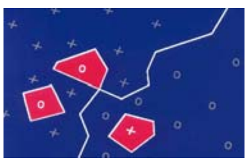
Advanced Machine Learning

Nonparametric methods

Eric Xing

Lecture 2, August 10, 2009



Reading:

© Eric Xing @ CMU, 2006-2009

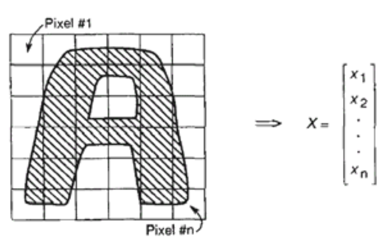
Univariate prediction without using a model: good or bad?

- Nonparametric Classifier (Instance-based learning)
  - Nonparametric density estimation
  - K-nearest-neighbor classifier
  - Optimality of kNN
- Spectrum clustering
  - Clustering
  - Graph partition and normalized cut
  - The spectral clustering algorithm
- Very little “learning” is involved in these methods

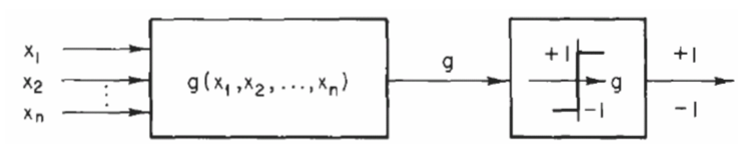
© Eric Xing @ CMU, 2006-2009

Classification

- Representing data:

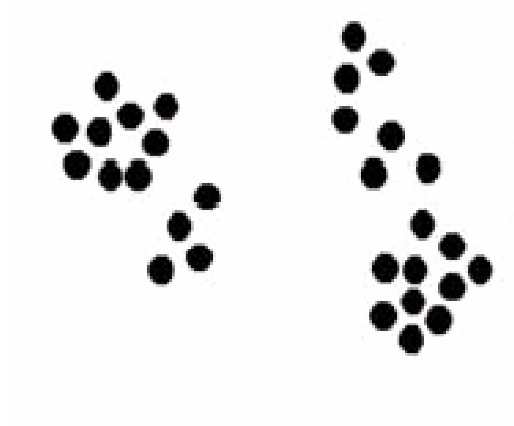


- Hypothesis (classifier)



© Eric Xing @ CMU, 2006-2009

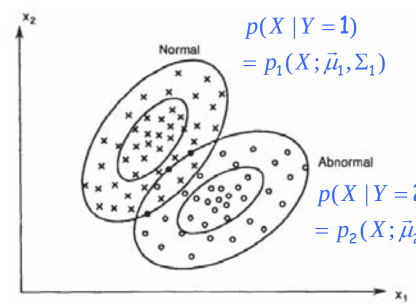
Clustering



© Eric Xing @ CMU, 2006-2009

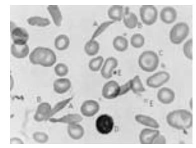
Decision-making as dividing a high-dimensional space

- Classification-specific Dist.: P(X|Y)



$$p(X \mid Y=1) = p_1(X; \vec{\mu}_1, \Sigma_1)$$

$$p(X \mid Y=2) = p_2(X; \vec{\mu}_2, \Sigma_2)$$



- Class prior (i.e., "weight"): P(Y)

© Eric Xing @ CMU, 2006-2009

The Bayes Rule

- What we have just did leads to the following general expression:

$$P(Y \mid X) = \frac{P(X \mid Y)\, p(Y)}{P(X)}$$

This is Bayes Rule

Bayes, Thomas (1763) An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London,* **53:370-418**

© Eric Xing @ CMU, 2006-2009

The Bayes Decision Rule for Minimum Error

- The *a posteriori* probability of a sample

$$P(Y=i\mid X)=\frac{p(X\mid Y=i)P(Y=i)}{p(X)}=\frac{\pi_i p_i(X\mid Y=i)}{\sum_i \pi_i p_i(X\mid Y=i)}\equiv q_i(X)$$

