Homework 8 Solutions

1 Clustering and informative projections

1. *Suboptimality of Lloyd's algorithm.*

   (a) The optimal solution has centers at $-9, 0, 9$, with a cost of 4.

   (b) With this initialization, Lloyd's algorithm converges to the centers $-10, -8, 6$, with a cost of 56.

2. *Projections*

   (a) $u_1, u_2 \in \mathbb{R}^p$ and $U = \begin{pmatrix} | & | \\ u_1 & u_2 \\ | & | \end{pmatrix}$. So we the following dimensions.

   - U: $p \times 2$
   - U^T: $2 \times p$
   - UU^T: $p \times p$
   - $u_1 u_1^T$: $p \times p$

   (b) We can collect the mappings into two groups:

   - $x \mapsto (u_1 \cdot x, u_2 \cdot x)$ and $x \mapsto U^T x$ are the same projection onto the u_1, u_2 directions
   - $x \mapsto (u_1 \cdot x)u_1 + (u_2 \cdot x)u_2$ and $x \mapsto UU^T x$ are the same reconstruction from the above projection

3. *Eigenvectors*

   (a) We can work out

$$\mathbb{E}[X \cdot u] = \mathbb{E}[X] \cdot u = \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix} \cdot \frac{1}{\sqrt{3}} \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} = \frac{6}{\sqrt{3}} = 2\sqrt{3}$$

$$\text{var}(X \cdot u) = u^T \text{cov}(X) u = \frac{1}{3} \begin{pmatrix} 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} 5 & -3 & 0 \\ -3 & 5 & 0 \\ 0 & 0 & 4 \end{pmatrix} \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} = \frac{1}{3}(5 - 3 - 3 + 5 + 4) = \frac{8}{3}$$

   (b) + (c) By multiplying out each of the vectors with $\text{cov}(X)$, we can find which ones are the eigenvectors. We see that $\begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}$ satisfies

$$\begin{pmatrix} 5 & -3 & 0 \\ -3 & 5 & 0 \\ 0 & 0 & 4 \end{pmatrix} \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 4 \end{pmatrix} = 4 \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}$$

   Thus this is an eigenvector with eigenvalue 4. We see $\frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}$ satisfies

$$\begin{pmatrix} 5 & -3 & 0 \\ -3 & 5 & 0 \\ 0 & 0 & 4 \end{pmatrix} \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix} = \frac{1}{\sqrt{2}} \begin{pmatrix} 2 \\ 2 \\ 0 \end{pmatrix} = \frac{2}{\sqrt{2}} \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}$$

Thus this is an eigenvector with eigenvalue 2. We see $\frac{1}{\sqrt{2}}\begin{pmatrix}1\\-1\\0\end{pmatrix}$ satisfies

$$\begin{pmatrix}5 & -3 & 0\\ -3 & 5 & 0\\ 0 & 0 & 4\end{pmatrix}\frac{1}{\sqrt{2}}\begin{pmatrix}1\\-1\\0\end{pmatrix} = \frac{1}{\sqrt{2}}\begin{pmatrix}8\\-8\\0\end{pmatrix} = \frac{8}{\sqrt{2}}\begin{pmatrix}1\\-1\\0\end{pmatrix}$$

Thus this is an eigenvector with eigenvalue 8.

(d) PCA would choose the two eigenvectors with the largest eigenvalues: $u_1 = \frac{1}{\sqrt{2}}\begin{pmatrix}1\\-1\\0\end{pmatrix}$ and $u_2 = \begin{pmatrix}0\\0\\1\end{pmatrix}$.

(e) The two dimensional projection of $x = (4, 0, 2)$ would be

$$\begin{pmatrix}x \cdot u_1\\ x \cdot u_2\end{pmatrix} = \begin{pmatrix}\frac{4}{\sqrt{2}}\\ 2\end{pmatrix} = \begin{pmatrix}2\sqrt{2}\\ 2\end{pmatrix}$$

(f) The three-dimensional reconstruction would be

$$(u_1 \cdot x)u_1 + (u_2 \cdot x)u_2 = 2\sqrt{2}u_1 + 2u_2 = \frac{2\sqrt{2}}{\sqrt{2}}\begin{pmatrix}1\\-1\\0\end{pmatrix} + 2\begin{pmatrix}0\\0\\1\end{pmatrix} = \begin{pmatrix}2\\-2\\2\end{pmatrix}$$

4. *Spectral decomposition*

(a) We can construct M from its spectral decomposition:

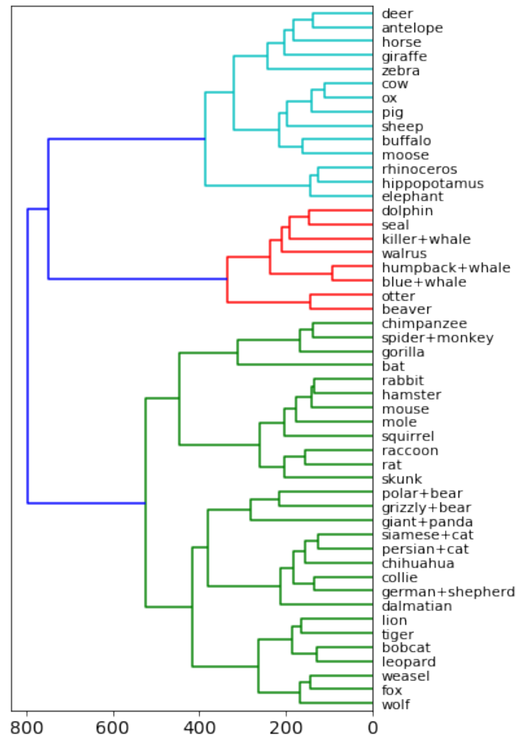
$$M = \begin{pmatrix}| & |\\ u_1 & u_2\\ | & |\end{pmatrix}\begin{pmatrix}\lambda_1 & 0\\ 0 & \lambda_2\end{pmatrix}\begin{pmatrix}— & u_1 & —\\ — & u_2 & —\end{pmatrix} = \frac{1}{5}\begin{pmatrix}2 & -1\\ 1 & 2\end{pmatrix}\begin{pmatrix}2 & 0\\ 0 & -1\end{pmatrix}\begin{pmatrix}2 & 1\\ -1 & 2\end{pmatrix} = \frac{1}{5}\begin{pmatrix}7 & 6\\ 6 & -2\end{pmatrix}$$

(b) If $Mu = \lambda u$ then $(M + 2I)u = (\lambda + 2)u$. Thus the eigenvalues of $M + 2I$ are $\lambda_1 + 2 = 4$ and $\lambda_2 + 2 = 1$.

(c) If $Mu = \lambda u$ then $M^2u = M(Mu) = M(\lambda u) = \lambda(Mu) = \lambda^2 u$. Thus the eigenvalues of M^2 are $\lambda_1^2 = 4$ and $\lambda_2^2 = 1$.

2 Programming problems

1. *Hierarchically clustering animals.*



2. *PCA on animals.* We can use the PCA package from scikit-learn to fit and transform the data to two dimension. Below is the result of this procedure. The embedding seems to be fairly reasonable. Many land predators are mapped closely together in the upper left-hand corner, while some sea mammals seem to also be mapped to the upper right-hand region.