

CS195-5 : Introduction to Machine Learning

Lecture 23

Greg Shakhnarovich

November 8, 2006

Announcements

Review: MDL

- Learning as compression: find the way to compress data most efficiently.
- Minimum Description length (MDL) principle:
 - Estimate parameters $\hat{\theta}_{\mathcal{M}}$ (ML, MAP etc.)
 - The two-stage MDL criterion for model selection:

$$\mathcal{M}^* = \operatorname{argmin}_{\mathcal{M}} DL \left(X_N \mid \hat{\theta}_{\mathcal{M}} \right) + DL \left(\hat{\theta}_{\mathcal{M}} \right).$$

- BIC is an asymptotic approximation for MDL, with $\hat{\theta}_{\mathcal{M}}$ estimated/transmitted with precision $1/\sqrt{N}$.

Review: information theory

- Entropy of a discrete RV X with multinomial pmf p , where $p_i = p(X = i)$:

$$H(p) = - \sum_{i=1}^M p_i \log p_i.$$

- Optimal code for X achieves, asymptotically, $H(p)$ bits per symbol.
- Cost of coding X with code optimal for $\hat{p} = q$ when the true pmf is p : the Kullback-Leibler divergence

$$D_{KL}(p||q) \triangleq \sum_{i=1}^m p_i \log \frac{p_i}{q_i}.$$

Plan for today

- A few properties of KL-divergence
- Proof of EM monotonically increasing likelihood.
- Unsupervised learning: clustering
 - k -means

Properties of KL-divergence

- $D_{KL}(p||q) \geq 0$ for any p, q
- $D_{KL}(p||q) = 0$ if and only if $p \equiv q$
- It's asymmetric:
 - If $p_i = 0, q_i \geq 0$

Properties of KL-divergence

- $D_{KL}(p||q) \geq 0$ for any p, q
- $D_{KL}(p||q) = 0$ if and only if $p \equiv q$
- It's asymmetric:
 - If $p_i = 0, q_i \geq 0 \Rightarrow 0 \cdot \log(0) \rightarrow 0$.

Properties of KL-divergence

- $D_{KL}(p||q) \geq 0$ for any p, q
- $D_{KL}(p||q) = 0$ if and only if $p \equiv q$
- It's asymmetric:
 - If $p_i = 0, q_i \geq 0 \Rightarrow 0 \cdot \log(0) \rightarrow 0$.
 - If $q_i = 0, p_i \geq 0$

Properties of KL-divergence

- $D_{KL}(p||q) \geq 0$ for any p, q
- $D_{KL}(p||q) = 0$ if and only if $p \equiv q$
- It's asymmetric:
 - If $p_i = 0, q_i \geq 0 \Rightarrow 0 \cdot \log(0) \rightarrow 0$.
 - If $q_i = 0, p_i \geq 0 \Rightarrow p_i \cdot \log(p_i/0) \rightarrow \infty$.

Properties of KL-divergence

- $D_{KL}(p||q) \geq 0$ for any p, q
- $D_{KL}(p||q) = 0$ if and only if $p \equiv q$
- It's asymmetric:
 - If $p_i = 0, q_i \geq 0 \Rightarrow 0 \cdot \log(0) \rightarrow 0$.
 - If $q_i = 0, p_i \geq 0 \Rightarrow p_i \cdot \log(p_i/0) \rightarrow \infty$.
- Continuous KL-divergence:

$$D_{KL}(p||q) = \int p(\mathbf{x}) \log \frac{p(\mathbf{x})}{q(\mathbf{x})} d\mathbf{x}.$$

Back to EM

- Recall: X are observed, Z are hidden; by chain rule

$$p(X, Z | \theta) = p(Z | X, \theta) p(X | \theta)$$

$$\log p(X, Z | \theta) - \log p(Z | X, \theta) = \log p(X | \theta)$$

- Now take expectation w.r.t. $p(Z | X, \theta^{old})$:

$$\underbrace{E_{p(Z | X, \theta^{old})} [\log p(X, Z | \theta)]}_{Q(\theta; \theta^{old})} - E_{p(Z | X, \theta^{old})} [\log p(Z | X, \theta)] = \log p(X | \theta)$$

