Foundations of Data Science
Lecture 7

Rumi Chunara, PhD

CS3943/9223

So Far…

- What is Data Science?
- Intro to R
- Data cleaning, sampling, processing
- Intro to ML – what is it
- Two Basic Algorithms
  - kNN
  - Linear Regression
- Time-series Analyses
  - Regression and lagged data in R
- Supervised Learning
  - Support Vector Machines
- Unsupervised Learning
  - K-means, PCA

Some Important Questions

- Is at least one of the predictors useful in predicting the response?
- Do all the predictors help to explain Y or is only a subset of the predictors useful?
- How well does the model fit the data?
- Given a set of predictor values, what response value should we predict, and how accurate is our prediction?

Evaluation

How can the performance of a system be evaluated?

Standard Methodology from Information Retrieval:

- Precision
- Recall
- F-measure (combination of Precision/Recall)

Recall

Measure of how much relevant information the system has extracted (coverage of system).

Basic idea:

$$\text{Recall} = \frac{\text{\# of correct answers given by system}}{\text{total \# of possible correct answers in text}}$$

Recall

Measure of how much relevant information the system has extracted (coverage of system).

Exact definition:

Recall = 1 if no possible correct answers

else:

$$\frac{\text{\# of correct answers given by system}}{\text{total \# of possible correct answers in text}}$$

Slide modified from Butt/Jurafsky/Martin

Precision

Measure of how much of the information the system returned is correct (accuracy).

Basic idea:

$$\text{Precision} = \frac{\text{\# of correct answers given by system}}{\text{\# of answers given by system}}$$

Slide from Butt/Jurafsky/Martin

Precision

Measure of how much of the information the system returned is correct (accuracy).

Exact definition:

$$\text{Precision} = \begin{cases} 1 & \text{if no answers given by system} \\ \dfrac{\#\text{ of correct answers given by system}}{\#\text{ of answers given by system}} & \text{else:} \end{cases}$$

Slide modified from Butt/Jurafsky/Martin

Evaluation

Every system, algorithm or theory should be **evaluated**, i.e. its output should be compared to the **gold standard** (i.e. the ideal output).
Suppose we try to find scientists…

Algorithm output:
O = {Einstein, Bohr, Planck, Clinton, Obama}
✓ ✓ ✓ ✗ ✗

Gold standard:
G = {Einstein, Bohr, Planck, Heisenberg}
✓ ✓ ✓ ✗

Precision:
What proportion of the output is correct?

$$\frac{|O \wedge G|}{|O|}$$

Recall:
What proportion of the gold standard did we get?

$$\frac{|O \wedge G|}{|G|}$$

Slide modified from Suchanek

Types of Errors

- False Positives
  - The system predicted **TRUE** but the value was **FALSE**
  - aka “False Alarms” or Type I error

- False Negatives
  - The system predicted **FALSE** but the value was **TRUE**
  - aka “Misses” or Type II error

Precision & Recall Exercise

What is the algorithm output, the gold standard ,the precision and the recall in the following cases?

1. Nostradamus predicts a trip to the moon
   for every century from the 15th to the 20th incl.
2. The weather forecast predicts that the next 3 days will be sunny. It does not say anything about the 2 days that follow. In reality, it is sunny during all 5 days.
3. On Elvis Radio ™ , 90% of the songs are by Elvis.
   An algorithm learns to detect Elvis songs.
   Out of 100 songs on Elvis Radio, the algorithm says that 20 are by Elvis (and says nothing about the other 80). Out of these 20 songs, 15 were by Elvis and 5 were not.

Slide from Suchanek

Unranked retrieval evaluation: Precision and Recall

- **Precision**: fraction of retrieved docs that are relevant = P(relevant|retrieved)
- **Recall**: fraction of relevant docs that are retrieved
  = P(retrieved|relevant)

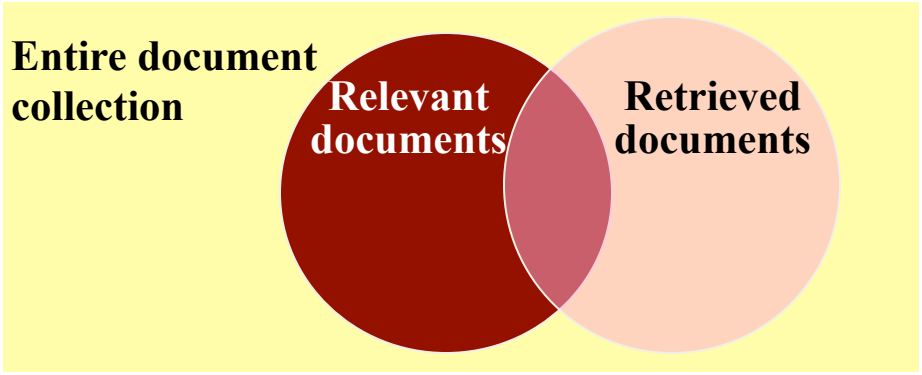
	Relevant	Nonrelevant
Retrieved	tp	fp
Not Retrieved	fn	tn

- Precision P = tp/(tp + fp)
- Recall R = tp/(tp + fn)

Summary: Precision/Recall

- Precision and recall are very key concepts
  - Definitely know these formulas, they are applicable everywhere (even real life)!
- F-Measure is a nice way to combine them to get a single number

Precision and Recall



	retrieved	not retrieved
irrelevant	retrieved & irrelevant	Not retrieved & irrelevant
relevant	retrieved & relevant	not retrieved but relevant

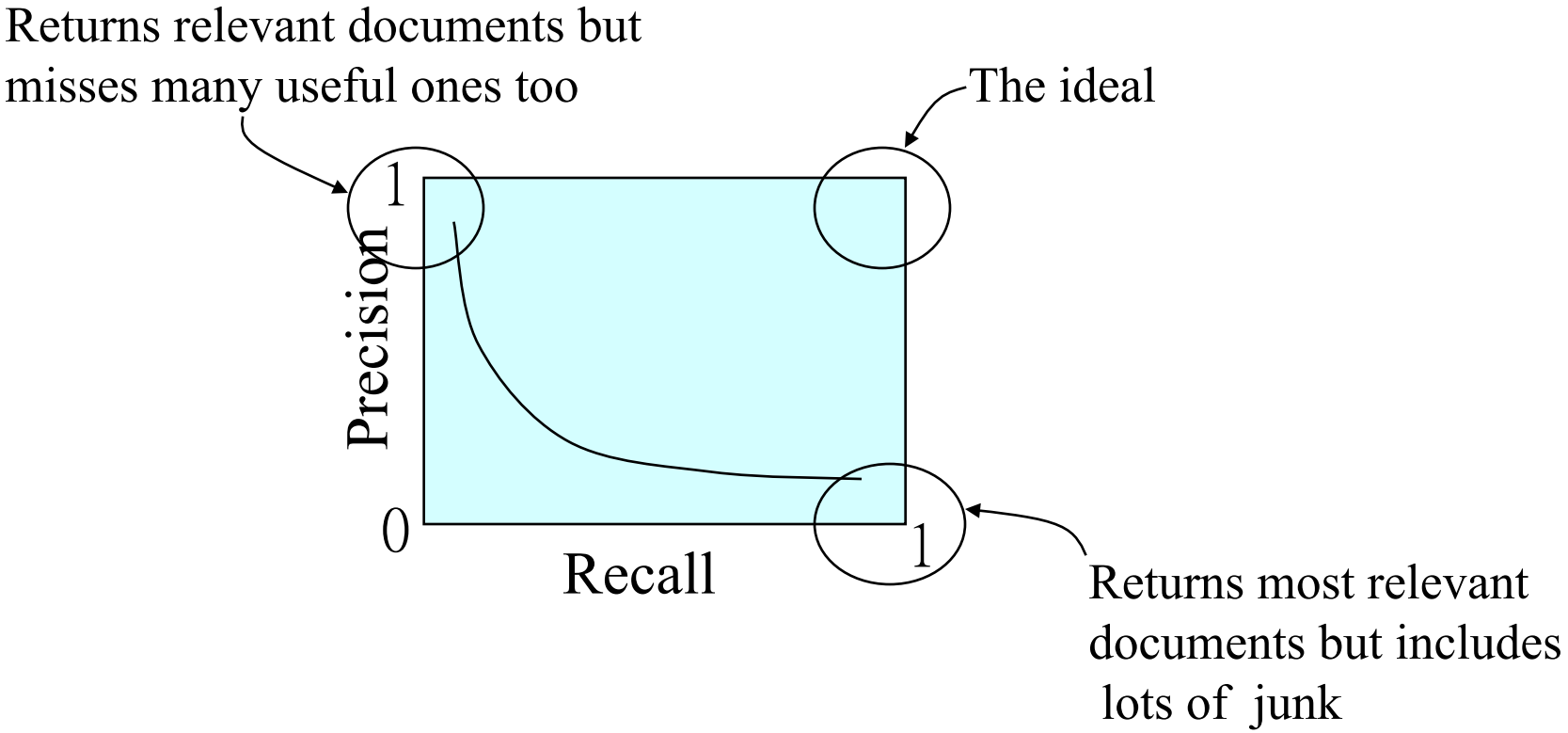
- Precision: How correct is the average answer provided by the system?
- Recall: How many correct pieces of information are retrieved?
- F1: Allows comparative evaluations

