CMSC 25025 / STAT 37601

HW2 - DUE: April 18, 2019

Student: Dimitrios Tanoglidis

Problem 1: PCA

Loss:

$$L(\mu, \lambda, V) = \sum_{i=1}^{n} ||x_i - \mu - V_k \lambda_i||^2 \tag{1}$$

Expanding this, we have:

$$\begin{aligned}
L(\mu, \lambda, V) &= \sum_{i=1}^{n} (x_i^T - \mu^T - \lambda_i^T V_k^T)(x_i - \mu - V_k \lambda_i) \\
&= \sum_{i=1}^{n} (x_i^T x_i - x_i^T \mu - x_i^T V_k \lambda_i - \mu^T x_i + \mu^T \mu + \mu^T V_k \lambda_i - \lambda_i^T V_k^T x_i + \lambda_i^T V_k^T \mu + \lambda_i^T V_k^T V_k \lambda_i) \\
&= \sum_{i=1}^{n} (x_i^T x_i - x_i^T \mu - x_i^T V_k \lambda_i - \mu^T x_i + \mu^T \mu + \mu^T V_k \lambda_i - \lambda_i^T V_k^T x_i + \lambda_i^T V_k^T \mu + \lambda_i^T \lambda_i) \quad (2)
\end{aligned}$$

where in the last line we used that $V_k^T V_k = I$.

Now, we take (and set equal to zero) the derivatives with respect to μ and λ_i.

To take this derivatives, we use the following matrix calculus identities:

$$\text{For } \alpha = y^T x, \quad \frac{\partial \alpha}{\partial z} = x^T \frac{\partial y}{\partial z} + y^T \frac{\partial x}{\partial z} \tag{3}$$

and

$$\text{For } \alpha = y^T A x, \quad \frac{\partial \alpha}{\partial z} = x^T A^T \frac{\partial y}{\partial z} + y^T A \frac{\partial x}{\partial z} \tag{4}$$

Using these, we have:

$$\begin{aligned}
\frac{\partial L}{\partial \lambda_i} &= \sum_{j=1}^{n} (-x_j^T V_k + \mu^T V_k - x_i^T V_k + \mu^T V_k + 2\lambda_i^T)\delta_{ij} = 0 \Rightarrow \\
\Rightarrow \quad & 2\lambda_i^T + 2\mu^T V_k - 2x_i^T V_k = 0 \Rightarrow \\
\Rightarrow \quad & \lambda_i^T = x_i^T V_k - \mu^T V_k \Rightarrow \lambda_i = V_k^T x_j - V_k^T \mu \Rightarrow \\
\Rightarrow \quad & \hat{\lambda}_i = V_k^T (x_i - \hat{\mu}) \qquad (5)
\end{aligned}$$

And:

$$
\begin{aligned}
\frac{\partial L}{\partial \mu} &= \sum_{i=1}^{n}(-x_i^T - x_i^T + \mu^T + \mu^T + \lambda_i^t v_k^T + \lambda_i^T V_k^T) = 0 \Rightarrow \\
\Rightarrow & \quad 2\sum_{i=1}^{n}(\mu^T - x_i^T + \lambda_i^T V_k^T) = 0 \Rightarrow \\
\Rightarrow & \quad \sum_{i=1}^{n}(\mu - x_i + V_k \lambda_i) = 0 \Rightarrow \\
\Rightarrow & \quad \sum_{i=1}^{n} \hat{m u} - \sum_{i=1}^{n} x_i + V_k \sum_{i=1}^{n} \hat{\lambda}_i = 0 \Rightarrow \\
\Rightarrow & \quad \hat{\mu} n - n\bar{x}_n + V_k \sum_{i=1}^{n} \hat{\lambda}_i = 0 \Rightarrow \\
\Rightarrow & \quad \hat{\mu} - \bar{x}_n + \frac{1}{n} V_k \sum_{i=1}^{n} \hat{\lambda}_i = 0 \qquad (6)
\end{aligned}
$$

Now using Eq. (5) in (6), we have:

$$
\begin{aligned}
& \hat{\mu} - \bar{x}_n + \frac{1}{n} V_k V_k^T \left(\sum_{i=1}^{n} x_i - n\hat{\mu} \right) = 0 \Rightarrow \\
& \hat{\mu} - \bar{x}_n + V_k V_k^T (\bar{x}_n - \hat{\mu}) = 0 \Rightarrow \\
& (V_k^T V_k - I)(\hat{\mu} - \bar{x}_n) = 0 \qquad (7)
\end{aligned}
$$

Eq. (7) has the **trivial** solution:

$$
\hat{\mu} - \bar{x}_n = 0 \Rightarrow \boxed{\hat{\mu} = \bar{x}_n = \frac{1}{n} \sum_{i=1}^{n} x_i} \qquad (8)
$$

Which, gives for $\hat{\lambda}_i$:

$$
\hat{\lambda}_i = V_k^T (x_i - \bar{x}_n) \qquad (9)
$$

Now, if the rank of $(V_k^T V_k - I) < d$, the solution is not unique, since we also have **non-trivial** solutions.