Likelihood of EM solution

$$\log p(X | \theta) = Q(\theta; \theta^{old}) - E_{p(Z | X, \theta^{old})} [\log p(Z | X, \theta)]$$

- Since $\theta^{new} = \operatorname{argmax}_{\theta} Q(\theta; \theta^{old})$, we have $Q(\theta^{new}; \theta^{old}) \geq Q(\theta^{old}; \theta^{old})$.
- Also,

$$\begin{aligned} & E_{p(Z | X, \theta^{old})} [\log p(Z | X, \theta^{old})] - E_{p(Z | X, \theta^{old})} [\log p(Z | X, \theta^{new})] \\ &= \int p(Z | X, \theta^{old}) \log \frac{p(Z | X, \theta^{old})}{p(Z | X, \theta^{new})} \\ &= D_{KL}(p(Z | X, \theta^{old}) \| p(Z | X, \theta^{new})) \geq 0. \end{aligned}$$

- So,

$$p(X | \theta^{new}) - p(X | \theta^{old}) \geq 0.$$

Unsupervised learning

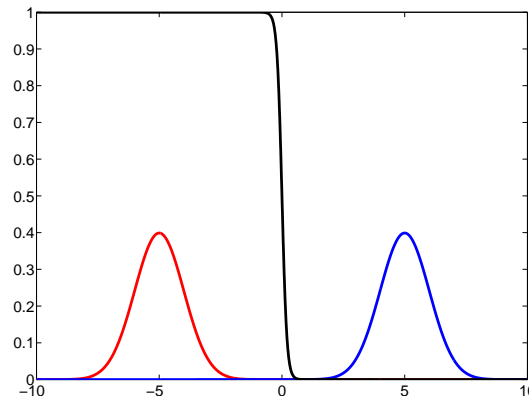
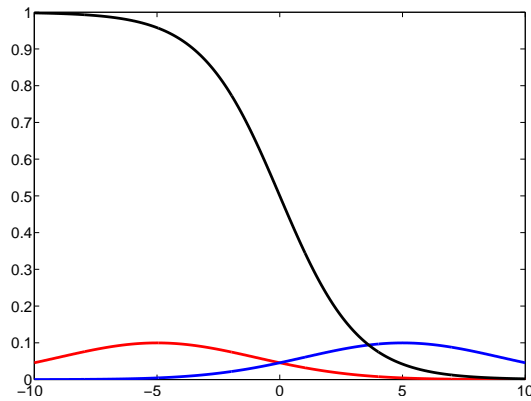
- Three main types of problems in unsupervised learning:
 - Density estimation: learning a density function from a few samples.
 - * Closed-form ML or MAP estimation for Gaussian, Bernoulli models; EM for mixture models.
 - Clustering: grouping *similar* training cases together.
 - Dimensionality reduction: learning to represent each training case using a small number of continuous variables from which the original data can be almost exactly reconstructed.
- For a clustering problem to be well-defined, we need to decide what “similar” means.

Clustering

- We have discussed mixture models as models for density estimation.
 - The goal has been to estimate $p(\mathbf{x}|\theta)$; the hidden variables \mathbf{z} were for convenience.
- What if we only care about \mathbf{z} , i.e. which component generated which \mathbf{x} ?
 - Clustering problem: assign each example to a group.

