ECE598: Information-theoretic methods in high-dimensional statistics Spring 2016

Lecture 12: Mutual Information Method: Continued

Lecturer: Yihong Wu *Scribe: Joseph Lubars, Mar 03, 2016*

12.1 Recap: Mutual Information Method

We have several equivalent definitions of mutual information from last class, capturing a measure of how far X and Y are from independence, or how much information about Y is provided by X:

$$\begin{aligned} I(X;Y) &= D(P_{XY}\|P_XP_Y) \\ &= D(P_{Y|X}\|P_Y|P_X) = \mathbb{E}_{x\sim P_X}[D(P_{Y|X=x}\|P_Y)] \\ &= \inf_{Q:X\perp Y \text{under } Q} D(P_{XY}\|Q_{XY}) \end{aligned}$$

Given the normal model $\theta \to X \to \hat{\theta}$, where θ generates the data X which generates an estimate $\hat{\theta}$, we can use the mutual information method to bound $I(\theta;\hat{\theta})$. In particular, as we saw last time, the following chain of inequalities always holds:

$$\min_{P_{\hat{\theta}|\theta}:\mathbb{E}\ell(\theta,\hat{\theta})\le R^*_\pi} I(\theta;\hat{\theta}) \le I(\theta;\hat{\theta}) \le I(\theta;X) \le \max_{P_\theta\in\mathcal{M}(\theta)} I(\theta;X)$$

We like to think of the left-most lower bound as the "cost" of an estimation task, which depends only on the prior and the loss function, but not on how the data is collected. We think of $\max_{P_\theta\in\mathcal{M}(\theta)} I(\theta;X)$ as the "capacity" of the model, which depends only on the model itself. Last lecture, we were able to compute the cost and capacity exactly for the Gaussian Location Model. In general, we may not be able to exactly compute the cost and capacity, so we will focus on methods for bounding them in this lecture.

12.2 Tensorization of Mutual Information

First, we would like to develop tools for bounding the mutual information of not just random variables, but random vectors as well. The chain rule for mutual information gives us an intuitive way to express the mutual information of a random vector as a sum of the mutual information of one-dimensional random variables:

Theorem 12.1 (Mutual Information Chain Rule)**.** *Let the random vector $X = (X_1,\dots,X_k)$ be jointly distributed with Y. Then:*

$$\begin{aligned} I(X;Y) &= I(X_1,X_2,\dots,X_k;Y) \\ &= I(X_1;Y) + I(X_2;Y|X_1) + \dots + I(X_k;Y|X^{k-1}) \end{aligned}$$

The proof of the chain rule follows from telescoping logs. For more information, see section 2.5 of [CT06]. In general, we cannot remove the conditioning and bound $I(X;Y)$ from above or below by $\sum_i I(X_i;Y)$. However, in some situations it is possible.

Example 12.1 (Tensorization in extremization problem)**.** Suppose $X = (X_1, \ldots, X_k)$ and $Y = (Y_1, \ldots, Y_k)$ are random vectors, and each coordinate of Y depends only on the corresponding coordinate of X:

$$\begin{aligned} X_1 &\to Y_1 \\ X_2 &\to Y_2 \\ &\vdots \\ X_k &\to Y_k \end{aligned}$$

Then the conditional distribution of Y given X factors:

$$P_{Y|X} = \prod_{i=1}^{k} P_{Y_i|X_i}$$

So long as the channels are decoupled like this, we have:

$$I(X;Y) \le \sum_{i=1}^{k} I(X_i, Y_i)$$

with equality if the X_i are independent from each other. Therefore, in particular:

$$\max_{P_X} I(X;Y) = \sum_{i=1}^{k} \max_{P_{X_i}} I(X_i, Y_i)$$

We can also consider a minimization problem for $I(X;Y)$. For example, if the coordinates of X are independent, i.e.:

$$P_X = \prod_{i=1}^{k} P_{X_i}$$

then we get can a lower bound on the mutual information:

$$I(X;Y) \ge \sum_{i=1}^{k} I(X_i, Y_i)$$

Equality holds when the coordinates of Y depend only on the corresponding coordinates of X, so $\min_{P_{Y|X}} I(X;Y)$ is achieved at the product of minimizers:

$$\min_{P_{Y|X}} I(X;Y) = \sum_{i=1}^{k} \min_{P_{Y_i|X_i}} I(X_i;Y_i)$$

In GLM, we could get nice bounds through the product structure. Otherwise, if there is no product structure, we would need to use the chain rule, which can be more difficult.

