ECE 6980
An Algorithmic and Information-Theoretic Toolbox for Massive Data

Instructor: Jayadev Acharya
Scribe: Xiao Xu
Lecture #4
6th September, 2016

Please send errors to xx243@cornell.edu and acharya@cornell.edu

We did a brief recap of the previous lecture. We then outline the three things we will discuss today:

- Basics of information theory
- Proof of Fano's Inequality
- A "simple" algorithm to learn "many" classes "almost" optimally

1 Basic Information Theory

1.1 Entropy

Definition 1. *The entropy of a discrete distribution P over $\mathcal{X}$ is defined as*

$$H(P) = \sum_{x \in \mathcal{X}} P(x) \log \left(\frac{1}{P(x)} \right) \tag{1}$$

Claim 2. *Let P be a discrete distribution over $\mathcal{X}$, then*

$$H(P) \leq \log |\mathcal{X}| \tag{2}$$

Proof. We use Jensen's inequality and the concavity of $\log(x)$ to prove the claim.

$$H(P) = \sum_{x \in \mathcal{X}} P(x) \log \left(\frac{1}{P(x)} \right) \leq \log \left(\sum_{x \in \mathcal{X}} P(x) \frac{1}{P(x)} \right) = \log |\mathcal{X}| \tag{3}$$

□

To understand entropy, we consider an example of distinguishing a number in a set. Suppose $\mathcal{X} = \{0, 1, 2, ..., 127\}$ and x is randomly chosen from $\mathcal{X}$ with equal probability. We would like to identify x by asking several Yes/No questions. The problem is what is the smallest number of questions we need to ask to find the exact value of x. The answer is $7 = \log(128)$ and we will use a binary search method to do this: firstly, we ask if $x \leq 64$, if yes, we ask the second question if $x \leq 32$, or otherwise, ask if $x \leq 96$ and keep doing this until we successfully identify the exact value of x. Actually, entropy H characterizes the shortest length we need to distinguish a random variable.

1.2 Joint Entropy

Definition 3. *We consider a joint discrete distribution P over $\mathcal{X} \times \mathcal{Y}$, then the joint entropy is defined as*

$$H(P) = \sum_{x,y} P(x,y) \log \left(\frac{1}{P(x,y)} \right) \tag{4}$$

Definition 4. *Suppose P is a joint distribution over $\mathcal{X} \times \mathcal{Y}$, the marginal distribution of P is defined as*

$$P_{\mathcal{X}}(x) = \sum_{y} P(x,y) \tag{5}$$

$$P_{\mathcal{Y}}(y) = \sum_{x} P(x,y) \tag{6}$$

Definition 5. *Suppose P is a joint distribution over $\mathcal{X} \times \mathcal{Y}$, we say P is a product distribution if*

$$P(x,y) = P_{\mathcal{X}}(x) \cdot P_{\mathcal{Y}}(y) \tag{7}$$

We consider the following example. Table 1 gives us some statistics of the weather in San Diego. Suppose $\mathcal{X} = \{\text{Sunny}, \text{Not Sunny}\}$, $\mathcal{Y} = \{\text{Hot}, \text{Cold}\}$.

	Hot	Cold
Sunny	30	125
Not Sunny	20	190

Table 1: Number of days of different weather

The question is, is the probability distribution of different kind of weather a product distribution? The answer is no since given $Y = \text{Hot}$ or Cold, the probability

$$\Pr(X = \text{Sunny} | Y = \text{Hot}) = \frac{3}{5} \neq \frac{25}{63} = \Pr(X = \text{Sunny} | Y = \text{Cold})$$

In fact, we can change the number in the table appropriately to make it a product distribution.

