INF4820: Algorithms for AI and NLP

Clustering

Milen Kouylekov & Stephan Oepen

Language Technology Group
University of Oslo

Oct. 2, 2014

Agenda

Yesterday

- Flat clustering
- k-Means

Today

- Bottom-up hierarchical clustering.
- How to measure the inter-cluster similarity (“linkage criterions”).
- Top-down hierarchical clustering.

Types of clustering methods (cont'd)

Hierarchical

- Creates a tree structure of hierarchically nested clusters.
- Topic of the this lecture.

Flat

- Often referred to as partitional clustering when assuming hard and disjoint clusters. (But can also be soft.)
- Tries to directly decompose the data into a set of clusters.

Flat clustering

- Given a set of objects $O = \{o_1, \ldots, o_n\}$, construct a set of clusters $C = \{c_1, \ldots, c_k\}$, where each object o_i is assigned to a cluster c_i.
- Parameters:
  - The cardinality k (the number of clusters).
  - The similarity function s.
- More formally, we want to define an assignment $\gamma : O \rightarrow C$ that optimizes some objective function $F_s(\gamma)$.
- In general terms, we want to optimize for:
  - High intra-cluster similarity
  - Low inter-cluster similarity

k-Means

Algorithm

Initialize: Compute centroids for k seeds.

Iterate:

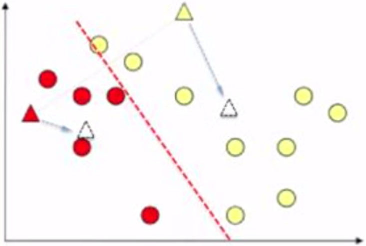
- Assign each object to the cluster with the nearest centroid.
- Compute new centroids for the clusters.

Terminate: When stopping criterion is satisfied.

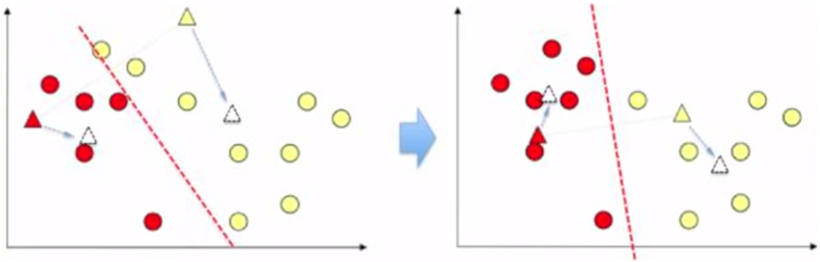
Properties

- In short, we iteratively reassign memberships and recompute centroids until the configuration stabilizes.
- WCSS is monotonically decreasing (or unchanged) for each iteration.
- Guaranteed to converge but not to find the global minimum.
- The time complexity is linear, $\mathrm{O}(kn)$.

