TD: Basic knowledge 2

1 Invariant representation

Question 1 : Which of the following representations are invariant to circular translations of $x \in \mathbb{R}^N$? Justify your answer.

1. $\Phi(x) = x(0)$
2. $\Phi(x) = \frac{1}{N}\sum_{i=1}^{N} x(i)$
3. $\Phi(x) = \frac{1}{N}\sum_{i=1}^{N} i \cdot x(i)$
4. $\Phi(x) = \frac{1}{N}\sum_{i=1}^{N} x(i) \cdot x(i-2)$

Let T be the circular translation of x by 1 steps: $(Tx)(i) = x(i+1 \mod N)$. with $x(0) = x(N)$

1. $\Phi(Tx) = (Tx)(0) = x(1) \neq x(0)$, so Φ is not invariant.
2. $\Phi(Tx) = \frac{1}{N}\sum_{i=1}^{N}(Tx)(i)$
   $= \frac{1}{N}\sum_{i=1}^{N} x(i+1 \mod N)$
   $= \frac{1}{N}\sum_{j=0}^{N-1} x(j) = \Phi(x)$, so Φ is invariant.
3. $\Phi(Tx) = \frac{1}{N}\sum_{i=1}^{N} i \cdot (Tx)(i)$
   $= \frac{1}{N}\sum_{i=1}^{N} i \cdot x(i+1 \mod N)$
   $= \frac{1}{N}\sum_{j=0}^{N-1}(j+1) \cdot x(j)$
   $= \frac{1}{N}\sum_{j=0}^{N-1} j \cdot x(j) + \frac{1}{N}\sum_{j=0}^{N-1} x(j) = \Phi(x) + \frac{1}{N}\sum_{j=0}^{N-1} x(j)$, so Φ is not invariant.
4. $\Phi(Tx) = \frac{1}{N}\sum_{i=1}^{N}(Tx)(i) \cdot (Tx)(i-2)$
   $= \frac{1}{N}\sum_{i=1}^{N} x(i+1 \mod N) \cdot x(i-1 \mod N)$
   $= \frac{1}{N}\sum_{j=0}^{N-1} x(j) \cdot x(j-2) = \Phi(x)$, so Φ is invariant.

2 Linear discriminant analysis (LDA)

The idea of Fisher is to find $w \in \mathbb{R}^N$ such that

- $|w^T(\mu_1 - \mu_2)|$ is large,
- $s_k(w) = \sum_{i:y_i=c_k}(w^T(x_i - \mu_k))^2$ is small.

This can be achieved by solving

$$\max_{w\in\mathbb{R}^N} J(w) = \frac{|w^T(\mu_1 - \mu_2)|^2}{s_1(w) + s_2(w)}$$

Question 2 : Derive an optimal w by solving $\frac{\partial J}{\partial w} = 0$.

We rewrite $J(w)$ using matrices:

$$(w^T(\mu_1 - \mu_2))^2 = w^T(\mu_1 - \mu_2)(\mu_1 - \mu_2)^T w = w^T A w$$

and

$$s_1(w)+s_2(w) = \sum_{k=1}^{2} \sum_{i:y_i=c_k} (w^T(x_i-\mu_k))^2 = w^T \left(\sum_{k=1}^{2} \sum_{i:y_i=c_k} (x_i - \mu_k)(x_i - \mu_k)^T \right) w = w^T B w$$

with $A = (\mu_1 - \mu_2)(\mu_1 - \mu_2)^T$ and $B = \sum_{k=1}^{2} \sum_{i:y_i=c_k} (x_i - \mu_k)(x_i - \mu_k)^T$.

Thus, we have

$$J(w) = \frac{w^T A w}{w^T B w}$$

We can then solve:

$$\begin{aligned}
&\frac{\partial J}{\partial w} = 0 \\
&\Leftrightarrow \frac{\partial}{\partial w} \left(\frac{w^T A w}{w^T B w} \right) = 0 \\
&\Leftrightarrow \frac{2Bw \cdot (w^T A w) - 2Aw \cdot (w^T B w)}{(w^T B w)^2} = 0 \\
&\Leftrightarrow Bw \cdot (w^T A w) = Aw \cdot (w^T B w) \\
&\Leftrightarrow Bw \cdot \lambda = Aw \qquad \text{with } \lambda = \frac{w^T A w}{w^T B w} = J(w) \text{ (scalar)} \\
&\Leftrightarrow \lambda B w = (\mu_1 - \mu_2)(\mu_1 - \mu_2)^T w \\
&\Leftrightarrow \lambda B w = \gamma(\mu_1 - \mu_2) \qquad \text{with } \gamma = (\mu_1 - \mu_2)^T w \text{ (scalar)} \\
&\Leftrightarrow w = \frac{\gamma}{\lambda} B^{-1}(\mu_1 - \mu_2) \\
&\Leftrightarrow w \propto B^{-1}(\mu_1 - \mu_2)
\end{aligned}$$

Question 3 : How is this w related to the $g_k(x) = \mathbb{P}(y = c_k|x)$ in LDA when $K = 2$?

