Midterm 1 Practice Questions

1. **Biased Coins**

   You have a box full of coins. There are two types of coins, $\mathcal{C}_1$ and $\mathcal{C}_2$. Coins of type $\mathcal{C}_1$ come up heads with probability 0.8 and coin of type $\mathcal{C}_2$ come up heads with probability 0.2. There are many more $\mathcal{C}_1$ coins in the box than $\mathcal{C}_2$ coins, in fact 90% of the coins are of type $\mathcal{C}_1$. You grab a coin at random from inside the box and flip it 10 times, getting five heads and five tails. Compute $p(D \mid \mathcal{C}_1)$, $p(D \mid \mathcal{C}_2)$. How probable is it that you have a coin of type $\mathcal{C}_1$, given these ten flips?

2. **Redundant Features in Naïve Bayes**

Suppose that we use a Naïve Bayes classifier to classify binary data with binary feature vectors $\boldsymbol{x}_n \in \{0, 1\}^D$. We'll classify them into two classes, $\mathcal{C}_1$ and $\mathcal{C}_2$. With Naïve Bayes and binary features, the class conditional distributions will be of the form of a product of Bernoulli distributions:

$$p(\boldsymbol{x} \mid \mathcal{C}_k) = \prod_{d=1}^{D} \mu_{kd}^{x_d} (1 - \mu_{kd})^{(1-x_d)},$$

where $x_d \in \{0, 1\}$, and $\mu_{kd} = p(x_d = 1 \mid \mathcal{C}_k)$. Assume also that the class priors are uniform, i.e., $p(\mathcal{C}_1) = p(\mathcal{C}_2) = \frac{1}{2}$.

(a) If $D = 1$ (i.e., there is only one feature), use the equations above to write out $\ln \frac{p(\mathcal{C}_1 \mid x)}{p(\mathcal{C}_2 \mid x)}$ for a single binary feature x.

(b) Now suppose we change our feature representation so that instead of using just a single feature, we use two redundant features (i.e., two features that always have the same value so that $x_1 = x_2$). Since they are the same, you can assume that $\mu_{k1} = \mu_{k2}$ also. With this feature representation, let's write $\hat{x} = x_1 \cdot x_2$, since there can only be two configurations of the x_1, x_2 pair, instead of four. What is $\ln \frac{p(\mathcal{C}_1 \mid \hat{x})}{p(\mathcal{C}_2 \mid \hat{x})}$ in terms of the value for $\ln \frac{p(\mathcal{C}_1 \mid x)}{p(\mathcal{C}_2 \mid x)}$ you calculated in part (a)?

(c) Does this seem like a bug or a feature? Why?

3. **Binomial Regression**

You've been hired by a startup to build a ratings system for restaurants. Users rate the restaurants on a scale of 0 to 10 (i.e., $t_n \in \{0, 1, \ldots, 10\}$) and you have a set of real-valued features for each restaurant, $\boldsymbol{x}_n \in \mathbb{R}^D$. Given the range of the t_n, it seems like a binomial distribution would be a good choice for building a regression model:

$$p(k \mid \rho) = \binom{10}{k} \rho^k (1-\rho)^{10-k},$$

where ρ parameterizes the distribution and takes values in $(0, 1)$, while k is the rating. Recall that $\binom{N}{K}$ is the binomial coefficient, i.e., $N!/(K!(N-K)!)$.

(a) We cook up some basis functions $\phi_j(\boldsymbol{x})$ and we plan to weight them using a set of weights $\boldsymbol{w}$ to determine ρ. However, $\phi(\boldsymbol{x})^\mathsf{T}\boldsymbol{w}$ can be negative and can be greater than one. How can we map it into the right space?

(b) Having figured out how to get a map into the right space, write down the log likelihood of a set of N data $\{t_n, \boldsymbol{x}_n\}_{n=1}^N$. You can ignore constants in the sum that don't depend on the inputs or $\boldsymbol{w}$.

(c) Compute the gradient of the log likelihood in terms of $\boldsymbol{w}$. Hint: the derivative of the logistic function is $\frac{d}{dz}\sigma(z) = \sigma(z)(1-\sigma(z))$.

4. **Hyperplanes and Discriminant functions**

Suppose we have the discriminant function $y(\boldsymbol{x}) = \boldsymbol{w}^\top \boldsymbol{x} + w_0$, and if $y(\boldsymbol{x}) \geq 0$ we assign $\boldsymbol{x}$ to $\mathcal{C}_1$, and if $y(\boldsymbol{x}) < 0$ we assign $\boldsymbol{x}$ to $\mathcal{C}_2$. Show that for any $\boldsymbol{x}_0, \boldsymbol{x}_1$ on the decision boundary $(\boldsymbol{x}_0 - \boldsymbol{x}_1)$ is perpendicular to the vector $\boldsymbol{w}$.

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

The Fisher Criterion is defined as

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

where

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

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

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

where

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

and

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

6. **Classification with Same-Mean Different-Variance Gaussians**

Consider the task of recognizing which of two Gaussian distributions a data point $\mathbf{x} = (x_1, x_2, \ldots, x_D)$ comes from. We will assume that the two distributions have exactly the same mean but different variances. Let the probability that $\mathbf{x}$ is in class C_i (where $i \in \{0, 1\}$) be given by

$$\Pr(\mathbf{x}|C_i) = \prod_{j=1}^{D} \mathcal{N}(x_j|\mu_j, \sigma_{ij})$$

Show that $P(C_0|x)$ can be written in the form

$$P(\mathrm{C}_0|x) = \frac{1}{1 + \exp(-\mathbf{w}^T \mathbf{y} + \theta)}$$

where y_i is an appropriate function of x_i, $y_i = g(x_i)$, and θ is some constant.

7. **Margin Distances**

Consider the hyperplane given by $\boldsymbol{w}^T\boldsymbol{x} + b = 0$. For an arbitrary data point $\boldsymbol{x}$, what is the distance between $\boldsymbol{x}$ and the hyperplane, in terms of $\boldsymbol{w}$ and b?