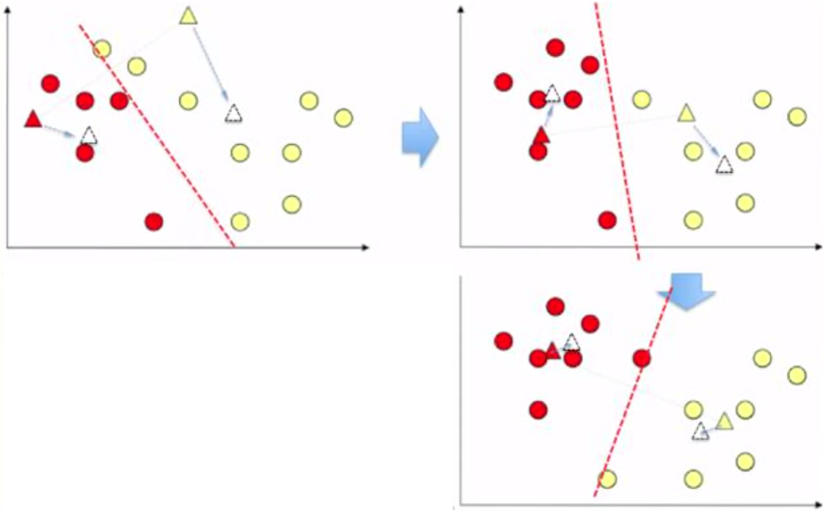
kMeans Example



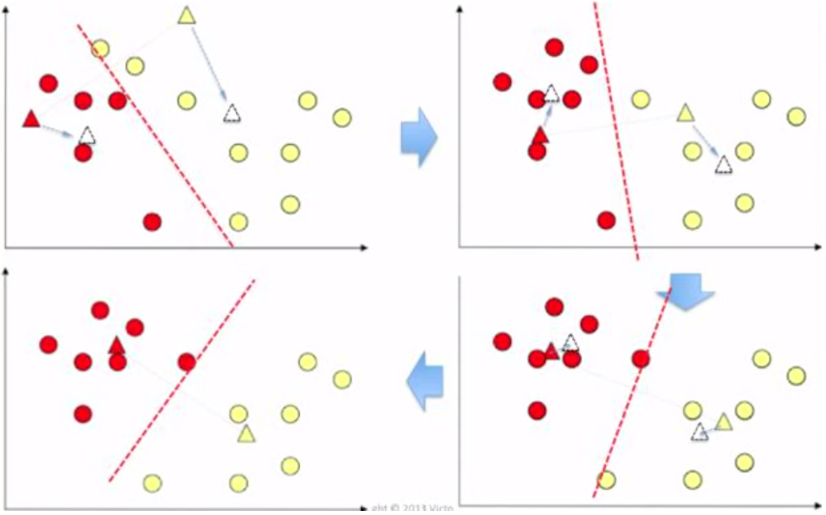
kMeans Example



kMeans Example



kMeans Example

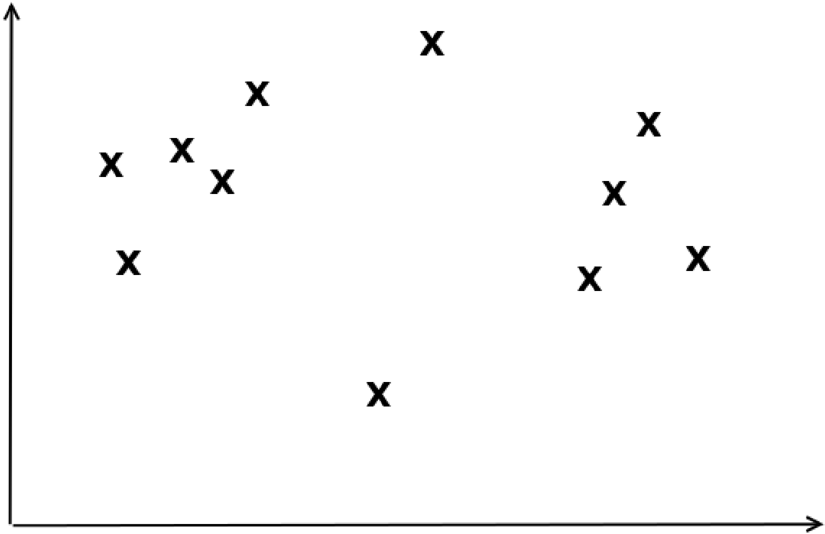


Comments on k-Means

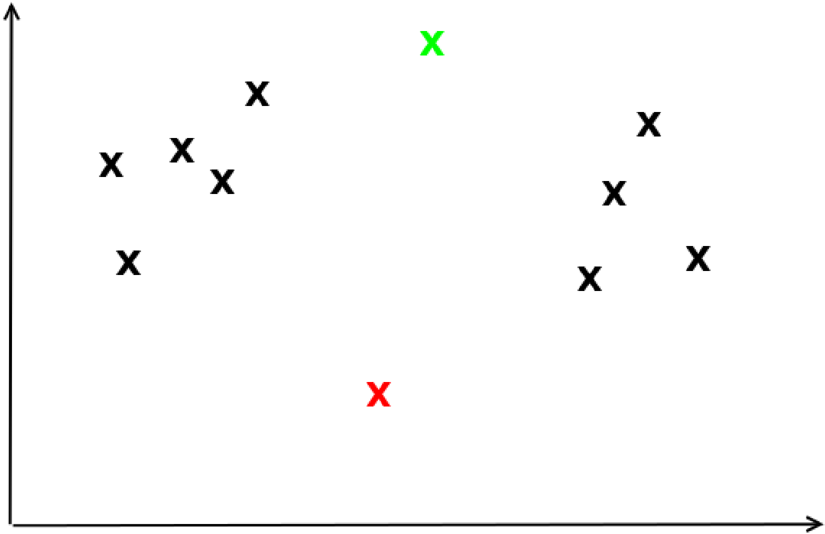
"Seeding"