- Bayes Test: $q_1(X) \overset{?}{\gtrless} q_2(X)$ $\quad \begin{cases} > & y=1 \\ < & y=2 \end{cases}$

$$\pi_1 P_1(X) \gtrless \pi_2 P_2(X) \qquad \boxed{\frac{P_1(X)}{P_2(X)}} \overset{?}{\underset{?}{\gtrless}} \frac{\pi_2}{\pi_1}$$

ℓ

- Likelihood Ratio:

$$\ell(X)=\frac{P_1(X)}{P_2(X)}$$

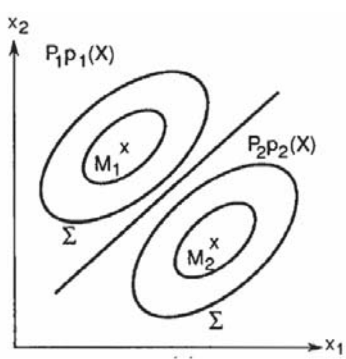
- Discriminant function:

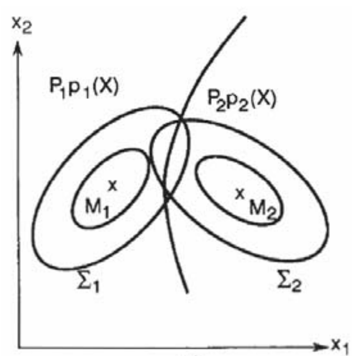
$$h(X)= \ln \ell(x) = \ln P_1(x) - \ln P_2(x) \gtrless \ln \pi_2 - \ln \pi_1 .$$

© Eric Xing @ CMU, 2006-2009

Example of Decision Rules

- When each class is a normal …





- We can write the decision boundary analytically in some cases … homework!!

© Eric Xing @ CMU, 2006-2009

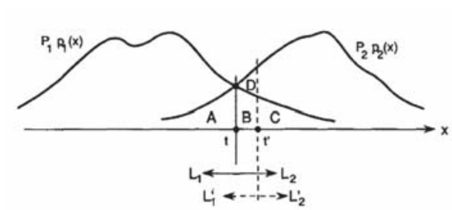
Bayes Error

- We must calculate the *probability of error*
  - the probability that a sample is assigned to the wrong class
- Given a datum X, what is the *risk*?

$$r(X) = \min[q_1(X), q_2(X)]$$

- The Bayes error (the expected risk):

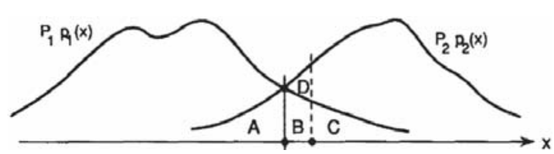
$$\begin{aligned}\epsilon &= E[r(X)] = \int r(x)p(x)dx \\ &= \int \min[\pi_i p_1(x), \pi_2 p_2(x)]dx \\ &= \pi_1 \int_{L_1} p_1(x)dx + \pi_2 \int_{L_2} p_2(x)dx \\ &= \pi_1\epsilon_1 + \pi_2\epsilon_2\end{aligned}$$



© Eric Xing @ CMU, 2006-2009

More on Bayes Error

- Bayes error is the lower bound of probability of classification error



- Bayes classifier is the theoretically best classifier that minimize probability of classification error
- Computing Bayes error is in general a very complex problem. Why?
  - Density estimation:
  - Integrating density function:

$$\epsilon_1 = \int_{\ln(\pi_1/\pi_2)}^{+\infty} p_1(x)dx \qquad \epsilon_2 = \int_{-\infty}^{\ln(\pi_1/\pi_2)} p_2(x)dx$$

© Eric Xing @ CMU, 2006-2009

Learning Classifier

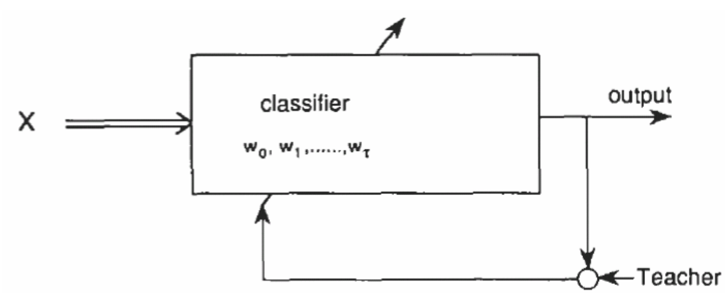
- The decision rule:

$$h(X) = -\ln p_1(X) + \ln p_2(X) \underset{<}{\overset{>}{}} \ln \frac{\pi_1}{\pi_2}$$