Precision/Recall Tradeoff

- You can increase recall by returning more docs.
- Recall is a non-decreasing function of the number of docs retrieved.
- A system that returns all docs has 100% recall!
- The converse is also true (usually): It's easy to get high precision for very low recall.
- Suppose the document with the largest score is relevant. How can we maximize precision?

15

Trade-off between Recall and Precision



Precision/Recall

- You can get high recall (but low precision) by retrieving all docs for all queries!
- Recall is a non-decreasing function of the number of docs retrieved

- In a good system, precision decreases as either the number of docs retrieved or recall increases
  - This is not a theorem, but a result with strong empirical confirmation

Difficulties in using precision/recall

- Should average over large document collection/query ensembles
- Need human relevance assessments
  - People aren't reliable assessors
- Assessments have to be binary
  - Nuanced assessments?
- Heavily skewed by collection/authorship
  - Results may not translate from one domain to another

Should we instead use the accuracy measure for evaluation?

- Given a query, an engine classifies each doc as "Relevant" or "Nonrelevant"
- The **accuracy** of an engine: the fraction of these classifications that are correct
  - (tp + tn) / (tp + fp + fn + tn)
- **Accuracy** is a commonly used evaluation measure in machine learning classification work
- Why is this not a very useful evaluation measure in IR?

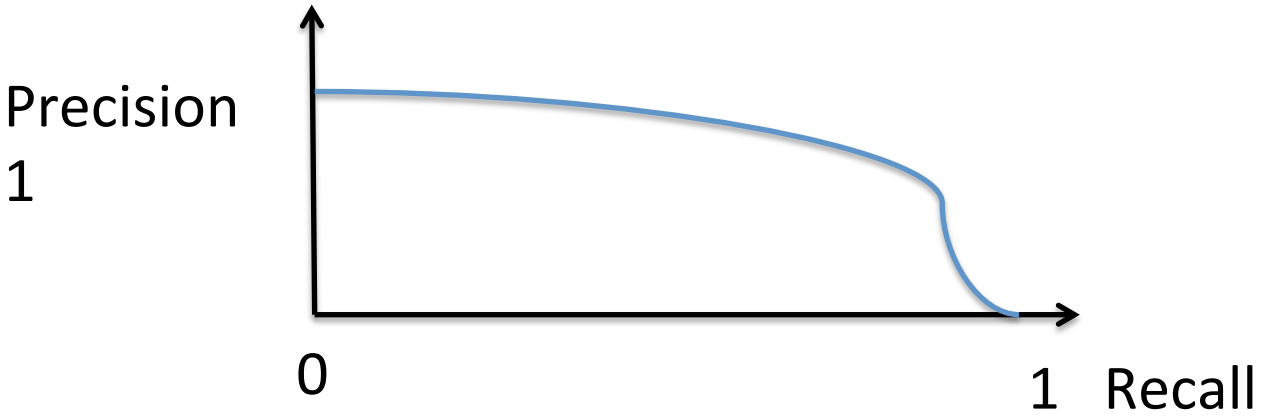
Why not just use accuracy?

- How to build a 99.9999% accurate search engine on a low budget….

- People doing information retrieval *want to find something* and have a certain tolerance for junk.

F1- Measure

You can't get it all...



The F1-measure combines precision and recall as the harmonic mean:

F1 = 2 * precision * recall / (precision + recall)

Slide from Suchanek

F-measure

Precision and Recall stand in opposition to one another. As precision goes up, recall usually goes down (and vice versa).

The F-measure combines the two values.

$$\text{F-measure} = \frac{(ß^2+1)PR}{ß^2 P+R}$$

- When ß = 1, precision and recall are weighted equally (same as F1).
- When ß is > 1, precision is favored.
- When ß is < 1, recall is favored.

Slide modified from Butt/Jurafsky/Martin

F: Why Harmonic Mean?

- Why don't we use a different mean of P and R as a measure?
  - e.g., the arithmetic mean
- The simple (arithmetic) mean is 50% for "return-everything" search engine, which is too high.
- Desideratum: Punish really bad performance on either precision or recall.
- Taking the minimum achieves this.
- But minimum is not smooth and hard to weight.
- F (harmonic mean) is a kind of smooth minimum.

23

F: Example

	relevant	not relevant	total
retrieved	20	40	60
not retrieved	60	1,000,000	1,000,060
total	80	1,000,040	1,000,120

- P = 20/(20 + 40) = 1/3
- R = 20/(20 + 60) = 1/4

$$F_1 = 2\frac{1}{\frac{1}{\frac{1}{3}}+\frac{1}{\frac{1}{4}}} = 2/7$$

24

Model Quality

There are typically other criteria used to measure the quality of models. e.g. for clustering models:

- Silhouette score
- Inter-cluster similarity (e.g. mutual information)
- Intra-cluster entropy

For regression models:

- Stability of the model (sensitivity to small changes)
- Compactness (sparseness or many zero coefficients)

Difficulties in Using Precision, Recall and F

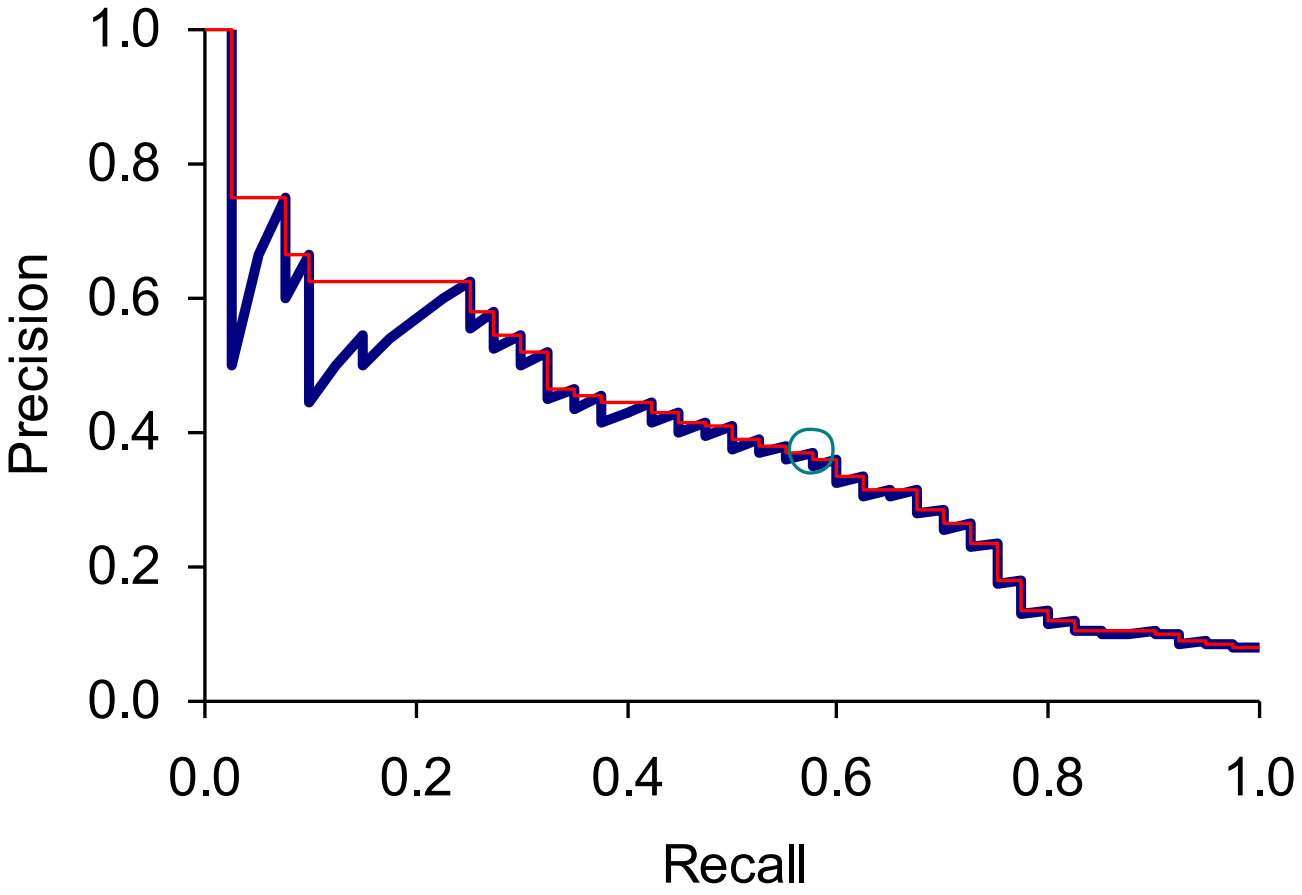
- We need relevance judgments for information-need-document pairs – but they are expensive to produce.
- For alternatives to using precision/recall and having to produce relevance judgments – see end of this lecture.

