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

EECS 281A / STAT 241A Statistical Learning Theory

Problem Set 2
Fall 2016

Issued: Thurs, September 8, 2016 **Due:** Tuesday, September 20, 2016
Note: Hand in hard copy at the start of class.

Relevant materials: Lectures plus course reader (Chap. 6 and 8).

Problem 2.1
Polynomial regression and model selection: Suppose that you are given a vector of responses $y \in \mathbb{R}^n$ and covariates $t \in \mathbb{R}^n$, and your goal is to fit a polynomial model of degree D to the data.

(a) Show how, with a suitable choice of features, this problem can formulated as a form of linear prediction.

(b) Write a routine to do a least-squares fit of a polynomial function of degree D, including a constant term, to data vectors y and t, each of length n. Using the data in `y.dat` and `t.dat`, fit polynomials of degree $D \in \{1, 2, \ldots, n-1\}$. Letting f_D denote the fitted polynomial of degree D, plot the the mean-squared error $R(D) = \frac{1}{n} \sum_{i=1}^{D} (y_i - f_D(t_i))^2$ versus the degree D.

(c) How does the MSE behave as a function of D, and why? What is special about the degree $n-1$ fit? What happens if you try to fit a polynomial of degree n? (Try to do so using a direct method, such as matrix inverse, to compute the least-squares fit.) Explain why.

(d) Using the new response vector $\widetilde{y} \in \mathbb{R}^n$ given in `yfresh.dat`, compute the average squared error $\widetilde{R}(D) = \frac{1}{n} \sum_{i=1}^{n} (\widetilde{y}_i - f_D(t_i))^2$ of your fits from (b). Why do you think that this plot is *qualitatively* different from the part from part (b)? What does this tell you how the fitted degree D should be chosen? (The problem of choosing the "right" degree is known as the *model selection problem*.)

(e) Using the MSE's $R(D)$ obtained in part (b), compute the adjusted quantities

$$F(D) = R(D) + \frac{\sigma^2 D \log n}{n} \qquad \text{with } \sigma = 0.25.$$

On the same plots, over the range $D \in \{2, \ldots, 9\}$, plot the functions $\widetilde{R}(D)$ and $F(D)$. How are the minimizing arguments of the two functions related? Why is this an interesting observation?

(f) (**BONUS:**) Consider the model selection rule

$$\widehat{D} = \arg \min_{D \in \{1,2,\ldots,n-1\}} F(D).$$

Prove something interesting and non-trivial about this rule as $n \to +\infty$, assuming that the data is drawn i.i.d. from a model of the form $y = f_{D^*}(t) + w$, where f_{D^*} is a fixed polynomial of degree D^* on the interval $[-1, 1]$, the covariate $t \sim \text{Uni}[-1, 1]$, and $w \sim N(0, \sigma^2)$, independent of t.

Problem 2.2

Exponential families and conjugate duality: Recall the one-dimensional exponential family $p_\eta(y) = h(y)e^{\eta y - A(\eta)}$, where $A(\eta) = \log\Big(\int_{\mathcal{Y}} h(y)e^{\eta y} dy\Big)$ (Think of the integral over $\mathcal{Y}$ as a sum for discrete random variables.)

(a) Prove that A is a convex function.

(b) The Kullback-Leibler divergence between distributions p_η and $p_{\widetilde{\eta}}$ is given by

$$D(p_\eta \,\|\, p_{\widetilde{\eta}}) = \mathbb{E}_\eta\Big[\log \frac{p_\eta(Y)}{p_{\widetilde{\eta}}(Y)}\Big],$$

where $\mathbb{E}_\eta$ denotes expectations under p_η. Express the KL divergence in terms of A and its derivative A'.

(c) The conjugate dual of A is defined as $A^*(t) := \sup_{\eta \in \mathbb{R}} \Big\{\eta t - A(\eta)\Big\}$. Compute the conjugate duals for:

   (i) Bernoulli random variable (logistic model)

   (ii) Gaussian random variable (as discussed in class)

   (iii) Poisson random variable

(d) Prove that the conjugate dual A^* is always a convex function.

Problem 2.3

Maximum likelihood and exponential families: Recall the one-dimensional exponential family from the previous problem, and suppose that we are given n i.i.d. samples $\{y_i\}_{i=1}^n$ samples.

(a) Give a simple expression for the maximum likelihood estimate $\widehat{\eta}$ that involves the inverse function of the derivative A' and the sample mean $\frac{1}{n}\sum_{i=1}^{n} y_i$. (The inverse function exists under suitable regularity conditions.)

(b) Use part (a) to compute closed-form estimates for the MLE in the { Gaussian, Poisson, Bernoulli } models.

(c) Suppose that we had an infinite number of samples ($n = \infty$), all drawn from the distribution p_{η^*} where η^* is the unknown true parameter. By taking suitable expectations, show that the MLE is given by

$$\widehat{\eta} = \arg\min_{\eta \in \mathbb{R}} \Big\{ A(\eta) - A(\eta^*) - A'(\eta^*)(\eta - \eta^*) \Big\}.$$

(d) Now assume that A is strictly convex. Prove that $\widehat{\eta} = \eta^*$ in the infinite data limit from (c).

Problem 2.4

Generalized linear models and stochastic gradient:

(a) Write out the stochastic gradient updates (using a single sample per round) for solving the GLM maximum likelihood problem.

(b) Make your updates explicit for the Poisson and logistic cases.

(c) Using the data in `Xone.dat` and `yone.dat`, use your code (with a step size decaying as $1/t$) to fit a logistic model to the data. Compute the probabilities $\mathbb{P}[y_i = 1 \mid x_i; \widehat{\theta}]$ based on your fitted vector $\widehat{\theta}$, and plot a histogram of these probabilities.

(d) Repeat the same for the data in `Xtwo.dat` and `ytwo.dat`. Comment on the differences between the histograms, and what it suggest about the accuracy of the fits.

(e) Via maximum likelihood, fit a 2-component Gaussian mixture model to each of the two data sets. **Note:** You are given the labels in `yone.dat` and `ytwo.dat`.

(f) Examine the connection between the fitted logistic vector $\widehat{\theta}$ from parts (c) and (d), and the fitted mean vectors in part (e). What does this tell you?