Gaussian mixture in the limit

- Suppose we fix $\Sigma_c = \sigma^2 \mathbf{I}$, for all $c = 1, \dots, k$ and set $\sigma^2 \rightarrow 0$.
- Suppose $c^* = \operatorname{argmin}_c \|\mathbf{x}_i - \mu_c\|$ (the closest mean).



$$\lim_{\sigma^2 \rightarrow 0} \gamma_{ic} = \begin{cases} 0 & \text{if } c \neq c^*; \\ 1 & \text{if } c = c^*. \end{cases}$$

- The responsibility become “winner take all”.

k -means clustering

1. Initialize k means μ_1, \dots, μ_k to random locations.
 - E.g., set to k randomly chosen *distinct* examples.
2. Repeat until no change in assignment:

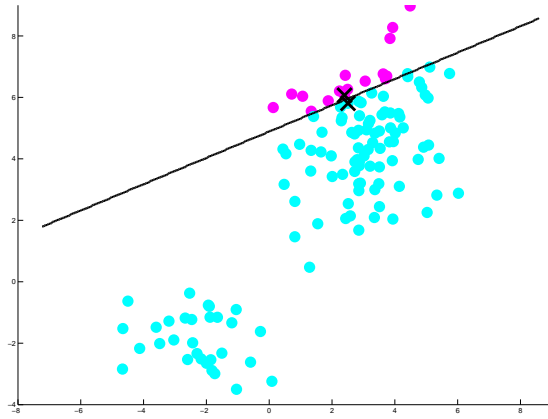
E-step: Assign each example to the closest mean:

$$y_i = \operatorname{argmin}_c \|\mathbf{x}_i - \mu_c\|.$$

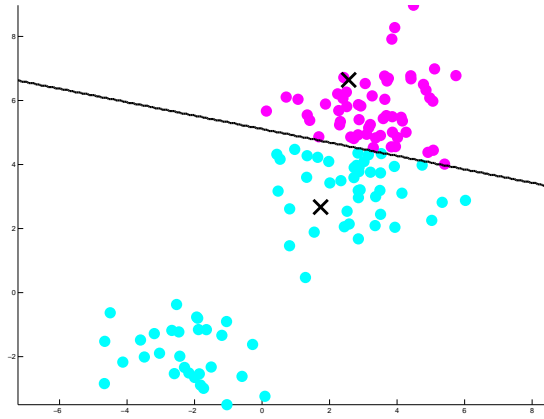
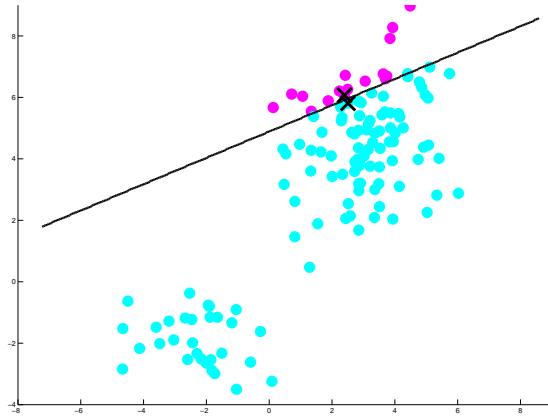
M-step: Reestimate each mean based only on examples assigned to it:

$$\text{Let } N_c = |\{\mathbf{x}_i : y_i = c\}|; \quad \mu_c = \frac{1}{N_c} \sum_{y_i=c} \mathbf{x}_i.$$

