ECE 6980

An Algorithmic and Information-Theoretic Toolbox for Massive Data

Instructor: Jayadev Acharya
Lecture #1

Scribe: Michael Luo, JA.
26th August, 2016

Please send errors to acharya@cornell.edu

1 Introduction

Definition 1. *A discrete distribution P over $\mathcal{X}$ is a function from a countable set $\mathcal{X}$ to $\mathbb{R}_+$, such that $\sum_{x\in\mathcal{X}} P(x) = 1$.*

We will study the problem of distribution estimation.

Problem. Given independent samples $X_1, \ldots, X_n$ from an unknown distribution P, output a distribution $\hat{P}$ such that $L(P, \hat{P})$ is *small.* Here $L(\cdot, \cdot)$ is a loss function, or distance measures between distributions.

In this lecture, we discuss some common statistical distance measures. We will restrict ourselves to discrete random variables over $\mathcal{X}$.

1.1 Total Variation/ℓ_1 distance

For a subset $A \subseteq \mathcal{X}$, let $P(A) = \sum_{x\in A} P(x)$ be the probability of observing an element in A.

Definition 2. *The* total variation distance *between P, and Q is*

$$d_{TV}(P, Q) = \sup_{A\subseteq\mathcal{X}} |P(A) - Q(A)|.$$

The TV distance is related to the ℓ_1 distance as follows:

Claim 3.

$$2 \cdot d_{TV}(P, Q) = |P - Q|_1 \overset{\text{def}}{=} \sum_{x\in\mathcal{X}} |P(x) - Q(x)|$$

Proof. Let A^* be the set such that $d_{TV}(P, Q) = P(A^*) - Q(A^*)$. Suppose there is an $x \in \mathcal{X} \setminus A$ such that $P(x) > Q(x)$, then $P(A^* \cup \{x\}) - Q(A^* \cup \{x\}) > P(A^*) - Q(A^*)$, a contradiction. Similarly, there is no $x \in A^*$ with $Q(x) > P(x)$. Therefore, $A^* = \{x \in \mathcal{X} : P(x) > Q(x)\}$. Hence,

$$\begin{align} |P - Q|_1 &\overset{\text{def}}{=} \sum_{x\in\mathcal{X}} |P(x) - Q(x)| \tag{1}\\ &= \sum_{x\in A^*} P(x) - Q(x) + \sum_{x\notin A^*} Q(x) - P(x) \tag{2}\\ &= P(A^*) - Q(A^*) + (1 - Q(A^*)) - (1 - P(A^*) \tag{3}\\ &= 2d_{TV}(P, Q). \tag{4}\end{align}$$

□

Interpretation as classification error

Consider the following problem. P, Q are two distributions we know. Nature chooses a distribution D, which is either P, or Q with equal probability, namely

$$D = \begin{cases} P & \text{w.p. } 0.5 \\ Q & \text{w.p. } 0.5. \end{cases}$$

We get to see a sample from D, and then make a (possibly randomized) prediction whether D is equal to P or Q.

Let $C(P|x)$ be the probability that we predict P upon observing x. Therefore, $C(Q|x) = 1 - C(P|x)$. Hence, by the Bayes rule

$$\begin{aligned} \Pr(error) &= \sum_{x \in \mathcal{X}} \left[\Pr(D = P) \cdot P(x)C(Q|x) + \Pr(D = Q) \cdot Q(x)C(P|x)\right] && (5) \\ &= \sum_{x \in \mathcal{X}} \left[\frac{1}{2} \cdot P(x)C(Q|x) + \frac{1}{2} \cdot Q(x) \cdot (1 - C(Q|x))\right] && (6) \\ &= \frac{1}{2} + \frac{1}{2} \sum_{x \in \mathcal{X}} C(Q|x)(P(x) - Q(x)). && (7) \end{aligned}$$

Therefore, to minimize the probability of error, we should make $C(Q|x) = 1$ for all elements such that $P(x) > Q(x)$. Therefore, we are left with a summation that equals the total variation distance. Let e^* be the least probability of error. Then

$$e^* = \frac{1}{2} - \frac{1}{2} d_{TV}(P, Q) = \frac{1}{2} - \frac{1}{4} |P - Q|_1. \qquad (8)$$

Let $Bern(p)$ be the Bernoulli random variable that takes value 1 with probability p, and value 0 with probability $1 - p$.

