

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

Nuno Vasconcelos

Fall 2004

**Due October 21, 2004**

1. In this problem we will consider the issue of linear regression and the connections between maximum likelihood and least squares solutions. Consider a problem where we have two random variables  $Z$  and  $\mathbf{X}$ , such that

$$z = f(\mathbf{x}, \theta) + \epsilon, \quad (1)$$

where  $f$  is a polynomial with parameter vector  $\theta$

$$f(\mathbf{x}, \theta) = \sum_{k=0}^K \theta_k x^k \quad (2)$$

and  $\epsilon$  a Gaussian random variable of zero mean and variance  $\sigma^2$ . Our goal is to estimate the best estimate of the function given an iid sample  $\mathcal{D} = \{(\mathcal{D}_x, \mathcal{D}_z)\} = \{(x_1, z_1), \dots, (x_n, z_n)\}$ .

a) Formulate the problem as one of least squares, i.e define  $\mathbf{z} = (z_1, \dots, z_n)^T$ ,

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

and find the value of  $\theta$  that minimizes

$$\|\mathbf{z} - \Phi\theta\|^2.$$

b) Formulate the problem as one of ML estimation, i.e. write down the likelihood function  $P_{Z|X}(z|x; \theta)$ , and compute the ML estimate, i.e. the value of  $\theta$  that maximizes  $P_{Z|X}(\mathcal{D}_z|\mathcal{D}_x; \theta)$ . Show that this is equivalent to a).

c) (The advantage of the statistical formulation is that makes the assumptions explicit. We will now challenge some of these.) Assume that instead of a fixed variance  $\sigma^2$  we now have a variance that depends on the sample point, i.e.

$$z_i = f(\mathbf{x}_i, \theta) + \epsilon_i, \quad (3)$$

where  $\epsilon_i \sim \mathcal{N}(0, \sigma_i^2)$ . This means that our sample is independent but no longer identically distributed. It also means that we have different degrees of confidence in the different measurements  $(z_i, x_i)$ . Redo b) under these conditions.

d) Consider the weighted least squares problem where the goal is to minimize

$$(\mathbf{z} - \Phi\theta)^T \mathbf{W} (\mathbf{z} - \Phi\theta)$$

where  $\mathbf{W}$  is a symmetric matrix. Compute the optimal  $\theta$  for this situation. What is the equivalent maximum likelihood problem? Rewrite the model (1), making explicit all the assumptions that lead to the new problem. What is the statistical interpretation of  $\mathbf{W}$ ?

e) The  $L_2$  norm is known to be prone to large estimation error if there are *outliers* in the training sample. These are training examples  $(z_i, x_i)$  for which, due to measurement errors or other extraneous causes,  $|z_i - \sum \theta_k x_i^k|$  is much larger than for the remaining examples (the *inliers*). In fact, it is known that a single outlier can completely derail the least squares solution, an highly undesirable behavior. It is also well known that other norms lead to much more robust estimators. One of such distance metrics is the  $L_1$ -norm

$$L_1 = \sum_i |z_i - \sum_k \theta_k x_i^k|.$$

In the maximum likelihood framework, which is the statistical assumption that leads to the  $L_1$  norm? Once again, rewrite the model (1), making explicit all the assumptions that lead to the new problem. Can you justify why this alternative formulation is more robust? In particular, provide a justification for **i)** why the  $L_1$  norm is more robust to outliers, and **ii)** the associated statistical model (1) copes better with them.

2. In this problem we consider two popular estimators, the best linear unbiased estimator (BLUE) and the best linear mean squared error estimator (BLMSEE), and their relationship to the maximum likelihood estimator. In all cases we assume an iid sample  $\mathcal{D} = \{x_1, \dots, x_n\}$  drawn from a random variable  $x$  of mean  $\mu$  and variance  $\sigma^2$  (which does not have to be Gaussian).

a) Consider the set of linear estimators, i.e. estimators of the form

$$\hat{\mu} = \sum_{i=1}^n w_i x_i \tag{4}$$

where  $w_i \in R$ . Find the set of  $w_i$  that minimizes the variance of  $\hat{\mu}$ . What is the bias of this estimator? Is this a good estimator? Why?

b) Consider the set of linear unbiased estimators. What is the BLUE, i.e. estimator in this class that has minimum mean squared error? What is its variance?

c) What is the BLMSEE, i.e. the estimator of the form of (4) that achieves the minimum mean squared error in the absence of constraints on the bias? What are its bias and variance? (Hint: One possibility to solve this problem is to first find the estimator of smallest MSE given a constraint on the bias, e.g.