We have

$$g_k(x) = \mathbb{P}(y = c_k|x) = \frac{\mathbb{P}(x|y = c_k)\mathbb{P}(y = c_k)}{\mathbb{P}(x)}$$

Therefore:

$$\log(g_k(x)) = \log(\mathbb{P}(x|y = c_k)) + \log(\mathbb{P}(y = c_k)) - \log(\mathbb{P}(x))$$

In LDA, we assume that the conditional distribution $\mathbb{P}(x|y = c_k)$ is Gaussian with mean μ_k and covariance matrix Σ. Thus, we have:

$$\log(g_k(x)) = \log(\pi_k) - \frac{1}{2}(x - \mu_k)^T \Sigma^{-1}(x - \mu_k) + \text{constant}$$

where $\pi_k = \mathbb{P}(y = c_k)$ is the prior probability of class c_k.

We can rewrite this as:

$$\log(g_k(x)) = -\frac{1}{2}x^T\Sigma^{-1}x + x^T\Sigma^{-1}\mu_k + \log(\pi_k) - \frac{1}{2}\mu_k^T\Sigma^{-1}\mu_k + \text{constant}$$

The decision boundary is when $g_1(x) = g_2(x)$:

$$\begin{aligned}
& g_1(x) = g_2(x) \\
\Leftrightarrow\ & \log(g_1(x)) = \log(g_2(x)) \\
\Leftrightarrow\ & x^T\Sigma^{-1}\mu_1 + \log(\pi_1) - \frac{1}{2}\mu_1^T\Sigma^{-1}\mu_1 = x^T\Sigma^{-1}\mu_2 + \log(\pi_2) - \frac{1}{2}\mu_2^T\Sigma^{-1}\mu_2 \\
\Leftrightarrow\ & x^T\Sigma^{-1}(\mu_1 - \mu_2) = \frac{1}{2}(\mu_1^T\Sigma^{-1}\mu_1 - \mu_2^T\Sigma^{-1}\mu_2) + \log(\frac{\pi_2}{\pi_1}) \\
\Leftrightarrow\ & x^T\Sigma^{-1}(\mu_1 - \mu_2) = \text{constant}
\end{aligned}$$

This is a linear equation in x, which means that the decision boundary is a hyperplane. The normal vector of this hyperplane is $w' = \Sigma^{-1}(\mu_1 - \mu_2)$.

In the previous question, we defined $B = \sum_{k=1}^{2}\sum_{i:y_i=c_k}(x_i - \mu_k)(x_i - \mu_k)^T$, which is the covariance matrix of the data.

The optimal w obtained from Fisher's LDA is:

$$w \propto B^{-1}(\mu_1 - \mu_2)$$

The normal vector to the decision hyperplane in Gaussian LDA is:

$$w' = \Sigma^{-1}(\mu_1 - \mu_2)$$

3 Discrete Fourier transform

Question 4 : Let $x \in \mathbb{R}^N$ and $\phi_k(u) = e^{iw_k u}$, $w_k = \frac{2\pi k}{N}$. Is the following equation true ? Justify your answer.

$$||x||^2 = \sum_{k=0}^{N-1} |\langle x, \phi_k\rangle|^2$$

By definition, we have

$$\langle x, \phi_k\rangle = \sum_{u=0}^{N-1} x(u)\cdot\overline{\phi_k(u)} = \sum_{u=0}^{N-1} x(u)\cdot e^{-iw_k u} = \hat{x}(w_k) \qquad \text{(discrete Fourier transform of } x\text{)}$$

According to the Parseval's identity, we have

$$||x||^2 = \sum_{u=0}^{N-1} |x(u)|^2 = \frac{1}{N}\sum_{k=0}^{N-1} |\hat{x}(w_k)|^2$$

Thus, the true equation is

$$||x||^2 = \frac{1}{N}\sum_{k=0}^{N-1} |\langle x, \phi_k\rangle|^2$$