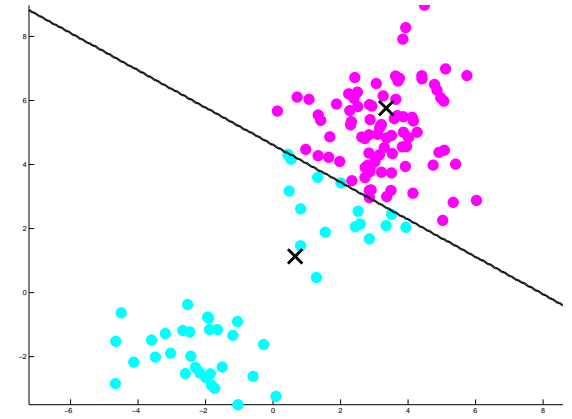
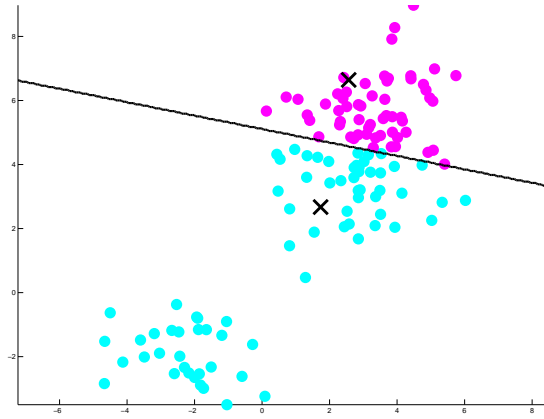
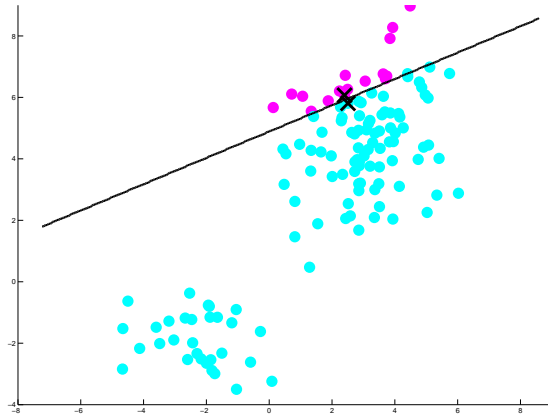
k -means: example



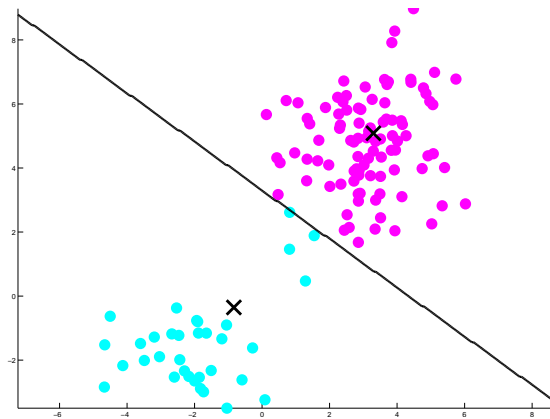
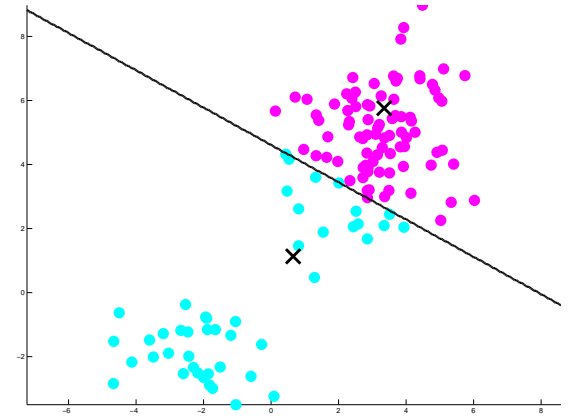
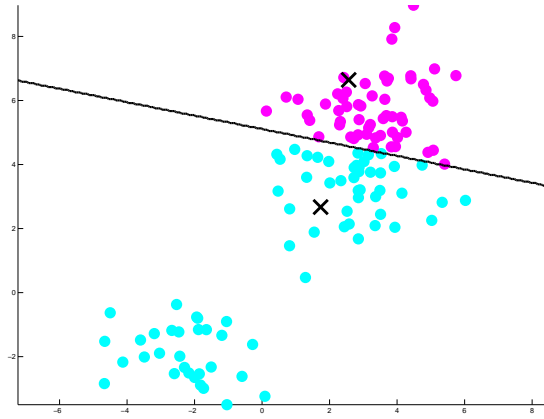
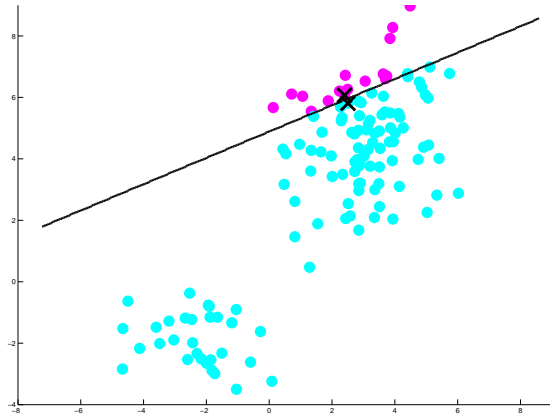
k -means: example



