarrow \chi^2_k$. But it avoids having to deal with the normalizing constant.

Now we move to the two-sample case. Let $F(t) = P(X \leq t)$ and $G(t) = Q(Y \leq t)$. Let $Z = F(Y)$. Then the cdf of Z is

$$H(z) = \mathbb{P}(Z \leq z) = \mathbb{P}(F(Y) \leq z) = \mathbb{P}(Y \leq R(z)) = G(R(z))$$

where $R(z) = F^{-1}(z)$. Under H_0, $Z \sim \text{Unif}(0,1)$. Now H has density

$$\rho(z) = \frac{q(F^{-1}(z))}{p(F^{-1}(z))}$$

and $\rho(z) = 1$ under H_0. Bera, Ghosh and Xiao (2013) suggest using the family

$$\rho_\theta(z) = c(\theta) \exp\left(\sum_{j=1}^{k} \theta_j \psi_j(x) \right).$$

Their test statistic is $m\overline{\psi}^T \overline{\psi}$ where

$$\overline{\psi}_j = \frac{1}{m} \sum_i \psi_j(V_i)$$

and $V_i = \hat{F}_n(Y_i)$. Bera, Ghosh and Xiao (2013) prove that the statistic again has a limiting χ^2_k distribution.

Zhou, Zheng and Zhang (arXiv:1509.03459) considered the high-dimensional case. The consider all one-dimensional projections of the data. There test is

$$T = \sqrt{\frac{nm}{n+m}} \sup_u T(u)$$

where the supremum is over the $d-1$-dimensional sphere and $T(u)$ is the Bera-Ghosh-Xiao statistic based on the one-dimensional data $u^T Y_i$. They also allow the parameter k to be chosen from the data. (In fact, they maximize the test over k.)

The limiting distribution of T under H_0 is complicated: it is the supremum of a Gaussian process. Tow get a practical test there are two possibilities. One is to use permutations. The other is based on a version of the bootstrap called *the multiplier bootstrap*. Their simulations suggest that this test works well. But it is unclear how it compares to the other tests.

5 Histogram Test

Under smoothness assumptions and compact support, Ingster (1987) showed that optimal tests can be obtained using histograms. Arias-Castro, Pelletier and Saligrama (2016) extended this to the multivariate case. Assume smoothness level β. For simplicity let $m = n$. Form a histogram with $N \approx n^{2/(4\beta+1)}$ bins. Set

$$T = \sum_j (C_j - D_j)^2$$

where C_j is the number of X_i's in bin j and let D_j is the number of Y_i's in bin j. We reject for T large. This test is, in theory, optimal. In fact, Ingster later showed that the test can be made adaptive to the degree of smoothness.

6 Sparsity

Let us write

$$X_i = (X_i(1), \ldots, X_i(d)), \quad Y_i = (Y_i(1), \ldots, Y_i(d)).$$

In some cases, we might suspect that P and Q only differ in a few features. In other words, there is sparsity. If so, the easiest thing is to do all the one-dimensional marginal tests and a Bonferroni correction. Let T_j be your favorite one dimensional test applied to the jth feature only. The the statistic to be $T = \vee_j T_j$. This test will have good power in the sparse case and it is very easy to compute.

7 Minimax Theory

What does it mean for a test to be optimal? Just as there is a theory for minimax estimation, there is also a theory for minimax testing. We discussed this a few weeks ago. I'll remind you of a few basic facts.

To keep it simple, suppose that $m = n$. We want to test $H_0 : P = Q$. Let $\mathcal{P}$ be a set of distributions and assume that $P, Q \in \mathcal{P}$.

Recall that a level α test is a function ϕ of the data taking values 0 or 1 such that $P(\phi = 1) \leq \alpha$ for every $P \in H_0$. Let Φ_n denote all level α tests. The minimax type II error, for a set of distributions $\mathcal{P}$ is

$$\beta_n(\epsilon) = \inf_{\phi \in \Phi_n} \sup_{P,Q} P^n(\phi = 0)$$

where the supremum is over all $P, Q \in \mathcal{P}$ such that $d(P, Q) > \epsilon$. Fix any small $\delta > 0$. We say that the minimax separation is ϵ_n if $\epsilon < \epsilon_n$ implies that $\beta_n(\epsilon) \geq \delta$.

If $\mathcal{P}$ is the β smoothness class and d is the L_2 distance between densities, then Arias-Castro, Pelletier and Saligrama (2016) show that

$$\epsilon_n \asymp \left(\frac{1}{n}\right)^{\frac{2\beta}{4\beta+d}}.$$

The minimax risk is achieved by the histogram test.

8 Discrete Distributions

Suppose that X_i and Y_i are discrete random variables taking values in $\{1, \ldots, d\}$. Let

$$C_j = \#\{X_i = j\}, \quad D_j = \#\{Y_i = j\}.$$

Let $C = (C_1, \ldots, C_d)$ and $D = (D_1, \ldots, D_d)$. These are multinomial and we can test $H_0 : P = Q$ using a likelihood ratio test or χ^2 test.

But when d is large, the usual tests might have poor power. Improved tests have been developed by [1] and Diakonikolas and Kane (2016). for example. Moreover, these tests are designed to have good power against alternatives with respect to total variation distance. For example, [1] propose the test statistic

$$T = \sum_j \frac{(C_j - D_j)^2 - (C_j + D_j)}{C_j + D_j}.$$

We reject when T is large. The prove that this test has good power as long as $\mathrm{TV}(P, Q) > d^{1/4}/\sqrt{n}$ which is the minimax bound.

References

[1] Siu-On Chan, Ilias Diakonikolas, Gregory Valiant, and Paul Valiant. Optimal algorithms for testing closeness of discrete distributions. In *Proceedings of the Twenty-Fifth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1193–1203. ACM, New York, 2014.

[2] Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander J. Smola. A kernel method for the two-sample problem. *CoRR*, abs/0805.2368, 2008.

[3] Pronoy K. Mondal, Munmun Biswas, and Anil K. Ghosh. On high dimensional two-sample tests based on nearest neighbors. *J. Multivariate Anal.*, 141:168–178, 2015.

[4] Paul R. Rosenbaum. An exact distribution-free test comparing two multivariate distributions based on adjacency. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 67(4):515–530, 2005.

[5] Bharath K. Sriperumbudur, Arthur Gretton, Kenji Fukumizu, Bernhard Schölkopf, and Gert R. G. Lanckriet. Hilbert space embeddings and metrics on probability measures. *J. Mach. Learn. Res.*, 11:1517–1561, 2010.