- Learning strategies
  - Generative Learning
  - Discriminative Learning
  - Instance-based Learning (Store all past experience in memory)
    - A special case of nonparametric classifier

© Eric Xing @ CMU, 2006-2009

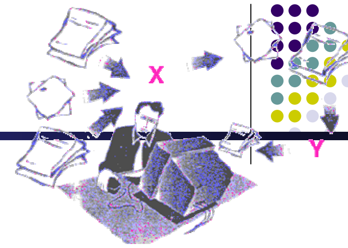
Supervised Learning



- K-Nearest-Neighbor Classifier:
  where the h(X) is represented by all the data, and by an algorithm

© Eric Xing @ CMU, 2006-2009

Recall: Vector Space Representation



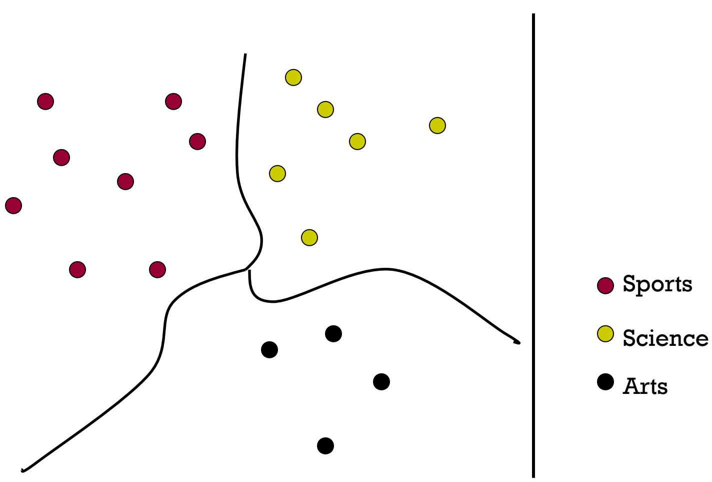
- Each document is a vector, one component for each term (= word).

	Doc 1	Doc 2	Doc 3	...
Word 1	3	0	0	...
Word 2	0	8	1	...
Word 3	12	1	10	...
...	0	1	3	...
...	0	0	0	...

- Normalize to unit length.
- High-dimensional vector space:
  - Terms are axes, 10,000+ dimensions, or even 100,000+
  - Docs are vectors in this space

