Lecture 7: Clustering

Advanced Business Analytics (CIS442D)
Simon Business School
2/15/2017

Some of the material is based on Chapter 14 in "**The analytics edge**" by Bertsimas, O`Hair, and Pulleyblank, and "**An Introduction to Statistical Learning**, with applications in R" by James, Witten, Hastie and Tibshirani with permission from the authors

Announcements

- Homework 6 will be posted tomorrow (last one)
- Homework 5 due tomorrow
- Tutorial
  - Monday
  - GitHub, clustering
- Python tool project due next week
  - Upload project and related files to GitHub
  - Update Solomon on the web address
- Analytics project due in two weeks

Today

- Boosting
  - Lecture 7 - boosting.ipynb
- Clustering
  - Lecture 7 - clustering.ipynb

Clustering

- Applications
- Practice exercise
- Algorithms
  - K-Means
  - Hierarchical clustering
  - Gaussian mixture models

Outline

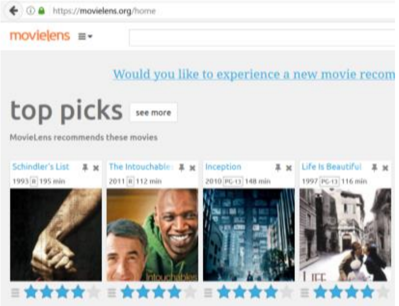
- **Applications**
- Practice exercise
- Algorithms
  - K-Means
  - Hierarchical clustering
  - Gaussian mixture models

Clustering

- Methods for discovering unknown subgroups in data
- Goal: discover patterns in data vs. make prediction

Dimensionality reduction

- MovieLens 1M Dataset
  - 3,900 movies
  - 6,040 users
  - 1,000,209 anonymous ratings
  - 6MB zipped, 24MB unzipped
- MovieLens 20M Dataset
  - 27,000 movies
  - 138,000 users
  - 20 million ratings
  - 190MB zipped, 875MB unzipped
- Cluster similar movies

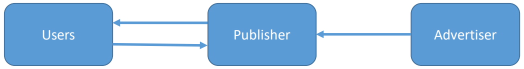


Online advertising

- Major method for generating revenue by websites
- Approximately 90% of Google's revenue
- US internet advertising revenues: 59.6B$ (2015)
- Key operational decision: matching users to ads
  - Affect click-through rate
  - Ads inventory optimization

The New York Times
SAMSUNG
SAMSUNG
SAMSUNG
Bring it home with up to $1500 off a kitchen package.*
SHOP NOW
Turmoil, From the Top Down, at the National

Online advertising (simplified setting)



- How to match users to ads?
- What data is available to the publisher?
- Market segmentation
  - Division of consumers into subgroups
  - Similarities based on web-browsing patterns (topics, frequency, time of day) and IP
- Many applications in marketing

Online advertising

Source: http://www.improvedigital.com/market-map/

Applications

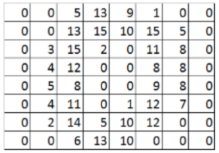
- Dimensionality reduction
- Market segmentation
- Fraud detection
- Improving predictions

Outline

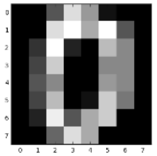
- Applications
- **Practice exercise**
- Algorithms
  - K-Means
  - Hierarchical clustering
  - Gaussian mixture models

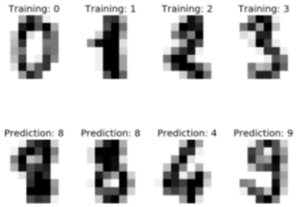
Practice exercise

- Convert images to digits
- Grayscale images
  - Pixel array representation 8x8
  - Each number represents intensity (typically, 0-255)



0	0	5	13	9	1	0	0
0	0	13	15	10	15	5	0
0	3	15	2	0	11	8	0
0	4	12	0	0	8	8	0
0	5	8	0	0	9	8	0
0	4	11	0	1	12	7	0
0	2	14	5	10	12	0	0
0	0	6	13	10	0	0	0