- We initialize the algorithm by choosing random *seeds* that we use to compute the first set of centroids.
- Many possible heuristics for selecting the seeds:
  - pick k random objects from the collection;
  - pick k random points in the space;
  - pick k sets of m random points and compute centroids for each set;
  - compute an hierarchical clustering on a subset of the data to find k initial clusters; etc..
- The initial seeds can have a large impact on the resulting clustering (because we typically end up only finding a local minimum of the objective function).
- Outliers are troublemakers.

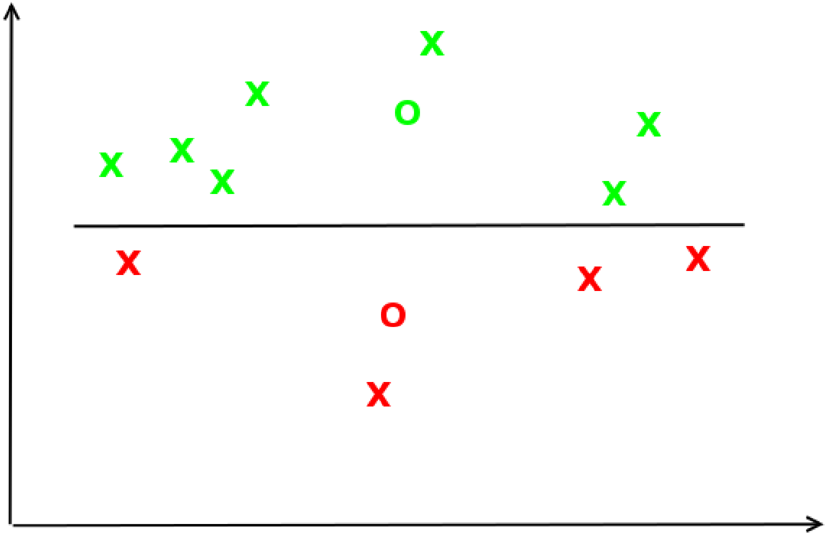
Initial Seed Choice



Initial Seed Choice



Initial Seed Choice



Hierarchical clustering

- Creates a tree structure of hierarchically nested clusters.
- Divisive (top-down): Let all objects be members of the same cluster; then successively split the group into smaller and maximally dissimilar clusters until all objects is its own singleton cluster.
- Agglomerative (bottom-up): Let each object define its own cluster; then successively merge most similar clusters until only one remains.

Agglomerative clustering

- Initially; regards each object as its own singleton cluster.
- Iteratively "agglomerates" (merges) the groups in a bottom-up fashion.
- Each merge defines a binary branch in the tree.
- Terminates; when only one cluster remains (the root).

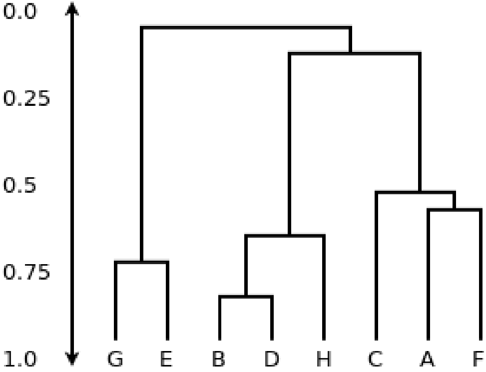
parameters: $\{o_1, o_2, \ldots, o_n\}$, sim

$C = \{\{o_1\}, \{o_2\}, \ldots, \{o_n\}\}$
$T = []$
do for $i = 1$ **to** $n - 1$
  $\{c_j, c_k\} \leftarrow \underset{\{c_j, c_k\} \subseteq C \,\wedge\, j \neq k}{\arg\max} \operatorname{sim}(c_j, c_k)$
  $C \leftarrow C \backslash \{c_j, c_k\}$
  $C \leftarrow C \cup \{c_j \cup c_k\}$
  $T[i] \leftarrow \{c_j, c_k\}$

- At each stage, we merge the pair of clusters that are most similar, as defined by some measure of inter-cluster similarity; sim.
- Plugging in a different sim gives us a different sequence of merges T.

