Statistics 212a - Information Theory and Statistics

Aditya Guntuboyina

28 June 2012

Information Theoretic Quantities

- Shannon entropy
- Kullback-Leibler Divergence or relative entropy
- f-divergences (or ϕ-divergences)
- Mutual information (Jensen-Shannon divergence) and its counterpart for f-divergences
- Bregman divergences
- Renyi divergences
- Fisher information
- Metric entropy

Shannon Entropy

- The Shannon entropy for a discrete random variable X is defined by $H(X) = -\sum_x \mathbb{P}(X = x) \log \mathbb{P}(X = x)$.
- $H(X)$ is the shortest expected codelength for X.
- Shannon entropy is central to the theory of MDL in statistics which seeks to give a principled way to do model selection.

Kullback-Leibler divergence

- The Kullback-Leibler divergence or relative entropy between two probability measures P and Q is defined by $D(P||Q) := \int p \log(p/q)$ where p and q denote densities of P and Q respectively.

- KL divergence has a close connection to binary hypothesis testing.

- It is used as a notion of distance between P and Q. For example, it is frequently as a loss function in statistical estimation problems.

f-divergences

- f-divergences are a general class of divergences (indexed by convex functions f) that include the KL divergence as a special case.
- Let $f : (0, \infty) \to \mathbb{R}$ be a convex function for which $f(1) = 0$. The f-divergence between two probability measures P and Q is defined by $D_f(P||Q) := \int qf(p/q)$.
- Every f-divergence can be viewed as a measure of distance between probability measures.

Important Special Cases

- $f(x) = x \log x$ gives KL divergence.
- $f(x) = |x - 1|/2$ gives total variation distance $V(P, Q) = \int |p - q|/2$.
- $f(x) = (\sqrt{x} - 1)^2$ gives the square of the Hellinger distance: $H^2(P, Q) := \int (\sqrt{p} - \sqrt{q})^2$.
- $f(x) = (x - 1)^2$ gives the chi-squared divergence $\chi^2(P||Q) := \int (p - q)^2/q$.

Properties and Applications of f-divergences

- Fundamentally linked to binary hypothesis testing. Each f-divergence can be viewed as the integrated Bayes risk in testing where the integral is with respect to a distribution on the prior (Liese, 2012)
- Linked to classification (Nguyen, Wainwright and Jordan, 2009)
- A recent paper uses a subclass of f-divergences for goodness of fit testing (Jager and Wellner, 2006).

Mutual Information

- The mutual information between two random variables Θ and X is defined as the KL divergence between their joint distribution and the product of their marginals.
- Mutual information appears quite often in statistics. For example, it gives a principled way of finding non-informative priors in Bayesian statistics.
- A theoretically sensible way of choosing a *non-informative* prior distribution for Θ (data being X) is to **maximize** $\int D(posterior||prior)d(marginal)$ over all priors. It turns out that this quantity is exactly the mutual information between Θ and X. (Bernardo, 1979)

Jensen-Shannon Divergence

- Another application of Mutual Information is in ICA. Given (data from) a random vector X, the goal is to find a square matrix A such that the components of AX are independent. Theoretically, a sensible idea is to choose A so that the mutual information between the components of AX is **minimized**.

- Consider the special case when Θ is uniformly distributed over a finite set $\{\theta_1, \ldots, \theta_N\}$. Suppose that the conditional distribution of X given $\Theta = \theta_i$ is P_i for $i = 1, \ldots, N$.

- It is easy to see that the mutual information equals $J := \sum_i D(P_i||\bar{P})/N$ where $\bar{P} := (P_1 + \cdots + P_N)/N$. This quantity is also known as the Jensen-Shannon divergence.

Jensen-Shannon Divergence and Multiple Testing

- J is clearly a variance-like quantity and it measures how close together or far away the probabilities $P_1, \ldots, P_N$ are.
- It is intuitively obvious therefore that J should be linked to the problem of multiple hypothesis testing where, based on an observation X, one needs to pick one of the hypotheses: $H_1 : X \sim P_1, \ldots, H_N : X \sim P_N$.
- This connection is a standard result in information theory called Fano's inequality.

Jensen-Shannon Divergence for f-divergences

- The Jensen-Shannon divergence, J, can be extended to f-divergences in the obvious way:
  $J_f := \inf_Q \sum_i D_f(P_i||Q)/N$.

- It turns out that J_f is also related to multiple hypothesis testing through an inequality that generalizes Fano's inequality to arbitrary f-divergences. (Gushchin, 2004 and Guntuboyina, 2011).

Bregman Divergences

- The Bregman divergences provide another class of divergences that are indexed by convex functions and include both the Euclidean distance and the KL divergence as special cases.

- Let ϕ be a differentiable strictly convex function. The Bregman divergence D_ϕ is defined by

$$D_\phi(x, y) := \phi(x) - \phi(y) - \langle x - y, \nabla\phi(y)\rangle$$

- The domain of ϕ is a space where convexity and differentiability make sense (e.g., whole or a subset of $\mathbb{R}^d$ or an L_p space).