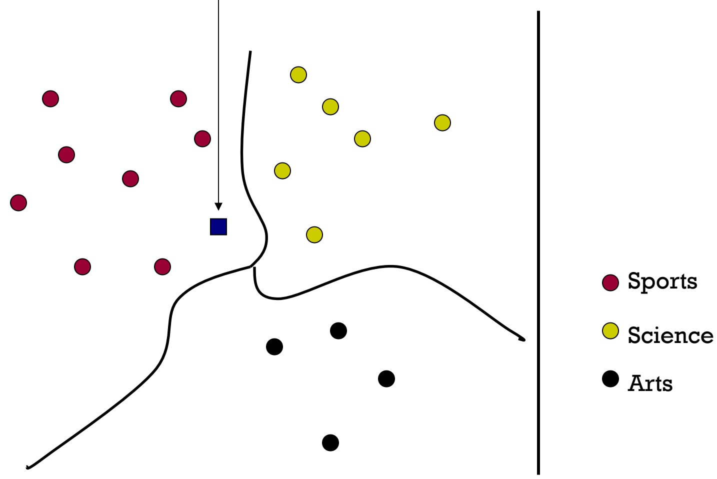
© Eric Xing @ CMU, 2006-2009

Classes in a Vector Space



© Eric Xing @ CMU, 2006-2009

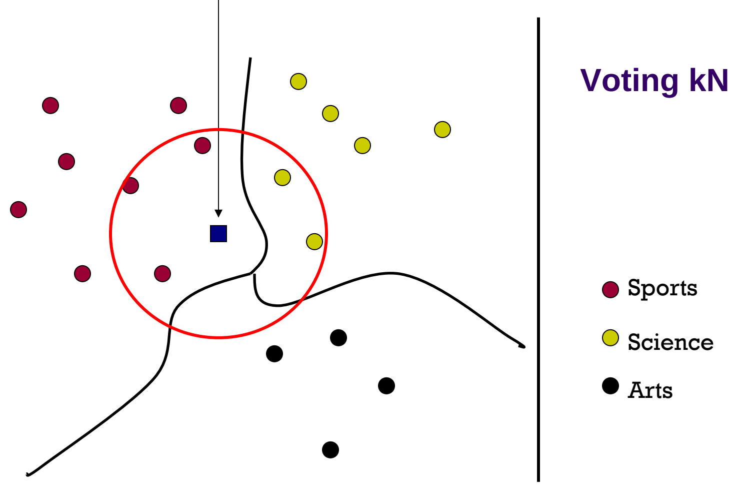
Test Document = ?



© Eric Xing @ CMU, 2006-2009

K-Nearest Neighbor (kNN) classifier

Voting kNN



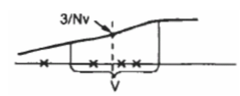
© Eric Xing @ CMU, 2006-2009

kNN Is Close to Optimal

- Cover and Hart 1967
- Asymptotically, the error rate of 1-nearest-neighbor classification is less than twice the Bayes rate [error rate of classifier knowing model that generated data]
- In particular, asymptotic error rate is 0 if Bayes rate is 0.

- Where does kNN come from?
  - Nonparametric density estimation

$$\hat{p}(X) = \frac{1}{N}\frac{k(X)}{V}$$



© Eric Xing @ CMU, 2006-2009

Nearest-Neighbor Learning Algorithm

- Learning is just storing the representations of the training examples in *D*.
- Testing instance *x*:
  - Compute similarity between *x* and all examples in *D*.
  - Assign *x* the category of the most similar example in *D*.
- Does not explicitly compute a generalization or category prototypes.
- Also called:
  - Case-based learning
  - Memory-based learning
  - Lazy learning

© Eric Xing @ CMU, 2006-2009

kNN is an instance of Instance-Based Learning

- What makes an Instance-Based Learner?
  - A distance metric
  - How many nearby neighbors to look at?
  - A weighting function (optional)
  - How to relate to the local points?

© Eric Xing @ CMU, 2006-2009

Euclidean Distance Metric

$$D(x, x') = \sqrt{\sum_i \sigma_i^2 (x_i - x_i')^2}$$

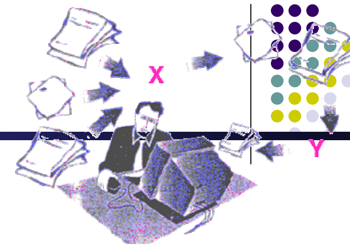
- Or equivalently,

$$D(x, x') = \sqrt{(x - x')^T \Sigma (x - x')}$$

- Other metrics:
  - L_1 norm: |x-x'|
  - L_∞ norm: max |x-x'| (elementwise …)
  - Mahalanobis: where Σ is full, and symmetric
  - Correlation
  - Angle
  - Hamming distance, Manhattan distance
  - …

© Eric Xing @ CMU, 2006-2009

Case Study: kNN for Web Classification

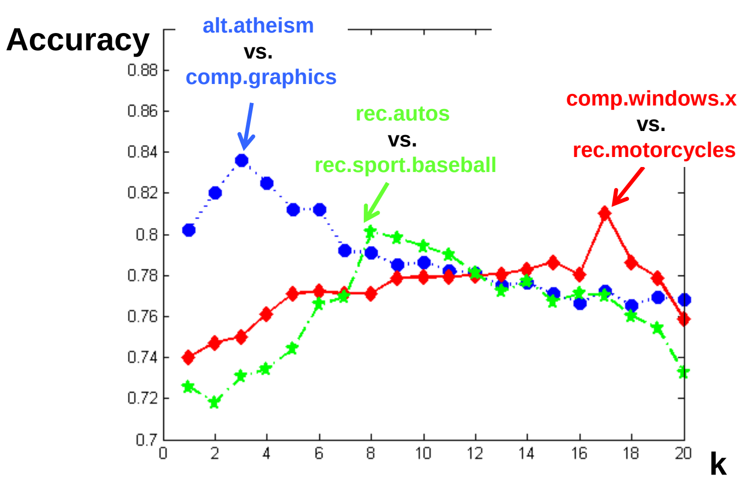


- Dataset
  - 20 News Groups (20 classes)
  - Download :(http://people.csail.mit.edu/jrennie/20Newsgroups/)
  - 61,118 words, 18,774 documents
  - Class labels descriptions

