Active Learning

Aaditya Ramdas

April 20, 2015

1 Normal Means

Suppose $X \sim N(\mu, \sigma^2) \in \mathbb{R}$. We want to test if $\mu = 0$ or $\mu > 0$, and say we know σ. Then noting that $Y = X/\sigma \sim N(0, 1)$ under the null, we suggest the test $I(X/\sigma > z_\alpha)$ which has error rate

$$P_1(X/\sigma \leq z_\alpha) = \Phi(z_\alpha - \mu/\sigma)$$

Hence everything depends on SNR μ/σ. Getting an observation from $X \sim N(\mu, \sigma)$ is exactly like getting an observation $Y \sim N(\mu/\sigma, 1)$. Similarly, suppose we get M measurements, calculating $\bar{X} \sim N(\mu, \sigma^2/M)$ just scales the variance down by M, and so is like getting one observation $Y \sim N(\sqrt{M}\mu/\sigma, 1)$.

So we can just study the problem of testing whether $\mu = 0$ for an observation $Y \sim N(\mu, 1)$ and translate results into other settings accordingly. For example, making one noisy observation with twice the variance corresponds to $Y \sim N(\mu, 2)$ is like observing $Y' \sim (\mu/\sqrt{2}, 1)$.

1.1 Problem Statement

We observe $X \sim N(\mu, I_n) \in \mathbb{R}^n$ i.e. $X_i \sim N(\mu_i, 1)$. We want to know whether $\mu = 0$ or not, and sometimes we even want to identify the nonzero locations of μ (eg: detecting stars in a noisy space snapshot with a cosmic haze corresponding to "Gaussian" white noise).

Remember this could be due to observing multiple measurements and averaging, and just rescaling the same question.

This is an extremely well studied problem, also because of connections to regression — more in the advanced statistics course.

1.2 Sparse Normal Means

Here we assume that μ is a sparse vector. Specifically, the number of nonzeros is like $n^{1-\beta}$ for $\beta \in (0,1)$. Then, if the nonzero components are of the order $\sqrt{2r\log n}$, the problem is impossible if $r < \beta$ and a procedure exists that drives FDP (false discovery proportion) and NDP (non-discovery proportion) to zero if $r > \beta$.

The most trivial test is when $\beta = 1$. Think of a 1-sparse Gaussian vector. With high probability, the maximum of n zero-centered Gaussians is $\sqrt{2\log n}$. If $\mu_i >> \sqrt{2\log n}$ for some i, then we can detect this, but if it is smaller then it will just get lost in the noise. So a decent test is just $I(\max_i X_i > \sqrt{2\log n})$. This is called the max-test.

One way to understand what it is doing is to convert each observation X_i into a p-value $p_i = P(N(0,1) > X_i)$. Then we have n different p-values, and we need to figure out if there is any signal or if all we are just seeing noise (they are uniformly distributed). If want to control type-1 error at α, we could look at the *smallest* p-value and reject it is smaller than α. What's wrong with this? Multiple hypothesis testing - we need to correct for the fact that we are testing many hypotheses simultaneously. To control the possibility of false discovery at α, we need to threshold the minimum p-value at α/n, because even if all p-values were uniform, then $P(\min_i p_i < \alpha/n) = 1 - \prod_i P(p_i > \alpha/n) = 1 - (1-\alpha/n)^n \approx 1 - e^{-\alpha} \approx \alpha$. This is called the **Bonferroni** correction, and is accurate only when there is a constant number of non-null hypotheses, and it is exactly like doing the max-test directly on the observations and it controls the "family wise error rate".

What if there are lots of non-null hypotheses? It is hard to identify which are true non-nulls, since some nulls get mixed up in them. People instead often like to control the False discovery rate (FDR) at level α. $FDR = \mathbb{E}FDP$, where FDP is the ratio of false discoveries to total discoveries.

Then When $\beta \in (0.75, 1)$, the **Benjamini-Hochberg** FDR procedure is provably optimal. This doesn't look just at the minimum p-value, but instead sorts the p-values and looks at whether $p_{(1)} \leq \alpha/n$ or $p_{(2)} \leq 2\alpha/n$ or ... — it finds the highest index k for which this is true, and rejects all p values smaller than $p_{(k)}$. One can show that this controls the false discovery rate at α. Additionally (I think), if every non-null was larger than $\sqrt{2\beta\log n}$ then I think you can show that it also finds exactly all the correct non-nulls (asymptotically).

When $\beta \in (0.5, 1)$, a test called **Higher Criticism** is optimal. It calculates

$$HC_\alpha = \sqrt{n}\frac{(\text{fraction significant at } \alpha) - \alpha}{\sqrt{\alpha(1-\alpha)}}$$

For example, if you had 250 independent tests, and you found 11 significant tests at level 0.05 without any corrections. Is this surprising? Not really, because we'd expect 12.5 tests to be signficant purely by chance (0.05 x 250).

Donoho and Jin's higher criticism doesn't stop there (idea from Tukey). It looks at the entire range of p-values (the whole distribution of the p-values) and finds if there is something odd

about this distribution at *any* level α.

$$HC = \sup_{\alpha} HC_{\alpha}$$

The test is rejected if HC is sufficiently large (found by analysis of HC under the null distribution). Matching lower bounds were also derived.