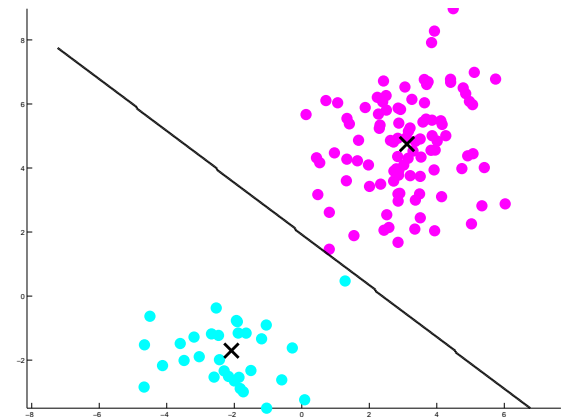
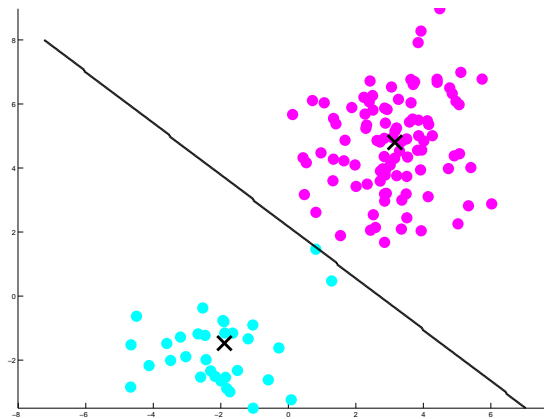
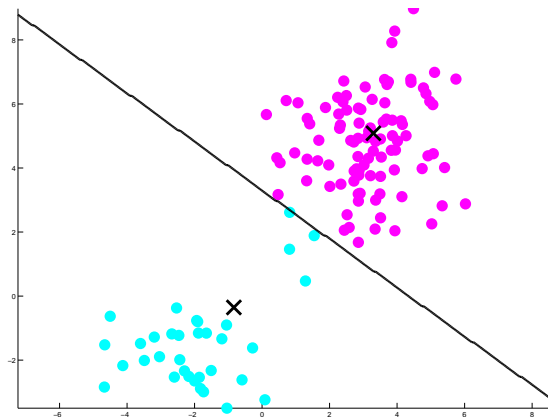
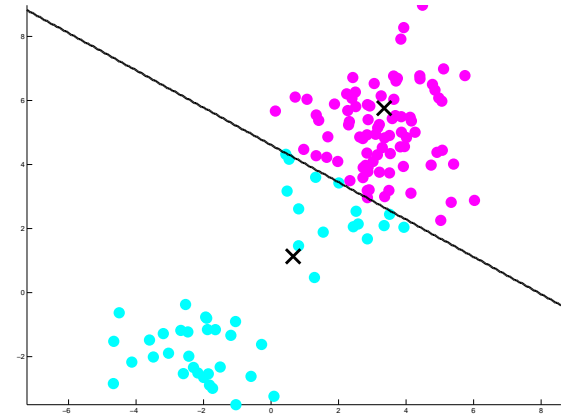
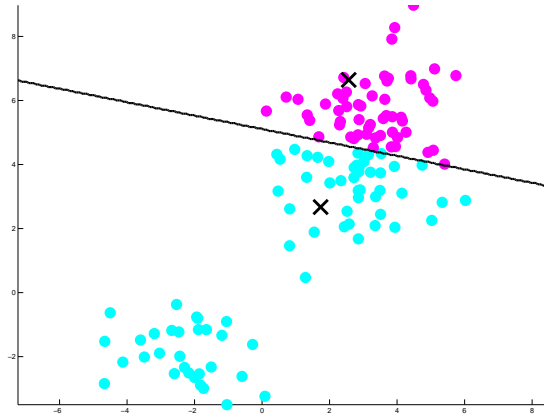
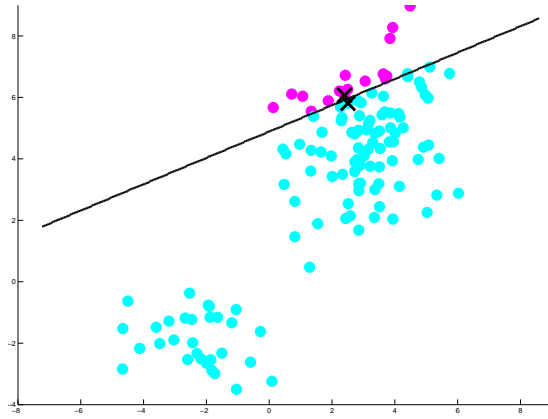
k -means: example