comp.graphics comp.os.ms-windows.misc comp.sys.ibm.pc.hardware comp.sys.mac.hardware comp.windows.x	rec.autos rec.motorcycles rec.sport.baseball rec.sport.hockey	sci.crypt sci.electronics sci.med sci.space
misc.forsale	talk.politics.misc talk.politics.guns talk.politics.mideast	talk.religion.misc alt.atheism soc.religion.christian

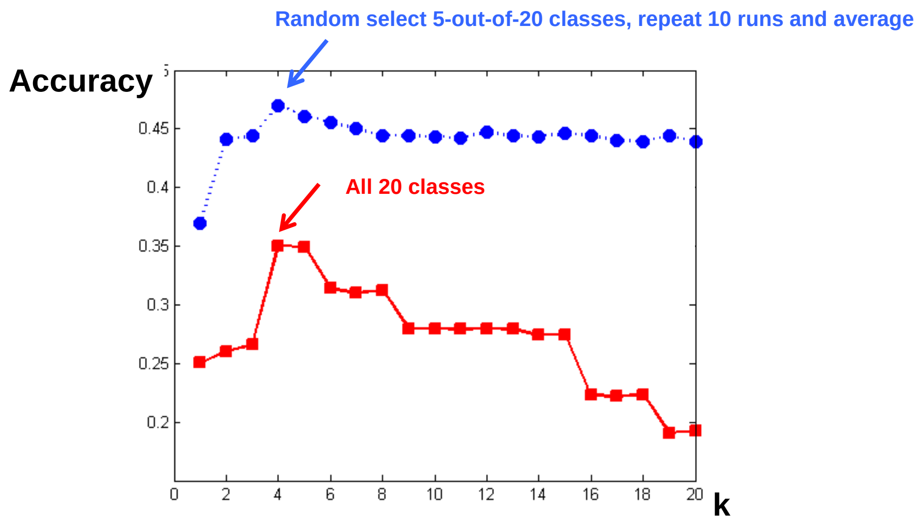
© Eric Xing @ CMU, 2006-2008

Results: Binary Classes



© Eric Xing @ CMU, 2006-2008

Results: Multiple Classes

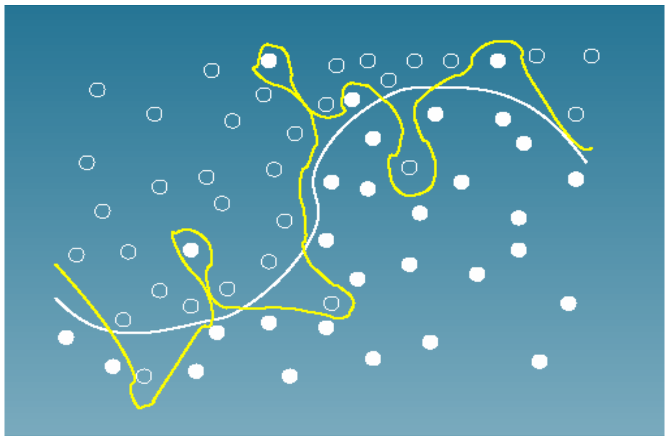


© Eric Xing @ CMU, 2006-2009

Is kNN ideal?

