Information Theory and Networks

Lecture 31: Information Theory and Estimation

Matthew Roughan
<matthew.roughan@adelaide.edu.au>
http://www.maths.adelaide.edu.au/matthew.roughan/Lecture_notes/InformationTheory/

School of Mathematical Sciences,
University of Adelaide

October 29, 2013

Part I

Information Theory and Estimation

Section 1

Connections

- Decoding is estimation
- Estimation requires information
  - sufficient statistics
  - Fano's inequality
  - Fisher information matrix
  - Cramer-Rao
- Maximum entropy

Section 2

Sufficient Statistics

Estimation

A common estimation problem

- We have a family probability distributions indexed by θ

$$\{f_\theta(x)\}$$

- Our goal is to take some samples $\{X_i\}$ and from these estimate (or infer) the particular $f_\theta(\cdot)$ from which they were drawn
- Typically we come up with an estimate $\hat{\theta}$
- Rather than use the raw data we often base the estimate on some statistics $T(X_1, \ldots .X_n)$ of the data, e.g., the mean and/or variances,
- There is a basic question about whether some set of statistics is sufficient for the estimation problem, or whether we should be using the raw data.

Data Processing Inequality

Definition

Random variables X, Y and Z are said to form a Markov chain in that order (denoted by $X \to Y \to Z$) if the conditional distribution of Z depends only on Y, i.e., Z is conditionally independent of X given Y.

Simple example: if $Z = g(Y)$ then $X \to Y \to Z$

Theorem (Data Processing Inequality)

If $X \to Y \to Z$ then

$$I(X;Y) \geq I(X;Z)$$

with equality iff $X \to Z \to Y$.

Simple example: $I(X;Y) \geq I(X;g(Y))$

Sufficient Statistics

A common estimation problem

- We have a family probability distributions indexed by θ

$$\{f_\theta(x)\}$$

- Assume we have samples $X_1, X_2, \ldots, X_n$, and statistic $T(X_1, X_2, \ldots, X_n)$, then

$$\theta \to \{X_1, X_2, \ldots, X_n\} \to T(X)$$

- The data processing inequality states that

$$I\left(\theta; \{X_1, X_2, \ldots, X_n\}\right) \geq I\Big(\theta; T(X_1, X_2, \ldots, X_n)\Big)$$

  for any distribution on θ.
  - No information is lost only if equality holds
  - So $\theta \to T(X_1, X_2, \ldots, X_n) \to \{X_1, X_2, \ldots, X_n\}$
- A statistic $T(X)$ is said to be sufficient for θ if it contains all the information in X about θ, i.e., we have equality above, i.e., $I(\theta; X) = I(\theta; T(X))$

Sufficient Statistic Example

- Let $X_i \in \{0, 1\}$ be IID Bernoulli RVs, with

$$\theta = P(X_i = 1)$$

- Given n samples $X_1, X_2, \ldots, X_n$ we take

$$T(X_1, X_2, \ldots, X_n) = \sum_{i=1}^{n} X_i$$

  Thus $\theta \to \{X_i\} \to T$

- Then

$$P\Big((X_1, X_2, \ldots, X_n) = (x_1, x_2, \ldots, x_n) \Big| T = k\Big) = \begin{cases} \frac{1}{\binom{n}{k}}, & \text{if } T = k \\ 0, & \text{otherwise.} \end{cases}$$

  essentially this means that given T, all sequences with a given number of 1s are equally likely.

- Thus $\theta \to T \to \{X_i\}$ and hence T is a sufficient statistic for θ

Minimal Sufficient Statistics

Definition (Minimal Sufficient Statistic)

A statistic $T(X)$ is a minimal sufficient statistic relative to $\{f_\theta(x)\}$ if it is a function of every other sufficient statistic $U(X)$.

In terms of the data processing inequality this means that

$$\theta \to T(X) \to U(X) \to X$$

Hence a minimal sufficient statistic maximally compresses the information about θ present in the sample X.

Section 3

Maximum Entropy Estimation

Laplace's principle of indifference

Definition (Laplace's principle of indifference)

If there are $n > 1$ possibilities for some event, and they are indistinguishable (except for their names) then each possibility should be assigned a equal probability $1/n$.

Often called the principle of insufficient reason.

Examples:

- What is the probability of a 6 on a dice?
- What is the probability of an Ace?

So this is the basic idea of probability that is often first presented to all students, from which we often develop more complicated ideas by counting and combinatorics.