12.3 Capacity as Information Radius

To start, let us consider another way of thinking about mutual information.

Theorem 12.2 (Another Representation of Mutual Information)**.**

$$I(X;Y) = \min_Q D(P_{Y|X}\|Q|P_X)$$

Proof. For any Q we have:

$$\begin{aligned} I(X;Y) &= D(P_{Y|X}\|P_Y|P_X) \\ &= \mathbb{E}\log\frac{P_{Y|X}}{Q}\frac{Q}{P_Y} \\ &= D(P_{Y|X}\|Q|P_X) - D(P_Y\|Q) \end{aligned}$$

We get the desired result by noting that $D(P_Y\|Q) \geq 0$ and optimizing over Q. In particular, we can bound the mutual information using a convenient choice of Q, as we will see in the next example:

Example 12.2 (GLM)**.** Suppose $X \sim P_\theta = \mathcal{N}(\theta, 1)$. Then, choosing the best possible Gaussian Q and applying the above bound, we have:

$$\begin{aligned} I(\theta, X) &\leq \mathbb{E}_\theta D(P_\theta\|Q) \\ &= \inf_{\mu\in\mathbb{R}, s\geq 0} D(\mathcal{N}(\theta,1)\|\mathcal{N}(\mu,S)) \\ &= \frac{1}{2}\log(1+\mathrm{Var}(X)) \end{aligned}$$

where the solution to the minimization problem comes from the well-known formula for Gaussian channel capacity [PW15, p. 28].

Geometric Interpretation

The above representation of mutual information has a nice geometric picture, as follows: Let $\mathcal{X}$ be some space, let $\ell : \mathcal{X}\times\mathcal{X} \to \mathbb{R}$ be a loss function, and let A be a subset of $\mathcal{X}$.

Definition 12.1 (Radius of a Set)**.** The radius of A is the smallest ball that covers A. Note that we do not require the center y of the ball to be contained in A:

$$\mathrm{rad}(A) \triangleq \inf_{y\in\mathcal{X}} \sup_{x\in A} \ell(x,y)$$

Definition 12.2 (Diameter of a Set)**.** The diameter of A is the largest loss between two points in A:

$$\mathrm{diam}(A) \triangleq \sup_{x,y\in A} \ell(x,y)$$

Remark 12.1. Note that $\mathrm{rad}(A) \leq \mathrm{diam}(A)$. If ℓ satisfies the triangle inequality, then we further have $\mathrm{rad}(A) \geq \frac{1}{2}\mathrm{diam}(A)$.

Nothing above required ℓ to be a valid metric. In fact, we will be examining the following case where ℓ is not symmetric and does not satisfy the triangle inequality:

- $A = \{P_\theta : \theta \in \Theta\} \triangleq \mathcal{P}$
- $\ell(P, Q) = D(P\|Q)$
- $\text{rad}(\mathcal{P}) = \inf_Q \sup_{P\in\mathcal{P}} D(P\|Q)$
- $\text{diam}(\mathcal{P}) = \sup_{P,Q\in\mathcal{P}} D(P\|Q)$

By bounding the radius of $\mathcal{P}$, we can now upper bound the capacity of $\mathcal{P}$.

Theorem 12.3 (Capacity Bounded by Radius)**.** *Suppose we have the model $\theta \to X$, where $\mathcal{P} = \{P_\theta\}$ is defined as above. Let $C(\mathcal{P})$ be the capacity of $\mathcal{P}$. Then:*

$$C(P) \le \text{rad}(\mathcal{P}) \le \text{diam}(\mathcal{P})$$

Proof. Using Theorem 12.2, we have:

$$\begin{aligned}
C(\mathcal{P}) &= \sup_{P_\theta\in\mathcal{M}(\theta)} I(\theta; X) \\
&= \sup_{P_\theta\in\mathcal{M}(\theta)} \inf_Q D(P_{X|\theta}\|Q|P_\theta) \\
&\le \inf_Q \sup_{P_\theta\in\mathcal{M}(\theta)} D(P_{X|\theta}\|Q|P_\theta) \\
&= \inf_Q \sup_{\theta\in\Theta} D(P_\theta\|Q) \\
&= \text{rad}(\mathcal{P}) \\
&\le \text{diam}(\mathcal{P}) = \sup_{\theta,\theta'\in\Theta} D(P_\theta\|P_{\theta'})
\end{aligned}$$