k -means: example



k -means: example



k -means as optimization

- What objective is optimized in k -means?

k -means as optimization

- What objective is optimized in k -means?
- The “ideal” objective: minimum squared

$$J^*(\mu) = \frac{1}{N} \sum_{i=1}^N \min_{c=1,\dots,k} (\mathbf{x}_i - \mu_c)^T (\mathbf{x}_i - \mu_c).$$

k -means as optimization

- What objective is optimized in k -means?
- The “ideal” objective: minimum squared

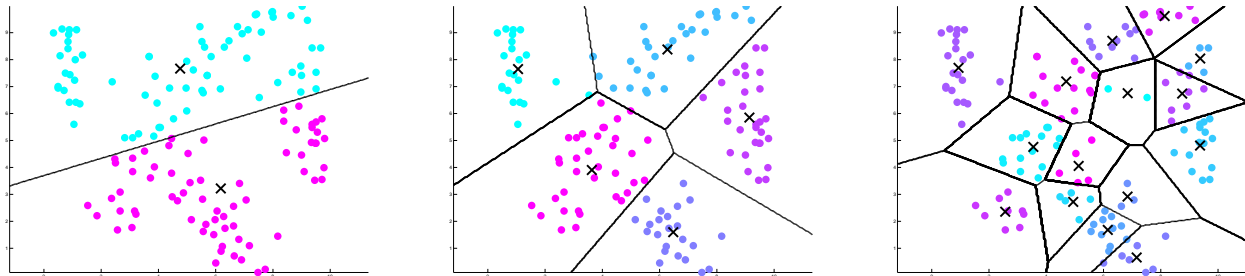
$$J^*(\mu) = \frac{1}{N} \sum_{i=1}^N \min_{c=1,\dots,k} (\mathbf{x}_i - \mu_c)^T (\mathbf{x}_i - \mu_c).$$

- That’s intractable; instead, k -means optimizes the upper bound

$$J(\mu, \{y_i\}) = \frac{1}{N} \sum_{i=1}^N (\mathbf{x}_i - \mu_{y_i})^T (\mathbf{x}_i - \mu_{y_i}).$$

Vector Quantization

- We can use the cluster mean as a *prototype* representing all the examples assigned to the cluster.
- *Vector quantization*: construct a *codebook* using k -means.



- Whenever need to transmit \mathbf{x} , transmit instead the closest codebook.
 - The bits to transmit: $\log kd$ once + $\log k$ for every message.

Setting k

- How can we set k ?

