CS 66: Machine Learning

Prof. Sara Mathieson

Spring 2019

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

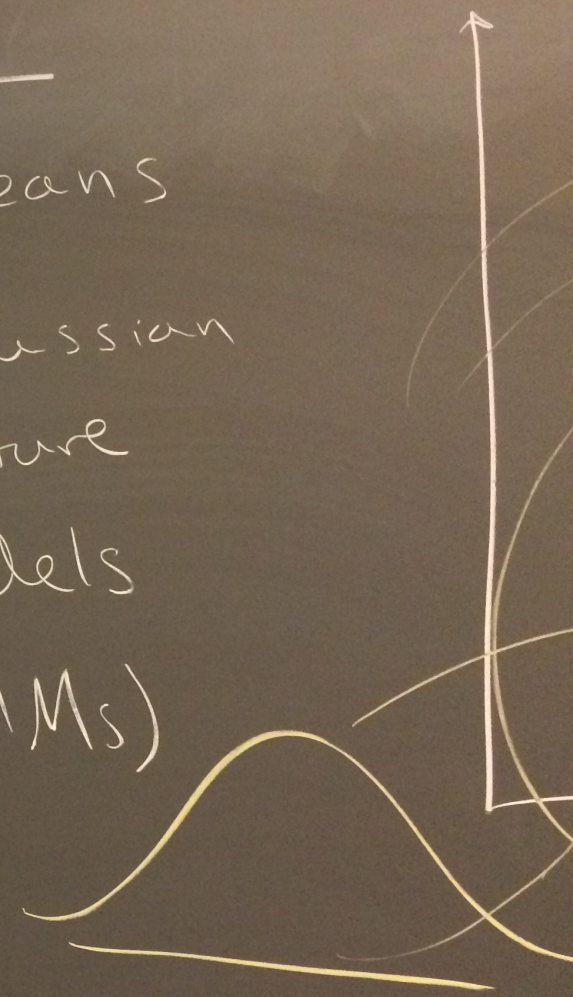
- Today in lab: work on proposals
- Goal: finish by end of lab (officially due Friday)
- Midterm 2 next Wed in lab (pick up a **study guide**!)

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

WCSS : within cluster sum of squares

$$J(\mathcal{C}) = WCSS = \sum_{k=1}^{K} \sum_{\vec{x}_i \in \mathcal{C}_k} \underbrace{\| \vec{x}_i - \vec{\mu}_k \|^2}_{\text{Euclidean distance}}$$

Goal

minimize

$\Rightarrow$ NP-hard (non-convex)

K-means "does well"

uares

$\|\mu_k\|^2$

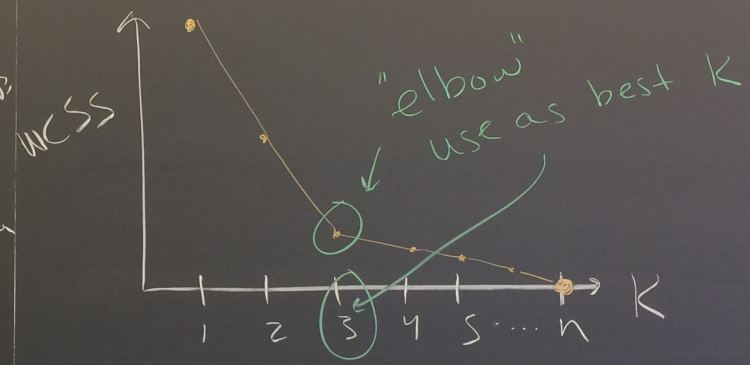
lean

nce

$\boxed{\text{E-step}}$ expected cluster membership, given current means.

$\boxed{\text{M-step}}$ given cluster membership, find means that maximize the likelihood of the data

How to choose K?