Claim 6. *If $P : \mathcal{X} \times \mathcal{Y}$ is a product distribution, then we have*

$$H(P) = H(P_{\mathcal{X}}) + H(P_{\mathcal{Y}}) \tag{8}$$

Proof.

$$\begin{aligned}
H(P) &= \sum_{x,y} P(x,y) \log\left(\frac{1}{P(x,y)}\right) \\
&= \sum_{x,y} P_{\mathcal{X}}(x) P_{\mathcal{Y}}(y) \log\left(\frac{1}{P_{\mathcal{X}}(x)} \frac{1}{P_{\mathcal{Y}}(y)}\right) \\
&= \sum_{x,y} P_{\mathcal{X}}(x) P_{\mathcal{Y}}(y) \log\left(\frac{1}{P_{\mathcal{X}}(x)}\right) + \sum_{x,y} P_{\mathcal{X}}(x) P_{\mathcal{Y}}(y) \log\left(\frac{1}{P_{\mathcal{Y}}(y)}\right) \\
&= \sum_{x} P_{\mathcal{X}}(x) \log\left(\frac{1}{P_{\mathcal{X}}(x)}\right) + \sum_{y} P_{\mathcal{Y}}(y) \log\left(\frac{1}{P_{\mathcal{Y}}(y)}\right) \\
&= H(P_{\mathcal{X}}) + H(P_{\mathcal{Y}})
\end{aligned} \tag{9}$$

□

Definition 7. *If X is a random variable from a distribution P over $\mathcal{X}$, we define the entropy of the random variable X as*

$$H(X) \triangleq H(P) \tag{10}$$

Similar to Claim 6, we also have the conclusion that if X, Y are independent r.v.s,

$$H(X,Y) = H(X) + H(Y) \tag{11}$$

More generally, we have the following claim.

Claim 8. *Consider two random variables X, Y, the following inequality holds:*

$$H(X,Y) \leq H(X) + H(Y) \tag{12}$$

Proof. According to the definition,

$$\begin{aligned}
H(X,Y) &= \sum_{x,y} P(x,y) \log\left(\frac{1}{P(x,y)}\right) \\
H(X) &= \sum_{x} P_X(x) \log\left(\frac{1}{P_X(x)}\right) = \sum_{x,y} P(x,y) \log\left(\frac{1}{P_X(x)}\right) \\
H(Y) &= \sum_{y} P_Y(y) \log\left(\frac{1}{P_Y(y)}\right) = \sum_{x,y} P(x,y) \log\left(\frac{1}{P_Y(y)}\right)
\end{aligned} \tag{13}$$

Thus, we have

$$\begin{aligned}
H(X) + H(Y) - H(X,Y) &= \sum_{x,y} P(x,y) \log\left(\frac{P(x,y)}{P_X(x) P_Y(y)}\right) \\
&= D(P || P_X \cdot P_Y) \geq 0
\end{aligned} \tag{14}$$

□

1.3 Conditional Entropy

Definition 9. *Consider two random variables* X, Y *defined on* $\mathcal{X}, \mathcal{Y}$ *respectively.* P *is the joint distribution. The conditional entropy of* X *given* Y *is defined as*

$$H(X|Y=y) = \sum_x P(X=x|Y=y)\log\left(\frac{1}{P(X=x|Y=y)}\right) \tag{15}$$

$$H(X|Y) = \sum_y P_Y(y)H(X|Y=y) = \sum_{x,y} P(x,y)\log\left(\frac{1}{P(X=x|Y=y)}\right) \tag{16}$$

Exercise. Show the chain rule of entropy:

$$H(X,Y) = H(Y) + H(X|Y) = H(X) + H(Y|X) \tag{17}$$

More generally, suppose $X_1, ..., X_n$ are n random variables, show that:

$$H(X_1, ...X_n) = H(X_1) + \sum_{i=2}^{n} H(X_i|X_1, ..., X_{i-1}) \tag{18}$$

Remark. Combine the chain rule of entropy and Claim 8 together, we can derive that

$$H(X|Y) \leq H(X) \tag{19}$$

Intuitively, when given Y, we get more information of X, then the uncertainty of X is smaller.