Dendrograms

- A hierarchical clustering is often visualized as a binary tree structure known as a *dendrogram*.
- A merge is shown as a horizontal line.
- The y-axis corresponds to the *similarity* of the merged clusters.



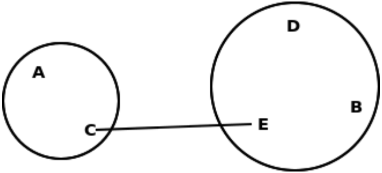
- We here assume dot-products of normalized vectors (self-similarity $= 1$).

Definitions of inter-cluster similarity

- How do we define the similarity between clusters?.
- In agglomerative clustering, a measure of cluster similarity $\mathrm{sim}(c_i, c_j)$ is usually referred to as a *linkage criterion*:
  - Single-linkage
  - Complete-linkage
  - Centroid-linkage
  - Average-linkage
- Determines which pair of clusters to merge in each step.

Single-linkage

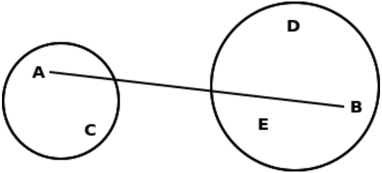
- Merge the two clusters with the minimum distance between any two members.
- Nearest-Neighbors.



- Can be computed efficiently by taking advantage of the fact that it's *best-merge persistent*:
  - Let the nearest neighbor of cluster c_k be in either c_i or c_j. If we merge $c_i \cup c_j = c_l$, the nearest neighbor of c_k will be in c_l.
  - The distance of the two closest members is a local property that is not affected by merging.
- Undesirable chaining effect: Tendency to produce 'stretched' and 'straggly' clusters.

Complete-linkage

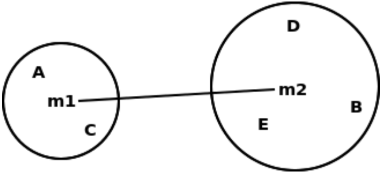
- Merge the two clusters where the maximum distance between any two members is smallest.
- Farthest-Neighbors.



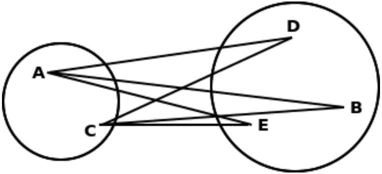
- Amounts to merging the two clusters whose merger has the smallest diameter.
- Preference for compact clusters with small diameters.
- Sensitive to outliers.
- Not best-merge persistent: Distance defined as the diameter of a merge is a non-local property that can change during merging.

Centroid-linkage

- Similarity of clusters c_i and c_j defined as the similarity of their cluster centroids $\vec{\mu}_i$ and $\vec{\mu}_j$.



- Equivalent to the average pairwise similarity between objects from different clusters:

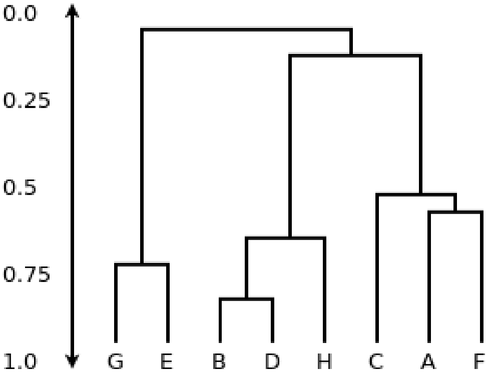


$$sim(c_i, c_j) = \vec{\mu_i} \cdot \vec{\mu_j} = \frac{1}{|c_i||c_j|} \sum_{\vec{x} \in c_i} \sum_{\vec{y} \in c_j} \vec{x} \cdot \vec{y}$$

- Not best-merge persistent.
- Not monotonic, subject to *inversions*: The combination similarity can increase during the clustering.

Monotinicity

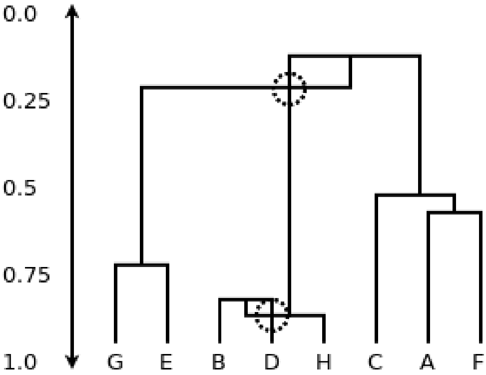
- A fundamental assumption in clustering: small clusters are more coherent than large.
- We usually assume that a clustering is monotonic;
- Similarity is *decreasing* from iteration to iteration.



