SMLDM HT 2014 - Part C Problem Sheet 3

1. Consider two univariate normal distributions $\mathcal{N}(\mu, \sigma^2)$ with known parameters $\mu_A = 10$ and $\sigma_A = 5$ for class A and $\mu_B = 20$ and $\sigma_B = 5$ for class B. Suppose class A represents the random score X of a medical test of normal patients and class B represents the score of patients with a certain disease. A priori there are 100 times more healthy patients than patients carrying the disease.

   (a) Find the optimal decision rule in terms of misclassification error (0-1 loss) for allocating a new observation x to either class A or B.

   Answer: The optimal decision for $X = x$ is to allocate to class

   $$\text{argmax}_{k \in \{A,B\}} \pi_k f_k(x).$$

   The patients should be classified as healthy iff

   $$\pi_A \frac{1}{\sqrt{2\pi}\sigma_A} \exp(-\frac{(x-\mu_A)^2}{2\sigma_A^2}) \geq \pi_B \frac{1}{\sqrt{2\pi}\sigma_B} \exp(-\frac{(x-\mu_B)^2}{2\sigma_B^2}),$$

   that is, using $\sigma_A = \sigma_B$, iff

   $$-2\sigma_A^2 \log(\pi_A/\pi_B) + (x-\mu_A)^2 \leq (x-\mu_B)^2.$$

   The decision boundary is attained for equality, that is if x fulfills

   $$2x(\mu_B - \mu_A) + \mu_A^2 - \mu_B^2 - 2\sigma_A^2 \log(\pi_A/\pi_B) = 0.$$

   For the given values, this implies that the decision boundary is at

   $$x = (50 \log 100 - 100 + 400)/(2 \cdot 10) \approx 26.51,$$

   that is all patients with a test score above 26.51 are classfied as having the disease.

   (b) Repeat (a) if the cost of a false negative (allocating a sick patient to group A) is $\theta > 1$ times that of a false positive (allocating a healthy person to group B). Describe how the rule changes as θ increases. For which value of θ are 84.1% of all patients with disease correctly classified?

   Answer: The optimal decision minimizes $\mathbb{E}(L(Y, \hat{Y}(x))|X = x)$. It is hence optimal to choose class A (healthy) over class B if and only if

   $$P(Y = A|X = x) \geq \theta P(Y = B|X = x).$$

   Using the same argument as above, the patients should be classified as healthy now iff (ignoring again the common denominator $\sum_{k \in \{A,B\}} \pi_k f_k(x)$),

   $$\pi_A \frac{1}{\sqrt{2\pi}\sigma_A} \exp(-\frac{(x-\mu_A)^2}{2\sigma_A^2}) \geq \theta \pi_B \frac{1}{\sqrt{2\pi}\sigma_B} \exp(-\frac{(x-\mu_B)^2}{2\sigma_B^2}).$$

   The decision boundary is now attained if x fulfills

   $$2x(\mu_B - \mu_A) + \mu_A^2 - \mu_B^2 - 2\sigma_A^2 \log(\pi_A/(\theta\pi_B)) = 0.$$

For increasing values of θ, patients with decreasingly smaller test scores are classified as having the disease.

84.1% of all patients carrying the disease are correctly classified if the decision boundary is at the 15.9%-quantile of the $\mathcal{N}(\mu_B, \sigma_B^2)$-distribution, which is at $q = 20 + 5\Phi^{-1}(0.159) \approx 15$. This decision boundary is attained if

$$15 = q = (50\log(100/\theta) - 100 + 400)/20,$$

which implies that for

$$\theta = 100\exp(-\frac{20q - 300}{50}) = 100,$$

approximately 84.1% of all patients are correctly classified as carrying the disease.

2. For a given loss function L, the risk R is given by the expected loss

$$R(\hat{Y}) = \mathbb{E}(L(Y, \hat{Y}(X))),$$

where $\hat{Y} = \hat{Y}(X)$ is a function of the random predictor variable X.

(a) Consider a regression problem and the squared error loss

$$L(Y, \hat{Y}(X)) = (Y - \hat{Y}(X))^2.$$

Derive the expression of $\hat{Y} = \hat{Y}(X)$ minimizing the associated risk.