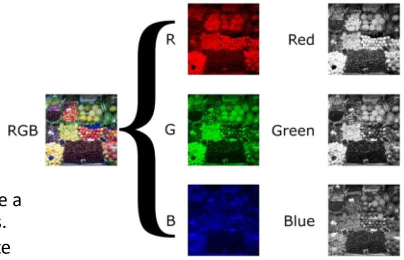




Source: http://scikit-learn.org

(1) Practice exercise

- Color images
  - Pixel array: 8x8x3
- References: [1], [2], [3]
- Exercise (in clustering.ipynb)
  1. Use the dataset digits to create a classifier that recognizes digits.
  2. Use cross validation to evaluate the quality of the classifier.



Source: Wikipedia

(2) Digit recognition

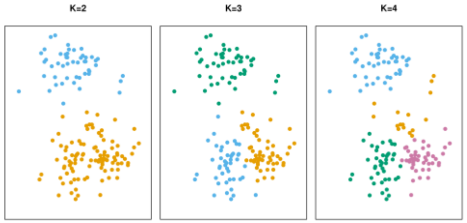
- Classification approach
  1. Prepare training data – images + labels
  2. Trained classifier (e.g., SVM)
- Clustering approach
  1. Group similar images to clusters
  2. Match clusters with labels

Outline

- Applications
- Practice exercise
- Algorithms
  - **K-Means**
  - Hierarchical clustering
  - Gaussian mixture models

K-Means

- Simple approach to clustering
- Specify the number of clusters
- Example: 2D dataset

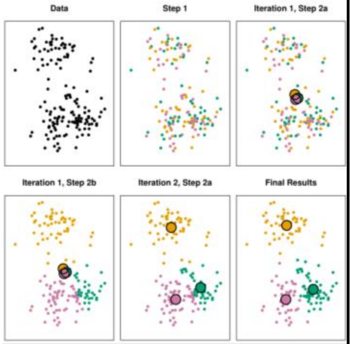


K-Means

- Input: data (*n* observations), *K* (the number of clusters)
- Output:
  - label for each observations (number between 1 to *K*)
  - *K* Centroids

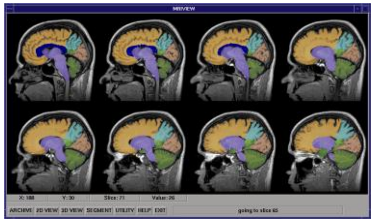
(3) K-Means

1. Randomly assign each observation to a cluster 1…K
2. Repeat
   a) For each cluster compute the cluster centroid (the average of vectors in the cluster)
   b) Assign each observation to the cluster with the closest centroid



(3.2) Image segmentation

- Image segmentation is the process of partitioning a digital image into multiple segments
- The goal of segmentation is to simplify and/or change the representation of an image into something that is more meaningful and easier to analyze [link]
- MRI imaging: surgical planning, post-surgical assessment, abnormality detection [link]



Source: cobre.mrn.org

K-Means – practical considerations

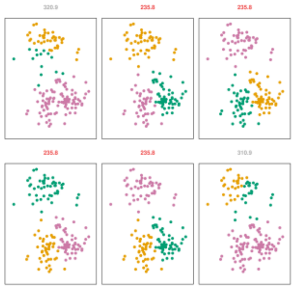
- Random outcome
  - Run multiple times and choose best
- Closest
  - Squared Euclidean distance between $\bar{x}_i$ and $\bar{x}_j$

$$\sum_{j=1}^{p}(x_{i,j} - x_{i',j})^2$$

  - Alternatives: link
- Categorical data
  - Distance: number of different attribute values

$$\sum_{j=1}^{p}1_{(x_{i,j} \neq x_{i',j})}$$

  - Centroids: Use mode instead of mean



K-Means – practical considerations

- Scaling
  - What happens to the distance if feature 1 is binary (M/F) and feature 2 is salary?

$$\sum_{j=1}^{p}(x_{i,j} - x_{i',j})^2 = (x_{i,1} - x_{j,1})^2 + (x_{i,2} - x_{j,2})^2$$

