8 Machine Learning and Bayesian Inference (jt796)

(*a*) Consider the **support vector machine** (SVM) for inputs $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$, $(\mathbf{x}, y) \in \mathcal{X} \times \{0, 1\}$, $\mathcal{X} \subset \mathbb{R}^d$. Let $(\mathbf{w}, b) \in \mathbb{R}^d \times \mathbb{R}$ denote the parameters which define the maximum-margin hyperplane returned by the SVM.

(*i*) The SVM classification function is given by the maximum-margin hyperplane:

$$f(\mathbf{x}) = \operatorname{sgn}(\langle \mathbf{w}, \mathbf{x} \rangle + b).$$

Express $\mathbf{w}$ in terms of the dual variables $\{\lambda_i\}_{i=1}^n$ associated with the margin violation constraints. Hence rewrite $f(\mathbf{x})$ in terms of λ_i and interpret how the hypothesis function classifies an unseen point $\mathbf{x}^*$. [4 marks]

(*ii*) What property of the hypothesis function allows the extension of the SVM to define a nonlinear decision boundary in the feature space $\mathcal{X}$? State the nonlinear extension of $f(\mathbf{x})$ and explain how this may improve classification performance. [4 marks]

(*iii*) Consider the SVM primal objective, where $C > 0$ is a constant:

$$\operatorname*{argmin}_{\mathbf{w}, \boldsymbol{\xi}} \left(\frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^n \xi_i \right).$$

Let $K_{ij} = \kappa(\mathbf{x}_i, \mathbf{x}_j)$ be the Gram matrix evaluated on training points, where $\kappa : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a valid kernel function. Assume we have subsumed the bias b into the kernel. Write the kernelised SVM objective in terms of $K, \boldsymbol{\alpha}$ and $\boldsymbol{\xi}$, where $\boldsymbol{\alpha}_i = \lambda_i y_i$. [3 marks]

(*iv*) Rewrite the SVM objective in terms of the hypothesis function applied to each example, $\mathbf{f} = (f(\mathbf{x}_1), \ldots, f(\mathbf{x}_n))^T$. [*Hint:* Relate $\mathbf{f}$ and $\boldsymbol{\alpha}$.] [3 marks]

(*b*) Consider modelling the data previously given as an underlying function contaminated with additive Gaussian noise, $y_i = f(\mathbf{x}_i) + \epsilon_i$, where $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$. Model the function f as a Gaussian process with zero mean, and covariance function $\kappa : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, i.e. $f \sim \mathsf{GP}(0, \kappa)$.
[*Hint:* Recall for a d-dimensional normally distributed random variable $\mathbf{z} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$, $p(\mathbf{z}) = \left((2\pi)^d \det \Sigma\right)^{-1/2} \exp\left(-\frac{1}{2}(\mathbf{z} - \boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{z} - \boldsymbol{\mu})\right)$.]

(*i*) Let $\mathbf{f}, \mathbf{y}, X$ denote the collection of function values, training labels and feature vectors, respectively. Write down the log-prior, $\log p(\mathbf{f}|X)$ in terms of the Gram matrix $\mathbf{K}_{ij} = \kappa(\mathbf{x}_i, \mathbf{x}_j)$. [2 marks]

(*ii*) Find the posterior over $\mathbf{f}$, $p(\mathbf{f}|X, \mathbf{y})$. Neglect the normalisation and compute the un-normalised log-posterior, neglecting constant terms. Compare against the SVM objective found in Part (*a*)(*iv*). [4 marks]