Problem 2: Bounds on the error probability

a, b are not negative numbers. Without loss of generality, let's assume that $a \leq b$, or $\min(a, b) = a$. So:

$$
a \leq b \Rightarrow a^2 \leq ab \Rightarrow \sqrt{a^2} \leq \sqrt{ab} \Rightarrow a \leq \sqrt{ab} \Rightarrow \boxed{\min(a, b) \leq \sqrt{ab}}. \qquad (10)
$$

Where we used the fact that a, b not negative when multiplying with a without changing the orientation of the inequality and also when taking the square root. The result is the same if we assume that $b \leq a$. Now let's use this to derive the bound of the error rate for a two category Bayes classifier.

The Bayes classifier classifies a vector of features, x as belonging to class $Y = 1, 2$ (can be generalized to classes C_i), according to the value of the posterior probabilities of the two classes. Namely:

$$\begin{aligned} P(Y=2|x) &\geq P(Y=1|x) \to \text{class 2} \quad (11)\\ P(Y=1|x) &\geq P(Y=2|x) \to \text{class 1} \quad (12)\end{aligned}$$

The Bayes error is the total probability of misclassification; namely the probability the vector x to belong in class $Y = 1$ in the region where it is classified as belonging in the class $Y = 2$ and the opposite. Denote these two regions as $\mathcal{R}_1, \mathcal{R}_2$. The Bayes error can be expressed as:

$$\begin{aligned} P(error) &= \int p(error, x)dx \\ &= \int_{\mathcal{R}_1} p(x, Y=2)dx + \int_{\mathcal{R}_2} p(x, Y=1)dx \\ &= \int_{\mathcal{R}_1} P(Y=2|x)p(x)dx + \int_{\mathcal{R}_2} P(Y=1|x)p(x)dx \end{aligned}$$

Now, the regions $\mathcal{R}_1, \mathcal{R}_2$, are defined according to the above inequalities: Region $\mathcal{R}_1$ is that where $P(Y = 2|x) \geq P(Y = 1|x)$ and the opposite. Thus, the Bayes error can be expressed in the compact form:

$$P(error) = \int \min\{P(Y=1|x)p(x), P(Y=2|x)p(x)\}dx \quad (13)$$

Using Bayes' theorem now, we can write:

$$\begin{aligned} P(Y=1|x)p(x) &= f(x|Y=1)P(Y=1) \quad (14)\\ P(Y=2|x)p(x) &= f(x|Y=2)P(Y=2) \quad (15)\end{aligned}$$

And rewrite the Bayes error as:

$$P(error) = \int \min\{f(x|Y=1)P(Y=1), f(x|Y=2)P(Y=2)\} \quad (16)$$

Using now that $\min(a, b) \leq \sqrt{ab}$ for non-negative numbers (like the probabilities), we get the bound:

$$\boxed{P(error) \leq \sqrt{P(Y=1)P(Y=2)} \int \sqrt{f(x|Y=1)f(x|Y=2)}dx} \quad (17)$$

And for $P(Y = 1) = P(Y = 2) = 1/2$, $P(error) \leq \frac{1}{2} \int \sqrt{f(x|Y = 1)f(x|Y = 2)}dx$.

Problem 3: Bayes rule, variances and priors

(a) The features vector will be classified as belonging to class $Y = 1$ when the posterior distributions satisfy the following inequality:

$$P(Y=1|x) \geq P(Y=2|x) \quad (18)$$

Which, from Bayes' theorem is, equivalently:

$$f(x|Y=1)P(Y=1) \geq f(x|Y=2)P(Y=2) \Rightarrow \mathcal{N}(\mu_1, \sigma^2)\pi_1 \geq \mathcal{N}(\mu_2, \sigma^2)\pi_2 \quad (19)$$

So we have:

$$\begin{aligned}
\exp\left[-\frac{1}{2\sigma^2}(x-\mu_1)^2\right]\pi_1 &\geq \exp\left[-\frac{1}{2\sigma^2}(x-\mu_2)^2\right]\pi_2 \Rightarrow \\
-\frac{1}{2\sigma^2}(x-\mu_1)^2+\log\pi_1 &\geq -\frac{1}{2\sigma^2}(x-\mu_2)^2+\log\pi_2 \Rightarrow \\
\frac{1}{2\sigma^2}\left[(x-\mu_1)^2-(x-\mu_2)^2\right] &\leq -\log\left(\frac{\pi_2}{\pi_1}\right) \Rightarrow \\
\left[(x-\mu_1)^2-(x-\mu_2)^2\right] &\leq 2\sigma^2\log\left(\frac{\pi_1}{\pi_2}\right) \Rightarrow \\
-(\mu_1-\mu_2)\cdot[2x-(\mu_1+\mu_2)] &\leq 2\sigma^2\log\left(\frac{\pi_1}{\pi_2}\right) \Rightarrow \\
2x-(\mu_1+\mu_2) &\geq \frac{2\sigma^2}{\mu_1-\mu_2}\log\left(\frac{\pi_1}{\pi_2}\right) \Rightarrow \\
x &\geq \frac{\mu_1+\mu_2}{2}+\frac{\sigma^2}{\mu_1-\mu_2}\log\left(\frac{\pi_1}{\pi_2}\right) \qquad (20)
\end{aligned}$$