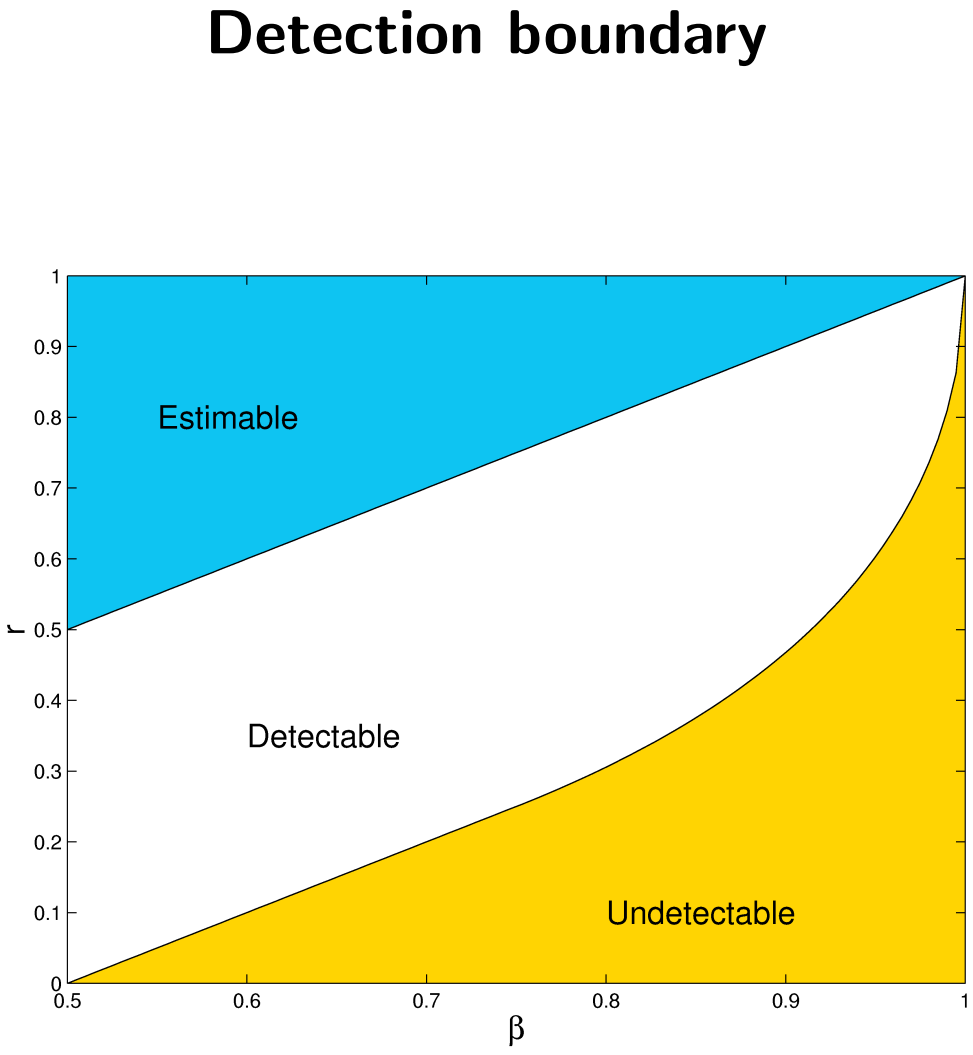

Figure 1: Donoho's graphical description of higher criticism's performance (upper and matching lower bounds) as a function of β, r.

1.3 Distilled Sensing

This is finally where active learning enters the picture. The idea is that we don't have to measure each coordinate at the original SNR. We can take a noisy snapshot, then zoom in on (say) half the coordinates, take another noisy snapshot, zoom in again on half of those coordinates, and so on.

The total "energy" used in this procedure must still be the same for everything to be fair, where energy is again defined in terms of number of measurements and the SNR of each measurement and so on. Haupt, Castro and Nowak showed in their AISTATS paper that this adaptive procedure can detect signals that no passive procedure can.

The model is as follows - the measurements happen in k rounds, and in each round $j = 1, ..., k$ we measure for all $i = 1, ..., n$

$$X_i^{(j)} = N\left(\sqrt{\phi_i^{(j)}}\mu_i, 1\right)$$

The procedure must satisfy $\sum_{i,j} \leq n$. The passive procedure uses $k = 1$ and $\phi_i^{(1)} = 1$ for all i. The active procedure splits up its energy more intelligently.

2 Active Classification under TNC

When the regression function is denoted $\eta(x) = P(Y = 1|X = x)$ for $\mathcal{X} = [0, 1]$ (say), let us assume that the joint distribution $P(X, Y)$ is such that the Bayes classifier $I(\eta(x) > 1/2)$ is coincidentally a threshold classifier at $t \in (0, 1)$.

Given n samples and labels according iid to $P(X, Y)$, how quickly can we identify the threshold? Answer: if

$$|\eta(x) - 1/2| \geq \lambda|x - t|^{k-1}$$

then, the regression function is said to satisfy $TNC(k)$ (sometimes used as two-sided conditions), and in such a case the passive minimax rate is $n^{-1/(2k-1)}$. The easiest case is when $k = 1$, and the rate is $1/n$.

It turns out that if you allow active sampling, then the optimal rate is $n^{-1/(2k-2)}$. When $k = 1$, the rate is exponentially fast. However, as $k \to \infty$, both active and passive will not be able to identify the threshold.

The rates for excess risk with respect to uniform distribution on x, is given by $n^{-k/(2k-1)}$ for passive and $n^{-k/(2k-2)}$ for active. Here as $k \to \infty$, both methods are fine.

3 Active Unsupervised Learning

Instead of classification, one can also consider performing other unsupervised learning tasks like active matrix completion. Sometimes, the gains are small, only log factors. But remember that even a constant factor gain of 5 means that we can work with 20 percent of the data, often shaving off computational load.

4 Announcement!

I am organizing (with Aarti Singh, Nina Balcan and Akshay Krishnamurthy) a workshop on active learning, with the aim of bridging theory and practice, at ICML 2015 in Lille, France. We have some great invited speakers, so if you want to learn more, please do attend (and submit your old/new work if you have any).