ECE 6980
Algorithmic and Information-Theoretic Methods in Data Science

Instructor: Jayadev Acharya — Lecture #0
Scribe: Ziteng Sun — 25th September, 2017

1 Scheffe Estimator

1 [Recap] Given M distribtuions, with $\frac{logM}{\epsilon^2}$ samples from p, we can find a distribution close to p.
2 $\mathcal{P}$ is a collection of distributions (e.g. all distributions over $[k]$). Let N_ϵ be the covering number of $\mathcal{P}$, which is the minimum value of M such that $\exists p_1, p_2, p_3, ..., p_M \in \mathcal{P}$ such that $\forall p \in \mathcal{P}, \exists p_j$ such that $d(p, p_j) \leq \epsilon$ (d is a metric). Then we can learn a distribution from $\mathcal{P}$ with $\frac{logN_\epsilon}{\epsilon^2}$ samples using Scheffe estimators.
3 $logN_\epsilon$ is called the metric entropy of $\mathcal{P}$ with respect to metric d.
4 For distributions ove $[k]$, $N_\epsilon \sim (\frac{c}{\epsilon})^k$, so the complexity of the learner is $\frac{logN_\epsilon}{\epsilon^2} \sim \frac{klog(k/\epsilon)}{\epsilon^2}$.

2 Learning a distribution with exponentially tiny error

As from previous lectures, we can learn a distribution over $[k]$ using $O(\frac{k}{\epsilon^2})$ samples with probability larger than 0.9. Using boosting technique, we can make the failure probability an arbitrary δ by adding an multiplicative factor of $\log\frac{1}{\delta}$. However, with the following analysis, we can show the complexity is $O(\frac{k+\log(1/\delta)}{\epsilon^2})$.

Theorem 1. *Mcdiarmid's Inequality*
Suppose $f : \mathcal{X}^n \to R$ is c- bounded difference, which means $|f(x_1, x_2, ..., x_n) - f(x_1, ..., x_i', ..., x_n)| \leq c, \forall x_1, ..., x_n, x_i' \in \mathcal{X}$. Further suppose $X_1, X_2, ..., X_n$ are independent, then we have:

$$\Pr\left(f(X_1, X_2, ..., X_n) - \mathbb{E}\left[f(X_1, X_2, ..., X_n)\right] > \epsilon\right) \leq \exp(-\frac{2\epsilon^2}{nc^2}) \tag{1}$$

For reference, see http://cs.nyu.edu/~rostami/ml/2007/ashish-mcdiarmid.pdf.

Let $f(x_1, x_2, ..., x_n) = \sum_{x=1}^{k} |\frac{N_x}{n} - \Pr(x)|$. Then $f(x_1, x_2, ..., x_n)$ is $\frac{2}{n}$- bounded difference. So we have:

$$\Pr\left(f(X_1, X_2, ..., X_n) - \mathbb{E}\left[f(X_1, X_2, ..., X_n)\right] > \epsilon\right) \leq \exp(-\frac{2(\epsilon/2)^2}{n(2/n)^2}) = \exp(-\frac{n\epsilon^2}{8}) \tag{2}$$

1 From previous lectures, with $O(\frac{k}{\epsilon^2})$ samples, $\mathbb{E}\left[f(X_1, X_2, ..., X_n)\right] \leq \epsilon/2$.
2 From Mcdiarmid's Inequality, with $O(\frac{\log(1/\delta)}{\epsilon^2})$ samples, we can make the failure probability less than δ.

Hence the total sample complexity is $O(\frac{\log(1/\delta)+k}{\epsilon^2})$. So for $\delta = e^{-k}$, the sample complexity will remain $O(\frac{k}{\epsilon^2})$.

3 Property Estimation

Let $f : \mathcal{P} \to R$ be a property of a distribution, which includes:

1 $\mathbb{E}[X]$
2 Entropy $\sum_x p(x) \log \frac{1}{p(x)}$
3 Mode
4 Support size
5 Distance to uniformity $\sum_x |p(x) - \frac{1}{k}|$
6 Number of heavy hitters
7 $\sum p_x^2$.

Definition 1. Learning the property of a distribution
Let p be a distribution over $[k]$ and $x_1, x_2, x_3, ..., x_n \sim p$ are independent samples. The goal is to find an estimator of the property $\hat{f}$ such that with probability > 0.9, we have $|\hat{f}(x_1, x_2, ..., x_n) - f(p)| < \epsilon$.

Definition 2. Symmetric Property
A property f is symmetric if $f(p_\sigma) = f(p) \forall \sigma \in S_k$ where S_k is all the permutation over $[k]$.

Next we consider the problem of entropy estimation. $H(p) = \sum_x p(x) \log(\frac{1}{p(x)})$ and the preformance of the empirical estimator.

Empirical estimatior of entropy:

$$\hat{H}(p) = H(\hat{p_n}) = \sum_x \frac{N_x}{n} log(\frac{n}{N_x}) \tag{3}$$

It can be shown that $H(p) \geq \mathbb{E}[H(\hat{p_n})]$. Now we consider the expectation of the bias and the variance of $\hat{H}$.

$$\begin{aligned}
H(p) - E[H(\hat{p_n})] &= \sum_x [p(x) \log \frac{1}{p(x)} - \mathbb{E}[\hat{p_n}(x)] \log \frac{1}{\hat{p_n}(x)}] && (4)\\
&= \sum_x \mathbb{E}\left[p(x) \log \frac{1}{p(x)} - \hat{p_n}(x) \log \frac{1}{\hat{p_n}(x)}\right] && (5)\\
&= \sum_x \mathbb{E}\left[(p(x) - \hat{p_n}(x)) \log \frac{1}{p(x)}\right] + \sum_x \mathbb{E}\left[\hat{p_n}(x) \log \frac{\hat{p_n}(x)}{p(x)}\right] && (6)\\
&= \sum_x \mathbb{E}\left[\hat{p_n}(x) \log \frac{\hat{p_n}(x)}{p(x)}\right] && (7)\\
&= \sum_x \mathbb{E}[D_{KL}(\hat{p_n}(x), p(x))] && (8)\\
&\leq \mathbb{E}\left[\sum_x \frac{(\hat{p_n}(x) - p(x))^2}{p(x)}\right] && (9)\\
&= \sum_x \frac{\mathbb{E}[(\hat{p_n}(x) - p(x))^2]}{p(x)} && (10)\\
&= \sum_x \frac{p(x)(1 - p(x))}{np(x)} && (11)\\
&= \frac{k-1}{n} && (12)
\end{aligned}$$

We can also show that $Var(\hat{p_n}(x)) < \frac{\log^2 n}{n}$.

Definition 3. Bias-Variance Decomposition of an Estimator
Suppose $\hat{z}$ is an estimator for a random variable z, then we have:

$$\mathbb{E}\left[(z-\hat{z})^2\right] = (z-\mathbb{E}\left[\hat{z}\right])^2 + \mathbb{E}\left[(\hat{z}-\mathbb{E}\left[\hat{z}\right])^2\right] \tag{13}$$

Hence if we want to estimate the entropy with high probability, we need $H(p) - E[H(\hat{p_n})] = O(\epsilon)$ and $Var[H(\hat{p_n})] = O(\epsilon^2)$. Hence we need $O(\frac{k}{\epsilon} + \frac{\log^2 k}{\epsilon^2})$ samples