10-704: Information Processing and Learning **Spring 2012**

Lecture 3: Fano's, Differential Entropy, Maximum Entropy Distributions

Lecturer: Aarti Singh *Scribes: Min Xu*

3.1 Review

In last class, we covered the following topics:

- **Gibbs inequality** states that $D(p \,||\, q) \geq 0$
- **Data processing inequality** states that if $X \to Y \to Z$ forms a Markov Chain, then $I(X, Y) \geq I(X, Z)$. It implies that if $X \to Z \to Y$ is also a Markov Chain, then $I(X, Y) = I(X, Z)$
- Let θ parametrize a family of distributions, let Y be the data generated, and let $g(Y)$ be a statistic. Then $g(Y)$ is a **sufficient statistic** if, for all prior distributions on θ, the Markov Chain $\theta \to g(Y) \to Y$ holds, or, equivalently, if $I(\theta, Y) = I(\theta, g(Y))$.
- **Fano's Inequality** states that, if we try to predict X based on Y, then the probability of error is lower bounded by

$$p_e \geq \frac{H(X \,|\, Y) - 1}{\log |\mathcal{X}|} \qquad \text{(weak version)}$$

$$H(X \,|\, Y) \leq H(p_e) + p_e \log(|\mathcal{X}| - 1) \qquad \text{(strong version)}$$

  Note that we use the notation $H(p_e)$ as short-hand for entropy of a Bernoulli random variable with parameter p_e.

3.2 Clarification

The notation $I(X, Y, Z)$ is not valid, we only talk about mutual information between a pair of random variables (or pair of sets of random variables). In previous lecture, we used $I(X, (Y, Z))$ to denote $I(X; Y, Z)$, the mutual information between X and (Y, Z). We then used the following decomposition:

$$\begin{aligned} I(X; Y, Z) &= I(X, Y) + I(X, Z \,|\, Y) \\ &= I(X, Z) + I(X, Y \,|\, Z) \end{aligned}$$

Note that $I(X; Y, Z) \neq I(X, Y; Z)$ in general and that it is very different from conditional mutual informations $I(X, Y \,|\, Z)$.

On a second note, when using Venn-diagram to remember inequalities about information quantities, if we have 3 or more variables, some regions of the Venn-diagram can be negative. For instance, in figure 3.2, the central region, $I(X, Y) - I(X, Y \,|\, Z)$, could be negative (recall that in a homework problem you showed that $I(X, Y)$ can be less than $I(X, Y \,|\, Z)$).

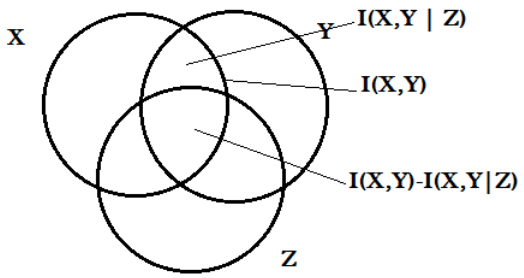

Figure 3.1: Venn-diagram for information quantities

3.3 More on Fano's Inequality

Note 1: Fano's Inequality is often applied with the decomposition $H(X \mid Y) = H(X) - I(X,Y)$ so that we have

$$p_e \geq \frac{H(X) - I(X,Y) - 1}{\log |\mathcal{X}|}.$$

Further, if X is uniform, then $H(X) = \log |\mathcal{X}|$ and Fano's Inequality becomes

$$p_e \geq 1 - \frac{I(X,Y) + 1}{\log |\mathcal{X}|}.$$

Note 2: From Fano's Inequality, we can conclude that there exist a estimator $g(Y)$ of X with $p_e = 0$ if and only if $H(X \mid Y) = 0$. If $H(X \mid Y) = 0$, then X is a deterministic function of Y, that is, there exist a deterministic function g such that $g(Y) = X$. Therefore, $p_e = 0$ for that estimator. If $p_e = 0$, then the strong version of Fano's Inequality implies that $H(X \mid Y) = 0$ as well.

3.3.1 Fano's Inequality is Sharp

We will show that there exist random variables X, Y for which Fano's Inequality is sharp. Let $\mathcal{X} = \{1, ..., m\}$, let p_i denote $p(X = i)$. Assume that $p_1 \geq p_2, ..., p_m$ and that $Y = \emptyset$.

Under this setting, it can be shown that the optimal estimate is $\hat{X} = 1$ and the error of probability $p_e = 1 - p_1$.

Because $Y = \emptyset$, Fano's inequality becomes $H(X) \leq H(E) + p_e \log(m-1)$.

