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

EECS 281A / STAT 241A Statistical Learning Theory

Problem Set 6

Fall 2012

Issued: Tues. Oct. 30, 2012 **Due:** Tues. Nov. 13, 2012

Reading: Chapters 14, 15, 26; Sampling methods chapter (just after Chap. 20)

Problem 6.1

Factor analysis: Consider the factor analysis model

$$Y = \mu + \Lambda X + W$$

where $X \sim N(0, I)$ is a d-dimensional Gaussian; $\mu \in \mathbb{R}^n$ is a mean vector; $W \sim N(0, \sigma^2 I)$ is an n-dimensional Gaussian; and $\Lambda \in \mathbb{R}^{n \times d}$ is the factor matrix. From the course website, you can download the ASCII format files `Y.dat` and `Lambda.dat`, containing an observation vector $y \in \mathbb{R}^{121}$ and a 121×5 factor matrix Λ (i.e., $d = 5$ and $n = 121$).

(a) Assuming that $\mu = 0$ and $\sigma^2 = 0.25$, compute the conditional mean vector $\mathbb{E}[X|y]$ and covariance matrix $\text{cov}[X|y]$. What does this estimate tell you about which factors were most heavily involved in generating y?

(b) What is the relation between $\mathbb{E}[X|y]$ and the MAP estimate of X given $Y = y$?

(c) *Optional:* How do you think that the MAP estimate of X given $Y = y$ would change if X had i.i.d. Laplacian entries (e.g., with density $p(x_i) \propto \exp(-|x_i|)$)?

Problem 6.2

Model selection for curve-fitting Suppose that we are interested in fitting curves to noisy data; in particular, consider the polynomial regression model linking the response variable $y \in \mathbb{R}$ to the covariate $x \in \mathbb{R}$ via

$$y = \sum_{k=1}^{D} \beta_k x^k + w, \tag{1}$$

where $w \sim N(0,1)$ is Gaussian noise. One model selection problem is that of choosing the appropriate degree D of this polynomial fit, which we explore in this problem.

(a) The course website has two ASCII files `Ymodel.dat` and `Xmodel.dat`, containing samples $\{x_i, y_i\}_{i=1}^n$ with $n = 100$. For $d = 1, 2, \ldots, 10$, fit the model (1) to the data by minimizing the least squares loss $L(\beta) = \frac{1}{2n}\sum_{i=1}^n \left\{y_i - \sum_{k=1}^d \beta_k (x_i)^k\right\}^2$. For which choice of d is this cost function smallest? On the same figure, plot the original data and the models fitted to the data for $d = 1, 2, 3, 4$.

(b) Show that the AIC method, when applied to this problem, reduces to choosing the degree $\widehat{d}$ that minimizes $L(\widehat{\beta}[d]) + d/n$, where $\widehat{\beta}[d]$ is the fitted set of parameters of the polynomial with degree d. Implement this model selection criterion for this data set, where d ranges over $\{1, 2, \ldots, 10\}$. What $\widehat{d}$ is chosen by the procedure?

(c) Given the model $\widehat{\beta} = \widehat{\beta}[\widehat{d}]$ chosen in part (b) and a new observed covariate x, one can generate a predicted response $\hat{y}$ as

$$\hat{y} \quad = \quad \sum_{k=1}^{\widehat{d}} \widehat{\beta}_k x^k.$$

The course website also contains two ASCII files `Ynew.dat` and `Xnew.dat` with $m = 500$ new samples. Using the samples in `Xnew.dat`, generate predictions $\hat{y}_i, i = 1, \ldots, m$, and then compute the prediction error $\sum_{i=1}^m (\hat{y}_i - y_i)^2$.

(d) Repeat part (c) for using the full model fit $\widehat{\beta}[10]$ with all $D = 10$ parameters. Is the prediction error of the full model higher/lower than your fitted model?

Problem 6.3

Accept/reject sampling. Suppose that we want to sample from a random variable X with density

$$p_X(x) \quad = \quad \begin{cases} cx(1-x) & \text{for } x \in [0,1] \\ 0 & \text{otherwise.} \end{cases},$$

where $c > 0$ is an appropriate constant.

(a) Suppose that you have a block-box routine to draw samples Y from a uniform distribution on $[0, 1]$. Describe how to perform accept-reject sampling to generate samples $X \sim p_X$.

(b) Suppose that you standardize your sampler so that, conditioned on the event $Y = 1/2$, it accepts with probability $1/2$. Implement this version of the sampler, and plot the histogram of $n = 10000$ randomly drawn samples. Hand in this histogram, and your code. (Be sure to document your code, explaining what you are doing.)

(c) Let T denote the random number of uniform samples that must be drawn in order for the algorithm to output one sample. Compute $\mathbb{E}[T]$, and compare this theoretical mean to the empirical mean over your $n = 10000$ samples. (Note: Your code will need to record the number of uniform samples that were generated for each of the $n = 10000$ samples in (b)).

Problem 6.4

Cautionary tale about importance sampling: Suppose that we wish to estimate the normalizing constant $Z(p)$ of a Gaussian density $p(\cdot) \sim \mathcal{N}(0, \sigma_p^2)$. Given i.i.d. samples $y_1, \ldots, y_n$ from a standard normal $q(\cdot) \sim \mathcal{N}(0, 1)$, consider the importance sampling estimate

$$\widehat{Z} = \frac{1}{n} \sum_{i=1}^{n} \frac{p^*(y_i)}{q(y_i)} \qquad \text{where} \quad p^*(y) = \exp(-\frac{1}{2\sigma_p^2} y^2).$$

(a) Show that $\widehat{Z}$ is an unbiased estimator of Z_p.

(b) Letting $f(y) = p^*(y)/q(y)$, show that $\text{var}(\widehat{Z}) = \frac{\text{var}(f(Y))}{n}$ whenever $\text{var}(f(Y))$ is finite. For what values of σ_p^2 is this variance actually finite?