Exercise Sheet #9

due on Tuesday, 19 January 2021, 12 noon sharp

1. **EXAMple:** The lecture introduced Fisher's linear discriminant analysis. Fisher's LDA is a relatively simple algorithm, but there is an even simpler one: To find a projection $\boldsymbol{w} \in \mathbb{R}^d$ that "separates" a dataset $X = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N] \in \mathbb{R}^{D \times N}$ into two classes $\mathscr{C}_1, \mathscr{C}_2$ (i.e. index sets with $\mathscr{C}_1 \cup \mathscr{C}_2 = [1, \ldots, N]$ and $\mathscr{C}_1 \cap \mathscr{C}_2 = \varnothing$), we can consider the class means

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

and choose a projection $\boldsymbol{w} \in \mathbb{R}^D$ such that the distance of projected classes $m_1 - m_2 = \boldsymbol{w}^\intercal(\boldsymbol{m}_1 - \boldsymbol{m}_2)$ is maximized.[1] Of course, this can be made arbitrarily large by increasing $\|\boldsymbol{w}\|$, so we have to constrain $\|\boldsymbol{w}\| = 1$ with a Lagrange multiplier λ: Find this "optimal" projection $\boldsymbol{w}$ by computing the gradient of

$$u(\boldsymbol{w}) = m_1 - m_2 + \lambda(1 - \|w\|^2),$$

setting it to zero, and using the constraint $\|w\| = 1$ to find the value of λ.

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. Consider data as in Exercise 1 above, with class means $\boldsymbol{m}_i$ as above and unnormalized covariances

$$S_i := \sum_{n \in \mathscr{C}_i} (\boldsymbol{x}_n - \boldsymbol{m}_i)(\boldsymbol{x}_n - \boldsymbol{m}_i)^\intercal.$$

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[2] 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_09.ipynb`

[1] This reason that this discriminant is not used in practice is that it does not necessarily separate the classes well, unless the the data covariance aligns with the vector $\boldsymbol{m}_1 - \boldsymbol{m}_2$.

[2] 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.