

Stat 992: Lecture 39

Selected homework solutions.

Moo K. Chung `mchung@stat.wisc.edu`

May 2, 2004

1. *Expectation-maximization.* We augment the data with with *latent* (unobserved) data X^m such that the complete data $X^c = (X, X^m)$. The density of X^c is denoted as

$$X^c = (X, X^m) \sim f(x^c) = f(x, x^m).$$

The conditional density for the missing data X^m is

$$f(x^m|x, \theta) = \frac{f(x, x^m|\theta)}{f(x|\theta)}.$$

Let $R(\theta|\theta_0, X) = \mathbb{E}[\ln f(X^m|X, \theta)|X, \theta_0]$. It can be shown that

$$\theta_0 = \arg \max_{\theta} R(\theta|\theta_0, x) \quad (1)$$

Problem 46. The standard proof.

$$\begin{aligned} & R(\theta|\theta_0, X) - R(\theta_0|\theta_0, X) \\ &= \mathbb{E}[\ln f(X^m|X, \theta)|X, \theta_0] - \mathbb{E}[\ln f(X^m|X, \theta_0)|X, \theta_0] \\ &= \mathbb{E}\left[\ln \frac{f(X^m|X, \theta)}{f(X^m|X, \theta_0)} \middle| X, \theta_0\right] \\ &\leq \ln \mathbb{E}\left[\frac{f(X^m|X, \theta)}{f(X^m|X, \theta_0)} \middle| X, \theta_0\right] \\ &= \ln \int \frac{f(x^m|x, \theta)}{f(x^m|x, \theta_0)} f(x^m|x, \theta_0) dx^m = 0. \end{aligned}$$

The inequality is due to Jensen's inequality which states for a convex function g and random variable Z , $\mathbb{E}g(Z) \geq g(\mathbb{E}X)$. Tulaya Limpiti note that $\ln t \leq t - 1$ for all $t > 0$ which again uses the concavity of log function. So without using Jensen's inequality,

$$\begin{aligned} & \mathbb{E}\left[\ln \frac{f(X^m|X, \theta)}{f(X^m|X, \theta_0)} \middle| X, \theta_0\right] \\ &\leq \mathbb{E}\left[\frac{f(X^m|X, \theta)}{f(X^m|X, \theta_0)} \middle| X, \theta_0\right] - 1 = 0 \end{aligned}$$

Jensen's inequality is a basis of almost all well known inequalities. For instance, Hölder's inequality can be derived from Jensen with a proper choice

of convex function. Given composite function $f = h \circ g$, we can find conditions that grantee the convexity of f in terms of h and g . For instance, f convex if h nondecreasing convex and g convex; f convex if h nonincreasing convex and g concave. Read Chapter 3 of *Convex Optimization* by S. Boyd and L. Vandenberghe, 2004. Combridge University Press or any convex analysis text book.

2. *Entropy.* The concept of entropy comes from thermodynamics and then C.B. Shannon introduced it for analyzing the complexity of information. See *A Mathematical Theory of Communication*, The Bell System Technical Journal. 1948. After this, Kullback used it in defining the distance between two random variables. Kullback-Leibler divergence between two probability density f and g is defined as

$$H(f, g) = \mathbb{E}_f \ln \frac{f(X)}{g(X)} = \int \log \frac{f(x)}{g(x)} f(x) .dx$$

Applying Jensen's inequality, it can be shown that $H(f, g) \geq 0$ and $H(f, f) = 0$. This is usually called *relative entropy* in machine learning community. One can define the dependency of two random variables using the relative entropy. Given two distributions f_1 and f_2 and their joint distribution f , The *mutual information* of f_1 and f_2 is defined as $I(f_1, f_2) = H(f, f_1 f_2)$. If $f = f_1 f_2$, two distributions are independent so the mutual information will vanish. This concept has been widely used as a similarity measure in image analysis.

Problem 50. Given two random variable $X \sim N(0, \sigma_1^2)$ and $Y \sim N(0, \sigma_2^2)$ with correlation $\rho(X, Y) = \rho$, compute the mutual information.

3. Solution to **Problem 49.** (Tulaya Limpiti). Given zero mean unit variance random field of the form $Y \propto K_H * W$, prove $\text{Cov} \dot{Y} = (HH')^{-1}/2$. The proof will follow the isotropic case. The kernel is

given by

$$K_H(x) = \frac{1}{(2\pi)^{n/2}|H|} \exp[-x'(HH')^{-1}x/2]$$

assuming $H = H'$ invertible. It can be shown that

$$K_H^2(x) = \frac{1}{(4\pi)^{n/2}|H|} K_{H/\sqrt{2}}(x).$$

Since $\partial_x(x'Ax) = (A + A')x/2$,

$$\partial_x K_H(x) = -K_H(x)(HH')^{-1}x.$$

Hence

$$\begin{aligned} \mathbf{Cov}\dot{Y} &= \frac{\int \partial_x K_H (\partial_x K_H)' dx}{\int K_H^2 dx} \\ &= \frac{\int K_H^2 (HH')^{-1} x x' (HH')^{-1} dx}{\int K_H^2 dx} \\ &= (HH')^{-1} \left[\int K_{H/\sqrt{2}}(x) x x' dx \right] (HH')^{-1}. \end{aligned}$$

If we let $x = Hu/\sqrt{2}$, $dx = 2^{-n/2}|H|du$ and $K_{H/\sqrt{2}}(x) = 2^{n/2}|H|^{-1}K_I(u)$. Hence

$$K_{H/\sqrt{2}}(x) dx = K_I(u) du$$

and using the previous result on isotropic kernel, the above expression simplifies to

$$\mathbf{Cov}\dot{Y} = \frac{1}{2}(HH')^{-1}.$$

See Worsley's paper to see application of this identity.