Data Literacy

Exercise Sheet #8

due on Monday, 20 December 2021, 10am sharp

1. **EXAMple:** Consider a dataset $X \in \mathbb{R}^{N \times D}$ consisting of N data points of D features each.

   (a) Assume that an orthonormal basis $U = \{\boldsymbol{u}_i\}_{i=1,\dots,D}$, $\boldsymbol{u}_i^\intercal \boldsymbol{u}_j = \delta_{ij}$ is given. Show that the *square reconstruction error*

   $$J = \frac{1}{N} \sum_{n=1}^{N} \|\boldsymbol{x}_n - \tilde{\boldsymbol{x}}_n\|^2 = \frac{1}{N} \sum_{n=1}^{N} \sum_{j=1}^{D} \left[\boldsymbol{x}_n - \sum_{i=1}^{M} a_{ni} \boldsymbol{u}_i - \sum_{i=M+1}^{D} b_i \boldsymbol{u}_i \right]_j^2$$

   is minimized by the choice $a_{ni} = \boldsymbol{x}_n^\intercal \boldsymbol{u}_i$, and $b_i = \bar{\boldsymbol{x}}^\intercal \boldsymbol{u}_i$.

2. **Theory Question:** Fisher's choice of univariate projection for linear discriminant analysis was introduced in the lecture. However, the actual *discriminant* itself was not derived. This latter step is the goal of this exercise. To find a projection $\boldsymbol{w} \in \mathbb{R}^d$ that "separates" a dataset $X = [\boldsymbol{x}_1, \dots, \boldsymbol{x}_N] \in \mathbb{R}^{D \times N}$ into two classes $\mathscr{C}_1, \mathscr{C}_2$ (i.e. index sets with $\mathscr{C}_1 \cup \mathscr{C}_2 = [1, \dots, N]$ and $\mathscr{C}_1 \cap \mathscr{C}_2 = \emptyset$), we consider the class means and unnormalized covariances

   $$\boldsymbol{m}_i := \frac{1}{N_i} \sum_{n \in \mathscr{C}_i} \boldsymbol{x}_n \qquad \text{(where } N_i := |\mathscr{C}_i|\text{), and } S_i := \sum_{n \in \mathscr{C}_i} (\boldsymbol{x}_n - \boldsymbol{m}_i)(\boldsymbol{x}_n - \boldsymbol{m}_i)^\intercal.$$

   LDA finds a projection $\boldsymbol{w} \in \mathbb{R}^D$. The projected points $x_n = \boldsymbol{w}^\intercal \boldsymbol{x}_n$ then have means $m_i = \boldsymbol{w}^\intercal \boldsymbol{m}_i$ and variance $s_i^2 = \frac{1}{N_i} \boldsymbol{w}^\intercal S_i \boldsymbol{w}$. We can assume[1] that this defines two separate Gaussian distributions $p(x_n \mid n \in \mathscr{C}_i) = \mathcal{N}(x_n; m_i, s_i^2)$. We can then consider a generative Gaussian mixture model that assigns prior probability $p(n \in \mathscr{C}_i) = N_i/(N_1 + N_2)$ to each class, then draws x_n by first drawing class i from $p(n \in \mathscr{C}_i)$, then drawing x_n from $p(x_n \mid n \in \mathscr{C}_i)$.

   One way to motivate a discriminant (i.e. a rule to predict class 1 or 2 from x_n) is thus to predict, at x_n, the class with higher posterior probability. The discriminant then lies at the point where the log odds cancel, i.e. where

   $$\log \frac{p(\mathscr{C}_1 \mid x_n)}{p(\mathscr{C}_2 \mid x_n)} = \log \frac{p(x_n \mid n \in \mathscr{C}_1) p(n \in \mathscr{C}_1)}{p(x_n \mid n \in \mathscr{C}_2) p(n \in \mathscr{C}_2)} = 0 \tag{1}$$

   (a) Solve Equation 1 above for $x \in \mathbb{R}$, by plugging in the definitions of the terms from above.

   (b) You will find that you arrive at a quadratic equation with two solutions $x_{1,2}$. Which one of these two is the discriminant?

3. **Practical Question:** You can find this week's practical exercise in `Exercise_08.ipynb`

[1] A frequently cited motivation for this assumption is that each scalar $x_n = \boldsymbol{w}^\intercal \boldsymbol{x}_n$ is a sum of d random variables. Thus, the Central Limit Theorem suggests that the x_n are indeed approximately Gaussian.