Note: In fact, if $\mathcal{P}$ is convex, then we have equality in the third step, which would give us $C(\mathcal{P}) = \text{rad}(\mathcal{P})$. This is a result of Kemperman (cf. [PW15, Theorem 4.5]).

Example 12.3 (GLM, bounded mean)**.** Let $\mathcal{P} = \{P_\theta\} = \{\mathcal{N}(\theta, n^{-1}) : |\theta| \le \delta\}$. We can compute the radius of P_θ, taking $Q \sim \mathcal{N}(0, n^{-1})$:

$$\begin{aligned}
\text{rad}(\mathcal{P}) &= \inf_Q \sup_{|\theta|\le\delta} D\left(\mathcal{N}(\theta, n^{-1})\|Q\right) \\
&\le \sup_{|\theta|\le\delta} D\left(\mathcal{N}(\theta, n^{-1})\|\mathcal{N}(0, n^{-1})\right) \\
&= \sup_{|\theta|\le\delta} \frac{n}{2}\theta^2 \\
&= \frac{n\delta^2}{2}
\end{aligned}$$

We have used the fact that the KL divergence between two normal distributions with mean u and v and identical variance σ^2 is $\frac{1}{2\sigma^2}|u-v|^2$. We can also compute the diameter quite easily:

$$\begin{aligned}\operatorname{diam}(\mathcal{P}) &= \sup_{\theta,\theta' \in [\pm\delta]} D(\mathcal{N}(\theta, n^{-1}) \| \mathcal{N}(\theta', n^{-1})) \\ &= \frac{n}{2} \sup_{\theta,\theta' \in [\pm\delta]} |\theta - \theta'|^2 \\ &= 2n\delta^2\end{aligned}$$

Note that in this case, using the diameter instead of the radius only loses a factor of 4.

Now, we can proceed to the more general bounded GLM:

Theorem 12.4 (Bounded GLM)**.** *Let $X \sim P_\theta = \mathcal{N}(\theta, \frac{1}{n} I_p)$. Let $\ell(\theta, \theta') = \|\theta - \theta'\|_2^2$ (quadratic loss), and let $\Theta = B_2(0, \rho) \subset \mathbb{R}^p$. Then:*

$$R^* \asymp \frac{p}{n} \wedge \rho^2$$

Remark 12.2. The interpretation is that if ρ^2 is small and either we do not have enough samples or dimension is very high so that $\frac{p}{n}$ is smaller than ρ^2, then we should discard all the your data and declare zero as the estimate, because data do not provide better resolution than the prior information.

Proof. (Upper bound) We are already done here. Using $\overline{X}$ as an estimator, we have from previous lectures that (up to constant factors for these bounds):

$$R^* \leq \frac{p}{n}$$

Using 0 as an estimator, we just showed:

$$R^* \leq \rho^2$$

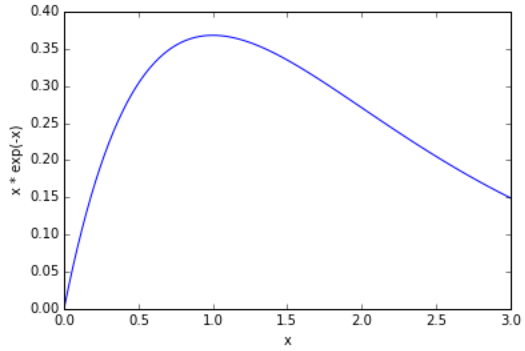
Therefore $R^* \leq \frac{p}{n} \wedge \rho^2$.