- This assumpion holds true for all our clustering criterions except for centroid-linkage.

Inversions — a problem with centroid-linkage

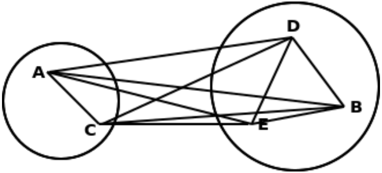
- Centroid-linkage is non-monotonic.
- We risk seeing so-called inversions:
- similarity can increase during the sequence of clustering steps.
- Would show as crossing lines in the dendrogram.



- The horizontal merge bar is lower than the bar of a previous merge.

Average-linkage (1:2)

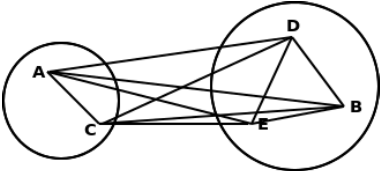
- AKA group-average agglomerative clustering.
- Merge the clusters with the highest average pairwise similarities in their union.



- Aims to maximize coherency by considering all pairwise similarities between objects within the cluster to merge (excluding self-similarities).
- Compromise of complete- and single-linkage.
- Monotonic but not best-merge persistent.
- Commonly considered the best default clustering criterion.

Average-linkage (2:2)

- Can be computed very efficiently if we assume (i) the *dot-product* as the similarity measure for (ii) *normalized* feature vectors.



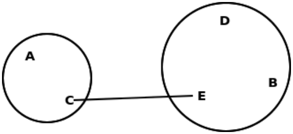
- Let $c_i \cup c_j = c_k$, and $sim(c_i, c_j) = W(c_i \cup c_j) = W(c_k)$, then $W(c_k) =$

$$\frac{1}{|c_k|(|c_k|-1)} \sum_{\vec{x} \in c_k} \sum_{\vec{y} \neq \vec{x} \in c_k} \vec{x} \cdot \vec{y} = \frac{1}{|c_k|\,(|c_k|-1)} \left(\left(\sum_{\vec{x} \in c_k} \vec{x} \right)^2 - |c_k| \right)$$

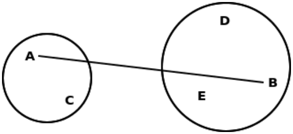
- The sum of vector similarities is equal to the similarity of their sums.

Linkage criterions

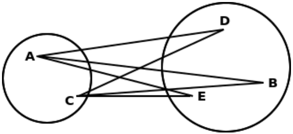
Single-link



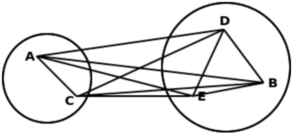
Complete-link



Centroid-link

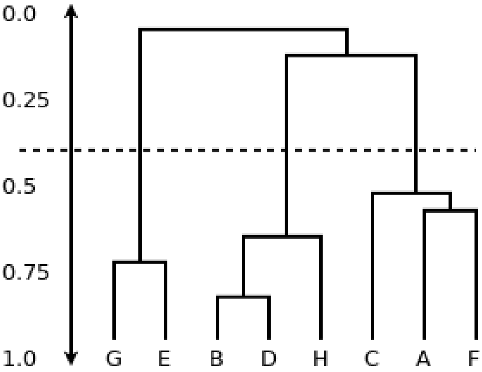


Average-link



Cutting the tree

- The tree actually represents *several partitions*;
- one for each level.
- If we want to turn the nested partitions into a single flat partitioning. . .
- we must cut the tree.



- A cutting criterion can be defined as a threshold on e.g. combination similarity, relative drop in the similarity, number of root nodes, etc.