The main Bregman Divergence result

- For example, take $\phi(x) = ||x||^2$ on $\mathbb{R}^d$ which gives the Euclidean distance and $\phi(x) = \sum_j p_j \log p_j$ on the simplex in $\mathbb{R}^d$ which gives the KL divergence.

- Let X be a random quantity taking values in the domain of ϕ and satisfying certain assumptions. Then $\mathbb{E}D_\phi(X, a)$ is minimized over a in the domain of ϕ at $a = \mathbb{E}X$. Moreover, this property characterizes Bregman divergences.

- For example: (a) $\mathbb{E}(X - a)^2$ is minimized when $a = \mathbb{E}X$ and (b) $\sum_i D(P_i||Q)$ is minimized when $Q = (P_1 + \cdots + P_N)/N$.

- A consequence is that the posterior mean is the Bayes estimator when the loss function is a Bregman divergence.

Renyi Divergences

- These are another generalization of the KL divergence.
- The Renyi divergence between two probability distributions P and Q is $D_\alpha(P||Q) := \left(\log\left(\int p^\alpha q^{1-\alpha}\right)\right)/(\alpha - 1)$. When $\alpha = 1$, by a continuity argument, D_α is defined as KL divergence.
- $D_{1/2}(P||Q) := -2\log\int\sqrt{pq}$ is called Bhattacharyya divergence (closely related to Hellinger distance). This quantity is smaller than KL and, as a result, it is sometimes easier to derive risk bounds with $D_{1/2}$ as the loss function as opposed to KL.

Fisher Information

- Arguably, the most fundamental quantity in classical (parametric) theoretical statistics.
- For a family of probability densities $p_\theta(x)$, the Fisher information is defined by

$$I(\theta) = \int \left(\frac{\partial}{\partial \theta} \log p_\theta(x) \right)^2 p_\theta(x) dx$$

- The Cramer-Rao lower bound says that the inverse of the Fisher information is a lower bound on the variance of any unbiased estimator of θ. A related result is the Van Trees inequality (Gill and Levit, 1995).

Fisher Information (continued)

Fisher information is also key in the theory of efficiency:

The official dogma on estimation is this: good estimators converge to the right thing and have limiting normal distributions. Moreover, the variance of the limiting distribution can't be smaller than a quantity defined by the Fisher Information. The estimators that achieve the asymptotic lower bound are called efficient. Maximum likelihood estimators are efficient.

The dogma is not quite correct, but much of it can be rescued in slightly altered form. - David Pollard (from his book *Asymptopia*).

Things go wrong via superefficiency (e.g, Hodges's example).

Fisher Information (continued)

- Fisher information is an information-theoretic quantity. It is considered as the information that the data contains about the parameter θ. Moreover, it is closely related to KL divergence. For sufficiently regular parametric models $\{P_\theta\}$, the KL divergence $D(P_{\theta^*}||P_\theta)$ behaves approximately like a quadratic form with matrix $I(\theta^*)/2$.

- Superefficiency can also be studied via the KL loss (Barron and Hengartner, 1998).

- Fisher information also appears in Bayesian statistics through the Jeffrey's prior which can be shown to asymptotically solve Bernardo's non-informative prior problem (Clarke and Barron, 1993).

Metric Entropy

- For a subset $\mathcal{F}$ of a metric space $(\mathcal{X}, \rho)$, the ϵ-covering number $M(\mathcal{F}, \epsilon; \rho)$ is defined as the smallest number of closed balls of radius ϵ whose union contains $\mathcal{F}$.

- The quantity $H_\epsilon(\mathcal{F}) := \log M(\mathcal{F}, \epsilon; \rho)$ is called the ϵ-entropy of $\mathcal{F}$. This notion is due to Kolmogorov and is related to Shannon entropy (Cover, Gacs and Gray, 1989).

- ϵ-entropy appears prominently in nonparametric estimation (for an overview, see the chapter by Nikouline and Solev in the book: Kolmogorov's Heritage in Mathematics).

Metric Entropy (continued)

- In nonparametric estimation problems, especially when answering questions related to overfitting, one needs to somehow measure the space of functions that are being fit to the data. Metric entropy is a convenient way of doing this.

- Metric entropy comes up in characterizations of minimax rates of convergence (Yang and Barron, 1999) and also in the study of rates of convergence of sieved and penalized likelihood estimators (see, e.g., Van de Geer's book).

- In this class, we study examples of metric entropy calculations (for smooth functions, convex sets and convex functions) and then focus on the above applications.

Tentative List of Topics

- f-divergences: properties, connections to testing, inequalities, connections to classification and application to goodness of fit testing.
- Fano's inequality and extensions to f-divergences
- Mutual information, Bernardo's non-informative prior problem and solution via Jeffrey's prior.
- Mutual information based ICA
- Minimum Description Length Principle
- Bregman Divergence and the mean-minimization property
- Fisher information: Cramer-Rao inequality, Van Trees inequality, Efficiency, Superefficiency, Bernstein-Von Mises theorem.
- Metric entropy: examples, minimax lower bounds, upper bounds, rates of convergence of sieved and penalized estimators.