CS 229: Lecture 7 Notes

Scribe: Hirsh Jain
Lecturer: Angela Fan

Overview

Overview of today's lecture:

- Hypothesis Testing
- Total Variation Distance
- Pinsker's Inequality
- Application of Pinsker's Inequality to Coin Tossing

Hypothesis Testing

Hypothesis Testing broadly fits into the framework of inference, where we have a hypothesis that we set as a null hypothesis and an alternative hypothesis. For example, for coin tosses, our null hypothesis may be that we have a fair coin which is $\mathrm{Ber}(0.5)$ and our alternative may be that we have a biased coin which is $\mathrm{Ber}(p)$ for some $p \neq 0.5$. Given samples from an unknown distribution, we would like to figure out which of our hypothesis is correct.

Moreover, there are two types of error:

- **Type 1**: False positive error. This refers to the rejection of a true null hypothesis.
- **Type 2**: False negative error. This refers to the failure to reject a false null hypothesis.

We want to find the true error, or the sum of the two errors.

Total Variational Distance

We frequently want to find the distance between two distributions. There are many ways to do this, and one of them is called **total variational distance**, or TVD. This is defined on two distributions A, B as:

$$\mathrm{TVD}(A, B) = \sup_{S \subset \Omega} |A(S) - B(S)|$$

where $A(S)$ is the probability that A assigns to subset S, and $B(S)$ is the same. Intuitively, this refers to the largest difference in probability that distributions A and B will assign to the same event. We can rewrite as

$$\sup_{S \in \Omega} \left\{ \sum_{x \in S} A(x) - \sum_{x \in S} B(x) \right\}$$

Suppose we define $S_{\max} = \{x | A(x) \geq B(x)\}$ Then, assuming finite distributions, it is clear that TVD can also be defined as

$$\sum_{x \in S_{\max}} A(x) - B(x) = - \sum_{x \notin S_{max}} A(x) - B(x)$$

where the equality holds because $\sum_x A(x) = \sum_x B(x) = 1$. Given the equality of the two above, we can take the average and it will also be equal, so we have:

$$\begin{aligned} \mathrm{TVD}(A, B) &= \sum_{x \in S_{\max}} A(x) - B(x) = - \sum_{x \notin S_{max}} A(x) - B(x) \\ &= \frac{1}{2} \sum_{x \in S_{\max}} |A(x) - B(x)| + \sum_{x \notin S_{max}} |A(x) - B(x)| \\ &= \frac{1}{2} \sum_{x} |A(x) - B(x)| \end{aligned}$$

where the signs and the placement of the absolute values follows from the definitions of $S_{\max}$. Note, however, that this is precisely the 1-norm, so we have:

$$\mathrm{TVD}(A, B) = \frac{1}{2} ||A - B||_1$$

Now, it's important to ask a basic question: why would we use this? There are some pros and some cons.

1. Pro: Symmetric. Unlike KL-divergence, we know that $\mathrm{TVD}(A, B) = \mathrm{TVD}(B, A)$.

2. Con: $TVD(A^2, B^2)$ can be equal to $TVD(A, B)$, where the former is defined as the distribution on two samples from each distribution. Example: $A = Ber(p)$ and $B = Ber(p)$ – $TVD(A, B) = TVD(A^2, B^2)$. This is bad because we would like to believe that two samples will give us more information about the true nature of the underlying distribution, but it does not always. [1]

[1] This should not be interpreted to mean that for any A, B, this is true. It isn't. This is simply a statement that such distributions A, B exist.

3. Pro: $1-TVD(A,B)$ is equal to the sum of the false positive error and the false negative error.

4. Pro: $\lim_{n\to\infty} TVD(A^n, B^n) = 1$. This is counterintuitive, given 2), but tells us that we will get (exponentially) more information about the true distributions with more sampling, even if having 2 samples vs 1 sample does not tell us anything new.

We will prove 4. from the list above. The folloiwing claim reflects the fact that total variation distance goes to 1 exponentially quickly.

Claim. Suppose we have X, Y such that $\text{TVD}(X,Y) = \delta$. We want to prove that for all $k \in \mathbb{N}$:

$$1 - 2e^{-k\delta^2/2} \leq \text{TVD}(X^k, Y^k)$$

Proof. By the definition of TVD , there exists a subset S such that given samples $x \sim X, y \sim Y$, we have: $P(x \in S) - P(y \in S) = \delta$. We also define $P(y \in S) = p \iff P(x \in S) = p + \delta$.

Given k samples of X, we know that the probability of any sample being in S is $p+\delta$. Thus, in expectation, $(p+\delta)k$ of the samples will be in S. Similarly, given k samples of Y, pk will be in S in expectation.