© Eric Xing @ CMU, 2006-2009

Is kNN ideal? … more later



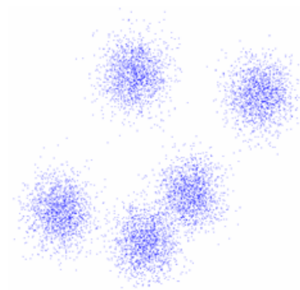
© Eric Xing @ CMU, 2006-2009

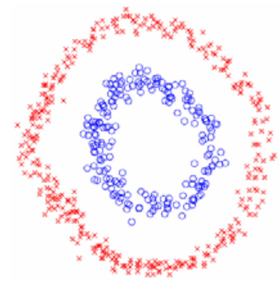
Effect of Parameters

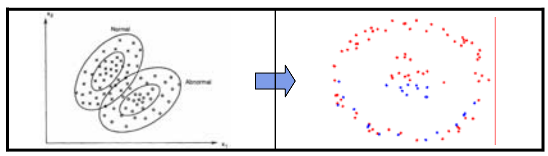
- Sample size
  - The more the better
  - Need efficient search algorithm for NN
- Dimensionality
  - Curse of dimensionality
- Density
  - How smooth?
- Metric
  - The relative scalings in the distance metric affect region shapes.
- Weight
  - Spurious or less relevant points need to be downweighted
- K

© Eric Xing @ CMU, 2006-2009

Clustering



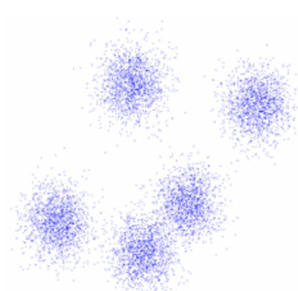




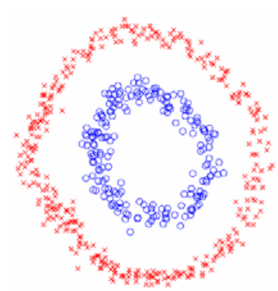
© Eric Xing @ CMU, 2006-2009

Data Clustering

- Two different criteria
  - Compactness, e.g., k-means, mixture models
  - Connectivity, e.g., spectral clustering



Compactness

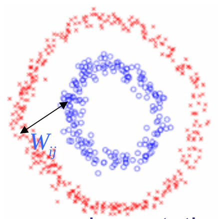


Connectivity

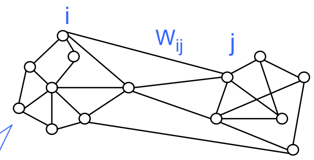
© Eric Xing @ CMU, 2006-2009

Graph-based Clustering

- Data Grouping



$W_{ij} = f(d(x_i, x_j))$



G = {V,E}

- Image sigmentation

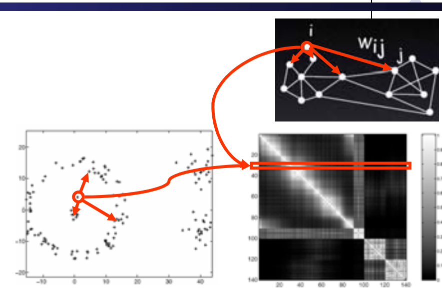
- Affinity matrix: $W = [w_{i,j}]$
- Degree matrix: $D = \mathrm{diag}(d_i)$
- Laplacian matrix: $L = D - W$
- (bipartite) partition vector:

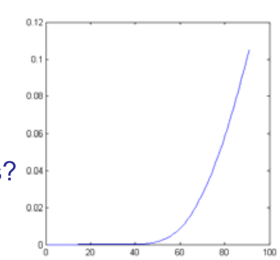
$$x = [x_1, \ldots, x_N]$$
$$= [1, 1, \ldots 1, -1, -1, \ldots -1]$$

© Eric Xing @ CMU, 2006-2009

Affinity Function

$$W_{i,j} = e^{\frac{-\|X_i - X_j\|_2^2}{\sigma^2}}$$





- Affinities grow as σ grows →
- How the choice of σ value affects the results?
- What would be the optimal choice for σ?

