Intelligent Data Analysis

Problem Set 2 **Due 3/24/2016**

1. Let μ and σ^2 be the mean and variance of a population, respectively. Show that $s^2 = \frac{1}{N}\sum_{t=1}^{N}\left(x^t - \mu\right)^2$ is an unbiased estimate of the variance for a sample $\chi = \left\{x^t\right\}_{t=1}^{N}$. That is, prove that $E[s^2] = \sigma^2$.

2. Let **A** be an *n* by *n* real symmetric positive definite matrix. Find a unit vector **x** ($\|\mathbf{x}\| = 1$) so as to maximize $\mathbf{x}^T\mathbf{A}\mathbf{x}$. What is the maximal value of $\mathbf{x}^T\mathbf{A}\mathbf{x}$? Hint: use Lagrangian to solve the constrained optimization problem.

3. Derive the formula of $E\left[\theta \mid \chi\right]$ in Chapter 3, pp 24. First, use Bayes' rule to obtain $p\left(\theta \mid \chi\right) = \frac{p(\chi \mid \theta)p(\theta)}{\int p(\chi \mid \phi)p(\phi)d\phi} = \alpha\prod_{t=1}^{N} p(x^t \mid \theta)p(\theta)$, where α is a normalization factor that depends on $\chi = \left\{x^t\right\}_{t=1}^{N}$ but is independent of θ,

$$p\left(x^t \mid \theta\right) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{(x^t-\theta)^2}{2\sigma^2}\right], \text{ and } p\left(\theta\right) = \frac{1}{\sqrt{2\pi}\sigma_0}\exp\left[-\frac{(\theta-\mu_0)^2}{2\sigma_0^2}\right].$$

(a) Show that $p\left(\theta \mid \chi\right) = \beta\exp\left[-\frac{1}{2}\left[\left(\frac{N}{\sigma^2}+\frac{1}{\sigma_0^2}\right)\theta^2 - 2\left(\frac{1}{\sigma^2}\sum_{t=1}^{N}x^t + \frac{\mu_0}{\sigma_0^2}\right)\theta\right]\right]$, where β is some constant. Note that integration is not necessary and you don't need to show what β is.

(b) According to (a), show that $p\left(\theta \mid \chi\right)$ is a normal density given by

$$p\left(\theta \mid \chi\right) = \frac{1}{\sqrt{2\pi}\sigma_N}\exp\left[-\frac{(\theta-\mu_N)^2}{2\sigma_N^2}\right], \text{ in which}$$

$$\frac{1}{\sigma_N^2} = \frac{N}{\sigma^2}+\frac{1}{\sigma_0^2} \text{ and } \frac{\mu_N}{\sigma_N^2} = \frac{N}{\sigma^2}m + \frac{\mu_0}{\sigma_0^2}, \text{ where } m = \frac{1}{N}\sum_{t=1}^{N}x^t.$$

(c) From (b), solve μ_N and σ_N^2 to obtain

$$\mu_N = E[\theta \mid \chi] = \frac{\frac{N}{\sigma^2}}{\frac{N}{\sigma^2}+\frac{1}{\sigma_0^2}} m + \frac{\frac{1}{\sigma_0^2}}{\frac{N}{\sigma^2}+\frac{1}{\sigma_0^2}} \mu_0 \text{ and } \sigma_N^2 = \frac{\sigma_0^2\sigma^2}{N\sigma_0^2+\sigma^2}.$$

4. Consider the error function of the ridge regression as follows:

$$E(w_2, w_1, w_0) = \sum_{t=1}^{N} \left(r^t - w_2 (x^t)^2 - w_1 x^t - w_0 \right)^2 + \lambda (w_2^2 + w_1^2),$$

where λ is a positive number. Derive the optimal condition in matrix form **Aw**=**b** that minimizes *E*. Note that **A** is a 3 by 3 matrix and $\mathbf{w} = (w_2, w_1, w_0)^T$.

5. Program Assignment This goal of this assignment is study model selection for a regression problem. Suppose one sample is described by $\chi = \{x^t, r^t\}_{t=1}^N$, in which $r^t = f(x^t) + \varepsilon^t$, where *f* is a deterministic function and $\varepsilon^t \sim N(0,1)$. Function *f* is given by $f(x) = 3\sin(3.14x) + 4$, where *x* is randomly drawn from [0,1].

   Generate your own data set.

   (a) Five samples are taken, each containing 20 cases. Note that the five samples have the same set of inputs $\{x^t\}_{t=1}^{20}$, but the corresponding responses might be different. Plot a sample of data along with *f*, as shown in Chapter 4, pp 17 (a). Plot five polynomial fits, namely, $g_i(\cdot)$, of order 1, 3, and 5. For each case, plot the average of the five fits, namely, $\bar{g}(\cdot)$. See Chapter 4, pp 17, dotted lines in (b), (c), (d).

   (b) In the same setting as that of (a), using one hundred models instead of five, plot bias^2, variance, and error for polynomials of order 1 to 5. See Chapter 4, pp 15 and 20.