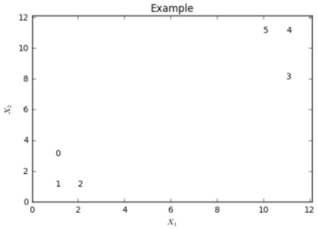
  - Consider scaling features (columns): subtract mean and divide by standard deviation (of each column)
- Which value of K to use?
  - Domain specific
  - Explore different values and study the result

Outline

- Applications
- Practice exercise
- Algorithms
  - K-Means
  - **Hierarchical clustering**
  - Gaussian mixture models

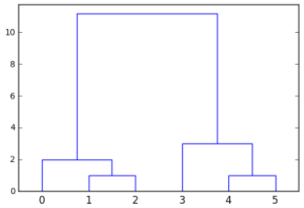
Hierarchical clustering

- Popular alternative to K-Means
- No need to specify K in advance
- Algorithm
  1. Compute distance between each pair of **observations**
  2. Begin with **n** distinct cluster of size 1
  3. Repeat
     1. Compute distance between each pair of **clusters**
     2. Merge clusters with minimal distance
  4. Create a dendrogram
  5. Cut the tree



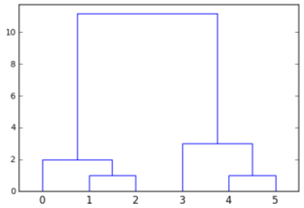
Dendrogram

- Observations are represented on the x-axis
- Distances between clusters are represented on the y-axis
- Examples:
  - The distance between observations 4 and 5 is 1
  - The distance between the cluster {3} and {4,5} is 3 (single linkage)
  - The distance between the clusters {0,1,2} and {3,4,5} is 11.2



Dendrogram – cont.

- Cut the tree at height 6:
  - {0,1,2} and {3,4,5}
- Cut the tree at height 2.5:
  - {0,1,2}, {3}, {4,5}
- Where to cut?
  - Application dependent
  - Greater distances between clusters are better (cut across long vertical lines)
- How many iterations until algorithm stops?



Dissimilarity measures (distances)

- Distance between observations: $d(\bar{x}_i, \bar{x}_j)$
  - Euclidean distance
  - Correlation
  - Etc.
- Distance between clusters (linkage): $d(U, V)$
  - Complete – maximal distance between observations in clusters $\max_{\bar{x}_i \in U, \bar{x}_j \in V} d(\bar{x}_i, \bar{x}_j)$
  - Single – minimal distance between observations in clusters $\min_{\bar{x}_i \in U, \bar{x}_j \in V} d(\bar{x}_i, \bar{x}_j)$
  - Average – average distance between observations in clusters$\frac{1}{|U||V|} \sum_{\bar{x}_i \in U, \bar{x}_j \in V} d(\bar{x}_i, \bar{x}_j)$
  - Centroid – distance between the centroids of each cluster

(4) Hierarchical clustering in Python

- Simple example
- Clustering movies
  - MovieLens 1M Dataset
  - Find groups of similar movies

Outline

- Applications
- Practice exercise
- Algorithms
  - K-Means
  - Hierarchical clustering
  - **Gaussian mixture models ***

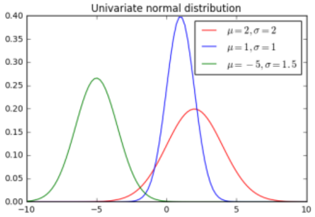
Gaussian mixture model

- Probabilistic approach to clustering
- Gaussian/normal
  - Centroids are the means of normal random variables
  - Other parameters define the shape of the distribution
  - Assignment to cluster according to the highest probability
- Mixture model
  - Combine multivariate normal distribution to form a complex distribution
- EM algorithm – extremely useful in unsupervised learning

Fitting the model

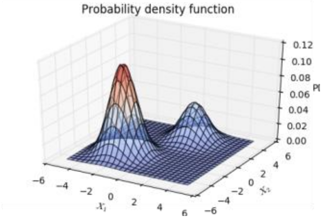
- EM Algorithm (high level)

