9 Machine Learning and Bayesian Inference (sbh11)

(*a*) You have a training set $\mathbf{s} = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_m, y_m)\}$ and wish to fit a function to it using linear regression based on the function

$$f_{\mathbf{w}}(\mathbf{x}) = w_0 + \sum_{i=1}^{n} w_i x_i.$$

You have reason to believe that the zero-mean additive noise in your training data has a Laplace density

$$p_{\mu,b}(\epsilon) = \frac{1}{2b} \exp\left(\frac{-|\epsilon - \mu|}{b}\right)$$

where the mean μ and b are parameters. Provide a full derivation of the *Maximum Likelihood* training algorithm for this scenario. Clearly state any assumptions you make. You may make use of the fact that if ϵ has Laplace density with parameters μ, b then $\epsilon + a$ has Laplace density with parameters $\mu + a, b$. [8 marks]

(*b*) You have reason to believe that the Dirichlet density is a suitable prior on the weight vector $\mathbf{w}$, so

$$p_{\boldsymbol{\alpha}}(\mathbf{w}) = A(\boldsymbol{\alpha}) \prod_{i=0}^{n} w_i^{\alpha_i - 1}$$

where $\boldsymbol{\alpha}$ is a vector of parameters and $A(\boldsymbol{\alpha})$ is a suitable function. Explain in detail how the error derived in Part (*a*) needs to be modified to obtain a *Maximum A Posteriori* training algorithm. [5 marks]

(*c*) Which of the parameters introduced in Parts (*a*) and (*b*), excluding $\mathbf{w}$, need to be chosen as part of the training process? [1 mark]

(*d*) Describe two ways in which the task in Part (*c*) can be achieved, and explain why the error measures derived should not be optimised in a single step with respect to both the parameters and $\mathbf{w}$ simultaneously. [6 marks]