26

Evaluating ranked results

- Evaluation of ranked results:
  - The system can return any number of results
  - By taking various numbers of the top returned documents (levels of recall), the evaluator can produce a *precision-recall curve*

A precision-recall curve

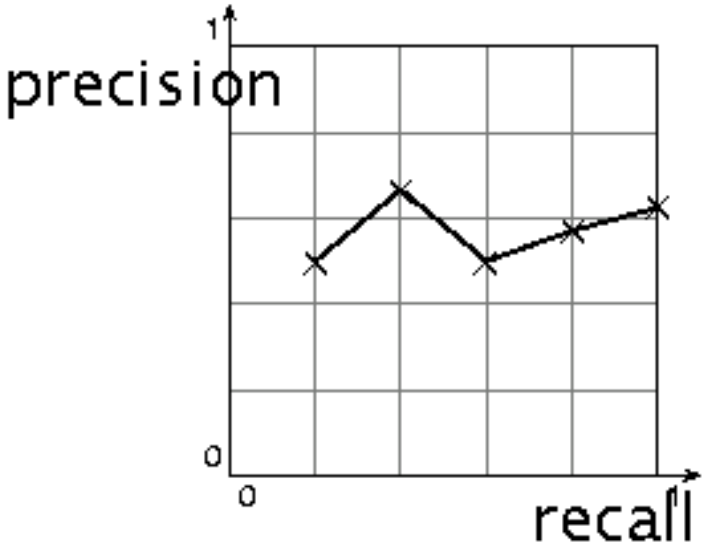


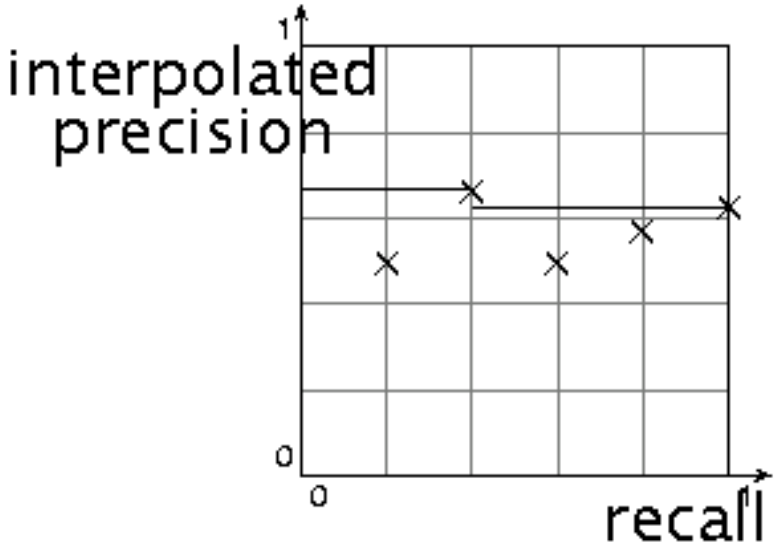
Averaging over queries

- A precision-recall graph for one query isn't a very sensible thing to look at
- You need to average performance over a whole bunch of queries.
- But there's a technical issue:
  - Precision-recall calculations place some points on the graph
  - How do you determine a value (interpolate) between the points?

Interpolated precision

- Idea: If locally precision increases with increasing recall, then you should get to count that…





Evaluation

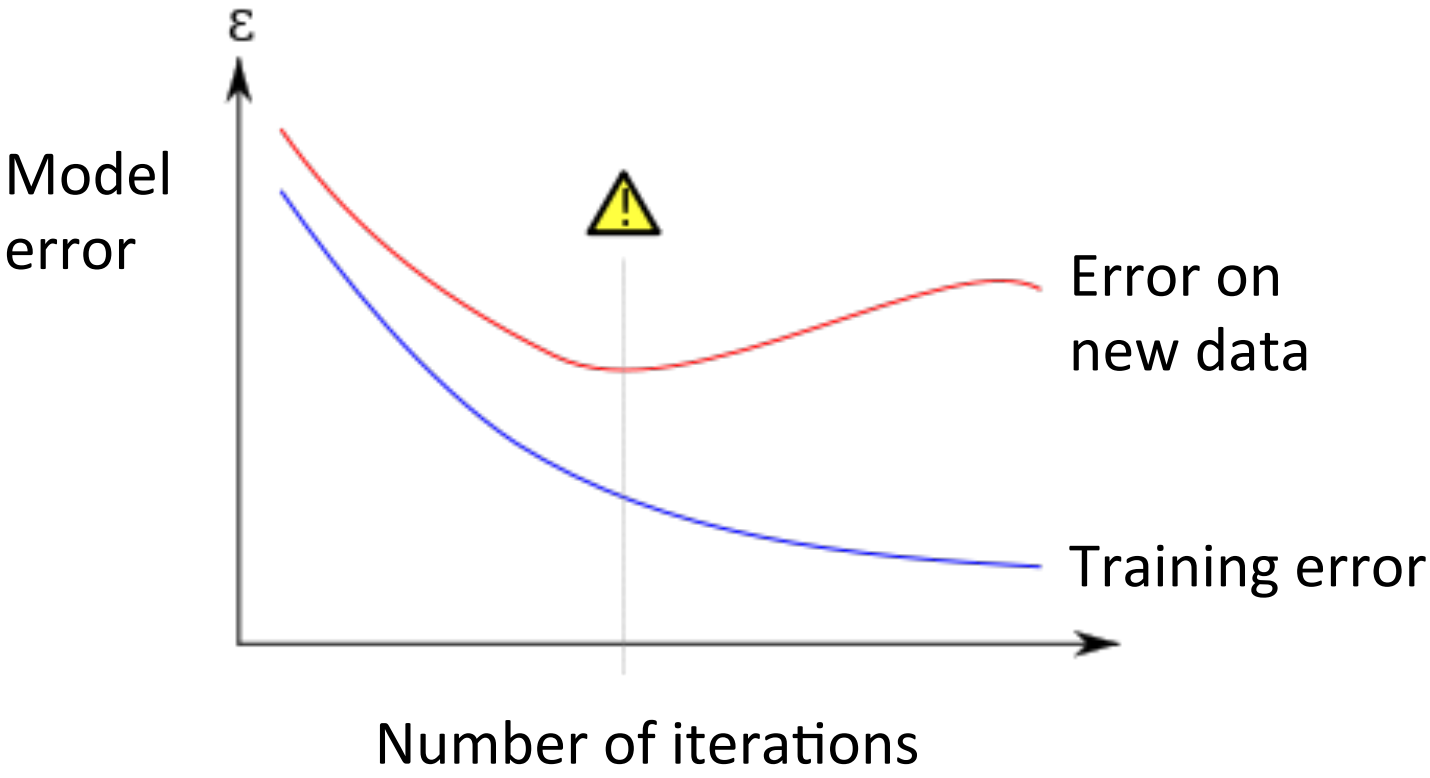
- Graphs are good, but people want summary measures!
  - Precision at fixed retrieval level
    - Precision-at-*k*: Precision of top *k* results
    - Perhaps appropriate for most of web search: all people want are good matches on the first one or two results pages
    - But: averages badly and has an arbitrary parameter of *k*
  - 11-point interpolated average precision
    - The standard measure in the early TREC competitions: you take the precision at 11 levels of recall varying from 0 to 1 by tenths of the documents, using interpolation (the value for 0 is always interpolated!), and average them
    - Evaluates performance at all recall levels

Over-fitting

- Your model should ideally fit an **infinite sample** of the type of data you're interested in.
- In reality, you only have a **finite set** to train on. A good model for this subset is a good model for the infinite set, up to a point.
- Beyond that point, the model quality (measured on new data) starts to **decrease**.
- Beyond that point, the model is **over-fitting** the data.

