SMLDM HT 2014 - MSc Problem Sheet 2

1. In lectures we derived the M step updates for a mixture of Gaussians, for the mixing proportions and cluster means, assuming the common covariance $\sigma^2 I$ is fixed and known. What happens to the algorithm if we set σ^2 to be very small? How does the resulting algorithm as $\sigma^2 \to 0$ relate to K-means?

   Answer: In the E step, the posterior probabilities are:

   $$q(z_i = k) \propto \pi_k f(x_i|\phi_k, \sigma^2) \propto \pi_k \exp(-\frac{1}{2\sigma^2}\|x_i - \phi_k\|_2^2) = \pi_k \exp(-\frac{1}{2}\|x_i - \phi_k\|_2^2)^{1/\sigma^2}$$

   When σ^2 is small, σ^{-2} is very large, so that the exponentiated term will be dominated by the k such that ϕ_k is closest to x_i by Euclidean distance. Thus,

   $$q(z_i = k) = \begin{cases} 1 & \text{for } k \text{ such that } \|x_i - \phi_k\| < \|x_i - \phi_c\| \text{ for all } c \neq k. \\ 0 & \text{otherwise} \end{cases}$$

   If there is another ϕ_c at same distance to x_i, $q(z_i)$ will spread probability mass equally among all such components. This looks exactly like the cluster assignment step of K-means. The M step is exactly the mean update step, thus K-means can be understood as an EM algorithm for a mixture of Gaussians with infinitesimally small σ^2.

2. In lectures we derived the M step updates for a mixture of Gaussians, for the mixing proportions and cluster means, assuming the common covariance $\sigma^2 I$ is fixed and known. If σ^2 is in fact not know and to be learnt as well, derive an M step update for σ^2.

   Answer: Differentiating the free energy with respect to $\nu = \sigma^2$ (you can also differentiate with respect to σ or σ^2, just bit more algebra),

   $$\begin{aligned} \nabla_\nu \mathcal{F}(\theta, q) &= \sum_{i=1}^{n}\sum_{k=1}^{K} q(z_i = k)\nabla_\nu \left(-\frac{p}{2}\log(2\pi/\nu) - \nu\frac{1}{2}\|x_i - \phi_k\|_2^2\right) \\ &= \sum_{i=1}^{n}\sum_{k=1}^{K} q(z_i = k)\left(\frac{p}{2}\frac{1}{\nu} - \frac{1}{2}\|x_i - \phi_k\|_2^2\right) = 0 \\ \sigma^2 &= \frac{\sum_{i=1}^{n}\sum_{k=1}^{K} q(z_i = k)\|x_i - \phi_k\|_2^2}{np} \end{aligned}$$

3. 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.

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

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

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

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

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 $\hat{Y}(x)$,

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

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 $X = x$.

5. 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

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

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

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

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.

6. 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 group 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.