HW 3 Solutions

1. *Bayes rule, variances, and priors.* (10 points) Let $f(x|Y=k) \sim N(\mu_k, \sigma^2), k=1,2$ be a two category one-dimensional problem with $P(Y=1)=\pi_1, P(Y=2)=\pi_2$.

   (a) Compute the decision boundary assuming $\mu_1 > \mu_2$ and compute the Bayes loss.

$$\begin{aligned} & \pi_1 f(x|Y=1) \geq \pi_2 f(x|Y=2) \\ \Leftrightarrow \quad & (x-\mu_2)^2 - (x-\mu_1)^2 \geq 2\sigma^2 \log \frac{\pi_2}{\pi_1} \\ \Leftrightarrow \quad & x \geq \frac{\mu_1+\mu_2}{2} + \frac{\sigma^2(\log \pi_2 - \log \pi_1)}{\mu_1 - \mu_2} \end{aligned}$$

   The decision boundary is the point $x = \frac{\mu_1+\mu_2}{2} + \frac{\sigma^2(\log \pi_2 - \log \pi_1)}{\mu_1-\mu_2}\}$.

$$\begin{aligned} \text{Bayes loss} &= \int_{R_1} \pi_2 \phi(x;\mu_2,\sigma^2)dx + \int_{R_2} \pi_1 \phi(x;\mu_1,\sigma^2)dx \\ &= \pi_2 \Phi\left(\frac{\mu_2-\mu_1}{2\sigma} - \frac{\sigma(\log \pi_2 - \log \pi_1)}{\mu_1-\mu_2}\right) + \pi_1 \Phi\left(\frac{\mu_2-\mu_1}{2\sigma} + \frac{\sigma(\log \pi_2 - \log \pi_1)}{\mu_1-\mu_2}\right), \end{aligned}$$

   where $\phi(x;\mu,\sigma^2)$ is the density of $N(\mu,\sigma^2)$, $\Phi(x)$ is the CDF of standard normal distribution.

   (b) Write the probability of error using the CDF of the normal distribution. Show that the error goes to zero as $\sigma \to 0$.

   As $\sigma \to 0$,

$$\text{Bayes loss} \to \pi_2 \Phi(-\infty) + \pi_1 \Phi(-\infty) = 0.$$

   (c) For fixed $\sigma$ what happens to the decision boundary as $\pi_1 \to 0$? For small $\pi_1$ what is a very simple classification rule that can guarantee low error rate?

   When $\sigma$ is fixed and $\pi_1 \to 0$, the decision boundary $x^* = \frac{\mu_1+\mu_2}{2} + \frac{\sigma^2(\log \pi_2 - \log \pi_1)}{\mu_1-\mu_2} \to +\infty$. Thus a simple rule that classifies all points into category 2 can guarantee low error rate.

   (d) Assume a classification problem with $K$ classes. Let $L_{k,l}$ be the cost of choosing class $l$ when class $k$ is true. Let $p(x,y)$ be the joint distribution of $X, Y$. The expected loss for a decision function $h$ is

$$L(h) = \sum_{k=1}^{K} \sum_{l=1}^{K} \int_{h(x)=l} p(x,k) L_{k,l} dx.$$

Show that the decision rule with lowest loss is given by the generalize Bayes rule:

$$h_B(x) = \text{argmin}_{j=1,\dots,K} \sum_{k=1}^{K} L_{k,j} p(k|x).$$

Hint: Use the loss of an individual example defined as $L(h, x) = \sum_{k=1}^{K} L_{k,h(x)} p(k|x)$.

$$\begin{aligned} L(h) &= \sum_{k=1}^{K} \sum_{l=1}^{K} \int_{h(x)=l} p(x,k) L_{k,l} dx \\ &= \sum_{k=1}^{K} \sum_{l=1}^{K} \int_{h(x)=l} p(x,k) L_{k,h(x)} dx \\ &= \sum_{k=1}^{K} \int p(x,k) L_{k,h(x)} dx \\ &= \int L(h,x) p(x) dx, \end{aligned}$$

where

$$L(h,x) = \sum_{k=1}^{K} L_{k,h(x)} p(k|x).$$

Thus, the function $L(h)$ is minimized by minimizing $L(h,x)$ for any specific $x$, which gives the result:

$$h_B(x) = \text{argmin}_{j=1,\dots,K} \sum_{k=1}^{K} L_{k,j} p(k|x).$$

(e) One way to avoid the degenerate situation from item 1c is to change the cost function. Set $L_{2,1}$ to be the cost of choosing class 1 when class 2 is true, and $L_{1,2}$ the cost of choosing class 2 when class 1 is true. Write the expected loss in this situation. Recompute the decision boundary. What values on $L_{i,j}$ would you assign to remedy the problem of small $\pi_1$.

$$\text{Loss} = \pi_2 L_{2,1} \int_{h(x)=1} \phi(x; \mu_2, \sigma^2) + \pi_1 L_{1,2} \int_{h(x)=2} \phi(x; \mu_1, \sigma^2).$$

By the result in (d), the classification area for category 1 is

$$R_1 = \{x : x \geq \frac{\mu_1 + \mu_2}{2} + \frac{\sigma^2(\log L_{2,1}\pi_2 - \log L_{1,2}\pi_1)}{\mu_1 - \mu_2}\}$$

We can set $L_{1,2} \gg L_{2,1}$ to avoid the problem caused by small $\pi_1$.

