ECE 299: STATISTICAL LEARNING THEORY

MAXIM RAGINSKY

Homework 3 Solutions

Assigned March 21, 2011; due March 30, 2011

The problem of *density estimation* is posed as follows. We obtain an i.i.d. sample $X_1, \ldots, X_n$ 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

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

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 $\mathcal{H}$:

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

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

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

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 $0 < c_- < c_+ < \infty$:*

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

Then

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

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

$$f_k = \arg\min_{h \in \mathcal{H}_k} D(f\|h).$$

You will prove this theorem in several steps.

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

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

Solution. Note that for any $h \in \mathcal{H}_k$

$$D(f\|h) = \mathbb{E}\left[\log\frac{f(X)}{h(X)}\right] \equiv \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n \log\frac{f(X_i)}{h(X_i)}\right].$$

Let $\widehat{f}_n$ be the empirical risk minimizer. Then

$$\widehat{f}_n = \arg\min_{h\in\mathcal{H}_k} \frac{1}{n}\sum_{i=1}^n \log\frac{f(X_i)}{h(X_i)}$$

and therefore

$$\frac{1}{n}\sum_{i=1}^n \log\frac{f(X_i)}{\widehat{f}_n(X_i)} \leq \frac{1}{n}\sum_{i=1}^n \log\frac{f(X_i)}{f_k(X_i)}. \tag{1}$$

Then we can write

$$\begin{aligned}
&D(f\|\widehat{f}_n) - D(f\|f_k)\\
&= \underbrace{D(f\|\widehat{f}_n) - \frac{1}{n}\sum_{i=1}^n \log\frac{f(X_i)}{\widehat{f}_n(X_i)}}_{T_1} + \underbrace{\frac{1}{n}\sum_{i=1}^n \log\frac{f(X_i)}{\widehat{f}_n(X_i)} - \frac{1}{n}\sum_{i=1}^n \log\frac{f(X_i)}{f_k(X_i)}}_{T_2} + \underbrace{\frac{1}{n}\sum_{i=1}^n \log\frac{f(X_i)}{f_k(X_i)} - D(f\|f_k)}_{T_3}
\end{aligned}$$

Because of (1), $T_2 \leq 0$, while

$$T_1 + T_3 \leq 2\sup_{h\in\mathcal{H}_k}\left|\frac{1}{n}\sum_{i=1}^n \log\frac{f(X_i)}{h_(X_i)} - \mathbb{E}\left[\log\frac{f(X)}{h(X)}\right]\right|.$$

Moreover, because $\log(1/u) = -\log u$, we can write

$$\left|\frac{1}{n}\sum_{i=1}^n \log\frac{h(X_i)}{f_(X_i)} - \mathbb{E}\left[\log\frac{h(X)}{f(X)}\right]\right|$$

for any h. Putting everything together, we get the desired bound.

(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

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

with probability at least $1-\delta$.

Solution. First, we will show that $\Delta_n(X^n)$ has bounded differences. Choose an arbitrary n-tuple $x_1, \ldots, x_n$ and an arbitrary x_i'; let $x_{(i)}^n$ denote x^n with x_i replaced by x_i'. Then, because both

f and all h are bounded between c_- and c_+, we get

$$\begin{aligned}
&\left|\Delta_n(x^n) - \Delta_n(x^n_{(i)})\right| \\
&\leq \frac{1}{n} \sup_{h \in \mathcal{H}_k} \left|\log \frac{h(x_i)}{f(x_i)} - \log \frac{h(x_i')}{f(x_i)}\right| \\
&\leq \frac{2}{n} \log \frac{c_+}{c_-}.
\end{aligned}$$

This holds for all i and all x^n and x_i', so $\Delta_n(X^n)$ has bounded differences with $c_1 = \ldots = c_n = (2/n)\log(c_+/c_-)$. Therefore, McDiarmid's inequality says that for any $t > 0$

$$\mathbb{P}\left(\Delta_n(X^n) \geq \mathbb{E}\Delta_n(X^n) + t\right) \leq \exp\left(-\frac{nt^2}{2\log^2(c_+/c_-)}\right).$$

Choosing $t = \log(c_+/c_-)\sqrt{\frac{2\log(1/\delta)}{n}}$ and using the fact that $\mathbb{E}\Delta_n(X^n) \leq 2\mathbb{E}R_n(\mathcal{L}_f(X^n))$, we get the desired result.

(3) Prove that

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

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

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

Define the set

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

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

Solution. The idea is to use the contraction principle twice. First, let us consider the function $\phi(u) = \log(u)$. Its derivative on the interval $[c_-/c_+, c_+/c_-]$ is bounded in magnitude by c_+/c_-, so

$$|\phi(u) - \phi(v)| \leq \frac{c_+}{c_-}|u - v|, \qquad \forall u, v \in \left[\frac{c_-}{c_+}, \frac{c_+}{c_-}\right].$$

Hence, ϕ is (c_+/c_-)-Lipschitz on the interval $[c_-/c_+, c_+/c_-]$. Therefore, letting $\mathcal{M}_f$ denote the class of all functions of the form $h(x)/f(x)$, $h \in \mathcal{H}_k$, we have by the contraction principle that

$$R_n(\mathcal{L}_f(X^n)) \leq \frac{c_+}{c_-} R_n(\mathcal{M}_f(X_n)).$$

Now consider the functions $\phi_i(u) = \frac{u}{f(X_i)}$, $i = 1, \ldots, n$. Then

$$|\phi_i(u) - \phi_i(v)| = \frac{1}{f_i(X_i)}|u - v| \leq \frac{1}{c_-}|u - v|.$$

It is easy to see that $\mathcal{M}_f(X^n) = \phi \circ \mathcal{H}_k(X^n)$, and by the generalized contraction principle we have

$$R_n(\mathcal{M}_f(X^n)) \leq \frac{1}{c_-} R_n(\mathcal{H}_k(X^n)).$$

Therefore,

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

Finally, since $\mathcal{H}_k(X^n) \subset \mathrm{conv}(\mathcal{H}(X^n))$, we have

$$R_n(\mathcal{H}_k(X^n)) \leq R_n(\mathrm{conv}(\mathcal{H}(X^n))) = R_n(\mathcal{H}(X^n)).$$

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