CS181 Practice Questions: Model Selection and Linear Classification

1. **Coin Flipping**

   You have 2 biased coins: one that comes up heads with probability 0.8 and another that comes up with heads with probability 0.2. You select a coin randomly and you observe 5 heads out of 10 coin flips. Calculate $p(D)$ for the models that represent the first and second coin respectively. Which coin are you more likely to have selected?

2. **Model Selection**

Suppose you had three models, M_1, M_2, M_3, each increasing in complexity. For example, you could imagine that the models represented unregularized polynomial regression, with M_1 linear regression, M_2 quadratic regression, and M_3 cubic regression. Within the context of Bayesian model selection, come up with a way to penalize the complexity of a model so you did not always choose M_3. Additionally, explain why, in many cases, Bayesian model selection will recover the simplest model to explain the data without explicit penalization.

3. **Convex Hulls and Linear Separability**

Define the convex hull of a set of data points ($\{\boldsymbol{x}_i\}$) as the set

$$\left\{ \sum_i \alpha_i \boldsymbol{x}_i \text{ such that } \alpha_i \geq 0 \text{ and } \sum_i \alpha_i = 1 \right\}$$

Additionally, we say that two sets of points $\{\boldsymbol{x}_i\}$ and $\{\boldsymbol{x}'_j\}$ are linearly separable if there exists a vector $\boldsymbol{w}$ and w_0 such that $\boldsymbol{w}^\mathsf{T}\boldsymbol{x}_i + w_0 > 0$ for all points in the first set and $\boldsymbol{w}^\mathsf{T}\boldsymbol{x}'_j + w_0 < 0$ for all points in the second set. Show that if two sets of points $\{\boldsymbol{x}_i\}$ and $\{\boldsymbol{x}'_j\}$ are linearly separable, their convex hulls do not intersect.

4. **Perceptron Algorithm**

Consider the perceptron algorithm, which is a binary classification algorithm that finds the best linear hyperplane to separate the basis-transformed input values. The error function that is minimized is 0 when the algorithm correctly labels a data point and otherwise:

$$E_p(\boldsymbol{w}) = -\sum_{n \in M} \boldsymbol{w}^{\mathsf{T}} \boldsymbol{\phi}(\boldsymbol{x}_n) t_n, \tag{1}$$

where we sum over the mislabeled values and $t_n = 1$ if the correct classification is $\mathcal{C}_1$ and $t_n = -1$ if the correct classification is $\mathcal{C}_2$. Derive the stochastic gradient descent relation to optimize the weight vector for this error function.

5. **Thresholded Discriminant Functions**

Suppose we have the discriminant function $y(\boldsymbol{x}) = \boldsymbol{w}^\mathsf{T}\boldsymbol{x} + w_0$, but that rather than assigning $\boldsymbol{x}$ to $\mathcal{C}_1$ when $y(\boldsymbol{x}) \geq 0$ and to $\mathcal{C}_2$ otherwise (as in Bishop 4.1.1), we instead assign $\boldsymbol{x}$ to $\mathcal{C}_1$ when $y(\boldsymbol{x}) \geq \eta$ for some η and to $\mathcal{C}_2$ otherwise. Do we gain any generality by moving to this thresholded decision rule? Why or why not?

6. **Maximizing Separation Between Classes (Bishop 4.4)**

Suppose, as in Fisher's Discriminant Analysis, that we want to find the vector $\boldsymbol{w}$ that maximizes the distance between the means of two classes $\mathcal{C}_1, \mathcal{C}_2$ that are projected onto it. That is, we want to maximize

$$\boldsymbol{w}^\mathsf{T}(\boldsymbol{m}_2 - \boldsymbol{m}_1), \tag{Bishop 4.2.2}$$

where $\boldsymbol{m}_k = \frac{1}{N_k} \sum_{\boldsymbol{x} \in \mathcal{C}_k} \boldsymbol{x}$.

(a) Show that by maximizing the criterion above subject to the constraint that $\boldsymbol{w}^\mathsf{T}\boldsymbol{w} = 1$, we find that $\boldsymbol{w}_{\max} \propto (\boldsymbol{m}_2 - \boldsymbol{m}_1)$. That is, $\boldsymbol{w}_{\max} = \alpha(\boldsymbol{m}_2 - \boldsymbol{m}_1)$ for some α.

(b) Geometrically, what is the interpretation of $\boldsymbol{w}_{\max}$?

7. **Fisher Criterion in Matrix Form (Bishop 4.5)**

The Fisher Criterion is defined as

$$J(\boldsymbol{w}) = \frac{(m_2 - m_1)^2}{s_1^2 + s_2^2}, \tag{Bishop 4.2.5}$$

where

$$\begin{aligned} m_k &= \boldsymbol{w}^\mathsf{T} \boldsymbol{m}_k \\ \boldsymbol{m}_k &= \frac{1}{N_k} \sum_{\boldsymbol{x} \in \mathcal{C}_k} \boldsymbol{x} \\ s_k^2 &= \sum_{\boldsymbol{x} \in \mathcal{C}_k} (\boldsymbol{w}^\mathsf{T} \boldsymbol{x} - m_k)^2 \end{aligned}$$

Show that we can write $J(\boldsymbol{w})$ in matrix form as

$$J(\boldsymbol{w}) = \frac{\boldsymbol{w}^\mathsf{T} \boldsymbol{S}_B \boldsymbol{w}}{\boldsymbol{w}^\mathsf{T} \boldsymbol{S}_W \boldsymbol{w}},$$

where

$$\boldsymbol{S}_B = (\boldsymbol{m}_2 - \boldsymbol{m}_1)(\boldsymbol{m}_2 - \boldsymbol{m}_1)^\mathsf{T}$$

and

$$\boldsymbol{S}_W = \sum_{\boldsymbol{x} \in \mathcal{C}_1} (\boldsymbol{x} - \boldsymbol{m}_1)(\boldsymbol{x} - \boldsymbol{m}_1)^\mathsf{T} + \sum_{\boldsymbol{x} \in \mathcal{C}_2} (\boldsymbol{x} - \boldsymbol{m}_2)(\boldsymbol{x} - \boldsymbol{m}_2)^\mathsf{T}$$