Hando

ITER 1

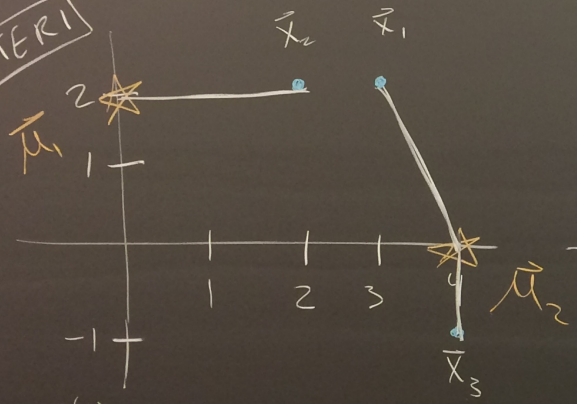
2

1

-1

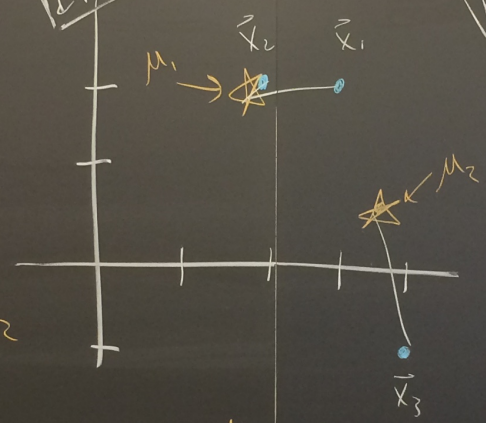
Handout 17

ITER 1



$C_1^{(1)} = \{\vec{x}_2\}$, $C_2^{(1)} = \{\vec{x}_1, \vec{x}_3\}$

ITER 2



$\mu_1 = \begin{bmatrix} 2 \\ 2 \end{bmatrix}$

$\mu_2 = \begin{bmatrix} 3.5 \\ 0.5 \end{bmatrix}$

END

$\mu_1 = \begin{bmatrix} 2.5 \\ 2 \end{bmatrix}$

$\mu_2 = \begin{bmatrix} 4 \\ -1 \end{bmatrix}$

$\{\vec{x}_1, \vec{x}_2\} = C_1^{(2)}$

$\{\vec{x}_3\} = C_2^{(2)}$

②

③

④

② yes (monotonic)

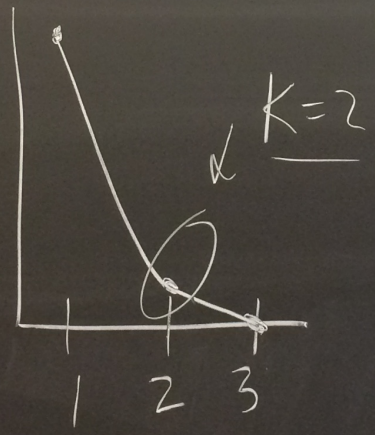
③ $K=1$ $\quad \vec{\mu} = \begin{bmatrix} 3 \\ 1 \end{bmatrix} = \begin{bmatrix} (3+2+4)/3 \\ (2+2-1)/3 \end{bmatrix}$

$WCSS(1) = (\sqrt{2})^2 + (1)^2 + (\sqrt{5})^2 \approx 8$

$WCSS(2) = \left(\frac{1}{2}\right)^2 + \left(\frac{1}{2}\right)^2 + 0^2 = \frac{1}{2}$

$WCSS(3) = 0$

④ ??



GMM

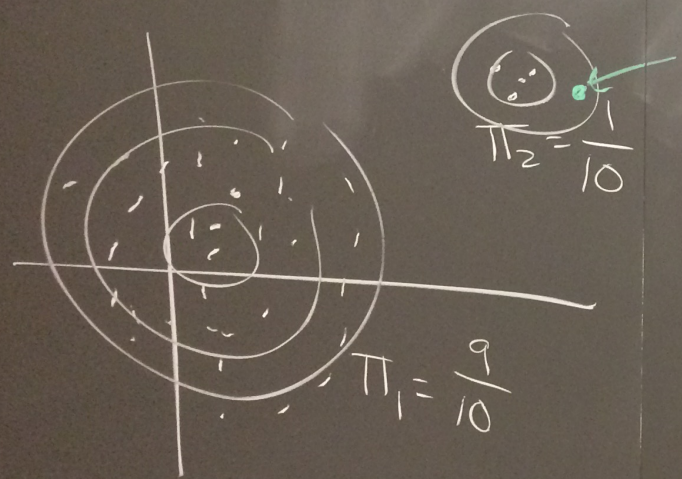
Outline for April 17

- 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

- π_k = cluster "size" (# of datapoints)