Now, we can apply the Chernoff bound to see that:

$$P(\text{at least } \left(p + \frac{\delta}{2}\right) k \text{ components of Y are in S}) < e^{\frac{-k\delta^2}{2}}$$
$$P(\text{at most } \left(p + \frac{\delta}{2}\right) k \text{ components of X are in S}) < e^{\frac{-k\delta^2}{2}}$$

Let S' be the set of k-tuples that contain more than $\left(p + \frac{\delta}{2}\right) k$ components of S. Then, we can bound:

$$\text{TVD}(X^k, Y^k) \geq P(X^k \in S') - P(Y^k \in S') > (1 - e^{\frac{-k\delta^2}{2}}) - e^{\frac{-k\delta^2}{2}} = 1 - 2e^{\frac{-k\delta^2}{2}}$$

which gives us the desired result. □

Pinsker's Inequality

Pinsker's Inequality states that:

$$D(P||Q) \geq \frac{1}{2\ln 2} ||P - Q||_1^2$$

where $D(P||Q)$ is the KL Divergence of P and Q, defined as:

$$D(P||Q) = \sum_x p(x) \log \left(\frac{p(x)}{q(x)} \right)$$

We can rewrite Pinsker's Inequality:

$$D(P||Q) \geq \frac{1}{2\ln 2} ||P - Q||_1^2 \iff \text{TVD}(P, Q) \leq \frac{1}{2}\sqrt{2\ln(2) \cdot D(P||Q)}$$

Proof: We're going to start with the case where P and Q are Bernoulli. This can then be used to prove any other case. Define P and Q as:

$$P = \begin{cases} 1 & \text{w.p. } p \\ 0 & \text{w.p. } 1-p \end{cases} \qquad Q = \begin{cases} 1 & \text{w.p. } q \\ 0 & \text{w.p. } 1-q \end{cases}$$

We assume without loss of generality that $p \geq q$. We can write out the KL-divergence and TVD explicitly:

$$D(p||q) = p \log \frac{p}{q} + (1-p) \log \left(\frac{1-p}{1-q} \right)$$
$$\text{TVD}(p, q) = ||(p, 1-p) - (q, 1-q)||_1 = 2(p-q)$$

We can define

$$f(p, q) = p \log \frac{p}{q} + (1-p) \log \left(\frac{1-p}{1-q} \right) - \frac{(2(p-q))^2}{2\ln 2}$$

We can take the derivative of this function:

$$\frac{\delta f(p, q)}{\delta q} = -\frac{p-q}{\ln 2} \left(\frac{1}{q(1-q)} - 4 \right)$$

Note that $p - q$ is always positive and $\ln 2$ is positive. Moreover, $q \cdot (1-q) \leq \frac{1}{4}$ for all q. Thus, the term inside the parentheses is positive, and the negative in front of the expression makes the derivative ≤ 0, and equal to 0 when $p = q$. From this, we can conclude that for $p > q$, the function must be positive, as it is zero at $p = q$ and decreasing. Thus:

$$\begin{aligned} f(p, q) = p \log \frac{p}{q} + (1-p) \log \left(\frac{1-p}{1-q} \right) - \frac{1}{2\ln 2}(2(p-q))^2 \geq 0 \\ \iff p \log \frac{p}{q} + (1-p) \log \left(\frac{1-p}{1-q} \right) \geq \frac{1}{2\ln(2)}(2(p-q))^2 \\ \iff D(P||Q) = \frac{1}{2\ln(2)} ||P - Q||_1^2 \end{aligned}$$

as desired. Moreover, we can prove the non-binary case via a reduction to the binary case.

Applications of Pinsker's Inequality to Coin Tossing

Note that $D(P^m||Q^m) = mD(P||Q)$. We can prove this using the chain rule of KL Divergence:

$$\begin{aligned}
D(P(X,Y)||Q(X,Y)) &= \sum_{x,y} p(x,y)\log\frac{p(x,y)}{q(x,y)} \\
&= \sum_{x,y} p(x)p(y|x)\log\frac{p(x)p(y|x)}{q(x)q(y|x)} \\
&= \sum_{x} p(x)\log\frac{p(x)}{q(x)}\sum_{y} p(y|x) + \sum_{x} p(x)\sum_{y} p(y|x)\log\frac{p(y|x)}{q(y|x)} \\
&= D(P_x||Q_x) + \sum_{x} p(x)D(p_y||q_y\ |X=x) \\
&= D(P_x||Q_x) + D(P_y||Q_y\ |X) \\
&= D(P_x||Q_x) + D(P_y||Q_y) \\
&= 2\cdot D(P||Q)
\end{aligned}$$

where $D(P_y||Q_y\ |X) = D(P_y||Q_y)$ follows from X being independent of Y. Now, we can iteratively apply this procedure to determine that $D(P^m||Q^m) = mD(P||Q)$ as desired.