Definition 10. *The mutual information of two r.v.s* X, Y *is defined as*

$$\begin{aligned} I(X;Y) &= H(X) - H(X|Y) \\ &= H(Y) - H(Y|X) \\ &= H(X) + H(Y) - H(X,Y) \end{aligned} \tag{20}$$

Intuitively, $I(X;Y)$ characterizes the information provided by Y(or X) to reduce the uncertainty of X(or Y) and is always non-negative.

2 Multiway Classification and Fano's Inequality

2.1 Multiway Classification

Suppose there are M different distributions $P_1, ..., P_M$. Consider the following steps:

1. Randomly choose a distribution P_X, $X \sim U[M]$,
2. Observe Y from distribution P_X,
3. Using the outcome Y to predict $\tilde{X}$.

For the process described above, we have the following claim:

Claim 11.

$$I(X;Y) \geq \Pr(\mathit{correct}) \cdot \log(M-1) - \log 2 \tag{21}$$

Proof. Define

$$Z = \begin{cases} 0, & \text{if } X \neq \tilde{X} \\ 1, & \text{if } X = \tilde{X} \end{cases} \tag{22}$$

It is obvious that $H(Z|X, \tilde{X}) = 0$. Thus, using the chain rule of entropy, we can get

$$H(X, Z|\tilde{X}) = H(X|\tilde{X}) + H(Z|X, \tilde{X}) = H(X|\tilde{X}) \tag{23}$$

On the other hand, we have

$$\begin{aligned} H(X, Z|\tilde{X}) &= H(Z|\tilde{X}) + H(X|Z, \tilde{X}) \\ &\leq H(Z) + \Pr(Z=1)H(X|\tilde{X}, Z=1) + \Pr(Z=0)H(X|\tilde{X}, Z=0) \\ &\leq \log 2 + \Pr(Z=0)\log(M-1) \end{aligned} \tag{24}$$

The last inequality holds because $H(X|\tilde{X}, Z=1) = 0$ and

$$H(X|\tilde{X}, Z=0) = H(X|\tilde{X}, X \neq \tilde{X}) \leq \log(M-1)$$

Thus, we can get

$$H(X|\tilde{X}) \leq \log 2 + \Pr(\mathit{error})\log(M-1) \tag{25}$$

Since $H(X) = \log M$, we have

$$I(X;\tilde{X}) \geq \Pr(\mathit{correct}) \cdot \log(M-1) - \log 2 \tag{26}$$

Consider the probability model, we have

$$X \to Y \to \tilde{X}$$

Using data processing inequality, we get the conclusion that

$$I(X;Y) \geq I(X;\tilde{X}) \geq \Pr(\mathit{correct}) \cdot \log(M-1) - \log 2 \tag{27}$$

□

We use this result to prove Fano's inequality.

2.2 Fano's Inequality

Theorem 12 (Fano's inequality)**.** *Suppose there are M different distributions $P_1, ..., P_M$ s.t.*

$$D(P_i||P_j) \leq \beta, \forall i, j$$

For the multiway classification problem defined in section 2.1, the following inequality holds:

$$\Pr(correct) \cdot \log(M-1) - \log 2 \leq \beta \tag{28}$$

Proof. For the multiway classification problem, it is not hard to find that

$$\Pr(X = j) = \frac{1}{M} \tag{29}$$

$$\Pr(Y = y) = \frac{1}{M} \sum_j P_j(y) = \bar{P}(y) \tag{30}$$

Using the result in Claim 11, we know that if $I(X;Y) \leq \beta$, the statement is true. Consider

$$\begin{aligned} I(X;Y) &= H(X) - H(X|Y) \\ &= \sum_{j,y} \Pr(X = j, Y = y) \log \left(\frac{\Pr(X = j | Y = y)}{\Pr(X = j)} \right) \\ &= \sum_{j,y} \Pr(X = j, Y = y) \log \left(\frac{\Pr(X = j, Y = y)}{\Pr(X = j)\Pr(Y = y)} \right) \\ &= \sum_{j,y} \frac{1}{M} P_j(y) \log \left(\frac{P_j(y)}{\frac{1}{M} \sum_j P_j(y)} \right) \\ &= \frac{1}{M} \sum_j D(P_j || \bar{P}) \end{aligned} \tag{31}$$