$$bias(\hat{\mu}) = \alpha \tag{5}$$

and then minimize over  $\alpha$  to obtain the best overall estimator. This will also help you on the next question.)

d) From a), b), and c), what can you say about the bias/variance trade-off of the linear estimator (4)?

e) Assuming that  $x$  is Gaussian, compare the different estimators above to the maximum likelihood estimator. What are the advantages and disadvantages of each?

f) Repeat e) assuming that  $x$  is a Bernoulli random variable with parameter  $p$ , i.e.  $x \in \{0, 1\}$  with

$$P_X(x) = p^x(1-p)^{1-x}.$$

g) Repeat e) assuming that  $x$  is an exponential random variable with parameter  $\lambda$ , i.e.

$$P_X(x) = \lambda e^{-\lambda x}.$$

**3.**

a) Problem 3.5.17 in DHS.

b) What is the ML estimate for  $\theta$  in this problem? What is the MAP estimate for  $\theta$  in this problem? Do you see any advantage in favoring one of the estimates in favor of the others? How does that relate to the uniform prior that was assumed for  $\theta$ ?

**4.** Consider problem **3** of the previous assignment, i.e. a random variable  $X$  such that  $P_X(k) = \pi_k$ ,  $k \in \{1, \dots, N\}$ ,  $n$  independent observations from  $X$ , a random vector  $\mathbf{C} = (C_1, \dots, C_N)^T$  where  $C_k$  is the number of times that the observed value is  $k$  (i.e.  $\mathbf{C}$  is the histogram of the sample of observations). We have seen that,  $\mathbf{C}$  has multinomial distribution

$$P_{C_1, \dots, C_N}(c_1, \dots, c_N) = \frac{n!}{\prod_{k=1}^N c_k!} \prod_{j=1}^N \pi_j^{c_j}.$$

In this problem we are going to compute MAP estimates for this model. Notice that the parameters are probabilities and, therefore, not every prior will be acceptable here (since  $\pi_j > 0$  and  $\sum_j \pi_j = 1$  for the prior to be valid). One distribution over vectors  $\pi = (\pi_1, \dots, \pi_N)^T$  that satisfies this constraint is the Dirichlet distribution

$$P_{\Pi_1, \dots, \Pi_N}(\pi_1, \dots, \pi_N) = \frac{\Gamma(\sum_{j=1}^N u_j)}{\prod_{j=1}^N \Gamma(u_j)} \prod_{j=1}^N \pi_j^{u_j - 1}$$

where the  $u_j$  are a set of *hyperparameters* (parameters of the prior) and

$$\Gamma(x) = \int_0^\infty e^{-t} t^{x-1} dt$$

the Gamma function.

a) Derive the MAP estimator for the parameters  $\pi_i$ ,  $i = 1, \dots, N$  using the Dirichlet prior.

b) Compare this estimator with the ML estimator derived in the previous assignment. What is the use of this prior equivalent to, in terms of the ML solution?

c) What is the effect of the prior as the number of samples  $n$  increases? Does this make intuitive sense?

d) In this problem and problem **3** we have seen two ways of avoiding the computational complexity of computing a fully Bayesian solution: i) to rely on a *non-informative* prior, and ii) to rely on an informative prior and compute the MAP solution. Qualitatively, what do you have to say about the results obtained with the two solutions? What does this tell you about the robustness of the Bayesian framework?

5. In this problem we explore the exponential family and conjugate priors. The exponential family is the family of densities of the form

$$P_{\mathbf{X}|\theta}(\mathbf{x}|\theta) = f(\mathbf{x})g(\theta)e^{\phi(\theta)^T u(\mathbf{x})}$$

with

$$[g(\theta)]^{-1} = \int f(\mathbf{x})e^{\phi(\theta)^T u(\mathbf{x})} d\mathbf{x}.$$

a) Show that, for a density in this family, the likelihood of a sequence  $\mathcal{D} = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$  is

$$P_{\mathbf{T}|\theta}(\mathcal{D}|\theta) \propto \prod_{i=1}^n f(\mathbf{x}_i) \exp \left\{ \phi(\theta)^T \sum_{i=1}^n u(\mathbf{x}_i) \right\}.$$

What is the normalization constant?