Answer: We have

$$\begin{aligned} R &= \mathbb{E}\left(\left(Y - \widehat{Y}(X)\right)^2\right) \\ &= \int \mathbb{E}\left(\left(Y - \widehat{Y}(X)\right)^2 \middle| X = x\right) f_X(x)\, dx \end{aligned}$$

so minimizing the risk can be achieved by minimizing for any x

$$\begin{aligned} & \mathbb{E}\left(\left(Y - \widehat{Y}(X)\right)^2 \middle| X = x\right) \\ = \; & \mathbb{E}\left(Y^2 \middle| X = x\right) - 2\widehat{Y}(x)\,\mathbb{E}\left(Y \middle| X = x\right) + \widehat{Y}(x)^2. \end{aligned}$$

This is clearly minimized for the conditional mean:

$$\widehat{Y}(X) = \mathbb{E}(Y|X).$$

(b) What if we use the ℓ_1 loss instead?

$$L(Y, \hat{Y}(X)) = |Y - \hat{Y}(X)|.$$

Answer: As before, we want to find $\hat{Y}(x)$ to minimize

$$\mathbb{E}\left(|Y - \hat{Y}(x)| \middle| X = x\right)$$

Differentiating the expression with respect to $\hat{Y}(x)$,

$$\mathbb{E}\left(\text{sign}(Y - \hat{Y}(x))\big|X = x\right) = -\mathbb{P}(Y < \hat{Y}(x)|X = x) + \mathbb{P}(Y > \hat{Y}(x)|X = x)$$

which occurs when $\mathbb{P}(Y < \hat{Y}(x)|X = x) = \mathbb{P}(tY > \hat{Y}(x)|X = x) = .5$, i.e. at the median conditional on $X = x$.

3. Consider applying LDA to a two-class dataset. We will verify some of the claims in the lectures. We use the notation in the lectures.

   (a) Show that $\Sigma^{-1}(\mu_1 - \mu_2)$ spans the one-dimensional discriminant subspace. What is the corresponding generalized eigenvalue?

   Answer: We can write $\bar{x} = \frac{n_1}{n}\mu_1 + \frac{n_2}{n}\mu_2$. Hence

$$\mu_1 - \bar{x} = \frac{n_2}{n}(\mu_1 - \mu_2)$$
$$\mu_2 - \bar{x} = \frac{n_1}{n}(\mu_2 - \mu_1)$$

   Plugging these into B, we see that it can be written as $c(\mu_1 - \mu_2)(\mu_1 - \mu_2)^\top$ where $c = \frac{n_1 n_2}{n^2}$. Thus it is easy to verify now that $\Sigma^{-1}(\mu_1 - \mu_2)$ solves the generalized eigenvector equation, with eigenvalue $\frac{n_1 n_2}{n^2}(\mu_1 - \mu_2)^\top \Sigma^{-1}(\mu_1 - \mu_2)$.

   (b) Explain why to predict the class of a data vector x it is sufficient to look at its projection onto the subspace spanned by $\Sigma^{-1}(\mu_1 - \mu_2)$.

   Answer: The projection onto the subspace spanned by $\Sigma^{-1}(\mu_1 - \mu_2)$ is simply

$$z = (\mu_1 - \mu_2)^\top \Sigma^{-1} x / \|\Sigma^{-1}(\mu_1 - \mu_2)\|.$$

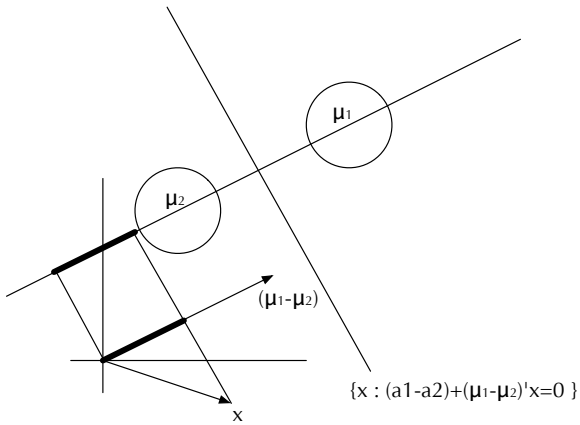
   The decision rule of LDA is to predict class 1 if

$$0 < (a_1 - a_2) + (\mu_1 - \mu_2)^\top \Sigma^{-1} x = (a_1 - a_2) + \|\Sigma^{-1}(\mu_1 - \mu_2)\| z,$$

   which can be computed from z.

   (c) In the case where the within-class covariance is $\Sigma = I$, explain the geometry of the decision rule of LDA with the help of a diagram.

   Answer:



4. Show that under a Naïve Bayes model, the Bayes classifier $\hat{Y}(x)$ minimizing the total risk for the $0-1$ loss function has a linear discriminant function of the form

$$\hat{Y}(x) = \arg\max_{k=1,2} \quad \alpha_k + \beta_k^\top x.$$

and find the values of α_k, β_k. (Use notation from lecture slides).