2. *Bernoulli mixtures:* (20 points)

(a) Assume your observations are binary in $R^d$, i.e. $X = (X_1, \dots, X_d), X_j \in \{0, 1\}, j = 1, \dots, d$. We assume the $d$ variables are independent with joint distribution $P(X_1 = x_1, \dots, X_d = x_d) = P(X_1 = x_1) \cdots P(X_d = x_d)$, and $P(X_j = x) = p_j^x (1 - p_j)^{1-x}$. Given a sample $X_1, \dots, X_n$ from this distribution write the log-likelihood, write the score equation for each parameter $p_j, j = 1, \dots, d$ and write the maximum likelihood estimate $\widehat{p}_j$.

Log-likelihood: $\ell(X, p_1, \dots, p_d) = \sum_{i=1}^d \sum_{j=1}^d [X_{ij} \log p_j + (1 - X_{ij}) \log(1 - p_j)]$.
Score equation $\partial \ell / \partial p_j = \sum_{i=1}^n [X_{ij}/p_j - (1 - X_{ij})/(1 - p_j)] = 0$.
Writing $S_j = \sum_{i=1}^n X_{ij}$, we get $(1 - p_j) S_j = (n - S_j) p_j$ so that $\widehat{p}_j = S_j / n$.

(b) Assume you have a Bernoulli model for each of $C$ classes, with probabilities $P(X_j = 1 | Y = c) = p_{c,j}, c = 1, \dots, C, j = 1, \dots, d$ and $\pi_c = P(Y = c)$. Write out the Bayes classifier in terms of these models, show that it is a linear classifier $\arg\max_c h_c(x)$, with $h_c(x) = W_c x + b_c$. Write out $W_c$ and $b_c$ in terms of the parameters of the model.

The Bayes classifier can be written as:

$$\widehat{Y} = \arg\max_c \log P(X, Y = c) = \arg\max_c \sum_j \left[ X_j \log \frac{p_{c,j}}{1 - p_{c,j}} + \log(1 - p_{c,j}) \right] + \log \pi_c.$$

So setting

$$W_c = [\log \frac{p_{c,j}}{1 - p_{c,j}}]_{j=1}^d, \quad b_c = \log \pi_c + \sum_{j=1}^d \log(1 - p_{c,j}),$$

we have the desired result.

(c) Assume now that the data $X_1, \dots, X_n$ come from a mixture model with $M$ components and each component a product of $d$ independent Bernoulli variables as in (a):

$$f(x; \theta) = \sum_{m=1}^M \pi_m f_m(x; \theta_m), \quad f_m(x; \theta_m) = \prod_{j=1}^d p_{j,m}^{x_j} (1 - p_{j,m})^{(1 - x_j)},$$

where $\theta = (\pi_1, \dots, \pi_M, \theta_1, \dots, \theta_M)$ and $\theta_m = (p_{1,m}, \dots, p_{d,m})$.
Derive the details of the EM algorithm for this model.

i. Write out the precise formula for computing the responsibilities $w_{mi}, m = 1, \dots, M, i = 1, \dots, n$ in terms of the current estimates $\widehat{\pi}_m, \theta_m, m = 1, \dots, M$.
$w_{mi} = \frac{\pi_m f_m(X_i; \theta_m)}{\sum_{r=1}^M \pi_r f_r(X_i; \theta_r)}$.

ii. In computing the responsibilities you would be computing ratios involving the $f_m$'s. If $d$ is large you are multiplying lots of small numbers on a computer and quickly reaching the rounding error of the computer. You may be dividing by very small numbers, which is risky on the computer. Instead it is convenient to write:

$$f_m(X_i;\theta_m) = \exp[\sum_j X_{ij}\log p_{j,m} + (1-X_{ij})\log(1-p_{j,m})].$$

For each data point, how would you use this expression for the different clusters to guarantee that the ratio you compute always has a value less than 1 in the numerator and a value greater than 1 and less than $M$ in the denominator.

For each example $X_i$ we compute

$$g_{mi} = \sum_j X_{ij}\log p_{j,m} + (1-X_{ij})\log(1-p_{j,m}) + \log\pi_m.$$

Let $g_i^* = \max_m g_{mi}$, then

$$w_{mi} = \frac{\exp[g_{mi} - g_i^*]}{\sum_{r=1}^M \exp[g_{ri} - g_i^*]}.$$

The numerator is less than or equal to one depending on whether the maximum is achieved at $m$ or not. And in the denominator the term corresponding to $m'$ for which $g_{ri} = g_i^*$ is equal to 1.

iii. Once you have the responsibilities. Write the formula for computing the new estimates $\widehat{\pi}_m^{new}, p_{j,m}^{new}$.

Just like the EM fo Gaussian mixtures: $\widehat{\pi}_m^{new} = \frac{1}{n}\sum_{i=1}^n w_{mi}$. Using the formulation in the slides, we need to minimize $\sum_{i=1}^n w_{mi}\log f_m(X_i,\theta_m)$. This is the same computation as in (a) with weights $w_{mi}$ for each example instead of weight 1. So $\widehat{p}_{j,m} = \frac{\sum_{i=1}^n w_{mi}X_{ij}}{\sum_{i=1}^n w_{mi}}$.