b) It has been shown that, apart from certain irregular cases, the exponential family is the only family of distributions for which there is a conjugate prior. Show that

$$P_{\theta}(\theta) = \frac{g(\theta)^n e^{\phi(\theta)^T \nu}}{\int g(\theta)^n e^{\phi(\theta)^T \nu} d\theta}$$

is a conjugate prior for the exponential family and compute the posterior distribution  $P_{\theta|\mathbf{T}}(\theta|\mathcal{D})$ . Denoting  $\mathbf{s} = \sum_{i=1}^n u(\mathbf{x}_i)$  as *the sufficient statistic*, compare the posterior with prior density. What is the result of “propagating” the prior through the likelihood function?

c) Consider the following table. For each row **i**) show that the likelihood function on the right column belongs to the exponential family, **ii**) show that the prior on the left column is a conjugate prior for the likelihood function on the right column, **iii**) compute the posterior  $P_{\theta|\mathbf{T}}(\theta|\mathcal{D})$ , and **iv**) interpret the meaning of the sufficient statistic and the “propagation” discussed in **b**).

| Likelihood                              | $P_{\mathbf{T} \theta}(\mathcal{D} \theta)$                                       | Prior        | $P_{\theta}(\theta)$  |
|---|---|--------------|---|
| <b>Bernoulli</b>                        | $\prod_{i=1}^n \theta^{x_i} (1 - \theta)^{1-x_i}$                                 | <b>Beta</b>  | $P_{\theta}(\theta; \alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1} (1 - \theta)^{\beta-1}$ |
| <b>Poisson</b>                          | $\prod_{i=1}^n \frac{e^{-\theta} \theta^{x_i}}{x_i!}$                             | <b>Gamma</b> | $P_{\theta}(\theta; \alpha, \beta) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} \theta^{\alpha-1} e^{-\beta\theta}$                          |
| <b>Exponential</b>                      | $\prod_{i=1}^n \theta e^{-\theta x_i}$  | <b>Gamma</b> | $P_{\theta}(\theta; \alpha, \beta) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} \theta^{\alpha-1} e^{-\beta\theta}$                          |
| <b>Normal</b> ( $\theta = 1/\sigma^2$ ) | $\prod_{i=1}^n \sqrt{\frac{\theta}{2\pi}} \exp\{-\frac{\theta}{2}(x_i - \mu)^2\}$ | <b>Gamma</b> | $P_{\theta}(\theta; \alpha, \beta) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} \theta^{\alpha-1} e^{-\beta\theta}$                          |

Table 1: In the case of the normal distribution,  $\mu$  is assumed known, the parameter is the precision  $\theta = 1/\sigma^2$ .

d) Repeat the steps of **c**) for the distributions of problem 4., i.e. the **multinomial** as the likelihood function and the **Dirichlet** as the prior.

6. **(computer)** (Note: This week’s computer problem requires more computation time than the previous ones - about 10h by our estimates. For that reason, I will give you two weeks to finish it up, i.e. it will be repeated in homework 4. Do not, however, take this as a reason to not start working on it right away. You may simply not have time to finish if leave it to the days prior to the deadline.)

This week we will continue trying to classify our cheetah example. Once again we use the decomposition into  $8 \times 8$  image blocks, compute the DCT of each block, and zig-zag scan. We also continue to assume that the class-conditional densities are multivariate Gaussians of 64 dimensions. The goal is to understand the benefits of a Bayesian solution. For this, using the training data in `TrainingSamplesDCT_new_8.mat` we created 4 datasets of size given by the table below. They are available in the file `TrainingSamplesDCT_subsets_8.mat`

| Dataset         | <i>cheetah</i> examples | <i>grass</i> examples |
|-----------------|-------------------------|-----------------------|
| $\mathcal{D}_1$ | 75                      | 300                   |
| $\mathcal{D}_2$ | 125                     | 500                   |
| $\mathcal{D}_3$ | 175                     | 700                   |
| $\mathcal{D}_4$ | 225                     | 900                   |

We start by setting up the Bayesian model. To simplify things a bit we are going to cheat a little. With respect to the class-conditional,

$$P_{\mathbf{x}|\mu, \Sigma} = \mathcal{G}(\mathbf{x}, \mu, \Sigma).$$

we assume that we know the covariance matrix (like Bayes might) but simply replace it by the sample covariance of the training set,  $\mathcal{D}$ , that we are working with (and hope he doesn't notice)<sup>1</sup>. That is, we use

$$\Sigma = \frac{1}{N} \sum_{i=1}^N \left( \mathbf{x}_i - \frac{1}{N} \sum_{i=1}^N \mathbf{x}_i \right) \left( \mathbf{x}_i - \frac{1}{N} \sum_{i=1}^N \mathbf{x}_i \right)^T$$