So, we only need to prove that $D(P_i||\bar{P}) \leq \beta$. Since

$$\begin{aligned} \sum_{j=1}^{M} D(P||Q_j) &= \sum_x P(x) \log \left(\frac{P^M(x)}{\prod_{j=1}^{M} Q_j(x)} \right) \\ &= M \sum_x P(x) \log \left(\frac{P(x)}{(\prod_{j=1}^{M} Q_j(x))^{1/M}} \right) \\ &\leq M \sum_x P(x) \log \left(\frac{P(x)}{\frac{1}{M}(\sum_{j=1}^{M} Q_j(x))} \right) \\ &= MD \left(P \middle|\middle| \frac{1}{M} \sum_{j=1}^{M} Q_j(x) \right) \end{aligned} \tag{32}$$

The inequality comes from convexity of $\exp(\cdot)$:

$$\begin{aligned}\left(\prod_{j=1}^{M} Q_j(x)\right)^{1/M} &= \exp\left(\frac{1}{M}\sum_{j=1}^{M}\log(Q_j(x))\right)\\ &\geq \frac{1}{M}\sum_{j=1}^{M}\exp(\log(Q_j(x))) \\ &= \frac{1}{M}\sum_{j=1}^{M} Q_j(x)\end{aligned} \tag{33}$$

Thus,

$$D(P_i||\bar{P}) \leq \frac{1}{M}\sum_{j} D(P_i||P_j) \leq \beta$$

Thus, $I(X;Y) \leq \beta$ and then we get the conclusion. □

3 Learning Distributions

Definition 13. *Consider a collection of distributions $\mathcal{P}$ and a distance measure $d : \mathcal{P} \times \mathcal{P} \to \mathbb{R}$, define an $\varepsilon-$cover of $\mathcal{P}$ as a set of distributions $P_1, P_2, ..., P_N \in \mathcal{P}$, s.t. $\forall P \in \mathcal{P}$, there exists $1 \leq i \leq N$ s.t. $d(P, P_i) < \varepsilon$.*

Claim 14. *For any collection of distributions $\mathcal{P}$, we use the total variation distance as the distance measure, i.e. $d = d_{TV}$. Let N_ε be the smallest size of the $\varepsilon-$cover of $\mathcal{P}$. Then for any distribution $P \in \mathcal{P}$, we need only*

$$\frac{\log(N_\varepsilon)}{\varepsilon^2} \tag{34}$$

samples to learn $\hat{P}$ s.t. $d_{TV}(\hat{P}, P) < \varepsilon$ with probability at least $3/4$.

To prove this claim, we first introduce the problem of finding the closest distribution. Consider a collection of distributions $\mathcal{P}$ and N distributions $P_1, P_2, ..., P_N \in \mathcal{P}$. Suppose there is another distribution $P \in \mathcal{P}$ and we observe n samples $X_1, ..., X_n$ from P. Our goal is to output the closest distribution to P among $\{P_i\}_1^N$ based on the distance measure $d = d_{TV}$.

Theorem 15. *With*

$$\frac{C\log(N)}{\varepsilon^2} \tag{35}$$

samples, with probability at least $3/4$ we can learn P_j s.t.

$$d_{TV}(P, P_j) \leq 8\Delta + O(\varepsilon) \tag{36}$$

where $\Delta = \min_j d_{TV}(P, P_j)$

In the next lecture, we will show how to prove this theorem and therefore prove the previous claim. Also, we will give a "simple" algorithm to learn distributions optimally.