Consider the following set-up for a coin tossing problem. Let 1 = Heads, 0 = Tails, and define P and Q as:

$$P = \begin{cases} 1 & \text{w.p. } \frac{1}{2}-\epsilon \\ 0 & \text{w.p. } \frac{1}{2}+\epsilon \end{cases} \qquad Q = \begin{cases} 1 & \text{w.p. } \frac{1}{2} \\ 0 & \text{w.p. } \frac{1}{2} \end{cases}$$

Let $x = (x_1, x_2, \ldots, x_m)$ be a number of coin flips. We can map x to either P or Q as a prediction on which distribution they came from. encoding P as 0 and Q as 1. This gives us a function A:

$$A : (x_1, \ldots, x_m) \to \{0, 1\}$$

We want to find the value of m (i.e. the number of samples) we need to ensure that A predicts between P and Q with > 90% probability. Explicitly, we want to find m such that:

$$P_{x\in P^m}[A(x) = 0] \geq \frac{9}{10} \text{ and } P_{x\in Q^m}[A(x) = 1] \geq \frac{9}{10}$$

Equivalently, when taking the expectation over all m-tuples in P^m and Q^m, we want:

$$E_{x\in P^m}[A(x)] \leq \frac{1}{10} \text{ and } E_{x\in Q^m}[A(x)] \geq \frac{9}{10} \implies E_{x\in Q^m}[A(X)] - E_{x\in P^m}[A(X)] \geq \frac{8}{10}$$

Lemma. $\widetilde{P}, \widetilde{Q}$ discrete on U, then given $f : U \to [0, B]$,

$$|E_{\widetilde{P}}[f(x)] - E_{\widetilde{Q}}[f(x)]| \leq \frac{B}{2} ||\widetilde{P} - \widetilde{Q}||_1$$

Proof. We can rewrite the left using expected value, and the law of the unconscious statistician (LOTUS) [2]:

$$\begin{aligned}
|E_{\widetilde{P}}[f(x)] - E_{\widetilde{Q}}[f(x)]| &= |\sum_x \widetilde{p}(x) f(x) - \sum_x \widetilde{q}(x) f(x)| \\
&= |\sum_x f(x)(\widetilde{p}(x) - \widetilde{q}(x))| \\
&= \left|\sum_x (\widetilde{p}(x) - \widetilde{q}(x)) \left(f(x) - \frac{B}{2}\right) + \frac{B}{2}\left(\sum_x \widetilde{p}(x) - \widetilde{q}(x)\right)\right| \\
&\leq \sum_x |\widetilde{p}(x) - \widetilde{q}(x)| \left|f(x) - \frac{B}{2}\right| \\
&\leq \frac{B}{2} ||\widetilde{P} - \widetilde{Q}||_1
\end{aligned}$$

Now, we can use this lemma: let $\widetilde{P} = P^m, \widetilde{Q} = Q^m, f = A$, so we have

$$||P^m - Q^m||_1 \geq 2||E_{x \in Q^m} A(X) - E_{x \in P^m} A(x)| \implies ||P^m - Q^m||_1 \geq 2 \cdot \frac{8}{10} = \frac{8}{5}$$

Now, using Pinsker's Lemma, we have that:

$$m \cdot D(P||Q) = D(P^m||Q^m) \geq \frac{1}{2\ln(2)} \cdot \left(\frac{8}{5}\right)^2 \implies m \geq \frac{1}{2\ln(2) \cdot D(P||Q)} \cdot \left(\frac{8}{5}\right)^2$$

Thus, it remains to bound $D(P||Q)$:

$$\begin{aligned}
D(P||Q) &= \left(\frac{1}{2} - \epsilon\right) \log\left(\frac{\frac{1}{2} - \epsilon}{\frac{1}{2}}\right) + \left(\frac{1}{2} + \epsilon\right) \log\left(\frac{\frac{1}{2} + \epsilon}{\frac{1}{2}}\right) \\
&= \frac{1}{2} \log\left((1 - 2\epsilon)(1 + 2\epsilon) + \epsilon \log\left(\frac{1 + 2\epsilon}{1 - 2\epsilon}\right)\right) \\
&\leq \frac{\epsilon}{\ln 2} \ln\left(1 + \frac{4\epsilon}{1 - 2\epsilon}\right) \\
&\leq \frac{4\epsilon^2}{\ln 2} \cdot \frac{1}{1 - 2\epsilon}
\end{aligned}$$

[2] https://en.wikipedia.org/wiki/Law_of_the_unconscious_statistician

where the last inequality uses the fact that $\ln(1+x) \leq e^x$. Now, if we assume that $\epsilon < \frac{1}{4}$, we can write:

$$D(P||Q) \leq \frac{8\epsilon^2}{\ln 2}$$

Finally, combining this with the above inequality, we have a bound on m:

$$m \geq \frac{1}{2\ln(2) \cdot D(P||Q)} \cdot \left(\frac{8}{5}\right)^2 \geq \frac{4}{25\epsilon^2}$$

which can be shown to be upto constants by the Chernoff bound. $\square$