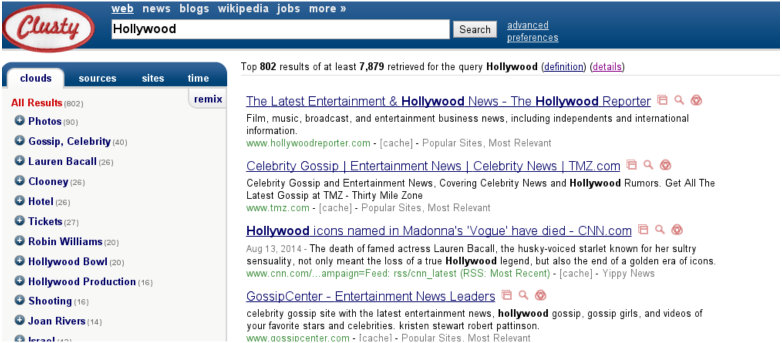
Divisive hierarchical clustering

Generates the nested partitions *top-down*:

- Start: all objects considered part of the same cluster (the root).
- Split the cluster using *a flat clustering algorithm* (e.g. by applying k-means for $k = 2$).
- Recursively split the clusters until only singleton clusters remain (or some specified number of levels is reached).

- Flat methods are generally very effective (e.g. k-means is *linear* in the number of objects).
- Divisive methods are thereby also generally more efficient than agglomerative, which are *at least quadratic* (single-link).
- Also able to initially consider the global distribution of the data, while the agglomerative methods must commit to early decisions based on local patterns.

- Group search results together by topic



Information Retrieval (2)

- Expand Search Query
- Who invented the light bulb?
- Word Similarity Clusters: invent, discover, patent, inventor innovator

News Aggregation

- Grouping news from different sources
- Useful for journalists, political analysts, private companies
- And not only news: Social Media: Twitter, Blogs

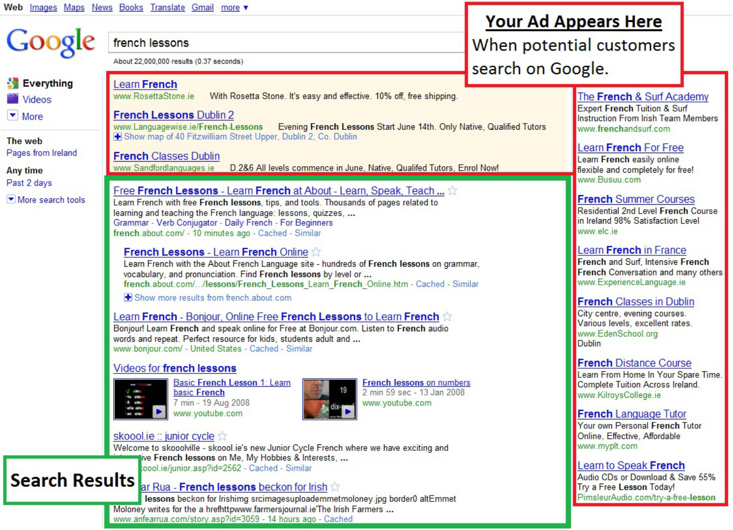
User Profiling

- Analyze user interests
- Propose interesting information/advertisement
- Spy on users
- NSA
- Weird conspiracy theory