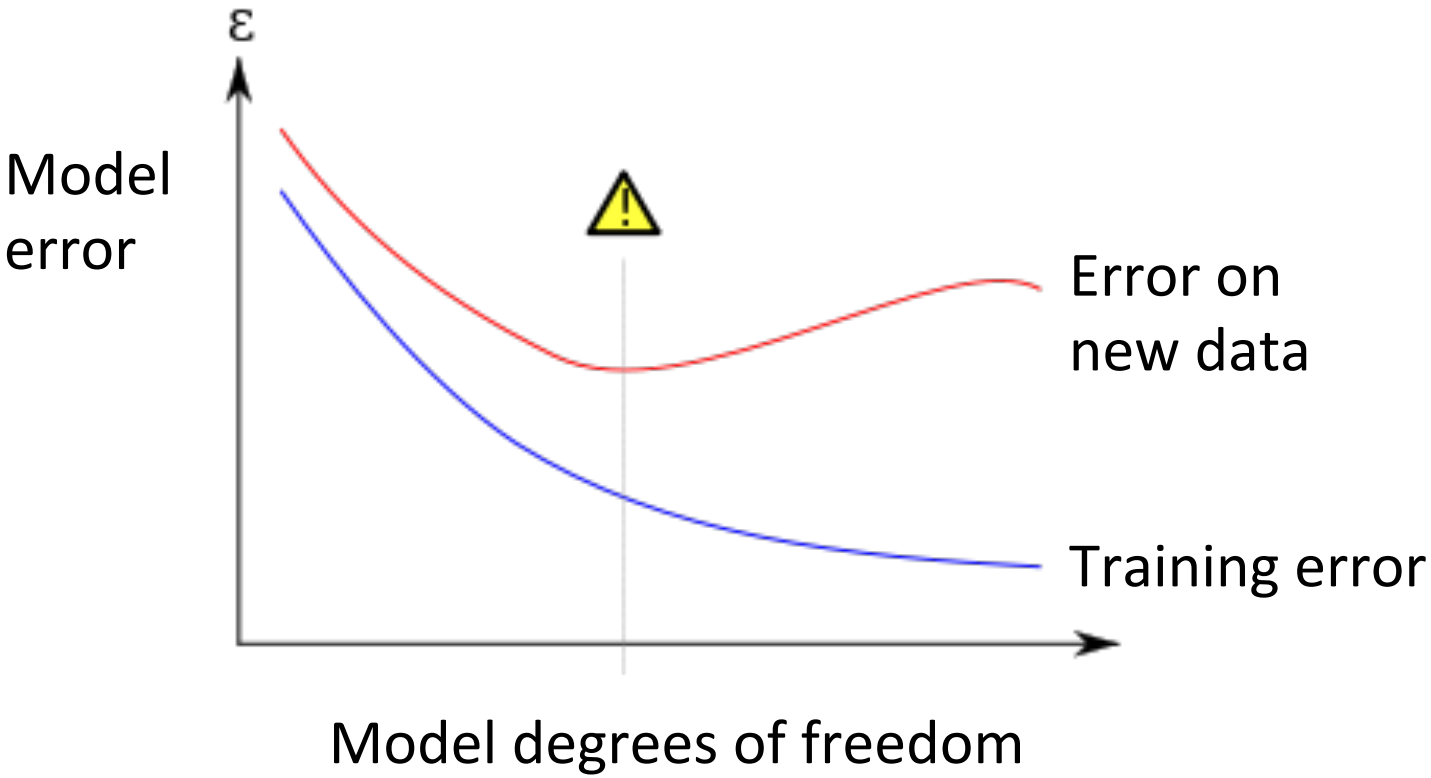
Over-fitting

Over-fitting during training



Over-fitting

Another kind of over-fitting



Cross-Validation

- Cross-validation involves **partitioning** your data into distinct **training** and **test** subsets.

- The test set **should never** be used to **train** the model.

- The test set is then used to **evaluate** the model after training.

LOOCV

- Leave one out cross validation
- Split observation set into two parts
- Instead of two sets of comparable size, a single observation (x_1,y_1) is used for the validation set, $MSE_1 = (y1-\hat{y}1)^2$
- Repeat the procedure on n-1 observations and

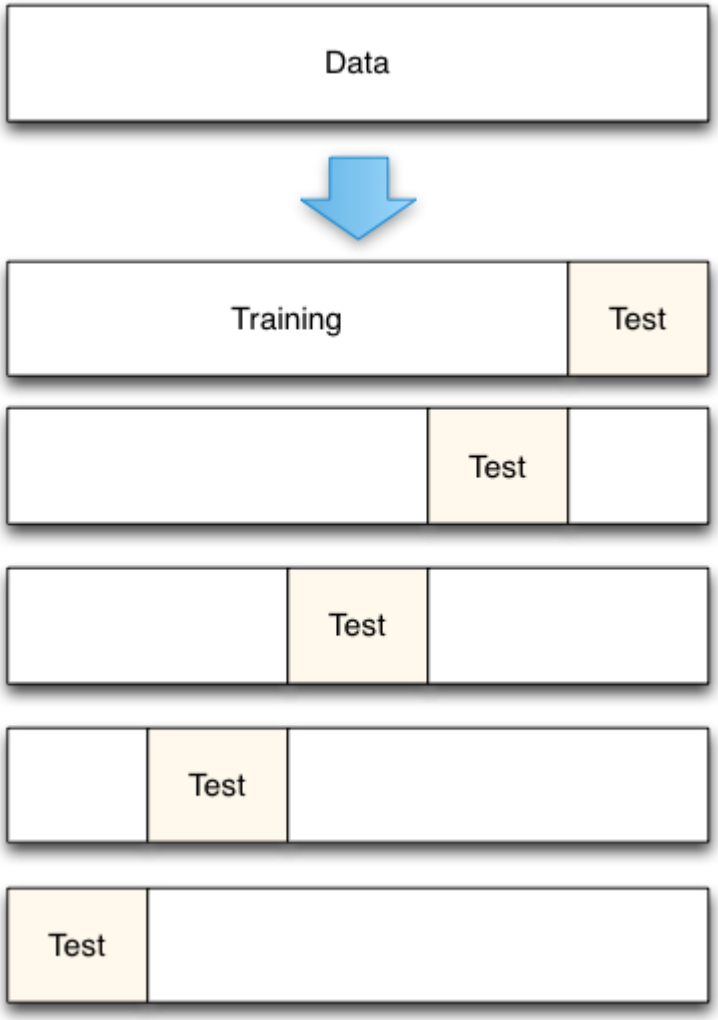
$$CV(n) = \frac{1}{n}\sum_{i=1}^{n} MSE_i$$

K-fold Cross-Validation

- To get more accurate estimates of performance you can do this k times.
- Break the data into k equal-sized subsets A_i
- For each i in 1,…,k do:
  - Train a model on all the other folds A_1,…, A_{i-1}, A_{i+1},…, A_k
  - Test the model on A_i
- Compute the **average performance** of the k runs