  $\pi_k = \frac{1}{K}$

- $\vec{\mu}_k$ = choose K data points to be means

- σ_k^2 = based on sample variance of points closest to each mean

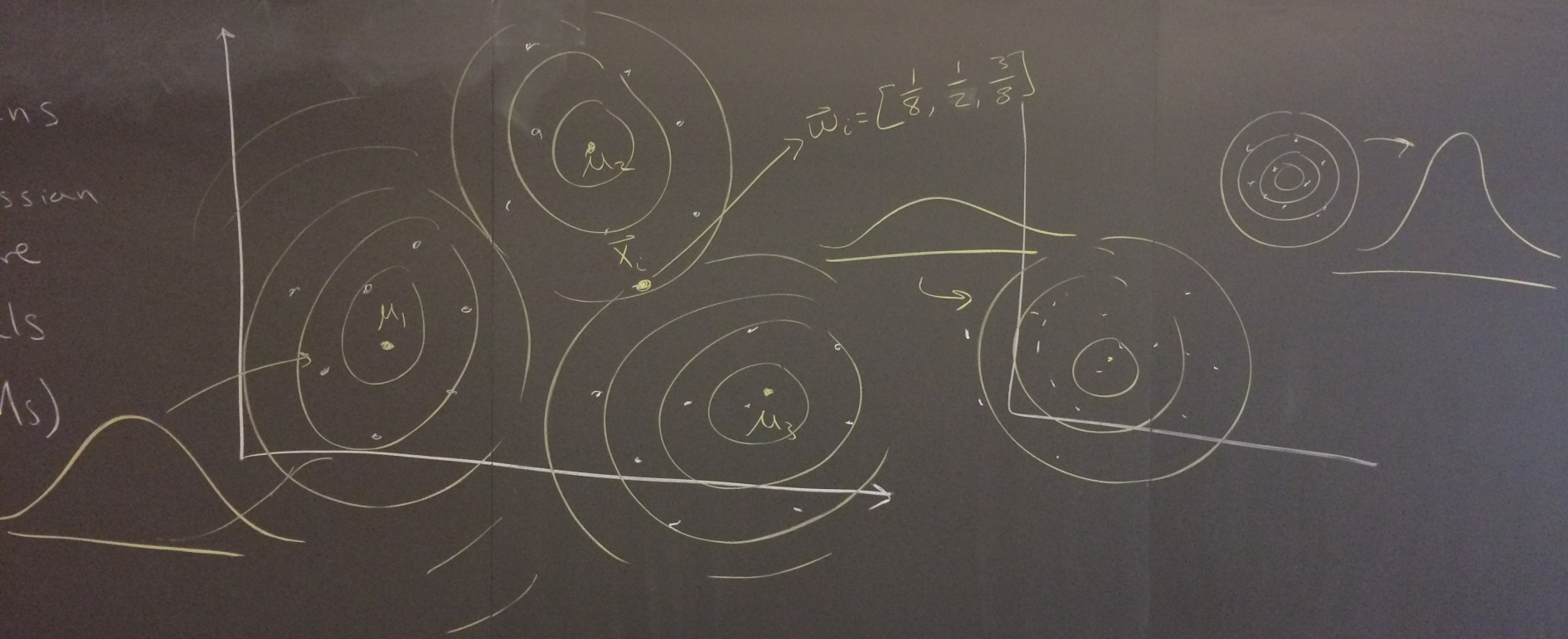
ns

ssian

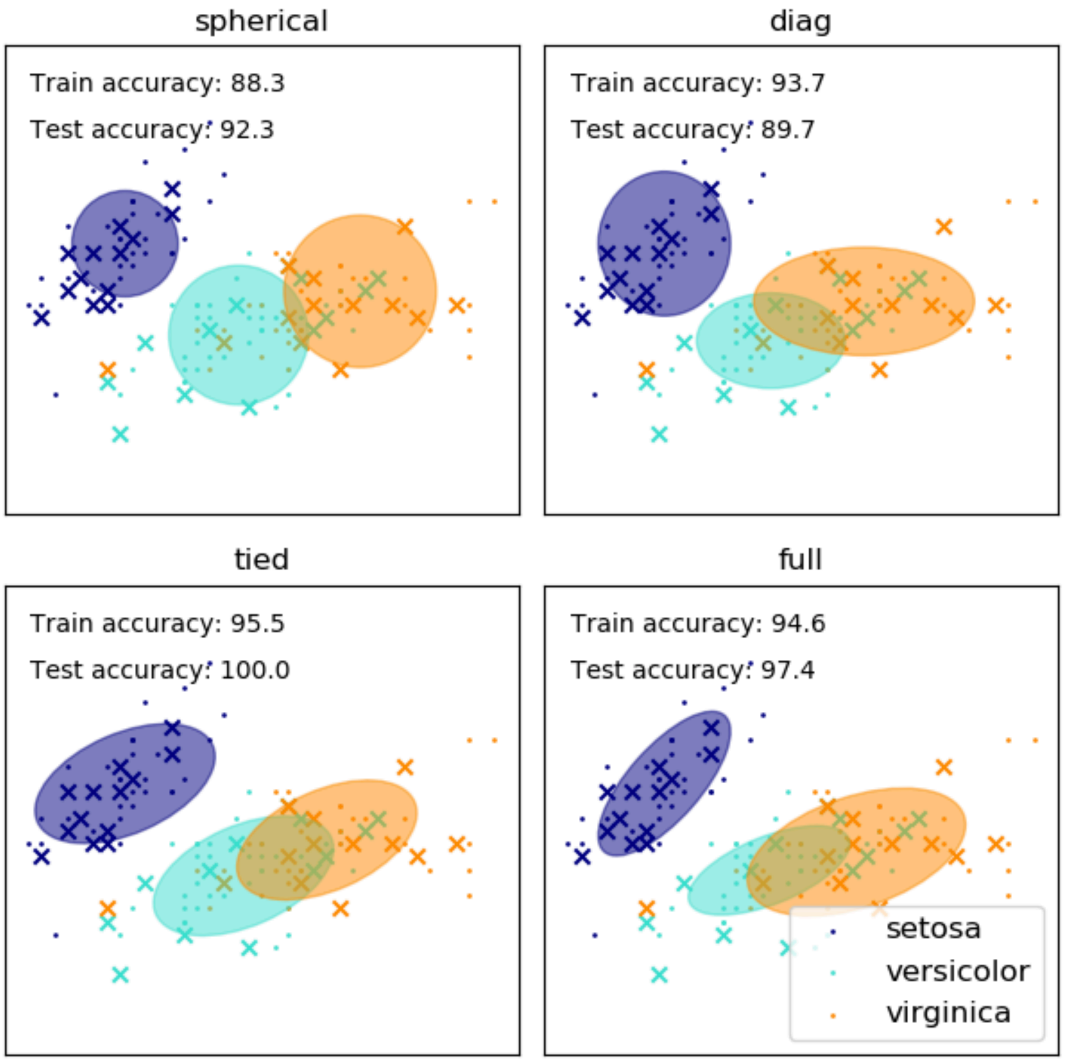
re

ls

Ms)



Example of different co-variance constraints on the Iris flower data



https://scikit-learn.org/stable/auto_examples/mixture/plot_gmm_covariances.html#sphx-glr-auto-examples-mixture-plot-gmm-covariances-py