Answer: The Bayes classifier is given in this discrete state space as

$$\hat{Y}(x) = \arg\max_{k=1,2} \quad \pi_k \mathbb{P}(X = x|Y = k) = \arg\max_{k=1,2} \quad \log \pi_k + \log \mathbb{P}(X = x|Y = k)$$

Now,

$$\begin{aligned}\log \mathbb{P}(X = x|Y = k) &= \sum_{j=1}^{p} x_{ij} \log \phi_{kj} + (1 - x_{ij}) \log(1 - \phi_{kj}) \\ &= \sum_{j=1}^{p} \log(1 - \phi_{kj}) + x_{ij} \log \frac{\phi_{kj}}{1 - \phi_{kj}}\end{aligned}$$

So that the discriminant functions are linear, with:

$$\alpha_k = \log \pi_k + \sum_{j=1}^{p} \log(1 - \phi_{kj})$$

$$\beta_{kj} = \log \frac{\phi_{kj}}{1 - \phi_{kj}}$$

for each k and j.

5. Suppose we have a two-class setup with classes -1 and 1, that is $\mathcal{Y} = \{-1, 1\}$ and a 2-dimensional predictor variable X. We find that the means of the two groups are at $\hat{\mu}_{-1} = (-1, -1)^\top$ and $\hat{\mu}_1 = (1, 1)^\top$ respectively. The a priori probabilities are equal.

   (a) Applying LDA, the covariance matrix is estimated to be, for some value of $0 \leq \rho \leq 1$,

   $$\hat{\Sigma} = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}.$$

   Find the decision boundary as a function of ρ.

   Answer: The constant $a_* = a_1 - a_{-1}$ is given by, using equal a priori probabilities,

   $$a_* = \hat{\mu}_1^T \hat{\Sigma}^{-1} \hat{\mu}_1 - \hat{\mu}_2^T \hat{\Sigma}^{-1} \hat{\mu}_2.$$

   Hence $a_* = 0$. The constant $b_* = b_1 - b_{-1}$ is on the other hand

   $$b_* = \hat{\Sigma}^{-1}(\mu_1 - \mu_{-1}) = \hat{\Sigma}^{-1}(2, 2)^\top = 2/(1 + \rho)(1, 1)^\top,$$

   Class 1 is chosen over class -1 for $x = (x^{(1)}, x^{(2)})^\top$ if and only if $a_* + b_*^\top x > 0$, that is iff

   $$\frac{2}{1 + \rho}(x^{(1)} + x^{(2)}) > 0.$$

   Equivalently, iff

   $$x^{(1)} + x^{(2)} > 0,$$

   which could have been guessed as the solution initially.

(b) Suppose instead that, we model each class with its own covariance matrix. We estimate the covariance matrices for group -1 as

$$\hat{\Sigma}_{-1} = \begin{pmatrix} 5 & 0 \\ 0 & 1/5 \end{pmatrix},$$

and for group 1 as

$$\hat{\Sigma}_{1} = \begin{pmatrix} 1/5 & 0 \\ 0 & 5 \end{pmatrix}.$$

Describe the decision rule and draw a sketch of it in the two-dimensional plane.

Answer: As in (a), the classification is class 1 if and only if

$$(x - \hat{\mu}_1)^T \hat{\Sigma}_1^{-1} (x - \hat{\mu}_1) \quad < \quad (x - \hat{\mu}_{-1})^T \hat{\Sigma}_{-1}^{-1} (x - \hat{\mu}_{-1}).$$

The difference with LDA in (a) is that $\hat{\Sigma}_1 \neq \hat{\Sigma}_{-1}$.

Let, as in the lectures, $a_k = \hat{\mu}_k^T \hat{\Sigma}_k^{-1} \hat{\mu}_k$ and similarly for b_k (for terms linear in x) and c_k (for terms quadratic in x) for $k = 1, 2$.

The constant $a_* = a_1 - a_{-1}$ is again 0. The term b_1 is now

$$b_1^T x = -2\hat{\mu}_1^T \hat{\Sigma}_1^{-1} x = -2(5x^{(1)} + x^{(2)}/5).$$

and

$$b_{-1}^T x = -2\hat{\mu}_{-1}^T \hat{\Sigma}_2^{-1} x = 2(x^{(1)}/5 + 5x^{(2)}).$$