Laplace's principle of indifference

Definition (Laplace's principle of indifference)

If there are $n > 1$ possibilities for some event, and they are indistinguishable (except for their names) then each possibility should be assigned a equal probability $1/n$.

- Note that the uniform distribution is the distribution with the maximum possible entropy
- So, why not see the principle of indifference as a special case of a larger rule of maximum entropy
- We'll need an analogue of entropy for continuous variates.

Differential Entropy

Definition (Differential Entropy)

The differential entropy $h(X)$ for a continuous RV X with support S and probability density function $f(x)$ is

$$h(X) = -\int_S f(x) \log f(x) \, dx$$

if this exists.

Examples:

- Uniform distribution: $U(0, a)$

$$h(X) = -\int_0^a \frac{1}{a} \log \frac{1}{a} \, dx = \log_2 a \text{ bits}$$

- Normal distribution: variance σ^2

$$h(X) = \frac{1}{2} \log_2 2\pi e \sigma^2 \text{ bits}$$

See [CT91, p.486-87] for a table of entropies for various other distributions.

Maximum Entropy

Definition (Maximum Entropy)

If there are $n > 1$ possibilities for some event, then each possibility should be assigned a probability consistent with maximising the entropy of the resulting distribution, consistent with any information we have about the distribution.

Philosophically, we are trying to impose the fewest additional assumptions on the distribution. We are aiming to avoid extracting information from thin air.

Information we might have:

- We know probabilities sum to 1
- We might know something like the mean or variance
- We might have some data

Maximum Entropy Distributions

Formally: maximise the entropy $h(f)$ over all probability densities f satisfying

1. $f(x) \geq 0$
2. $\displaystyle\int_S f(x)\,dx = 1$
3. $\displaystyle\int_S f(x) r_i(x)\,dx = \alpha_i$, for $i = 1, 2, \ldots, m$

The first two are just standard constraints on densities. The third implies certain "moment" constraints on the distribution.

Solution

Add Lagrange multiplier for each constraint, and maximise the functional

$$J\{f\} = \int g(f)\,dx = \int -f(x)\ln f(x) + \lambda_0 f(x) + \sum_{i=1}^{m} \lambda_i f(x) r_i(x)\,dx$$

Euler-Lagrange equation:

$$0 = \frac{\partial g}{\partial f} = -1 - \ln f(x) + \lambda_0 + \sum_{i=1}^{m} \lambda_i r_i(x)$$

Rearranging we get

$$f(x) = e^{\lambda_0 - 1 + \sum_{i=1}^{m} \lambda_i r_i(x)}$$

where the λ_i are (as yet) unknown Lagrange multipliers.

Example 1

Dice: we know there are 6 possibilities, but have not other information.

Maximising the entropy $H(X)$ corresponds to choosing the uniform distribution (as in the principle of indifference).

Example 2

Assume that we know $X \geq 0$ (which specifies is support $S = [0, \infty)$, and that we know it mean

$$\int_S f(x) x \, dx = \mu$$

Then we get the exponential distribution

$$f(x) = e^{\lambda_0 - 1 + \lambda_1 x} = A e^{-\lambda x}$$

We can calculate the constants by putting f back into the constraints

$$\begin{aligned} \int_0^\infty f(x) \, dx &= A \frac{1}{\lambda} \\ &= 1 \\ \int_0^\infty x f(x) \, dx &= A \frac{1}{\lambda^2} \\ &= \mu \end{aligned}$$

So $A = \lambda$ and $\lambda = 1/\mu$ so $f(x) = \frac{1}{\mu} e^{-x/\mu}$

Example 3

Assume that X has support $(-\infty, \infty)$, and we know its mean μ and variance σ^2.

- the exponent will be a quadratic
  - so the distribution is a Gaussian distribution
- Lagrange multipliers are chosen so that the mean and variance match

Applications

- Estimation:
  - suppose you have been told the mean and variance of a set of data
  - in absence of any other information, the maximum entropy estimate of the distribution from which the data was drawn is the normal distribution (with said mean and variance)
  - lots of other cases:
    - ★ spectral estimation
    - ★ traffic matrix estimation (max relative entropy)
- Physics:
  - see next lecture

Further reading I

Thomas M. Cover and Joy A. Thomas, *Elements of information theory*, John Wiley and Sons, 1991.

E.T. Jaynes, *Information theory and statistical methanics*, Physical Review **106** (1957), no. 4, 620–630.

______, *Information theory and statistical methanics. ii*, Physical Review **108** (1957), no. 2, 171–190.

David J. MacKay, *Information theory, inference, and learning algorithms*, Cambridge University Press, 2011.