We are, however, going to assume unknown mean and a Gaussian prior of mean  $\mu_0$  and covariance  $\Sigma_0$

$$P_{\mu}(\mu) = \mathcal{G}(\mu, \mu_0, \Sigma_0).$$

Regarding the mean  $\mu_0$ , we assume that it is zero for all coefficients other than the first (DC) while for the DC we consider two different strategies:

- *strategy 1*:  $\mu_0$  is smaller for the (darker) *cheetah* class ( $\mu_0 = 1$ ) and larger for the (lighter) *grass* class ( $\mu_0 = 3$ ).
- *strategy 2*:  $\mu_0$  is equal to half the range of amplitudes of the DCT coefficient for both classes ( $\mu_0 = 2$ );

For the covariance  $\Sigma_0$  we assume a diagonal matrix with  $(\Sigma_0)_{ii} = \alpha w_i$ . The mean  $\mu_0$  (for the two strategies) and the weights  $w_i$  are given in the files `Prior_1.mat` (strategy 1) and `Prior_2.mat` (strategy 2).

**a)** Consider the training set  $\mathcal{D}_1$  and strategy 1 (we will use this strategy until **d**). For each class, compute the covariance  $\Sigma$  of the class-conditional, and the posterior mean  $\mu_1$ , and covariance  $\Sigma_1$  of

$$P_{\mu|\mathbf{T}}(\mu|\mathcal{D}_1) = \mathcal{G}(\mu, \mu_1, \Sigma_1).$$

Next, compute the parameters of the predictive distribution

$$P_{\mathbf{x}|\mathbf{T}}(\mathbf{x}|\mathcal{D}_1)$$

---

<sup>1</sup>There are appropriate ways to set up covariance priors but things would get a little more complex, and we avoid them.

for each of the classes. Then, using ML estimates for the class priors, plug into the Bayesian decision rule, classify the cheetah image and measure the probability of error. All of the parameters above are functions of  $\alpha$ . Repeat the procedure for the values of  $\alpha$  given in the file `Alpha.mat`. Plot the curve of the probability of error as a function of  $\alpha$ . Can you explain the results?

**b)** For  $\mathcal{D}_1$ , compute the probability of error of the ML procedure identical to what we have used last week. Compare with the results of **a)**. Can you explain? See “what to hand in” below.

**c)** Repeat **a)** with the MAP estimate of  $\mu$ , i.e. using

$$P_{\mathbf{x}|\mathbf{T}}(\mathbf{x}|\mathcal{D}_1) = P_{\mathbf{x}|\mu}(\mathbf{x}|\mu_{MAP})$$

where

$$\mu_{MAP} = \arg \max_{\mu} P_{\mu|\mathbf{T}}(\mu|\mathcal{D}_1)$$

Compare the curve with those obtained above. Can you explain the results? See “what to hand in” below.

**d)** Repeat **a)** to **c)** for each of the datasets  $\mathcal{D}_i, i = 2, \dots, 4$ . Can you explain the results? See “what to hand in” below.

**e)** Repeat **a)** to **d)** under strategy 2 for the selection of the prior parameters. Comment the differences between the results obtained with the two strategies. See “what to hand in” below.

### Some Tips:

1. The prior for  $\mu$ 's are different for background and foreground. So in each file the  $\mu$ 's are in 'mu0\_BG' and 'mu0\_FG' for background and foreground respectively.
2. When displaying the curves of PE vs.  $\alpha$ , it is better to use 'log' mode in the  $\alpha$  axis. There are two commands in MATLAB that you should use, assuming the  $x$ -axis represents  $\alpha$ : 1) `set(gca, 'XScale', 'log')`; or, 2) `semilogX(...)`. Please refer to MATLAB help for details.
3. Remember to use a fast Gaussian evaluation inside the loop.

**What to hand in:** To minimize the number of pages on your report I would advise on handing in, for each dataset and each strategy, a single plot with the curves of classification error as a function of  $\alpha$  for: 1) solution based on the predictive equation, 2) MAP solution, and 3) ML solution. The latter will obviously be a constant (horizontal) line. Try to explain 1) the relative behavior of these three curves, 2) how that behavior changes from dataset to dataset, and 3) how all of the above change when strategy 1 is replaced by strategy 2.