Setting k

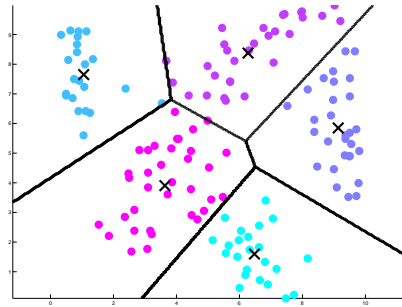
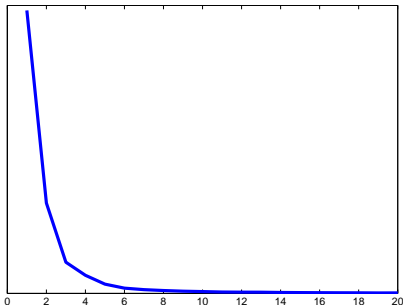
- How can we set k ? Cross-validation doesn't work (why?)

Setting k

- How can we set k ? Cross-validation doesn't work (why?)
- The relevant statistic: *within-class dissimilarity*

$$W_k = \sum_{c=1}^k \sum_{y_i=y_j=c} \|\mathbf{x}_i - \mathbf{x}_j\|^2.$$

- A popular (heuristic) strategy: look for an “elbow” in W_k

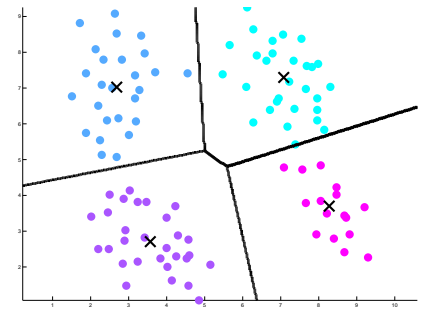
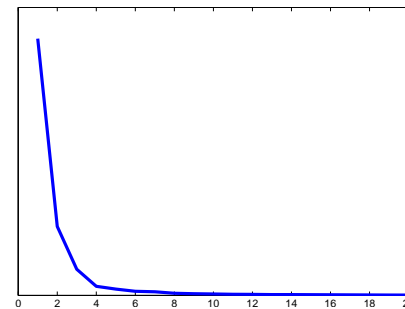
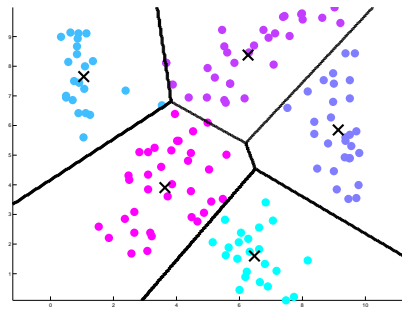
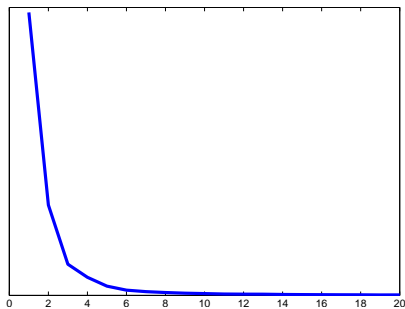


Setting k

- How can we set k ? Cross-validation doesn't work (why?)
- The relevant statistic: *within-class dissimilarity*

$$W_k = \sum_{c=1}^k \sum_{y_i=y_j=c} \|\mathbf{x}_i - \mathbf{x}_j\|^2.$$

- A popular (heuristic) strategy: look for an “elbow” in W_k



Mixture of Gaussians EM versus k -means

- k -means:
 - No probabilistic model \Rightarrow no estimated density.
 - faster to compute (only a single explanation for each data point).
 - Limited by the underlying assumption of spherical clusters
- We can bring back the covariance—get “hard EM”.
 - Still limited by the shape of the covariance (ellipsoid).
- Both EM and k -means depend on initialization (can get stuck in local optima).

Next time

Other clustering methods.