Exercise Show that the mean and variance of $Bern(p)$ is p and $p(1 - p)$ respectively.

Exercise See that (8) makes sense by taking

- $P = Bern(0)$, and $Q = Bern(1)$.
- $P = Bern(0)$, and $Q = Bern(0)$.

We now move on to another distance measure.

1.2 KL Divergence.

Definition 4. *The* KL divergence *between P, and Q is*

$$D(P||Q) \stackrel{\text{def}}{=} \sum_{x \in \mathcal{X}} P(x) \log \frac{P(x)}{Q(x)}.$$

To be a distance, we first should show that this is non-negative. This follows easily using Jensen's inequality and concavity of logarithms. We will use the following simple inequality.

Using basic calculus, we can show that for $x > -1$, $\log(1 + x) \le x$.

Therefore,

$$\begin{aligned} D(P||Q) &= \sum_{x\in\mathcal{X}} P(x)\log\frac{P(x)}{Q(x)} && (9)\\ &= -\sum_{x\in\mathcal{X}} P(x)\log\left(1+\frac{Q(x)-P(x)}{P(x)}\right) && (10)\\ &\geq -\sum_{x\in\mathcal{X}} P(x)\cdot\frac{Q(x)-P(x)}{P(x)} && (11)\\ &= -\sum_{x\in\mathcal{X}} Q(x)-P(x) = 0 && (12)\end{aligned}$$

Data Compression and KL divergence. We will be covering the basics of information theory in one of the lectures. We will show that a discrete random variable with distribution P can be compressed to about $\sum P(x)\log(1/P(x))$ bits in expectation, *using a scheme designed with the knowledge of* P. However, suppose we thought that the random variable is distributed according to Q, when in fact it was distributed according to P. The extra number of bits we use on average due to this misspecification will be exactly equal to the KL divergence between P and Q.

KL divergence will be helpful in proving some of the lower bounds on the sample complexity of problems we consider owing to the Pinsker's Inequality. We will prove it in the assignment.

Pinsker's Inequality. For discrete distributions P and Q,

$$|P-Q|_1^2 \leq 2\cdot D(P||Q). \tag{13}$$

One of the nice properties of KL divergence is additivity for independent random variables. Suppose $P = P_1\times\ldots\times P_m$ and $Q = Q_1\times\ldots\times Q_m$ are product distributions over $\mathcal{X} = \mathcal{X}_1\times\ldots\times\mathcal{X}_m$. For example, each P_i can be an independent toss from a Bernoulli random variable, and P is the outcome of m independent tosses. Therefore, $\mathcal{X}_i = \{0,1\}$, and $\mathcal{X} = \{0,1\}^m$.

Then, (show this)

$$D(P||Q) = \sum_{i=1}^{m} D(P_i||Q_i).$$

Finally, we discuss about chi-squared distance.

1.3 Chi-squared distance

There are a few different definitions of the chi-squared distance. We will use the following (asymmetric) definition.

Definition 5. *The* chi-squared *distance between* P *and* Q *is*

$$\chi^2(P,Q) = \sum_{x\in\mathcal{X}}\frac{(P(x)-Q(x))^2}{Q(x)}.$$

Expanding the expression, and using $\sum_{x\in\mathcal{X}} P(x) = \sum_{x\in\mathcal{X}} Q(x) = 1$,

$$\chi^2(P,Q) = \left[\sum_{x\in\mathcal{X}}\frac{(P(x)^2}{Q(x)}\right] - 1.$$

Therefore, using $\log x < x - 1$,

$$D(P||Q) = \sum_{x\in\mathcal{X}} P(x) \log \frac{P(x)}{Q(x)} \tag{14}$$

$$\leq \sum_{x\in\mathcal{X}} P(x) \cdot \left(\frac{P(x)}{Q(x)} - 1\right) = \chi^2(P, Q). \tag{15}$$

Relation between the measures

The following relation between distance measures will be useful:

$$\boxed{\frac{1}{2}|P - Q|_1^2 \leq D(P||Q) \leq \chi^2(P, Q)}$$

1.4 Concentration

To analyze the estimators we design, concentration inequalities will be useful.