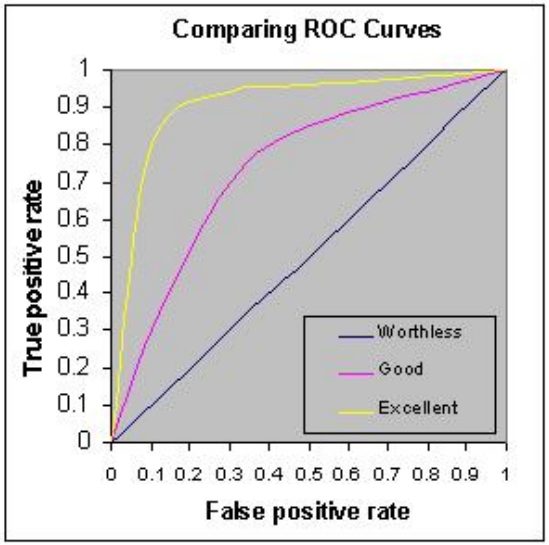
$$CV(k) = \frac{1}{k}\sum_{i=1}^{k} MSE_i$$

5-fold Cross-Validation



ROC curves

- It is common to plot classifier performance at a variety of settings or thresholds
- Receiver Operating Characteristic (ROC) curves plot true positives against false positives.
- The overall performance is calculated by the Area Under the Curve (AUC)



ROC

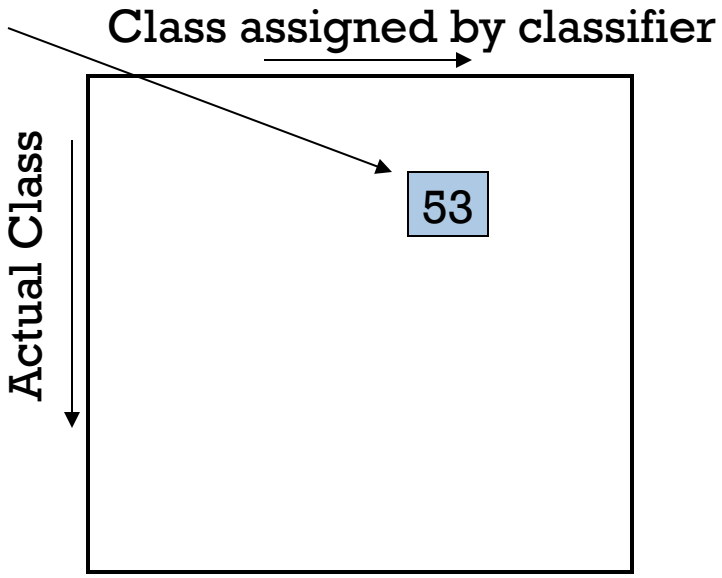
- An ROC curve represents a relation between sensitivity (RECALL) and specificity (NOT PRECISION). Sensitivity is the other name for recall but specificity is not PRECISION.

- So, if your problem involves kind of searching a needle in the haystack when for ex: the positive class samples are very rare compared to the negative classes, use a precision recall curve.

- Othwerwise use a ROC curve because a ROC curve remains the same regardless of the baseline prior probability of your positive class (the important rare class).

Good practice department: Make a confusion matrix

This (*i*, *j*) entry means 53 of the docs actually in class *i* were put in class *j* by the classifier.



- In a perfect classification, only the diagonal has non-zero entries
- Look at common confusions and how they might be addressed

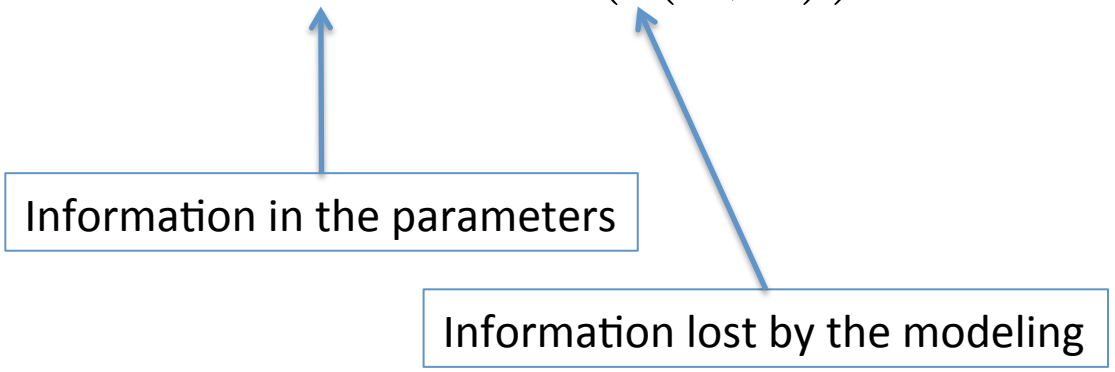
Goodness of Fit

- Another view of model performance.
- Measure the model likelihood of the unseen data. $l(x;\theta)$
- However, we've seen that model likelihood is likely to improve by adding parameters.
- Two information criteria measures include a cost term for the number of parameters in the model

Akaike Information Criterion

- Akaike Information Criterion (AIC) based on entropy
- The best model has the lowest AIC.
  - Greatest model likelihood
  - Fewest free parameters

$$AIC = 2k - 2\ln(l(x;\theta))$$



Cluster Validation

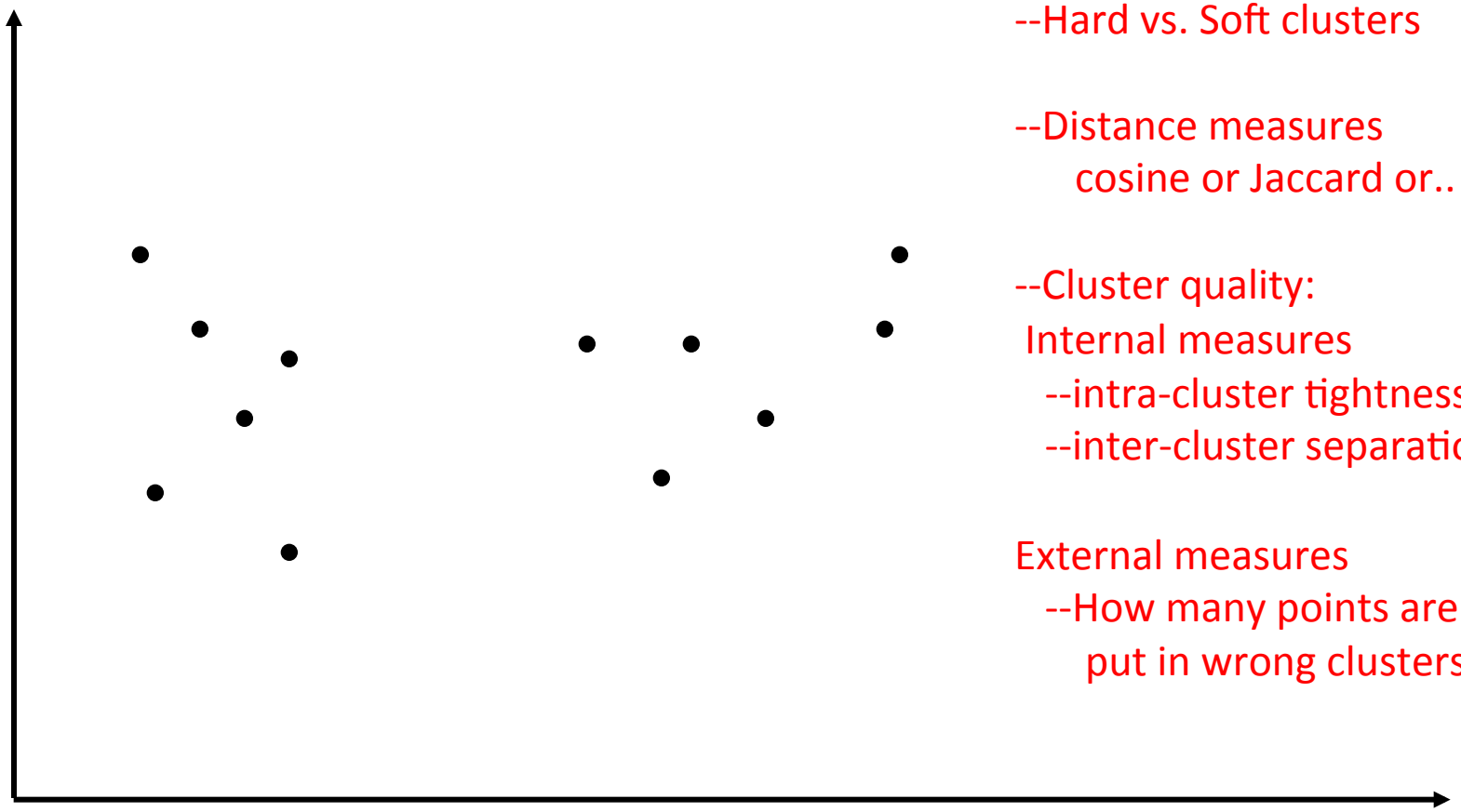
Cluster Evaluation

- "Clusters can be evaluated with "internal" as well as "external" measures
  - Internal measures are related to the inter/intra cluster distance
    - A good clustering is one where
      - » (Intra-cluster distance) the sum of distances between objects in the same cluster are minimized,
      - » (Inter-cluster distance) while the distances between different clusters are maximized
      - » Objective to minimize: F(Intra,Inter)
  - External measures are related to how representative are the current clusters to "true" classes. Measured in terms of purity, entropy or F-measure

