Midterm 1 Practice Solutions

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?

   The prior probabilities are $p(\mathcal{C}_1) = 0.9$ and $p(\mathcal{C}_2) = 0.1$. We compute:

$$p(D \mid \mathcal{C}_1) = \binom{10}{5}(0.8)^5(0.2)^{10-5}$$
$$p(D \mid \mathcal{C}_2) = \binom{10}{5}(0.2)^5(0.8)^{10-5}$$

   Each coin has identical likelihoods, so it is only the prior that matters: $p(\mathcal{C}_1 \mid D) = 0.9$.

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.

Because priors are equal:

$$\ln \frac{p(\mathcal{C}_1 \mid x)}{p(\mathcal{C}_2 \mid x)} = \ln \frac{p(x \mid \mathcal{C}_1)}{p(x \mid \mathcal{C}_2)}$$

So

$$\ln \frac{p(\mathcal{C}_1 \mid x)}{p(\mathcal{C}_2 \mid x)} = x \ln \mu_1 + (1 - x) \ln(1 - \mu_1) - x \ln \mu_2 - (1 - x) \ln(1 - \mu_2)$$

(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)?

$$\begin{aligned}
\ln \frac{p(\mathcal{C}_1 \mid \hat{x})}{p(\mathcal{C}_2 \mid \hat{x})} &= \ln \frac{p(x_1 \mid \mathcal{C}_1) p(x_2 \mid \mathcal{C}_1)}{p(x_1 \mid \mathcal{C}_2) p(x_2 \mid \mathcal{C}_2)} \\
&= 2 \left(\hat{x} \ln \mu_1 + (1 - \hat{x}) \ln(1 - \mu_1) - \hat{x} \ln \mu_2 - (1 - \hat{x}) \ln(1 - \mu_2) \right)
\end{aligned}$$

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

This is a bug because it is now more confident than it should be. These features are tightly coupled, but naïve Bayes assumes they are independent.

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?
This is a perfect use case for the logistic or sigmoid function $\sigma(z) = \frac{1}{1+e^{-z}}$.

(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 $\boldsymbol{w}$.
We have that the likelihood is:

$$p(\{t_n\} \mid \{\boldsymbol{x}_n\}, \boldsymbol{w}) = \prod_n \binom{10}{t_n} \sigma(\phi(x_n)^{\mathsf{T}}\boldsymbol{w})^{t_n} (1 - \sigma(\phi(x_n)^{\mathsf{T}}\boldsymbol{w}))^{10-t_n}$$

The log-likelihood is:

$$\ln p(\{t_n\} \mid \{\boldsymbol{x}_n\}, \boldsymbol{w}) = \sum_n \ln \binom{10}{t_n} + t_n \ln \sigma(\phi(\boldsymbol{x}_n)^{\mathsf{T}}\boldsymbol{w}) + (10 - t_n) \ln(1 - \sigma(\phi(\boldsymbol{x}_n)^{\mathsf{T}}\boldsymbol{w})),$$

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

$$\begin{aligned}
&\frac{d}{d\boldsymbol{w}} \ln p(\{t_n\} \mid \{\boldsymbol{x}_n\}, \boldsymbol{w}) \\
&= \sum_n \frac{t_n}{\sigma(\phi(\boldsymbol{x}_n)^{\mathsf{T}}\boldsymbol{w})} \sigma(\phi(\boldsymbol{x}_n)^{\mathsf{T}}\boldsymbol{w})(1 - \sigma(\phi(\boldsymbol{x}_n)^{\mathsf{T}}\boldsymbol{w}))\phi(\boldsymbol{x}_n) \\
&\quad + \frac{10 - t_n}{1 - \sigma(\phi(\boldsymbol{x}_n)^{\mathsf{T}}\boldsymbol{w})} (-\sigma(\phi(\boldsymbol{x}_n)^{\mathsf{T}}\boldsymbol{w})(1 - \sigma(\phi(\boldsymbol{x}_n)^{\mathsf{T}}\boldsymbol{w})))\phi(\boldsymbol{x}_n) \\
&= \sum_n t_n (1 - \sigma(\phi(\boldsymbol{x}_n)^{\mathsf{T}}\boldsymbol{w}))\phi(x_n) - (10 - t_n)\sigma(\phi(\boldsymbol{x}_n)^{\mathsf{T}}\boldsymbol{w})\phi(\boldsymbol{x}_n).
\end{aligned}$$

Further simplification is possible but unnecessary.

4. **Hyperplanes and Discriminant functions**

Suppose we have the discriminant function $y(\boldsymbol{x}) = \boldsymbol{w}^{\mathsf{T}}\boldsymbol{x} + w_0$, 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 $\boldsymbol{w}$.

