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

EECS 281A / STAT 241A Statistical Learning Theory

Problem Set 2
Fall 2012

Issued: Tues. Sep. 4, 2012 **Due:** Thurs. Sep. 13, 2012

Reading: Chapters 6 and 8

Problem 2.1
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 $\mathcal{L}(\theta) = \|y - X\theta\|_2^2$ 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 2.2
The course website contains a data set `classification2d.dat` of (x_i, y_i) pairs, where the x_i are 2-dimensional vectors and y_i is a binary label.

(a) Plot the data, using 0's and X's for the two classes. The plots in the following parts should be plotted on top of this plot.

(c) Write a program to fit a logistic regression model using stochastic gradient ascent (or IRLS if you prefer). Plot the line where the logistic function is equal to 0.5.

(d) Fit a linear regression to the problem, treating the class labels as real values 0 and 1. (You can solve the linear regression in any way you'd like, including solving the normal equations, using the LMS algorithm, or calling the built-in routines in Matlab or R). Plot the line where the linear regression function is equal to 0.5.

(e) The data set `testing.dat` is a separate data set generated from the same source. Test your fits from the previous parts on these data and compare the results.

Problem 2.3

The ridge regression estimate is defined as

$$\widehat{\theta} \in \arg\min_{\theta \in \mathbb{R}^d} \left\{ \|y - X\theta\|_2^2 + \lambda_n \|\theta\|_2^2 \right\}$$

where $\lambda_n > 0$ is a positive regularization weight.

(a) Can the ridge regression problem have multiple optimal solutions? Why or why not? Justify your answer.

(b) In a Bayesian model, the parameter θ is viewed as random, and equipped with a prior distribution π. The maximum a posteriori (MAP) estimate is obtained by maximizing the function $f(\theta) := \mathbb{P}(y \mid X, \theta)\, \pi(\theta)$. Explain how the ridge regression estimate can be recovered as a MAP estimate.

(c) Suppose that the matrix X is orthonormal. Give an explicit and easily computed expression for the ridge regression solution as a function (y, X, λ_n).

(d) If we replace the quantity $\|\theta\|_2^2$ with the ℓ_1-norm $\|\theta\|_1 = \sum_{j=1}^d |\theta_j|$, the resulting estimator is known as the Lasso. Assuming that X is orthonormal, give an explicit and easily computed expression for the Lasso solution as a function of (y, X, λ_n).

(e) Based on parts (c) and (d), which estimator (ridge or Lasso) is likely to lead to a sparser solution? Explain. (*Note:* A vector is sparse if it has a relatively small number $s \ll d$ of non-zero components.)

Problem 2.4

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.

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

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

(iii) $\text{Mult}(\theta)$—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.

(iv) the uniform distribution over the interval $[0, \eta]$.

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

(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 2.5

(ML/entropy, conjugacy and duality): Given a function $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$, the dual function is a new function $f^* : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$, defined as follows:

$$f^*(v) \quad = \quad \sup_{u \in \mathbb{R}^n} \big\{ v^T u - f(u) \big\}. \tag{1}$$

(Note that the supremum can be $+\infty$ for some $v \in \mathbb{R}^n$.)

(a) Given the cumulant generating function $A(\theta) = \log[1 + \exp(\theta)]$, for a Bernoulli variable, compute the dual function A^*. What is the link between this computation and maximum likelihood estimation? How is A^* related to binomial entropy? Compute the double dual A^{**}, and verify that $A^{**} = A$. How is computing A^{**} related to maximum entropy?

(b) Using the definition (1), prove that the dual function f^* is always convex: i.e., for all $\lambda \in [0, 1]$, $v, v' \in \mathbb{R}^n$, $f^*\big(\lambda v + (1-\lambda)v'\big) \leq \lambda f^*(v) + (1-\lambda) f^*(v')$.

(c) Given a function f, assume that it is differentiable on $\mathbb{R}^n$, and that it satisfies the duality relation $f^{**} = f$. Use definition (1) for f^* and $f^{**} = f$ to prove that $f(u) \geq f(w) + \nabla f(w)^T(u - w)$ for all $u, w \in \mathbb{R}^n$.

Hint: Each of parts (b) and (c) require proofs, but the arguments need not be very long.

Problem 2.6

Maximum entropy and exponential families For a discrete random variable $X \in \mathcal{X}$ with distribution $p(\cdot)$, the (Boltzmann-Shannon) entropy is given by $H(p) := -\sum_{x \in \mathcal{X}} p(x) \log p(x)$. (We assume that $0 \log 0 = 0$ in this expression). The entropy is a measure of the uncertainty associated with X. Although entropy can be defined more generally, for this problem assume that $|\mathcal{X}|$ is finite.

(a) Suppose that we are given a set of expectation constraints on $p(\cdot)$, say of the form $\sum_{x \in \mathcal{X}} p(x) T_\alpha(x) = \mu_\alpha$ for a collection of functions $\{\phi_1, T_2, \ldots, T_D\}$. (In practice, these constraints would be imposed by making observations.) Consider the maximum entropy problem of maximizing $H(p)$ subject to these expectation constraints, the non-negativity condition $p(x) \geq 0$ for all $x \in \mathcal{X}$, and the normalization constraint $\sum_{x \in \mathcal{X}} p(x) = 1$. Write out the Lagrangian associated with this constrained optimization problem.

(b) By computing stationary points of the Lagrangian, show that the optimal solution $\widehat{p}$ takes the form of an exponential family.