

8 Foundations of Data Science (DJW)

Fisher’s Iris dataset contains, among other things, measurements of `Petal.Length` and `Sepal.Length` for samples from each of three species of iris. Suppose we want to fit the model

$$\text{Petal.Length} = \alpha_s + \beta_s \text{Sepal.Length} + \text{Normal}(0, \sigma^2)$$

where s is the species.

Note: The $\text{Normal}(\mu, \sigma^2)$ distribution has density function

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/(2\sigma^2)}.$$

- (a) Explain what is meant by ‘linear model’, ‘feature’, and ‘orthogonal projection’. Rewrite the above model as a linear model made up of linearly independent features, and explain why they are linearly independent. [7 marks]
- (b) You are given a library function `proj(y, [e1, ..., en])`. It returns a list $[\lambda_1, \dots, \lambda_n]$ such that $\lambda_1 e_1 + \dots + \lambda_n e_n$ is the orthogonal projection of the vector y onto the subspace spanned by vectors $\{e_1, \dots, e_n\}$. Explain what is meant by the ‘least squares method’, and give pseudocode using `proj` to find the least squares estimators for α_s and β_s . [2 marks]
- (c) Explain how to compute the maximum likelihood estimators of α_s , β_s , and σ . In your answer, you should explain the relationship between the least squares method and maximum likelihood estimation. [5 marks]
- (d) We wish to know whether the β_s coefficients for the three species are noticeably different. Outline the Bayesian approach to answering this question. [6 marks]