User Profiling

- Facebook

- Google


Web Images Maps News Books Translate Gmail more
Google
french lessons
About 22,000,000 results (0.37 seconds)
Your Ad Appears Here
When potential customers search on Google.
Everything
Videos
More
The web
Pages from Ireland
Any time
Past 2 days
More search tools
Learn French
www.RosettaStone.ie With Rosetta Stone. It's easy and effective. 10% off, free shipping.
French Lessons Dublin 2
www.Languagewise.ie/French-Lessons Evening French Lessons Start June 14th. Only Native, Qualified Tutors
Show map of 40 Fitzwilliam Street Upper, Dublin 2, Co. Dublin
French Classes Dublin
www.Sandfordlanguages.ie D.2&6 All levels commence in June, Native, Qualifed Tutors, Enrol Now!
Free French Lessons - Learn French at About - Learn, Speak, Teach ...
Learn French with free French lessons, tips, and tools. Thousands of pages related to learning and teaching the French language: lessons, quizzes, ...
Grammar - Verb Conjugator - Daily French - For Beginners
french.about.com/ - 10 minutes ago - Cached - Similar
French Lessons - Learn French Online
Learn French with the About French Language site - hundreds of French lessons on grammar, vocabulary, and pronunciation. Find French lessons by level or ...
french.about.com/.../lessons/French_Lessons_Learn_French_Online.htm - Cached - Similar
Show more results from french.about.com
Learn French - Bonjour, Online Free French Lessons to Learn French
Bonjour! Learn French and speak online for Free at Bonjour.com. Listen to French audio words and repeat. Perfect resource for kids, students adult and ...
www.bonjour.com/ - United States - Cached - Similar
Videos for french lessons
Basic French Lesson 1: Learn basic French
7 min - 19 Aug 2008
www.youtube.com
French lessons on numbers
2 min 59 sec - 13 Jan 2008
www.youtube.com
skoool.ie :: junior cycle
Welcome to skooolville - skoool.ie's new Junior Cycle French where we have exciting and
ive French lessons on Me, My Hobbies & Interests, ...
oool.ie/junior.asp?id=2562 - Cached - Similar
ar Rua - French lessons beckon for Irish
lessons beckon for Irishimg srcimagesuploademmetmoloney.jpg border0 altEmmet
Moloney writes for the a hrefhttpwww.farmersjournal.ie'The Irish Farmers ...
www.anfearrua.com/story.asp?id=3059 - 14 hours ago - Cached
Search Results
The French & Surf Academy
Expert French Tuition & Surf Instruction From Irish Team Members
www.frenchandsurf.com
Learn French For Free
Learn French easily online flexible and completely for free!
www.Busuu.com
French Summer Courses
Residential 2nd Level French Course in Ireland 98% Satisfaction Level
www.elc.ie
Learn French in France
French and Surf, Intensive French French Conversation and many others
www.ExperienceLanguage.ie
French Classes in Dublin
City centre, evening courses. Various levels, excellent rates.
www.EdenSchool.org
Dublin
French Distance Course
Learn From Home In Your Spare Time. Complete Tuition Across Ireland.
www.KilroysCollege.ie
French Language Tutor
Your own Personal French Tutor Online, Effective, Affordable
www.myplt.com
Learn to Speak French
Audio CDs or Download & Save 55%
Try a Free Lesson Today!
PimsleurAudio.com/try-a-free-lesson

What we have learned so far

- Lisp is Great!
- Vector Space Modeling
  - Represent objects as vector of features
  - Calculate similarity between vectors

Two categorization tasks in machine learning

Classification

- Supervised learning, requiring labeled training data.
- Given some training set of examples with class labels, train a classifier to predict the class labels of new objects.

Clustering

- Unsupervised learning from unlabeled data.
- Automatically group similar objects together.
- No pre-defined classes: we only specify the similarity measure.
- General objective:
  - Partition the data into subsets, so that the similarity among members of the same group is high (homogeneity) while the similarity between the groups themselves is low (heterogeneity).

What is next

- Structured classification
  - sequences
  - labelled sequences
  - trees

Quiz (1)

- **Question 1**: What is the cosine similarity of the vectors:
  A: [4,0,0,1,12,0,8,0]
  B: [0,1,2,0,0,1,0,3]

Quiz (2)

- **Question 2**: Which Classifier runs faster on new data:
  A: Rocchio
  B: kNN

Quiz (3)

- **Question 3**: The classifier produced the following classification result :

	Classifier	Tag
Example1	B	A
Example2	B	B
Example3	A	A
Example4	A	B
Example5	A	A
Example6	A	A

- Calculate the precision,recall and F-Measure of class **A**

Quiz (4)

- **Question 4**: What is the main problem of the kMeans algorithm

Quiz (1)

- **Question 1**: What is the cosine similarity of the vectors:
  A: [4,0,0,1,12,0,8,0]
  B: [0,1,2,0,0,1,0,3]
- **Answer**: 0

Quiz (2)

- **Question 2**: Which Classifier runs faster on new data:
  A: Rocchio
  B: kNN
- **Answer**: Depends
- In general case Rocchio

Quiz (3)

- **Question 3**: The classifier produced the following classification result :

	Classifier	Tag
Example1	B	A
Example2	B	B
Example3	A	A
Example4	A	B
Example5	A	A
Example6	A	A

- Calculate the precision, recall and F-Measure of class **A**
- **Answer:** Precision $3/4 = 0.75$ Recall $3/4 = 0.75$

Quiz (4)

- **Question 4**: What is the main problem of the kMeans algorithm
- **Answer:** Sometimes it does not find the optimal solution