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?

   $$\text{For coin 1:} \qquad p(D) = \binom{10}{5}(.8)^5(.2)^5$$
   $$\text{For coin 2:} \qquad p(D) = \binom{10}{5}(.8)^5(.2)^5$$

   The terms are equal - you are uncertain about which coin you have.

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.

Choose the prior over the model space to assign higher probability to the simpler models. This is the same thing that we did to encourage less complex parameters in Bayesian linear regression; we put a prior over the weight vector assigning the highest probability to weight vectors that lie close to zero. This is a a phenomenon known as Bayesian Occam's Razor and is discussed on in Bishop (pg. 164, fig. 3.13). Briefly, since simpler models can explain a smaller subset of data than more complex models, and each model has to integrate to 1, the simpler models can put higher probabilities on the data that they do describe than the complex models, which must put some of their mass on the data that the simpler model cannot describe.

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.

By definition, we have a $\boldsymbol{w}$ and w_0 such that $\boldsymbol{w}^\mathsf{T}\boldsymbol{x}_i + w_0 > 0$ and $\boldsymbol{w}^\mathsf{T}\boldsymbol{x}'_k + w_0 < 0$. Assume for the sake of contradiction that the convex hulls do intersect. Thus, there exists some set of $\{\alpha_i\}$ and $\{\beta_j\}$ with $\alpha_i > 0$ and $\beta_j > 0$ such that $\sum_i \alpha_i \boldsymbol{x}_i = \sum_j \beta_j \boldsymbol{x}'_j$. Now, we have that

$$\begin{aligned}
\boldsymbol{w}^\mathsf{T}\boldsymbol{x}_i + w_0 > 0 &\implies \\
\sum_i \alpha_i \left(\boldsymbol{w}^\mathsf{T}\boldsymbol{x}_i + w_0 \right) > 0 &\implies \\
\sum_i \alpha_i \boldsymbol{w}^\mathsf{T}\boldsymbol{x}_i + w_0 > 0 &\implies \\
\boldsymbol{w}^\mathsf{T} \sum_i \alpha_i \boldsymbol{x}_i + w_0 > 0 &\implies \\
\boldsymbol{w}^\mathsf{T} \sum_j \beta_j \boldsymbol{x}'_j + w_0 > 0 &\implies \\
\sum_j \beta_j \boldsymbol{w}^\mathsf{T}\boldsymbol{x}'_j + w_0 > 0 &\implies \\
\sum_j \beta_j \left(\boldsymbol{w}^\mathsf{T}\boldsymbol{x}'_j + w_0 \right) > 0. &
\end{aligned}$$

But this last step is a contradiction because $\boldsymbol{w}^\mathsf{T}\boldsymbol{y}_i + w_0 < 0$ and $\beta_i \geq 0$.

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.

We note that the SGD is given by $\boldsymbol{w}^{(n+1)} = \boldsymbol{w}^{(n)} - \eta \nabla E_p(\boldsymbol{w}) = \boldsymbol{w}^{(n)} + \eta \boldsymbol{\phi}(\boldsymbol{x}_n) t_n$. Using this, we can iteratively improve the initial $\boldsymbol{w}^{(0)}$ until convergence.

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?

No, we don't. In particular, for any $\boldsymbol{x}$, we have that $y(\boldsymbol{x}) \geq \eta$ iff $y(\boldsymbol{x}) - \eta \geq 0$, and so by setting $\hat{w}_0 = w_0 - \eta$, we can equivalently predict $\mathcal{C}_1$ if $\boldsymbol{w}^\mathsf{T}\boldsymbol{x} + \hat{w}_0 \geq 0$, and $\mathcal{C}_2$ otherwise.

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 $\boldsymbol{w}_{\max}$?

(a) We can write down the Lagrangian as $L = \boldsymbol{w}^\mathsf{T}(\boldsymbol{m}_2 - \boldsymbol{m}_1) + \lambda(\boldsymbol{w}^\mathsf{T}\boldsymbol{w} - 1)$. Taking gradients wrt $\boldsymbol{w}$, we get $\nabla_{\boldsymbol{w}} L = \boldsymbol{m}_2 - \boldsymbol{m}_1 + 2\lambda\boldsymbol{w}$. Setting the gradient to zero and solving gives the result.

(b) Geometrically, $\boldsymbol{w}$ is parallel to a line drawn between the means $\boldsymbol{m}_1$ and $\boldsymbol{m}_2$.

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

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

For the numerator, we have

$$\begin{aligned} (m_2 - m_1)^2 &= (\boldsymbol{w}^\mathsf{T}(\boldsymbol{m}_2 - \boldsymbol{m}_1))^2 \\ &= \boldsymbol{w}^\mathsf{T}(\boldsymbol{m}_2 - \boldsymbol{m}_1)(\boldsymbol{m}_2 - \boldsymbol{m}_1)^\mathsf{T} \boldsymbol{w} \\ &= \boldsymbol{w}^\mathsf{T} \boldsymbol{S}_B \boldsymbol{w} \end{aligned}$$

For the denominator, we have

$$
\begin{aligned}
s_1^2 + s_2^2 &= \sum_{\boldsymbol{x} \in \mathcal{C}_1} (\boldsymbol{w}^\mathsf{T} \boldsymbol{x} - m_1)^2 + \sum_{\boldsymbol{x} \in \mathcal{C}_2} (\boldsymbol{w}^\mathsf{T} \boldsymbol{x} - m_2)^2 \\
&= \sum_{\boldsymbol{x} \in \mathcal{C}_1} (\boldsymbol{w}^\mathsf{T} (\boldsymbol{x} - \boldsymbol{m}_1))^2 + \sum_{\boldsymbol{x} \in \mathcal{C}_2} (\boldsymbol{w}^\mathsf{T} (\boldsymbol{x} - \boldsymbol{m}_2))^2 \\
&= \sum_{\boldsymbol{x} \in \mathcal{C}_1} \boldsymbol{w}^\mathsf{T} (\boldsymbol{x} - \boldsymbol{m}_1)(\boldsymbol{x} - \boldsymbol{m}_1)^\mathsf{T} \boldsymbol{w} + \sum_{\boldsymbol{x} \in \mathcal{C}_2} \boldsymbol{w}^\mathsf{T} (\boldsymbol{x} - \boldsymbol{m}_2)(\boldsymbol{x} - \boldsymbol{m}_2)^\mathsf{T} \boldsymbol{w} \\
&= \boldsymbol{w}^\mathsf{T} \boldsymbol{S}_W \boldsymbol{w}
\end{aligned}
$$