STAT161/261: Quiz Solutions

1. Basic PCA. Given 3 data points in 2-d space, $(1,1)$, $(2,2)$ and $(3,3)$:

   (a) The sample mean is
   $$\mathbf{m} = \frac{1}{3}\left[(1,1)+(2,2)+(3,3)\right] = (2,2).$$
   The data matrix with the mean removed is
   $$\tilde{\mathbf{X}} = \begin{bmatrix} -1 & -1 \\ 0 & 0 \\ 1 & 1 \end{bmatrix}.$$
   The fastest way to compute the PCA in this case is to see that
   $$\tilde{\mathbf{X}} = \begin{bmatrix} -1 \\ 0 \\ 1 \end{bmatrix} [1\ 1] = s\mathbf{u}\mathbf{v}^T,$$
   where $\mathbf{u}$ and $\mathbf{v}$ are unit vectors,
   $$\mathbf{u} = \frac{1}{\sqrt{2}}\begin{bmatrix} -1 \\ 0 \\ 1 \end{bmatrix}, \quad \mathbf{v} = \frac{1}{\sqrt{2}}\begin{bmatrix} 1 \\ 1 \end{bmatrix},$$
   and $s = 2$. The representation $\tilde{\mathbf{X}} = s\mathbf{u}\mathbf{v}^T$ is an SVD of $\tilde{\mathbf{X}}$, so the sample covariance matrix $\mathbf{S} = \tilde{\mathbf{X}}^T\tilde{\mathbf{X}}$ has eigenvalues
   $$\lambda_1 = s^2 = 4, \quad \lambda_2 = 0,$$
   with the eigenvector corresponding to λ_1 being $\mathbf{v}$.

   (b) In this case, the PoV using $K = 1$ PC is
   $$\text{POV} = \frac{\lambda_1}{\lambda_1 + \lambda_2} = 1.$$

   (c) Since the PoV= 1, the data variance is entirely explained with one component, so there is not reconstruction error.

2. The k-means algorithm will proceed as follows:

   - Initial cluster centers: $\mu_1 = (0,0)$, $\mu_2 = (0,2)$.
   - Membership: Cluster 1: $(0,0)$, $(3,0)$. Cluster 2: $(0,2)$
   - Updated cluster centers: $\mu_1 = (1.5,0)$, $\mu_2 = (0,2)$.
   - Membership: Cluster 1: $(0,0)$, $(3,0)$. Cluster 2: $(0,2)$

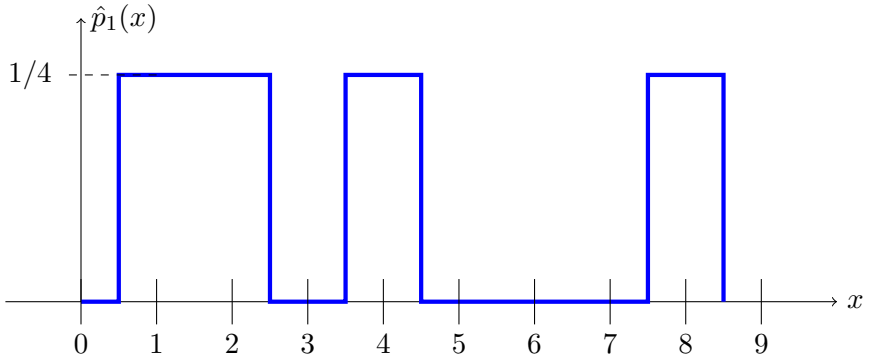

Figure 1: Non-parametric estimate with fixed bandwidth h.

The algorithm will now remain at the same point.

3. (a) The simplest method is to subtract the mean μ so that

$$y - \mu = w_0 + w_1 x + w_2 x^2 + \epsilon, \quad \epsilon \sim \mathcal{N}(0, 1),$$

Now the problem is a standard least squares problem. Define the matrix $\mathbf{X}$ and vector $\mathbf{b}$:

$$\mathbf{X} = \begin{bmatrix} 1 & x_1 & x_1^2 \\ \vdots & & \vdots \\ 1 & x_N & x_N^2 \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} y_1 - \mu \\ \vdots \\ y_N - \mu \end{bmatrix}.$$

Hence the model is

$$\mathbf{b} = \mathbf{X}\mathbf{w} + \epsilon,$$

and the LS solution is

$$\widehat{\mathbf{w}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{b}.$$

You do not need to simplify this beyond this point.

(b) Since the noise ϵ is zero mean, the LS estimate will be unbiased.

