MATH888: High-dimensional probability and statistics Fall 2021

Lecture 33 — Nov 22, 2021

Sebastien Roch, UW-Madison *Scribe: Liancheng Fang*

1 Overview

In today's lecture we continue our quick tour of information theory, which prepare us for the proof of Fano's method. More material can be found in [1].

2 Quick tour of information theory

We assume RVs are discrete, we use X to denote RV, x for a specific value, $\mathcal{X}$ for the set of all possible value, $p(x)$ for probability mass function at x. Convention: $0\log 0 = 0, 0\log\frac{0}{0} = 0$

Definition 1. $H(X) = -\sum_{x\in\mathcal{X}} p(x)\log p(x)$

Remark. *Under the assumption that $\mathcal{X}$ is finite, it holds that $H(x) \leq \log|\mathcal{X}|$, and the equality is achieved for uniform distribution only.*

Definition 2. $H(X|Y) = -\sum_{x,y} p(x,y)\log p(x|y)$

Remark. *Since $H(X|Y) = -\sum_y p(y)\sum_x p(x|y)\log p(x|y)$, conditional entropy can be understand as the expectation of entropy of conditional distribution of X given Y.*

Lemma 3. $H(X,Y) = H(X) + H(Y|X)$

Proof. $H(X,Y) = -\sum_{x,y} p(x,y)\log p(x,y) = -\sum_{x,y} p(x,y)\log p(x) - \sum_{x,y} p(x,y)\log p(y|x) = H(X) + H(Y|X)$. □

Lemma 4. *(**Chain rule**)* $H(X_1,...,X_n) = \sum_{i=1}^n H(X_i|X_{i-1},...,X_1)$

Proof. By induction. □

Definition 5. $I(X;Y) = KL(P_{XY}\|P_X P_Y)$

Remark. *$I(X;Y)$ is always non-negative, and equal to zero iff $X \perp\!\!\!\perp Y$.*

Lemma 6. $I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X)$

Proof. $I(X;Y) = \sum_{x,y} p(x,y)\log\frac{p(x,y)}{p(x)p(y)} = \sum_{x,y} p(x,y)\log p(x|y) - \sum_{x,y} p(x,y)\log p(x) = -H(X|Y) + H(X)$. □

Lemma 7. $H(X|Y) \leq H(X)$

Proof. This follows from $I(X;Y) = H(X) - H(X|Y) \geq 0$. □

Definition 8. ***(conditional mutual information)***

$$\begin{aligned} I(X;Y|Z) &:= H(X|Z) - H(X|Y,Z) \\ &= \sum_{x,y,z} p(x,y,z) \log \frac{p(x,y|z)}{p(x|z)p(y|z)} \end{aligned}$$

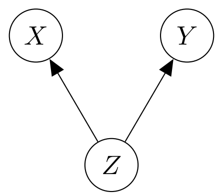
Lemma 9. $I(X;Y|Z) = 0 \iff X \perp\!\!\!\perp Y|Z$

Proof. $X \perp\!\!\!\perp Y|Z \iff p(x,y|z) = p(x|z)p(y|z), \forall x,y,z \iff I(X;Y|Z) = 0$ $\square$

Remark. *The conditional independence is equivalent to the following two factorization of joint density function:*

$$p(x,y,z) = p(z)p(x|z)p(y|z), \forall x,y,z \tag{1}$$

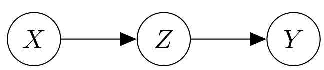
which means we first generate z from $p(z)$, then generate x, z from $p(x|z), p(y|z)$ respectively. In graphical model this is:



If we apply Bayesian rule on $p(x|z)$, we get

$$p(x,y,z) = p(x)p(z|x)p(y|z), \forall x,y,z \tag{2}$$

which means we first generate x from $p(x)$, then generate z from $p(z|x)$, finally generate y from $p(y|z)$. In graphical model this is:



Lemma 10. ***(Chain rule for MI)***

$$I(X_1, ..., X_n; Y) = \sum_{i=1}^{n} I(X_i; Y|X_{i-1}, ..., X_1)$$

Lemma 11. ***(Data processing inequality)*** *If $X \to Y \to Z$, then $I(X;Y) \geq I(X;Z)$. Equivalently, $H(X|Y) \leq H(X|Z)$*

Remark. *Interpretation: Consider Z as some function of Y, data processsing inequality tells that any function of Y does not contain more information about X than Y itself.*

Proof. Apply chain rule with $(X_1, X_2) = (Z, Y), (Y, Z)$ respectively, we get the following two inequalities:

$$I(X;Y,Z) = I(X;Z) + I(X;Y|Z) = I(X;Y) + I(X;Z|Y)$$

Since $I(X;Y|Z) \geq 0$ and $I(X;Z|Y) = 0$ by assumption, the proof completes. $\square$

Theorem 12. *(Fano's Inequality) Suppose $X \to Y \to \hat{X}$, and $P_e = \mathbb{P}(\hat{X} \neq X)$, then*

$$H(P_e) + P_e \log |\mathcal{X}| \geq H(X|Y)$$

where $H(P_e)$ is defined as the entropy of an indication RV, formally we define $H(P_e) := -P_e \log P_e - (1 - P_e)\log(1 - P_e)$.

Proof.

$$E = \begin{cases} 1 & \text{if } \hat{X} \neq X \\ 0 & \text{otherwise} \end{cases}$$

Apply chain rule with different order of E, X, we get

$$\begin{aligned} H(E, X|\hat{X}) &= H(X|\hat{X}) + H(E|X, \hat{X}) \\ &= H(E|\hat{X}) + H(X|\hat{X}, E) \end{aligned} \tag{3}$$

Note that $H(E|\hat{X}) \leq H(E) = H(P_e)$, $H(X|\hat{X}, E) = \mathbb{P}(E = 0)H(X|\hat{X}, E = 0) + \mathbb{P}(E = 1)H(X|\hat{X}, E = 1) \leq P_e \log(|\mathcal{X}| - 1)$, $H(E|X, \hat{X}) = 0$, plug these inequality into (3), we get

$$H(X|\hat{X}) \leq H(P_e) + P_e \log(|\mathcal{X}|) \tag{4}$$

Finally, by data process inequality, we have

$$H(X|\hat{X}) \geq H(X|Y)$$

Plug this into (4) complete the proof. □

Application to Fano's method

- $|\mathcal{X}| = M$
- J uniform at random in $[M]$
- $Z \sim P_{\theta_j}$ given $J = j$
- test $\psi(Z)$

Apply Fano's inequality with $J = X$, $Z = Y$, $\psi(X) = \hat{X}$.
Note that $H(Z, J) = H(J|Z) = H(J) - I(Z; J) = \log M - I(Z; J)$, and $H(P_e)$ term in Fano's inequality ≤ 1 since the indicator function can only take 2 value. From Fano we have

$$1 + Q[\psi(Z) \neq J] \log M \geq \log M - I(Z; J)$$

Rearrange terms we get

$$Q[\psi(Z) \neq J] \geq 1 - \frac{I(Z; J) + 1}{\log M}$$

References

[1] Cover, Thomas M., *Elements of information theory*, John Wiley & Sons, 1999.

[2] Wainwright, Martin J., *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*, Cambridge Series in Statistical and Probabilistic Mathematics, 2019.