So the decision boundary for a feature vector x to be classified as belonging to the class $Y=1$, is:

$$\boxed{x \geq \frac{\mu_1+\mu_2}{2}+\frac{\sigma^2}{\mu_1-\mu_2}\log\left(\frac{\pi_1}{\pi_2}\right)} \qquad (21)$$

(b) Let x_0 denoting the above decision boundary, $x_0=\frac{\mu_1+\mu_2}{2}+\frac{\sigma^2}{\mu_1-\mu_2}\log\left(\frac{\pi_1}{\pi_2}\right)$. The probability of error (probability of misclassification) is, as we explained in the previous problem (problem 2):

$$P(error)=\pi_1\int_{-\infty}^{x_0}\mathcal{N}(\mu_1,\sigma^2)dx+\pi_2\int_{x_0}^{+\infty}\mathcal{N}(\mu_2,\sigma^2)dx \qquad (22)$$

Now, by definition, the first integral is the CDF of the Gaussian:

$$\int_{-\infty}^{x_0}\mathcal{N}(\mu_1,\sigma^2)dx \equiv G(\mu_1,\sigma^2;x_0)=\frac{1}{2}\left[1+\operatorname{erf}\left(\frac{x_0-\mu_1}{\sigma\sqrt{2}}\right)\right] \qquad (23)$$

To compute the second integral, we can use the following trick:

$$\int_{-\infty}^{+\infty}\mathcal{N}(\mu_2,\sigma^2)dx=1\Rightarrow\int_{x_0}^{+\infty}\mathcal{N}(\mu_2,\sigma^2)dx=1-\int_{-\infty}^{x_0}\mathcal{N}(\mu_2,\sigma^2)dx \qquad (24)$$

But here the second integral is simply the CDF of $\mathcal{N}$ so, we can finally write:

$$\int_{x_0}^{+\infty}\mathcal{N}(\mu_2,\sigma^2)dx=1-G(\mu_2,\sigma^2;x_0)=\frac{1}{2}\left[1-\operatorname{erf}\left(\frac{x_0-\mu_2}{\sigma\sqrt{2}}\right)\right] \qquad (25)$$

The error probability is then:

$$\begin{aligned}
P(error) &= \pi_1 G(\mu_1,\sigma^2;x_0)+\pi_2[1-G(\mu_2,\sigma^2;x_0)] \qquad (26)\\
&= \frac{\pi_1}{2}\left[1+\operatorname{erf}\left(\frac{x_0-\mu_1}{\sigma\sqrt{2}}\right)\right]+\frac{\pi_2}{2}\left[1-\operatorname{erf}\left(\frac{x_0-\mu_2}{\sigma\sqrt{2}}\right)\right] \qquad (27)
\end{aligned}$$

For the limit $\sigma\to 0$ we can use that $\lim_{x\to+\infty}(x)=1$.
It is crucial to note that, from the definition of x_0:

$$x_0-\mu_1=\frac{\mu_2-\mu_1}{2}+\frac{\sigma^2}{\mu_1-\mu_2}\log\left(\frac{\pi_1}{\pi_2}\right)<0,\text{ since }\mu_1>\mu_2\text{ when }\sigma\to 0 \qquad (28)$$

and

$$x_0 - \mu_2 = \frac{\mu_1 - \mu_2}{2} + \frac{\sigma^2}{\mu_1 - \mu_2} \log\left(\frac{\pi_1}{\pi_2}\right) > 0, \text{ since } \mu_1 > \mu_2 \text{ when } \sigma \to 0 \tag{29}$$

So, the limit of the error when $\sigma \to 0$ is:

$$\lim_{\sigma \to 0} P(error) = \frac{\pi_1}{2}[1 + \mathrm{erf}(-\infty)] + \frac{\pi_2}{2}[1 - \mathrm{erf}(+\infty)] \tag{30}$$

But $\mathrm{erf}(+\infty) = 1$ and $\mathrm{erf}(-x) = -\mathrm{erf}(x)$, so :

$$\lim_{\sigma \to 0} P(error) = \frac{\pi_1}{2}[1 + (-1)] + \frac{\pi_2}{2}[1 - (+1)] = 0 \tag{31}$$