

**Homework Set Four**  
ECE 285  
Department of Computer and Electrical Engineering  
University of California, San Diego

Nuno Vasconcelos

Fall 2004

**Due October 28, 2004**

**1. Bayesian regression:** in last week's problem set we showed that various forms of linear regression by the method of least squares are really just particular cases of ML estimation under the model

$$\mathbf{z} = \Phi\theta + \epsilon$$

where  $\mathbf{z} = (z_1, \dots, z_n)^T$ ,  $\theta = (\theta_1, \dots, \theta_k)^T$

$$\Phi = \begin{bmatrix} 1 & \dots & x_1^K \\ \vdots & & \vdots \\ 1 & \dots & x_n^K \end{bmatrix}$$

and  $\epsilon = (\epsilon_1, \dots, \epsilon_n)^T$  is a normal random process  $\epsilon \sim \mathcal{N}(\mathbf{0}, \Sigma)$ . It seems only natural to consider the Bayesian extension of this model, an extension that has been the subject of some recent research under the denomination of *Gaussian processes*. For this, we simply extend the model considering a Gaussian prior

$$P_\theta(\theta) = \mathcal{G}(\theta, \mathbf{0}, \Gamma).$$

**a)** Given a training set  $\mathcal{D} = \{(\mathcal{D}_x, \mathcal{D}_z)\} = \{(x_1, z_1), \dots, (x_n, z_n)\}$ , compute the posterior distribution

$$P_{\theta|\mathbf{T}}(\theta|\mathcal{D})$$

and the predictive distribution

$$P_{z|\theta, x}(z|\theta, x).$$

**b)** Consider the MAP estimate

$$\theta_{MAP} = \arg \max_{\theta} P_{\theta|\mathbf{T}}(\theta|\mathcal{D}).$$

How does it differ from the weighted least squares estimate? What is the role of the terms that were not present in the latter? Is there any advantage in setting them to anything other than zero?

**c)** Consider the case in which prior covariance  $\Gamma$  is a diagonal matrix, not necessarily the identity. Suppose that you are told that  $K$ , i.e. the number of parameters in  $\theta$  or the degree of the polynomial  $\phi(x)^T\theta$ , is somewhere between 1 and 25. How would you set up  $\Gamma$  and why? Discuss the implications of your selection on the bias and variance of your MAP solution

$$z_{MAP} = \Phi(x)\theta_{MAP}.$$

**2.** Problem 3.5.20 in DHS

**3.** In this problem we will explore the unintuitive nature of high-dimensional spaces. In **a)** we study the volume of a sphere, while in **b)** we consider the probability mass of a Gaussian.

**a-i)** Consider a space of dimension  $n$ . It should not be very hard to convince yourself that the volume of an hypersphere of radius  $r$ , i.e. the set of points

$$\{\mathbf{x} | x_1^2 + \dots + x_n^2 \leq r^2\},$$

is of the form

$$V_n(r) = C_n r^n$$

where  $C_n$  is a constant that depends on  $n$ . Show that

$$C_n = \frac{\pi^{n/2}}{\Gamma(\frac{n}{2} + 1)}$$

where

$$\Gamma(n) = \int_0^\infty e^{-x} x^{n-1} dx.$$

Hint: you might want to consider that

$$\int e^{-r^2} d\mathbf{x} = \int e^{-r^2} \frac{dV_n(r)}{dr} dr.$$

**a-ii)** Show that

$$C_n = \frac{2\pi}{n} C_{n-2},$$

and determine  $C_1$  and  $C_2$ . Plot  $C_n$  for  $n = 1, \dots, 20$ . Comment your results, and how they may differ from what would be intuitively expected.

**a-iii)** Compute  $C_n$  when  $n = 2k$ , for an integer  $k$ . Find the value the  $k$  for which the volume of the sphere starts to decrease.

**b-i)** Consider a  $n$ -dimensional Gaussian distribution of zero mean and identity covariance

$$\mathcal{G}(\mathbf{x}) = \frac{1}{(2\pi)^{n/2}} e^{-\frac{1}{2}\mathbf{x}^T \mathbf{x}}.$$

Show that the point where the probability density is largest is  $\mathbf{x} = 0$ .

**b-ii)** Consider the hypersphere

$$S_{0.01}(\mathbf{x}) = \left\{ \mathbf{x} \mid \frac{\mathcal{G}(\mathbf{x})}{\mathcal{G}(\mathbf{0})} \leq \frac{1}{100} \right\}$$

where the probability drops to 0.01 of the maximum value. Show that the probability mass in the tails of the Gaussian, i.e. *outside* this sphere is

$$P_n = P[\chi^2(n) \geq 9.21]$$

where  $\chi^2(n)$  is the chi-squared distribution with  $n$  degrees of freedom. Plot this probability for  $n = 1, \dots, 20$ . What happens to the probability mass of the Gaussian as the dimension of the space increases?

**b-iii)** Consider an iid sample  $\mathbf{x}_i$  drawn from this Gaussian. If the sample has size  $k$  what is the expected number of points that will fall inside  $S_{0.01}$  when  $n = 8$  and  $n = 64$ ? Explain how this may affect the classification results of the computer problem in problem set **2**.

4. Extensive experimental evidence in the area of image compression has shown that the discrete cosine transform (DCT) of image patches is a very good approximation to their PCA. In homework set 2, problem 7, we saw that all but one of the DCT coefficients (features) have zero mean, and only one has non-zero mean. The latter is the so-called DC coefficient, because it results from projecting the image patch into the vector  $\mathbf{1} = (1, 1, \dots, 1)^T$  and, therefore, is proportional to the average (DC) value of the patch. In this problem we are going to explore the connection between the DCT and PCA to explain this fact. For this, we are going to assume that:

- an image patch is a collection of random variables  $\mathbf{X} = \{x_1, \dots, x_n\}$  which are identically distributed

$$P_{X_i}(x) = f(x), \forall i \in \{1, \dots, 64\}$$

where  $f(x)$  is a common probability density function.

- the pixels in the image patch are correlated, according to the *correlation coefficient*

$$\rho_{ij} = \frac{E[X_i X_j]}{\sqrt{E[X_i^2]} \sqrt{E[X_j^2]}}.$$

This obviously implies that we do not have an iid sample.

a) Consider the PCA of  $\mathbf{X}$ . Show that it is not affected by a change of variables of the type  $\mathbf{Z} = \mathbf{X} - \mu_x$  where  $\mu_x = E[\mathbf{X}]$ .

b) Given a), we can assume that  $\mathbf{X}$  has zero mean. We will do so for the remainder of the problem. Show that in the extreme of highly correlated pixel values, i.e. when

$$\rho_{ij} \rightarrow 1, \quad \forall i, j$$

the vector  $\mathbf{1}$  is the largest principal component. Since, neighboring image pixels do tend to be highly correlated, this helps explain why the DC coefficient is always present.

c) Let  $\Phi$  be the matrix whose columns  $\phi_i$  are the principal components and consider the set of coefficients (features)  $\mathbf{Z}$  resulting from the projection of an arbitrary image patch  $\mathbf{X}$  into these components, i.e.

$$\mathbf{z} = \Phi^T \mathbf{x}.$$

Consider the DCT coefficients  $z_i = \phi_i^T \mathbf{x}$ , noting that  $\mathbf{z}_1 = \mathbf{1}^T \mathbf{x}$  is the DC coefficient. Show that

$$E[z_i] = 0, \quad \forall i > 1,$$

i.e. that the remaining (AC) coefficients have zero mean.

5. Finish up the computer problem of assignment 3.