The first inequality is Markov's Inequality.

Theorem 6 (Markov's Inequality.). *Let X be a non-negative random variable. Then for $\alpha > 0$,*

$$\Pr(X \geq \alpha) \leq \frac{\mathbb{E}[X]}{\alpha}.$$

Proof. Let P be the distribution of X. By definition of expectation,

$$\mathbb{E}[X] = \int_{x\geq 0} x \cdot dP(x) \geq \int_{x\geq\alpha} x \cdot dP(x) \geq \int_{x\geq\alpha} \alpha \cdot dP(x) = \alpha \cdot \Pr(X \geq \alpha).$$

□

Chebyshev's inequality proves concentration of random variables around the mean in terms of the standard deviation.

Theorem 7 (Chebyshev's inequality). *Let X be a random variable with mean μ and variance $\mathbb{E}[(X - \mu)^2] = \sigma^2$. Then,*

$$\Pr(|X - \mu| > \alpha \cdot \sigma) \leq \frac{1}{\alpha^2}.$$

Proof. Let $Y = |X - \mu|$. Apply Markov's inequality on Y^2 and use $\mathbb{E}[Y^2] = \sigma^2$. □

Concentration is very useful when applied to sum, and averages of random variables. Suppose $X_1, \ldots, X_n$ are independent random variables from a distribution with mean μ, and variance σ^2. Let $\bar{X} = (X_1 + \ldots + X_n)/n$ be the sample mean. By Linearity of expectations, we have $\mathbb{E}[\bar{X}] = \mu$. By independence,

$$Var(\bar{X}) = \frac{1}{n^2} \sum_{i=1}^{n} Var(X_i) = \frac{\sigma^2}{n}. \tag{16}$$

Applying the Chebyshev's inequality,

$$\Pr\left(|\bar{X} - \mu| \geq \alpha \cdot \sqrt{\frac{\sigma}{n}}\right) \leq \frac{1}{\alpha^2} \tag{17}$$

1.5 Bernoulli Probability Estimation

We have enough machinery now to study the problem of estimating a Bernoulli distribution.

Problem. Given samples $X_1, \ldots$ from an unknown P, an unknown $Bern(p)$ random variable. The goal is to output a distribution $\hat{P}$ such that with probability at least 3/4, $d_{TV}(P, Q) \leq \varepsilon$.

Suppose we observe n samples $X_1, \ldots, X_n$, and let $\hat{p} = (X_1 + \ldots + X_n)/n$. We output the distribution $\hat{P} = Bern(\hat{p})$. Then,

$$\mathbb{E}[\hat{p}] = p,$$

and

$$\mathbb{E}[\hat{p}] = \frac{p(1-p)}{n}.$$

By Chebyshev's inequality,

$$\Pr\left(|p - \hat{p}| > 2 \cdot \sqrt{\frac{p(1-p)}{n}} \right) \leq \frac{1}{4}$$

Suppose, $n > 4p(1-p)/\varepsilon^2$, then with probability at least 3/4,

$$\Pr\left(|p - \hat{p}| > \varepsilon\right) \leq \frac{1}{4}. \tag{18}$$

Note that the TV distance between $Bern(p)$ and $Bern(q)$ is $|p - q|$.

Theorem 8. *With*

$$n \geq 4\frac{p(1-p)}{\varepsilon^2}$$

samples, we can estimate P to a total variation ε with probability at least 3/4.

In the next lecture, we will see why this might be the best we can do.