The quadratic terms are

$$x^T c_1 x = x^T \hat{\Sigma}_1^{-1} x = 5(x^{(1)})^2 + (x^{(2)})^2/5$$

and

$$x^T c_{-1} x = x^T \hat{\Sigma}_{-1}^{-1} x = (x^{(1)})^2/5 + 5(x^{(2)})^2.$$

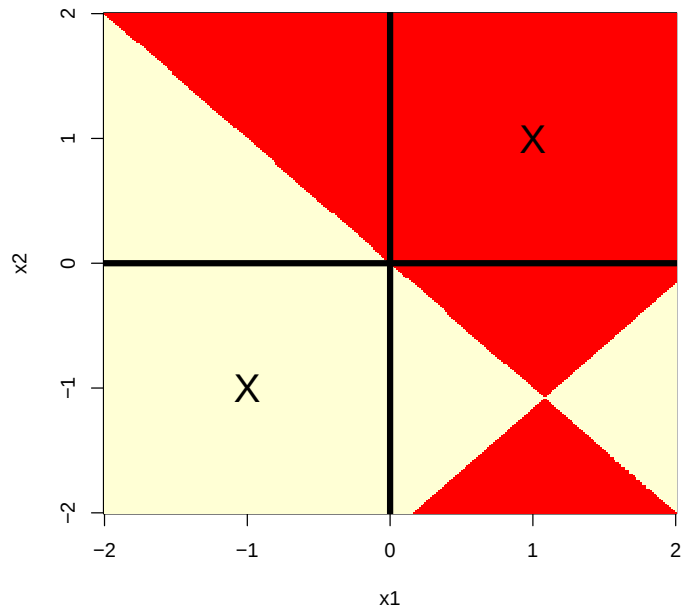
The observations x is thus classified as belonging to group 1 if and only if

$$5(x^{(1)})^2 + (x^{(2)})^2/5 - 2(5x^{(1)} + x^{(2)}/5) \quad < \quad (x^{(1)})^2/5 + 5(x^{(2)})^2 + 2(x^{(1)}/5 + 5x^{(2)}).$$

Bringing all terms to the left side and dividing by $5 - 1/5$, the classification is group 1 if and only if

$$(x^{(1)})^2 - (x^{(2)})^2 - \frac{\cdot 13}{6}(x^{(1)} + x^{(2)}) < 0.$$

Here, we thus obtain linear decision boundaries, even though we are using QDA (which typically produces quadratic decision boundaries). The decision boundaries are shown in the figure below, along with the group means.



6. (Challenging; submit this in Friday Week 5)

   Implement the EM algorithm from Problem Sheet 2 Question 4 on a dataset of handwritten digits of '0', to '9'.

   You can obtain the dataset from
   `http://www.stats.ox.ac.uk/\%7Eteh/teaching/smldmHT2014/usps.txt`
   Load the dataset using
   `X <- as.matrix(read.table("usps.txt"))`
   The dataset is binary and has 1000 rows and 64 columns. Each row gives a 8×8 binary image, and you can visualize the images using a command like
   `image(matrix(X[i,],8,8))`
   where $i \in \{1, \ldots, 1000\}$.

   In the E step of the algorithm, you will need to compute the the posterior probabilities $q(z_i = k)$ in a numerically stable manner. Specifically, the unnormalized posterior probabilities are:

$$\pi_k \prod_{j=1}^{p} (\phi_{kj})^{x_{ij}} (1 - \phi_{kj})^{1-x_{ij}}$$

   We compute the logarithms instead:

$$S_{ik} = \log \pi_k + \sum_{j=1}^{p} x_{ij} \log \phi_{kj} + (1 - x_{ij}) \log(1 - \phi_{kj})$$

   Then the posterior probabilities are computed as:

$$q(z_i = k) = \frac{e^{S_{ik} - \max_{k'} S_{ik'}}}{\sum_l e^{S_{il} - \max_{k'} S_{ik'}}}$$

while the marginal log probability is computed as:

$$\log p(x_i) = \log \sum_k e^{S_{ik}} = \left(\max_{k'} S_{ik'} \right) + \log \sum_k e^{S_{ik} - \max_{k'} S_{ik'}}$$

This is called the "log-sum-exp" trick.

In the M step, to prevent overfitting, where the estimated probabilities approach 0 or 1, we simply add small constants to the formulas:

$$\phi_{kj} = \frac{\alpha + \sum_{i=1}^n q(z_i = k) x_{ij}}{2\alpha + \sum_{i=1}^n q(z_i = k)} \qquad \pi_k = \frac{\beta + \sum_{i=1}^n q(z_i = k)}{K\beta + n}$$

You can try setting $\alpha = \beta = 1$. (We will understand this technique, called Laplace smoothing, in the coming lectures.)

(a) Run the EM algorithm, with 20 mixture components for 50 iterations. Plot the log likelihood as it increases over iterations of the algorithm. Has the algorithm converged? Display the learned means of each mixture component.

(b) Do the clusters generally correspond to the classes of handwritten digits?

(c) Do you get the same solution if you run the EM algorithm from different initial starting configurations?

(d) If you increase the number of components do the resulting log likelihood objective function generally increase or decrease?

(e) How many components do you think is sensible?