General issues in clustering

- Inputs/Specs
  - Are the clusters "hard" (each element in one cluster) or "Soft"
    - Hard Clustering=> partitioning
    - Soft Clustering=> subsets..
  - Do we know how many clusters we are supposed to look for?
    - Max # clusters?
    - Max possibilities of clusterings?
- What is a good cluster?
  - Are the clusters "close-knit"?
  - Do they have any connection to reality?
    - Sometimes we try to figure out reality by clustering…
- Importance of notion of distance
  - Sensitivity to outliers?

Clustering issues



--Hard vs. Soft clusters

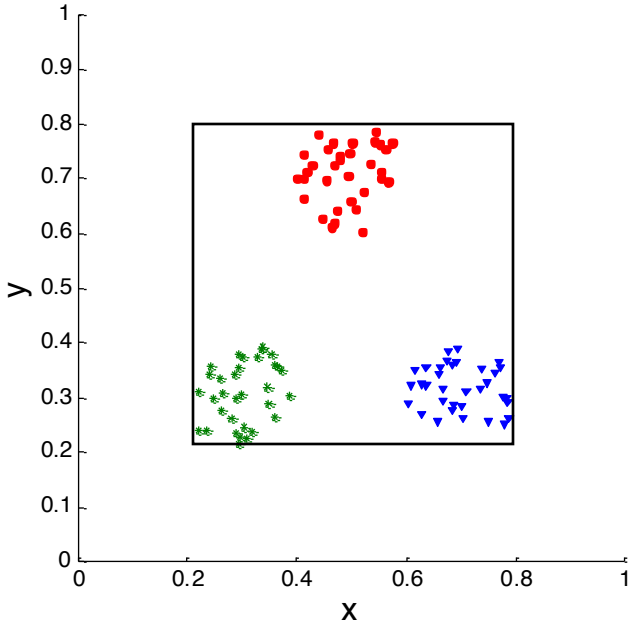
--Distance measures
cosine or Jaccard or..

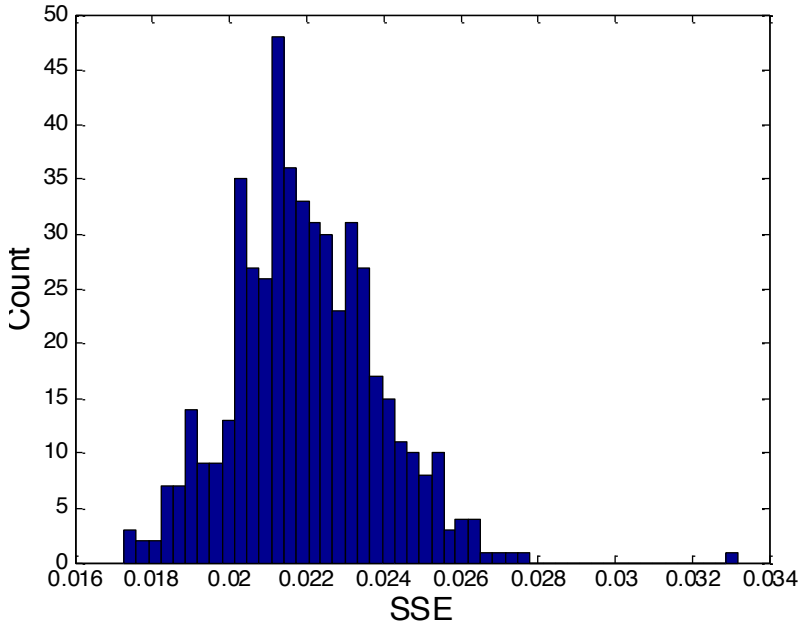
--Cluster quality:
Internal measures
--intra-cluster tightness
--inter-cluster separation

External measures
--How many points are put in wrong clusters.

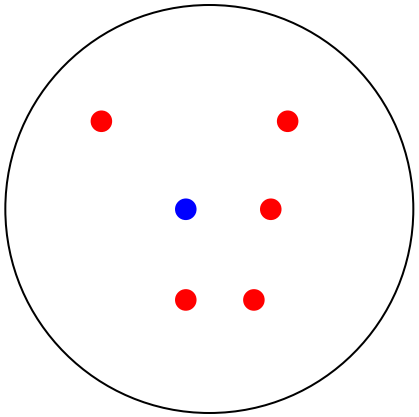
Statistical Framework for SSE

- Example
  - Compare SSE of 0.005 against three clusters in random data
  - Histogram shows SSE of three clusters in 500 sets of random data points of size 100 distributed over the range 0.2 – 0.8 for x and y values

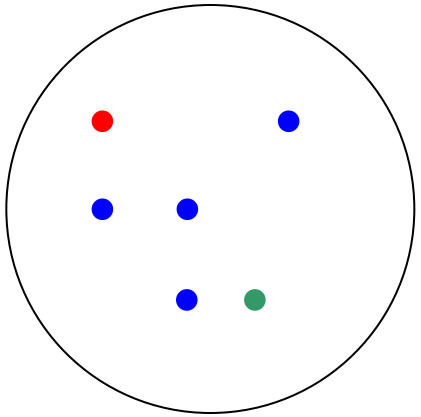




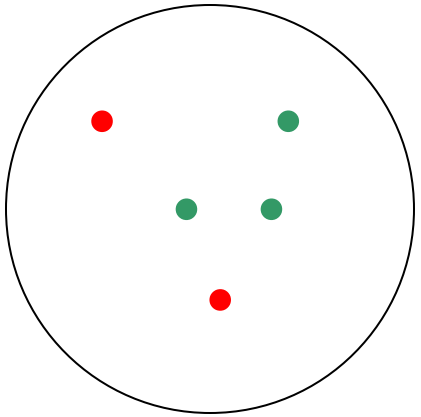
Purity example



Cluster I



Cluster II



Cluster III

Cluster I: Purity = 1/6 (max(5, 1, 0)) = 5/6

Cluster II: Purity = 1/6 (max(1, 4, 1)) = 4/6

Cluster III: Purity = 1/5 (max(2, 0, 3)) = 3/5

Overall
Purity
= weighted purity

Rand-Index: Precision/Recall based

Number of points	Same Cluster in clustering	Different Clusters in clustering
Same class in ground truth	A	C
Different classes in ground truth	B	D

$$RI = \frac{A + D}{A + B + C + D}$$

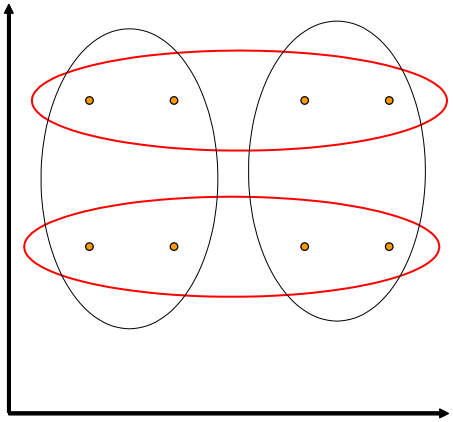
$$P = \frac{A}{A + B}$$

$$R = \frac{A}{A + C}$$

Inter/Intra Cluster Distances

Intra-cluster distance/tightness

- (Sum/Min/Max/Avg) the (absolute/ squared) distance between
  - All pairs of points in the cluster OR
  - Between the centroid and all points in the cluster OR
  - Between the "medoid" and all points in the cluster



