UC Berkeley
Department of Electrical Engineering and Computer Science
Department of Statistics

EECS 281A / STAT 241A Statistical Learning Theory

Problem Set 3
Fall 2011

Issued: Monday, October 10, 2011 **Due:** Monday, October 24, 2011

Reading: For this problem set: Chapters 8, 9.

Problem 3.1
For each of the following problems, write out the maximum likelihood problem based on n i.i.d. samples $X_1, \ldots, X_n$, and compute the maximum likelihood estimate $\widehat{\theta}$.

(a) Let $p(x;\mu) = \mu^x (1-\mu)^{1-x}$ be a Bernoulli distribution, and consider estimating μ.

(b) Let $X \sim \mathrm{Poi}(\lambda)$, and consider estimating the intensity parameter λ.

(c) Let $X \in \mathbb{R}^d$ be a zero-mean multivariate Gaussian, parameterized in canonical form in terms of a symmetric positive definite matrix $\Gamma \succ 0$ as $p(x;\Gamma) = \frac{1}{\sqrt{(2\pi)^d \det(\Gamma)^{-1}}} \exp\big(-\frac{1}{2} x^T \Gamma x\big)$, and consider estimating the matrix Γ.

Problem 3.2
Maximum a posteriori (MAP) and MLE: Suppose that we adopt a Bayesian perspective, and view the parameter $\theta \in \mathbb{R}^d$ as a random variable, say distributed according to the prior distribution $\theta \sim \pi(\cdot)$. Given n i.i.d. samples $\{X_1, \ldots, X_n\}$, the MAP estimate is defined as the maximizer of the (rescaled) posterior likelihood $\frac{1}{n} \log p(\theta \mid X_1, X_2, \ldots, X_n)$.

(a) Suppose that $(X_i \mid \theta)$ is Gaussian with mean θ and fixed (known) variance $\sigma^2 > 0$, and let the prior $\pi(\cdot)$ distribution of θ be normal $N(\theta_0, \tau^2)$, where $\theta_0 \in \mathbb{R}$ and $\tau^2 > 0$ are fixed, known parameters. Compute the MAP estimate of θ.

(b) Compute the maximum likelihood estimate of θ.

(c) What happens to the MAP estimate as the number of samples n goes to infinity?

Problem 3.3
Recall that a probability distribution in the exponential family takes the form

$$p(x;\eta) = h(x) \exp\{\eta^T T(x) - A(\eta)\}$$

for a parameter vector η, often referred to as the *natural parameter*, and for given functions T, A, and h.

(a) Determine which of the following distributions are in the exponential family, exhibiting the T, A, and h functions for those that are.

  (a) $N(\mu, I)$—multivariate Gaussian with mean vector μ and identity covariance matrix.

  (b) $\mathrm{Dir}(\alpha)$—Dirichlet with parameter vector $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_K)$.

(c) Mult(θ)—multinomial with parameter vector $\theta = (\theta_1, \theta_2, \ldots, \theta_K)$. Use the fact that $\theta_K = 1 - \sum_{k=1}^{K-1} \theta_k$ and express the distribution using a $(K-1)$-dimensional parameter η.

(d) the log normal distribution: the distribution of $Y = \exp(X)$, where $X \sim N(0, \sigma^2)$.

(e) the Ising model: an undirected graphical model $G = (V, E)$ involving a binary random vector X taking values in $\{0,1\}^n$ with distribution $p(x; \theta) \propto \exp\big\{ \sum_{s \in V} \theta_s x_s + \sum_{(s,t) \in E} \theta_{st} x_s x_t \big\}$.

(b) Recall that the function $A(\eta)$ has moment-generating properties—in particular, $\nabla_\eta A(\eta) = \mathbb{E}[T(X)]$. Demonstrate that this relationship holds for those examples that are in the exponential family in part (a).

Problem 3.4

The course homepage has a data set named "lms.dat" that contains twenty rows of three columns of numbers. The first two columns are the components of an input vector x and the last column is an output y value. (We will not use a constant term for this problem; thus the input vector and the parameter vector are both two dimensional.)

(a) Solve the normal equations for these data to find the optimal value of the parameter vector. (I recommend using MATLAB or R.)

(b) Find the eigenvectors and eigenvalues of the covariance matrix of the input vectors and plot contours of the cost function J in the parameter space. These contours should of course be centered around the optimal value from part (a).

(c) Initializing the LMS algorithm at $\theta = 0$ plot the path taken in the parameter space by the algorithm for three different values of the step size ρ. In particular let ρ equal the inverse of the maximum eigenvalue of the covariance matrix, one-half of that value, and one-quarter of that value.

Problem 3.5

(Properties of Kullback-Leibler divergence:) Given two probability distributions p and q (where the random variables take values in $\{0, 1, \ldots, k-1\}$), the Kullback-Leibler divergence is defined as $D(p\|q) = \sum_{x=0}^{k-1} p(x) \log \frac{p(x)}{q(x)}$.

(a) Show that $D(p \,\|\, q) \geq 0$ for all p, q, with equality if and only if $p = q$.

(b) Use part (a) to show that the $H(p) = -\sum_x p(x) \log p(x)$ satisfies $H(p) \leq \log k$ for all distributions p. When does equality hold?