Outline for April 17

• Practice K-means clustering
• Gaussian Mixture Models (GMM)
• Friday:
  – Finish GMMs
  – Hierarchical clustering algorithms
  – PCA and dimensionality reduction
• Mon/Wed next week: midterm review

• Today in lab: work on proposals
• Goal: finish by end of lab (officially due Friday)
• Midterm 2 next Wed in lab (pick up a study guide!)
Announcements

* Lab today: work on proposal (due Friday)
* Midterm 2 next Wed
* Hand back SVM pset (go over in lab)

Today

* K-means
* Gaussian Mixture Models (GMMs)
Outline for April 17

• Practice K-means clustering

• Gaussian Mixture Models (GMM)

• Friday:
  – Finish GMMs
  – Hierarchical clustering algorithms
  – PCA and dimensionality reduction

• Mon/Wed next week: midterm review
**K-means**

Goal: minimize

\[ J(c) = WCSS = \sum_{k=1}^{K} \sum_{x_i \in E_k} \| x_i - \bar{x}_k \|^2 \]

\[ WCSS: \text{within cluster sum of squares} \]

\[ \Rightarrow \text{NP-hard (non-convex)} \]

K-means "does well"
**E-step** expected cluster membership, given current means.

**M-step** given cluster membership, find means that maximize the likelihood of the data.

**How to choose K?**

Graph showing the "elbow" and using it as the best K.
Handout 17

ITER 1

\[ c_{1}^{(1)} = \frac{3}{3} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \cdot \frac{3}{3} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}, \quad C_{2}^{(1)} = \frac{3}{3} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}, \quad c_{2}^{(1)} = \frac{3}{3} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \]

ITER 2

\[ M_{1} = \begin{pmatrix} 3 & 3 \\ 3 & 3 \end{pmatrix} \]

\[ M_{2} = \begin{pmatrix} 3 & 3 \\ 3 & 3 \end{pmatrix} \]

END

\[ \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \cdot \begin{pmatrix} 2.5 \\ 2 \end{pmatrix} \]

\[ \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \cdot \begin{pmatrix} 4 \\ -1 \end{pmatrix} \]
2. Yes (monotonic)

3. $K = 1$
   \[ \bar{\mu} = \begin{bmatrix} \frac{3}{3} \\ \frac{1}{1} \end{bmatrix} = \begin{bmatrix} (3+2+4)/3 \\ (2+2-1)/3 \end{bmatrix} \]

   $\text{WCCSS(1)} = (\sqrt{2})^2 + (1)^2 + (\sqrt{5})^2 = 8$
   $\text{WCCSS(2)} = (\frac{1}{2})^2 + (\frac{1}{2})^2 + 0^2 = \frac{1}{2}$
   $\text{WCCSS(3)} = 0$

4. $K = 2$

Diagram:
- $1 \rightarrow 2 \rightarrow 3$
Outline for April 17

• Practice K-means clustering

• Gaussian Mixture Models (GMM)

• Friday:
  – Finish GMMs
  – Hierarchical clustering algorithms
  – PCA and dimensionality reduction

• Mon/Wed next week: midterm review
GMM
EM
algorithm

Initialization

\( \Pi_k = \text{cluster "size" (\# of datapoints)} \)

\( \Pi_k = \frac{1}{K} \)

\( \tilde{\mu}_k = \text{choose } K \text{ data points to be means} \)

\( \tilde{\sigma}_k^2 = \text{based on sample variance of points closest to each mean} \)
Example of different co-variance constraints on the Iris flower data