If we set X with the distribution $\{1 - p_e, \frac{p_e}{m-1}, ..., \frac{p_e}{m-1}\}$, then we have

$$\begin{aligned} H(X) &= \sum_{i=1}^{m} p_i \log \frac{1}{p_i} \\ &= (1 - p_e) \log \frac{1}{1 - p_e} + \sum_{i=2}^{m} \frac{p_e}{m-1} \log \frac{m-1}{p_e} \\ &= (1 - p_e) \log \frac{1}{1 - p_e} + p_e \log \frac{1}{p_e} + p_e \log(m-1) \\ &= H(p_e) + p_e \log(m-1) \end{aligned}$$

Thus we see that Fano's inequality is tight under this setting.

3.4 Differential Entropy

Definition 3.1 *Let X be a continuous random variable. Then the **differential entropy** of X is defined as:*

$$H(X) = -\int f(x) \log f(x) dx$$

Differential entropy is usually defined in terms of natural log, $\log \equiv \ln$*, and is measured in* nats.

Example:
For $X \sim Unif[0, a)$:

$$H(X) = -\int_0^a \frac{1}{a} \log \frac{1}{a} dx = -\log \frac{1}{a} = \log a$$

In particular, note that *differential entropy can be negative.*

Example:
Let $X \sim N(0, \sigma^2)$. Let $\phi(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\{-\frac{1}{2}\frac{x^2}{\sigma^2}\}$ denote the pdf of a zero-mean Gaussian random variable, then

$$\begin{aligned} H(X) &= -\int_{-\infty}^{\infty} \phi(x) \log \phi(x) dx \\ &= -\int_{-\infty}^{\infty} \phi(x) \{-\frac{x^2}{2\sigma^2} - \log \sqrt{2\pi\sigma^2}\} dx \\ &= \frac{1}{2} + \log \sqrt{2\pi\sigma^2} \\ &= \frac{1}{2} \ln(2\pi e \sigma^2) \end{aligned}$$

Definition 3.2 *We can similarly define **relative entropy** for a continuous random variable as*

$$D(f_1 \,||\, f_2) = \int f_1(x) \ln \frac{f_1(x)}{f_2(x)} dx$$

Example:
Let f_1 be density of $N(0, \sigma^2)$ and let f_2 be density of $N(\mu, \sigma^2)$.

$$\begin{aligned} D(f_1 \,||\, f_2) &= \int f_1 \ln f_1 - \int f_1 \ln f_2 \\ &= -\frac{1}{2} \ln 2\pi e\sigma^2 - \int f_1(x) \{-\frac{(x-\mu)^2}{2\sigma^2} - \ln \sqrt{2\pi\sigma^2}\} dx \\ &= -\frac{1}{2} \ln 2\pi e\sigma^2 + \mathbb{E}_{f_1} \frac{(x-\mu)^2}{2\sigma^2} + \frac{1}{2} \ln 2\pi\sigma \\ &= -\frac{1}{2} + \mathbb{E}_{f_1} \frac{(x^2 - 2x\mu + \mu^2)}{2\sigma^2} \\ &= -\frac{1}{2} + \left[\frac{\sigma^2 + \mu^2}{2\sigma^2}\right] = \frac{\mu^2}{2\sigma^2} \end{aligned}$$

3.5 Maximum Entropy Distributions

We will now derive the maximum entropy distributions under moment constraints. Here is a preview; the next lecture will greatly elaborate on this section.

To solve for the maximum entropy distribution subject to linear moment constraints, we need to solve the following optimization program:

$$\max_f \; H(f) \tag{3.1}$$

$$\text{s.t. } f(x) \geq 0 \tag{3.2}$$

$$\int f(x)dx = 1 \tag{3.3}$$

$$\int f(x)s_i(x)dx = \alpha_i \quad \text{for all } i = 1, ..., n \tag{3.4}$$

For example, if $s_1(x) = x$ and $s_2(x) = x^2$, then optimization 3.1 yields the maximum entropy distribution with mean α_1 and second moment α_2.

To solve Optimization 3.1, we form the Lagrangian:

$$L(\lambda, f) = H(f) + \lambda_0 \int f + \sum_{i=1}^{n} \lambda_i \int f s_i(x)$$

We note that taking "derivative" of $L(\lambda, f)$ with respect to f gives:

$$\frac{\partial}{\partial f} L(\lambda, f) = \left[-\ln f - 1 + \lambda_0 + \sum_{i=1}^{n} \lambda_i s_i(x) \right]$$

Setting the derivative equal to 0 and we get that the optimal f is of the form

$$f^*(x) = \exp\left(-1 + \lambda_0^* + \sum_{i=1}^{n} \lambda_i^* s_i(x) \right)$$

where λ^* is chosen so that $f^*(x)$ satisfies the constraints. Thus, the maximum entropy distribution subject to moment constraints belongs to the exponential family. In next lecture, we will see that the same is true with inequality constraints (i.e. bounds on the moments).