© Eric Xing @ CMU, 2006-2009

A Spectral Clustering Algorithm
Ng, Jordan, and Weiss 2003

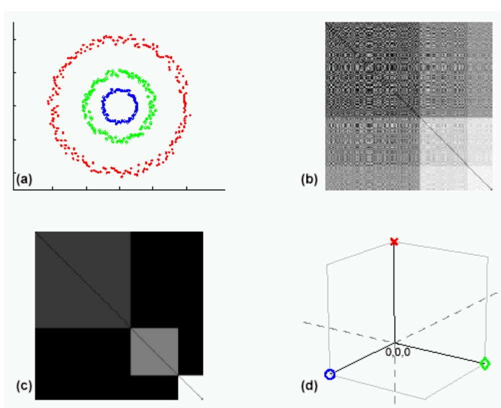
- Given a set of points $S=\{s_1,\ldots s_n\}$
- Form the affinity matrix $w_{i,j} = e^{\frac{-\|s_i - s_j\|_2^2}{\sigma^2}}, \quad \forall i \neq j, \quad w_{i,i} = 0$
- Define diagonal matrix $D_{ii} = \Sigma_\kappa a_{ik}$
- Form the matrix $L = D^{-1/2} W D^{-1/2}$
- Stack the k largest eigenvectors of L to for the columns of the new matrix X:

$$X = \begin{bmatrix} | & | & & | \\ x_1 & x_2 & \cdots & x_k \\ | & | & & | \end{bmatrix}$$

- Renormalize each of X's rows to have unit length and get new matrix Y. Cluster rows of Y as points in R^k

© Eric Xing @ CMU, 2006-2009

Why it works?



- K-means in the spectrum space !

© Eric Xing @ CMU, 2006-2009

More formally …

- Spectral clustering is equivalent to minimizing a generalized normalized cut

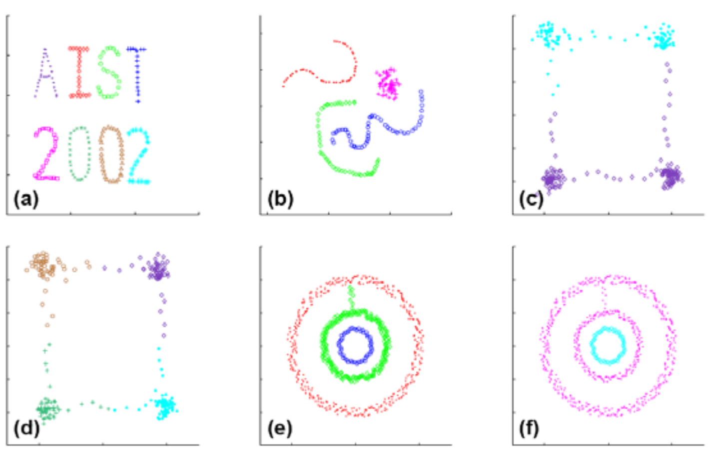
$$\min \ \mathrm{Ncut}(A_1, A_2 \ldots A_k) = \sum_{r=1}^{k} \left(\frac{\mathrm{cut}(A_r, \overline{A}_r)}{d_{A_r}} \right)$$

$$\min \ Y^{\mathrm{T}} D^{-1/2} W D^{-1/2} Y$$
$$\text{s.t.} \ \ Y^{\mathrm{T}} Y = I$$

$$Y = \begin{array}{cc} \text{segments} & \\ \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} & \text{pixels} \end{array}$$

© Eric Xing @ CMU, 2006-2009

Toy examples



Images from Matthew Brand (TR-2002-42)

© Eric Xing @ CMU, 2006-2009

Spectral Clustering

- Algorithms that cluster points using eigenvectors of matrices derived from the data
- Obtain data representation in the low-dimensional space that can be easily clustered
- Variety of methods that use the eigenvectors differently (we have seen an example)
- Empirically very successful
- Authors disagree:
  - Which eigenvectors to use
  - How to derive clusters from these eigenvectors

© Eric Xing @ CMU, 2006-2009

Summary

- ***Two nonparametric methods:***
  - ***kNN classifier***
  - ***Spectrum clustering***
- A nonparametric method does not rely on any assumption concerning the structure of the underlying density function.
- Good news:
  - Simple and powerful methods; Flexible and easy to apply to many problems.
  - kNN classifier asymptotically approaches the ***Bayes classifier,*** which is theoretically the best classifier that minimizes the probability of classification error.
  - Spectrum clustering optimizes the normalized cut
- Bad news:
  - High memory requirements
  - Very dependant on the scale factor for a specific problem.

© Eric Xing @ CMU, 2006-2009