The decision boundary is the set of $\boldsymbol{x}$ such that $y(\boldsymbol{x}) = 0 \implies \boldsymbol{w}^{\mathsf{T}}\boldsymbol{x} + w_0 = 0$. So, we have

$$\begin{aligned}
y(x_0) - y(x_1) &= (\boldsymbol{w}^{\mathsf{T}}\boldsymbol{x}_0 + w_0) - (\boldsymbol{w}^{\mathsf{T}}\boldsymbol{x}_1 + w_0) \\
&= \boldsymbol{w}^{\mathsf{T}}\boldsymbol{x}_0 - \boldsymbol{w}^{\mathsf{T}}\boldsymbol{x}_1 \\
&= \boldsymbol{w}^{\mathsf{T}}(x_0 - x_1) \implies \\
0 &= \boldsymbol{w}^{\mathsf{T}}(x_0 - x_1)
\end{aligned}$$

so the vectors are perpendicular.

5. **Fisher Criterion in Matrix Form**

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

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

For the numerator, we have

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

For the denominator, we have

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

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

$$P(C_0|x) = \frac{1}{1 + \exp(-\mathbf{w}^T\mathbf{y} + \theta)}$$

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

By Bayes' Theorem, we can rewrite $P(C_0|\mathbf{x})$ as

$$P(C_0|\mathbf{x}) = \frac{P(\mathbf{x}|C_0)P(C_0)}{P(\mathbf{x}|C_0)P(C_0) + P(\mathbf{x}|C_1)P(C_1)} = \frac{1}{1 + \dfrac{P(\mathbf{x}|C_1)}{P(\mathbf{x}|C_0)}\dfrac{P(C_1)}{P(C_0)}}$$

Writing out $P(\mathbf{x}|C_i)$ in terms of the Normal Distribution, we have

$$P(\mathbf{x}|C_i) = \frac{1}{(2\pi)^{D/2}\prod_{j=1}^{D}\sigma_{ij}} \exp\left(-\frac{1}{2}\sum_{j=1}^{D}\frac{(x_j-\mu_j)^2}{\sigma_{ij}^2}\right)$$

So we can write

$$\frac{P(\mathbf{x}|C_1)}{P(\mathbf{x}|C_0)} = \frac{\prod_{j=1}^{D}\sigma_{0j}}{\prod_{j=1}^{D}\sigma_{1j}} \exp\left(-\frac{1}{2}\sum_{j=1}^{D}\left((x_j-\mu_j)^2\left(\frac{1}{\sigma_{1j}^2} - \frac{1}{\sigma_{0j}^2}\right)\right)\right)$$

If we let

$$\theta = \ln\left(\frac{P(C_1)}{P(C_0)}\frac{\prod_{j=1}^{D}\sigma_{0j}}{\prod_{j=1}^{D}\sigma_{1j}}\right)$$
$$w_i = \left(\frac{1}{\sigma_{1i}^2} - \frac{1}{\sigma_{0i}^2}\right)$$
$$y_i = (x_i-\mu_i)^2/2$$

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?

First, observe that the hyperplane is orthogonal to $\boldsymbol{w}$. For two arbitrary points on the hyperplane $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$, we can see

$$\boldsymbol{w}^T(\boldsymbol{x}_1 - \boldsymbol{x}_2) = \boldsymbol{w}^T\boldsymbol{x}_1 - \boldsymbol{w}^T\boldsymbol{x}_2 = -b - (-b) = 0$$

We can scale $\boldsymbol{w}$ to $r\frac{\boldsymbol{w}}{\|\boldsymbol{w}\|_2}$, so it's some constant r multiplied by the unit vector. Call $\boldsymbol{x}_\perp$ the point on the hyperplane satisfying the following equation:

$$\boldsymbol{x}_\perp + r\frac{\boldsymbol{w}}{\|\boldsymbol{w}\|_2} = \boldsymbol{x}$$

Then, left-multiplying by $\boldsymbol{w}^T$, we can see

$$\boldsymbol{w}^T\boldsymbol{x}_\perp + r\frac{\boldsymbol{w}^T\boldsymbol{w}}{\|\boldsymbol{w}\|_2} = \boldsymbol{w}^T\boldsymbol{x}$$

Since $\boldsymbol{x}_\perp$ is on the hyperplane

$$-b + r\|\boldsymbol{w}\|_2 = \boldsymbol{w}^T\boldsymbol{x} \Rightarrow r = \frac{\boldsymbol{w}^T\boldsymbol{x} + b}{\|\boldsymbol{w}\|_2}$$

Therefore, the displacement between the hyperplane and $\boldsymbol{x}$ is given by $\boldsymbol{r}$. We can multiply by the sign of the label of $\boldsymbol{x}$ to make sure this value is always positive.