ECE 299: STATISTICAL LEARNING THEORY

MAXIM RAGINSKY

Homework 3

Assigned March 21, 2011; due March 30, 2011

The problem of *density estimation* is posed as follows. We obtain an i.i.d. sample $X_1, \dots, X_n$ of $\mathbb{R}^d$-valued random variables whose common distribution is unknown. Assuming that it has a well-defined probability density function (pdf) f, we would like to estimate it. We will measure the accuracy of an estimate $\widehat{f}$ by the *Kullback–Leibler divergence*

$$D(f\|\widehat{f}) = \int_{\mathbb{R}^d} f(x) \log \frac{f(x)}{\widehat{f}(x)} dx.$$

A popular strategy for density estimation is by *maximum likelihood* (ML). That is, if we know that f comes from some parametric class $\mathcal{H} = \{h_\theta : \theta \in \Theta\}$, then we let

$$\widehat{f}_n = h_{\widehat{\theta}_n}, \qquad \text{where } \widehat{\theta}_n = \operatorname*{arg\,min}_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^n \log \frac{1}{h_\theta(X_i)}.$$

However, $\mathcal{H}$ may not be rich enough to accurately approximate the unknown f. For this reason, we may consider more complicated classes of densities. For example, given a positive integer k, let $\mathcal{H}_k$ denote the class of all k-component mixtures over $\mathcal{H}$:

$$\mathcal{H}_k \triangleq \left\{ h = \sum_{j=1}^k c_j h_{\theta_j} : c_1, \dots, c_k \ge 0, \sum_{j=1}^k c_j = 1; \theta_1, \dots, \theta_k \in \Theta \right\}$$

Now consider performing ML estimation over $\mathcal{H}_k$:

$$\widehat{f}_n = \operatorname*{arg\,min}_{h \in \mathcal{H}_k} \frac{1}{n} \sum_{i=1}^n \log \frac{1}{h(X_i)}.$$

In this homework assignment, you will be asked to prove the following result:

Theorem 1. *Suppose that both the unknown density f and the base densities h_θ, $\theta \in \Theta$, are bounded between some strictly positive constants $0 < c_- < c_+ < \infty$:*

$$c_- \le f(x) \le c_+; \qquad c_- \le h_\theta(x) \le c_+, \forall \theta \in \Theta.$$

Then

$$D(f\|\widehat{f}_n) \le D(f\|f_k) + \frac{4c_+}{(c_-)^2} \mathbb{E} R_n(\mathcal{H}(X^n)) + 2 \log\left(\frac{c_+}{c_-}\right) \sqrt{\frac{2\log(1/\delta)}{n}}$$

with probability at least $1 - \delta$, where

$$f_k = \operatorname*{arg\,min}_{h \in \mathcal{H}_k} D(f\|h).$$

You will prove this theorem in several steps.

(1) **Uniform deviation bound.** Prove that

$$D(f\|\widehat{f}_n) - D(f\|f_k) \le 2 \sup_{h\in\mathcal{H}_k} \left| \frac{1}{n}\sum_{i=1}^n \log\frac{h(X_i)}{f(X_i)} - \mathbb{E}\left[\log\frac{h(X)}{f(X)}\right]\right|.$$

(2) **From uniform deviations to Rademacher averages**. Let

$$\Delta_n(X^n) \triangleq \sup_{h\in\mathcal{H}_k} \left| \frac{1}{n}\sum_{i=1}^n \log\frac{h(X_i)}{f_(X_i)} - \mathbb{E}\left[\log\frac{h(X)}{f(X)}\right]\right|.$$

Let $\mathcal{L}_f$ be the class of all functions of the form $\log[h(x)/f(x)]$, $h\in\mathcal{H}_k$. Prove that

$$\Delta_n(X^n) \le 2\mathbb{E}R_n(\mathcal{L}_f(X^n)) + \log\left(\frac{c_+}{c_-}\right)\sqrt{\frac{2\log(1/\delta)}{n}}$$

with probability at least $1-\delta$.

(3) Prove that

$$R_n(\mathcal{L}_f(X^n)) \le \frac{c_+}{(c_-)^2} R_n(\mathcal{H}(X^n)).$$

Hint: You will need the following generalization of the contraction principle: Let $\mathcal{A}\subset\mathbb{R}^n$ be a bounded set. Let $\phi_1,\dots,\phi_n:\mathbb{R}\to\mathbb{R}$ be n functions, such that there exists some positive constant L so that for any two distinct $a=(a_1,\dots,a_n), a'=(a'_1,\dots,a'_n)\in\mathcal{A}$

$$\max_{1\le i\le n}\frac{|\phi_i(a_i)-\phi_i(a'_i)|}{|a_i-a'_i|} \le L.$$

Define the set

$$\phi\circ\mathcal{A} \triangleq \{(\phi_1(a_1),\dots,\phi_n(a_n)) : a=(a_1,\dots,a_n)\in\mathcal{A}\}.$$

Then $R_n(\phi\circ A) \le LR_n(\mathcal{A})$.

(4) Combine the results of the previous problems to finish the proof.