4. Nonparametric:

(a) This is shown in Fig. 1. It is relatively easy to plot since the non-parametric estimate will simply be four rectangles centered around the data points. The first two rectangles in this case are side-by-side so the estimate appears as three rectangles.

(b) This one is tedious to compute by hand. We first compute $d_2(x)$, the distance to the second closest point to x. This is plotted in the top figure in Fig. 2. Once this is computed, the density can be plotted as in the lower figure.

(c) In part (a), the probability estimate is

$$P(X \geq 2) = \int_2^\infty \widehat{p}_1(x)dx = \frac{1}{4}\left[0.5 + 1 + 1\right] = \frac{2.5}{4},$$

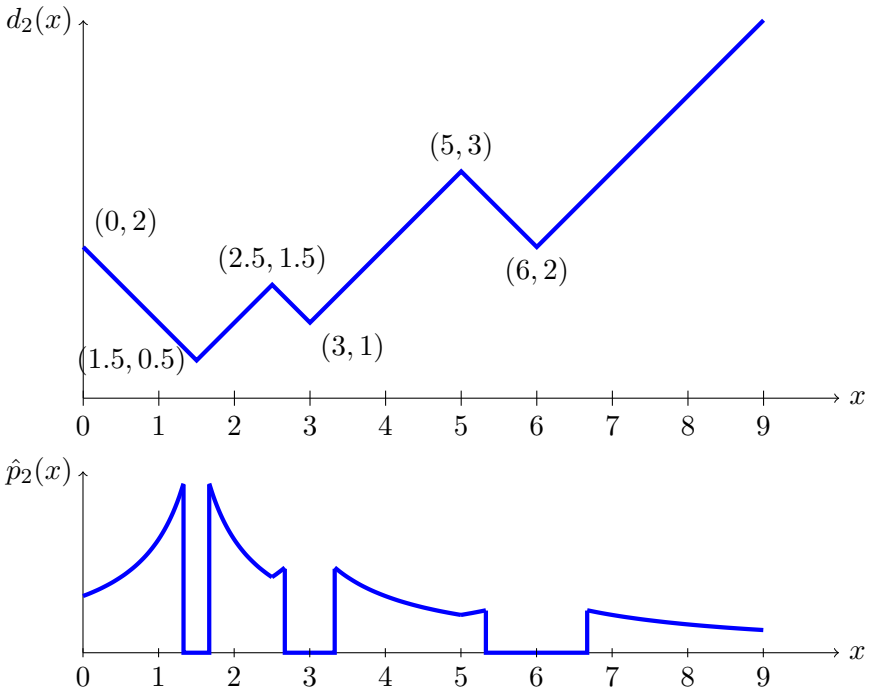

Figure 2: Top: Distance to second closest point $d_2(x)$. Bottom: The non-parametric density estimate with variable bandwidth.

where in the second to last step we simply took the area under the curve for the values $x \geq 2$. For the density estimate in part (b), note that for $x \geq 7$ (the last non-zero portion):

$$\widehat{p}_2(x) = \frac{1}{4(x-4)}.$$

Hence,

$$P(X \geq 2) = \int_2^\infty \widehat{p}_2(x)dx \geq \int_7^\infty \frac{1}{4(x-4)}dx = \infty,$$

since the integral doesn't converge. This occurs since $\widehat{p}_2(x)$ is not a proper density.

5. False. PCA may be poor in discriminating classes. This is why LDA is preferred.

6. True. As explained in class, this occurs since the EM can be considered as a type majorization-minimization algorithm.

7. True.

8. (a) The log likelihood function is

$$L(\theta) = \sum_{i=1}^{N} \ln p(x_i|\theta) = N \ln(\theta) - \sum_{i=1}^{N} x_i\theta.$$

Taking the derivative

$$\frac{\partial L(\theta)}{\partial \theta} = 0 \Leftrightarrow \frac{N}{\theta} = \sum_{i=1}^{N} x_i \Leftrightarrow \widehat{\theta} = \left[\frac{1}{N} \sum_{i=1}^{N} x_i \right]^{-1}.$$

Thus, the ML estimate is the inverse of the sample mean.

(b) In this case, we would maximize

$$J(\theta) = \ln p(\mathbf{x}|\theta) + \ln p(\theta) = N \ln(\theta) - \sum_{i=1}^{N} x_i \theta + \frac{(\theta - \mu)^2}{2\sigma^2} + \text{const},$$

where $\mu = 0$, $\sigma^2 = 1$ and the constant term does not depend on θ.