ELL784: Problem Set 2

September 16, 2022

1. Suppose we have observed a set of data points $x_1, x_2, ..., x_N$ drawn from a univariate Gaussian distribution with mean μ and variance σ^2. Write down the likelihood function for this data. By differentiating the log likelihood with respect to μ and σ^2, derive their respective maximum likelihood estimates. Also obtain the expected values of these estimates (see Bishop, Exercise 1.12).

2. A biased coin with probability of heads given by θ is tossed N times, and M heads are observed. (a) Write down the likelihood function, and obtain the maximum likelihood estimate for θ, as a function of N and M. (b) Now assume that the prior distribution on θ is given by the $Beta(2,2)$ distribution:

$$p(\theta) = 6\theta(1-\theta); 0 \leq \theta \leq 1 \tag{1}$$

   Use a Bayesian formulation to obtain the posterior distribution for θ as a function of N and M. Derive the maximum a posteriori (MAP) estimate for θ. How does this differ from the maximum likelihood estimate? What intuition does this give you about the choice of prior we used?

3. Suppose I have data points $\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N$, where each point is a vector, so for instance $\mathbf{x}_1 = (x_{11}, x_{12}, ..., x_{1D})$, where D is the dimensionality of the *input space*. Now suppose I use a set of M basis functions, $\phi_1(.), \phi_2(.), ..., \phi_M(.)$ to map my data points into a new M-dimensional *feature space*. My $N \times M$ design matrix in this feature space is given by:

$$\mathbf{\Phi} = \begin{pmatrix} \phi_1(\mathbf{x}_1) & \phi_2(\mathbf{x}_1) & ... & \phi_M(\mathbf{x}_1) \\ \phi_1(\mathbf{x}_2) & \phi_2(\mathbf{x}_2) & ... & \phi_M(\mathbf{x}_2) \\ ... & & & \\ \phi_1(\mathbf{x}_N) & \phi_2(\mathbf{x}_N) & ... & \phi_M(\mathbf{x}_N) \end{pmatrix} \tag{2}$$

   Also suppose that the target or output value for the n^{th} data point is given by t_n, and define $\mathbf{t} = (t_1, t_2, ..., t_N)^{\mathrm{T}}$. Show that the matrix $\mathbf{\Phi}(\mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi})^{-1}\mathbf{\Phi}^{\mathrm{T}}$ (the product of $\mathbf{\Phi}$ and its *Moore-Penrose pseudo-inverse*) orthogonally projects the vector $\mathbf{t}$ onto the space spanned by the columns of $\mathbf{\Phi}$. (This corresponds to algebraically proving the geometrical interpretation of the least-squares solution, which was discussed in class.)

4. Consider a data set containing K classes of data, denoted $C_1, C_2, ..., C_K$. Suppose I am given a $K \times K$ *loss matrix* L, such that L_{ij} denotes the loss incurred in classifying an object of class C_i into class C_j. Also suppose that I have the *reject option*, i.e., for some objects my classifier may refuse to classify them, and in such a case the loss incurred is λ. (a) Obtain a formulation for the optimal decision criterion, i.e., the one which minimises the expected loss, as a function of the class posterior probabilities from the classifier, L, and λ. (b) Suppose my loss matrix is given by $L_{ii} = 0; L_{ij} = 1; 1 \leq i, j \leq K; i \neq j$. Show that in this case the criterion reduces to a simple rejection threshold θ (see Bishop Figure 1.26). How does θ relate to λ?