(Lower bound) First, to make things simpler, we will consider the case where $p = 1$. Before, when obtaining a lower bound on minimax risk, we used a Gaussian prior. However, we cannot use such a prior in this case because the Gaussian distribution is not supported on a ball of radius ρ. Instead, we will choose a uniform prior $\pi \sim \text{Uniform}(-r, r)$, with $r < \rho$. As before, we have:

$$\min_{P_{\hat{\theta}|\theta}: \mathbb{E}\ell(\theta,\hat{\theta}) \leq D} I(\theta; \hat{\theta}) \leq I(\theta; \hat{\theta}) \leq I(\theta; X) \leq \operatorname{rad}(\{\mathcal{N}(\theta, \frac{1}{n}) : |\theta| \leq r\})$$

We already have that the radius above is bounded by $\frac{nr^2}{2}$. However, the cost $\mathfrak{C} = \min_{P_{\hat{\theta}|\theta}: \mathbb{E}\ell(\theta,\hat{\theta}) \leq D} I(\theta; \hat{\theta})$ is much harder to calculate. We will therefore use a trick called the Shannon lower bound to bound $\mathfrak{C}$. The Shannon lower bound says that the cost given a non-Gaussian prior is not too far away from the cost given a Gaussian prior, provided that the prior is fairly Gaussian-like:

$$\mathfrak{C} \geq \mathfrak{C} \mid_{\theta \sim \text{Gaussian}} - D(\text{unif}(-r, r) \| \mathcal{N}(0, \frac{r^2}{3}))$$



Note: The quantity $\frac{r^2}{3}$ above is the variance of the uniform distribution.

We have (from last lecture) that the cost given a Gaussian prior is $\frac{1}{2}\log\frac{r^2/3}{D}$. Furthermore, we have that $D(\text{unif}(-r,r)\|\mathcal{N}(0,\frac{r^2}{3})) = D(\text{unif}(-1,1)\|\mathcal{N}(0,\frac{1}{3})) = c_1$ is a constant that does not depend on r. Therefore, for some other constant c:

$$\begin{aligned}\mathfrak{C} &\geq \frac{1}{2}\log\frac{r^2/3}{D} - c_1 \\ &= \frac{1}{2}\log\frac{r^2 c}{D}\end{aligned}$$

To complete the lower bound, remember that $\frac{1}{2}\log\frac{r^2 c_2}{D} \leq \mathfrak{C} \leq \frac{nr^2}{2}$, so:

$$\begin{aligned}R^* \geq R^*_\pi &\geq cr^2\exp(-nr^2), \forall r \in [0,\rho] \\ &\geq \sup_{r\in[0,\rho]} cr^2\exp(-nr^2) \\ &\asymp \frac{1}{n}\wedge\rho^2\end{aligned}$$

To justify the last step, do a change of variables $x = nr^2$, so the expression becomes $\frac{1}{n}\sup_{0\leq x\leq n\rho^2} x\exp(-x)$. If we examine the function $x\exp(-x)$, we see that it achieves a global maximum of $\frac{1}{e}$ at $x = 1$. Therefore, if $x < 1$ we should choose $x\exp(-x)$, and if $x \geq 1$ we should choose $\frac{1}{e}$. This gives us:

$$\frac{1}{n}\sup_{0\leq x\leq n\rho^2} x\exp(-x) = \frac{1}{n}(n\rho^2 e^{-n\rho^2}\wedge\frac{1}{e})$$

Recap:

In order to get the upper bound on the minimax risk, we used the radius, which can be thought of as the maximum distance between a central estimate and any other point in the space of distributions. The lower bound on the minimax risk came from the Shannon lower bound, which is based on how different the selected prior distribution is from a Gaussian distribution.

To extend the lower bound to an arbitrary dimension p, start with a uniform prior over a ball of radius r, calculate its variance, and use the Shannon lower bound again.

□

References

[CT06] Thomas M. Cover and Joy A. Thomas. *Elements of information theory, 2nd Ed.* Wiley-Interscience, New York, NY, USA, 2006.

[PW15] Y. Polyanskiy and Y. Wu. Lecture notes on information theory. Feb 2015. http://www.ifp.illinois.edu/~yihongwu/teaching/itlectures.pdf.