Inter-cluster distance

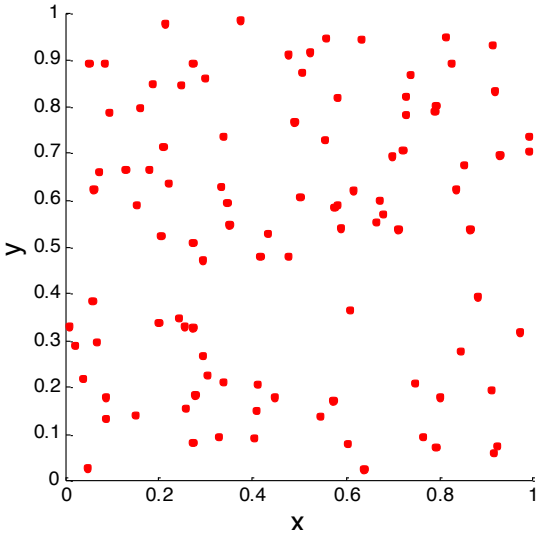
Sum the (squared) distance between all pairs of clusters

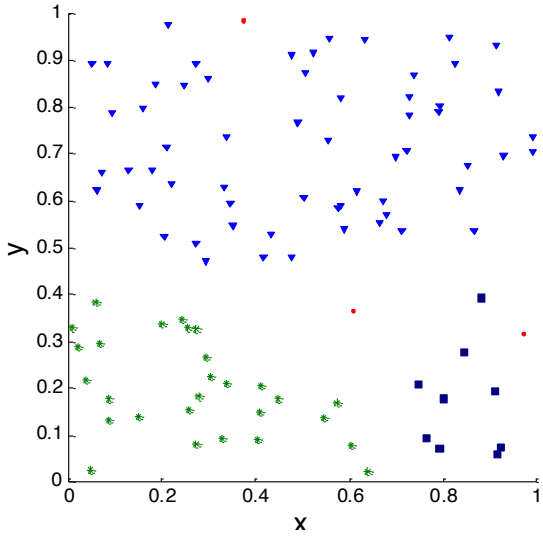
Where distance between two clusters is defined as:

- distance between their centroids/medoids
- Distance between the closest pair of points belonging to the clusters (single link)
  - (Chain shaped clusters)
- Distance between farthest pair of points (complete link)
  - (Spherical clusters)

Clusters found in Random Data

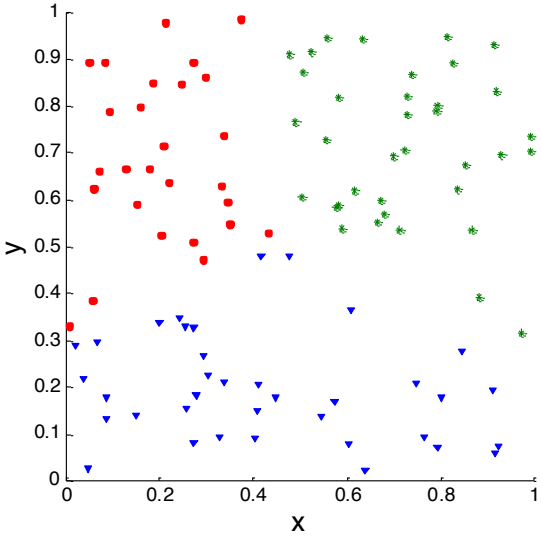
Random Points

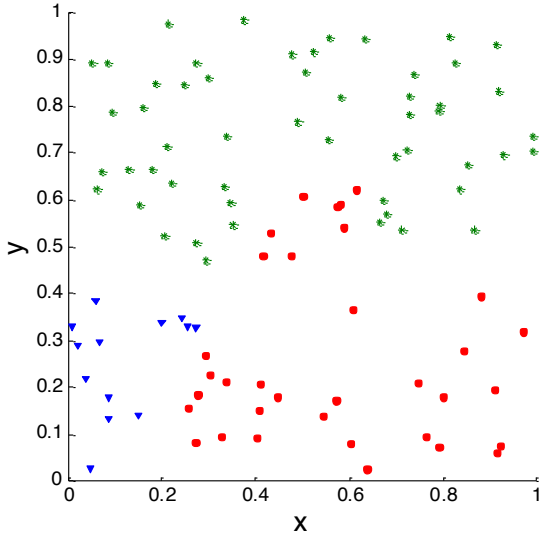




DBSCAN

K-means





Complete Link

Different Aspects of Cluster Validation

1. Determining the clustering tendency of a set of data, i.e., distinguishing whether non-random structure actually exists in the data.
2. Comparing the results of a cluster analysis to externally known results, e.g., to externally given class labels.
3. Evaluating how well the results of a cluster analysis fit the data *without* reference to external information.
   - Use only the data
4. Comparing the results of two different sets of cluster analyses to determine which is better.
5. Determining the ‘correct’ number of clusters.

For 2, 3, and 4, we can further distinguish whether we want to evaluate the entire clustering or just individual clusters.

Framework for Cluster Validity

- Need a framework to interpret any measure.
  - For example, if our measure of evaluation has the value, 10, is that good, fair, or poor?
- Statistics provide a framework for cluster validity
  - The more "atypical" a clustering result is, the more likely it represents valid structure in the data
  - Can compare the values of an index that result from random data or clusterings to those of a clustering result.
    - If the value of the index is unlikely, then the cluster results are valid
  - These approaches are more complicated and harder to understand.
- For comparing the results of two different sets of cluster analyses, a framework is less necessary.
  - However, there is the question of whether the difference between two index values is significant

Measures of Cluster Validity

- Numerical measures that are applied to judge various aspects of cluster validity, are classified into the following three types.
  - External Index: Used to measure the extent to which cluster labels match externally supplied class labels.
    - Entropy
  - Internal Index: Used to measure the goodness of a clustering structure *without* respect to external information.
    - Sum of Squared Error (SSE)
  - Relative Index: Used to compare two different clusterings or clusters.
    - Often an external or internal index is used for this function, e.g., SSE or entropy
- Sometimes these are referred to as criteria instead of indices
  - However, sometimes criterion is the general strategy and index is the numerical measure that implements the criterion.

External Validation

Algorithm 21.4: Algorithm for matching partitions and clusters

MatchPartitionCluster **(**P,C,$match$**)**:

1 **foreach** $p \in P$ **do**
2 $\quad match(p) \leftarrow \emptyset$
3 $\quad$ **foreach** $c \in C$ **do**
4 $\quad\quad overlap(p, c) \leftarrow \frac{|p \cap c|}{|p|}$
5 **while** $overlap \neq \emptyset$ **do**
6 $\quad (p_{max}, c_{max}) \leftarrow GetMaxOverlap(overlap)$
7 $\quad match(p_{max}) \leftarrow c_{max}$
8 $\quad overlap \leftarrow overlap - \{overlap(p_{max}, *), overlap(*, c_{max})\}$

Purity-Based Measure

- Purity

$$-\frac{|c_i \cap p_j|}{|c_i|} \qquad \max_j \rho_{ij} \qquad purity_C = \sum_r \frac{|c_i|}{|c|} purity_i,$$

- Precision/Recall/F-Measure

  prec(i,j), recall(i,j), $F(i,j) = \frac{2 \times prec(i,j) \times rec(i,j)}{prec(i,j) + rec(i,j)}$