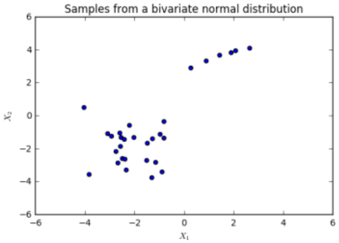
1. Initialize probability distributions $\mu_1, \mu_2, \mu_3, \sigma_1, \sigma_2, \sigma_3$
2. Compute probabilities $p_{i,k}$
3. Update $\mu_1, \mu_2, \mu_3, \sigma_1, \sigma_2, \sigma_3$
4. Return to 2



Multivariate normal distributions

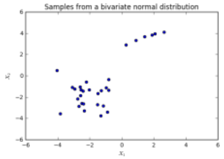
- Normal distributions can be defined for multi-dimensional data
- Defined by the mean and covariance matrix of each distribution

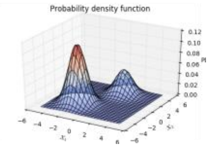




Fitting data to a mixture model

- Goal:
  - find the mean and covariance matrixes defining each cluster
- Input: data
- Output:
  1. Probability that each observation corresponds to a cluster
  2. Parameters defining each normal distribution
- EM algorithm
  - Iteratively evaluates (1) and (2)


Samples from a bivariate normal distribution


Probability density function

Takeaways

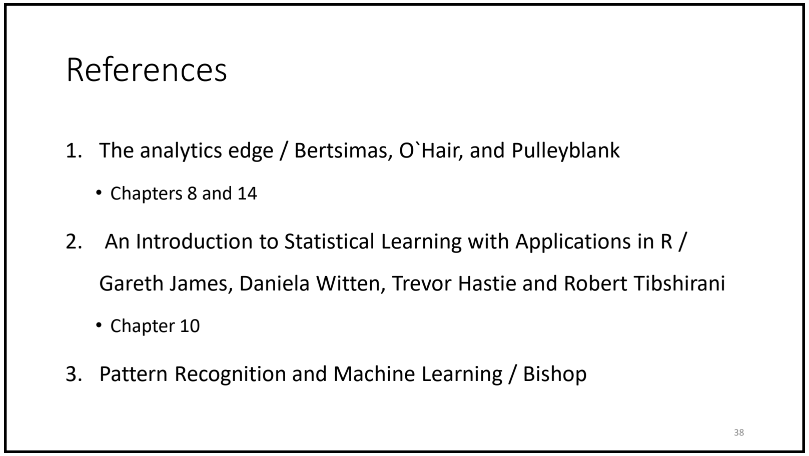
- Probabilistic approach to clustering
- Gaussian/normal
  - Centroids are the means of normal random variables
  - Other parameters define the shape of the distribution
  - Assignment to cluster according to the highest probability
- Mixture model
  - Combine multivariate normal distribution to form a complex distribution
  - Defined using the mean, covariance matrixes, and a weight for each distribution
- EM algorithm – extremely useful in unsupervised learning
- Python
  - Fitting a mixture model (scikit learn)
  - Computing probability distribution
  - Plotting 3-dimensional data

Supervised vs. unsupervised learning

- Supervised learning
  - The prediction objective is clear (label / response variables)
  - Assessing the quality of the results is straightforward (e.g., CV)
- Unsupervised learning
  - No simple goal
  - Exploratory in nature
  - Subjective, no way of checking the true answer (unsupervised)

Summary

- Clustering
  - Unsupervised learning
  - Applications
- Algorithms
  - K-Means
  - Hierarchical clustering
  - Gaussian mixture models
- Python tools for clustering



References

1. The analytics edge / Bertsimas, O`Hair, and Pulleyblank
   - Chapters 8 and 14
2. An Introduction to Statistical Learning with Applications in R / Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani
   - Chapter 10
3. Pattern Recognition and Machine Learning / Bishop