- Entropy

$$e_i = -\sum_q \rho_{ij} \log_2 \rho_{ij}.$$

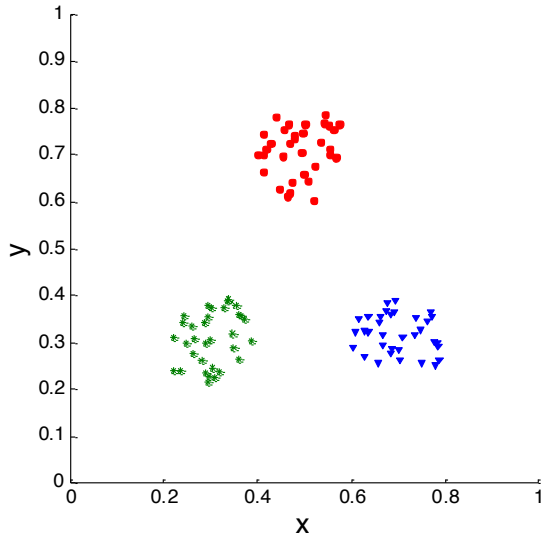
$$e_C = \sum_r \frac{|c_i|}{|c|} e_i,$$

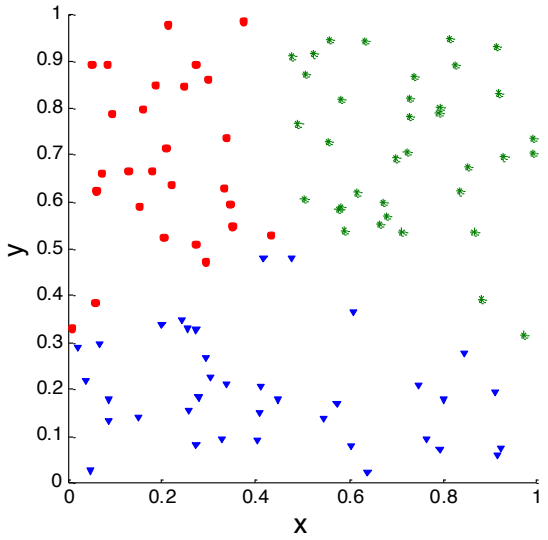
Measuring Cluster Validity Via Correlation

- Two matrices
  - Proximity Matrix
  - "Incidence" Matrix
    - One row and one column for each data point
    - An entry is 1 if the associated pair of points belong to the same cluster
    - An entry is 0 if the associated pair of points belongs to different clusters
- Compute the correlation between the two matrices
  - Since the matrices are symmetric, only the correlation between n(n-1) / 2 entries needs to be calculated.
- High correlation indicates that points that belong to the same cluster are close to each other.
- Not a good measure for some density or contiguity based clusters.

Measuring Cluster Validity Via Correlation

- Correlation of incidence and proximity matrices for the K-means clusterings of the following two data sets.



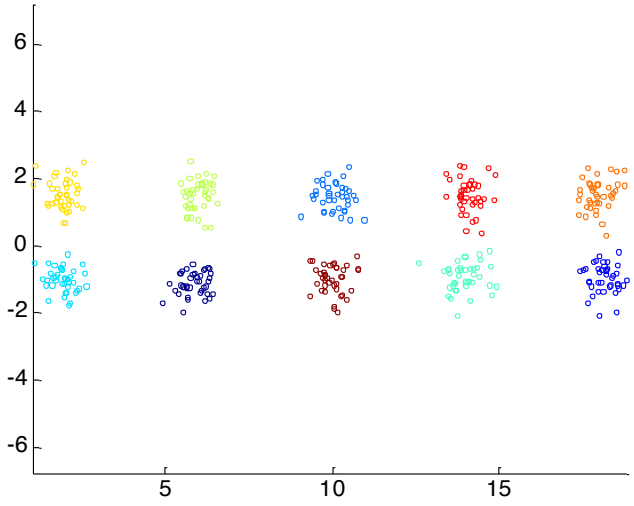


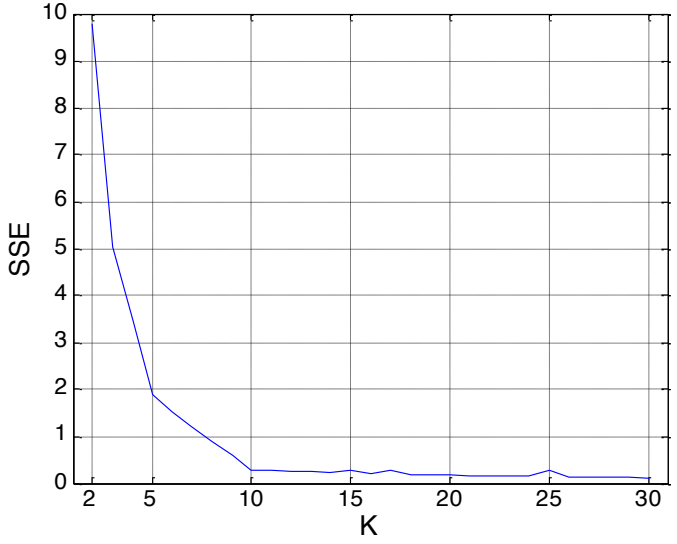
Corr = -0.9235

Corr = -0.5810

Internal Measures: SSE

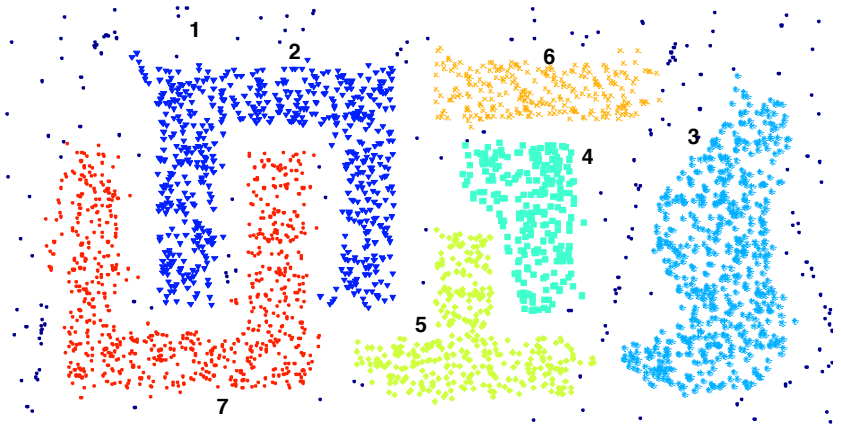
- Clusters in more complicated figures aren't well separated
- Internal Index: Used to measure the goodness of a clustering structure without respect to external information
  - SSE
- SSE is good for comparing two clusterings or two clusters (average SSE).
- Can also be used to estimate the number of clusters

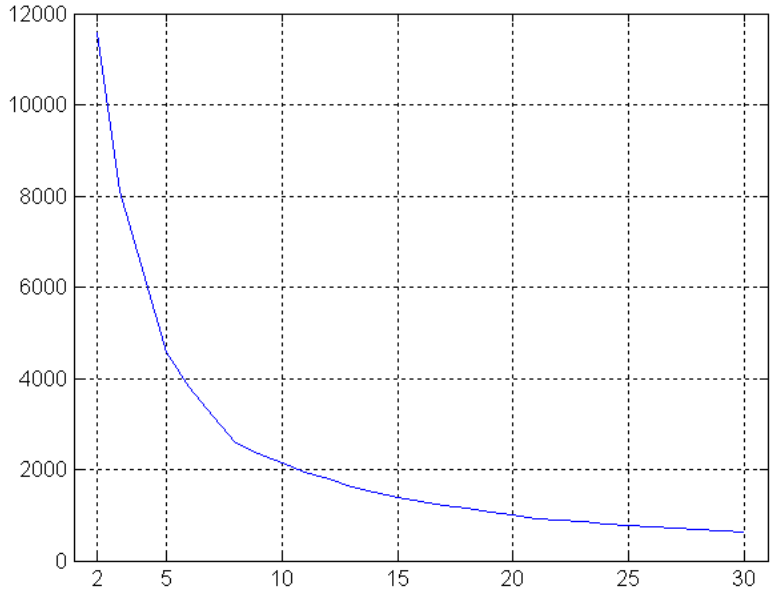




Internal Measures: SSE

- SSE curve for a more complicated data set





SSE of clusters found using K-means

Internal Measures: Cohesion and Separation

- Cluster Cohesion: Measures how closely related are objects in a cluster
  - Example: SSE
- Cluster Separation: Measure how distinct or well-separated a cluster is from other clusters
- Example: Squared Error
  - Cohesion is measured by the within cluster sum of squares (SSE)

$$\boldsymbol{WSS} = \sum_{\boldsymbol{i}} \sum_{\boldsymbol{x} \in \boldsymbol{C}_i} (\boldsymbol{x} - \boldsymbol{m}_{\boldsymbol{i}})^2$$

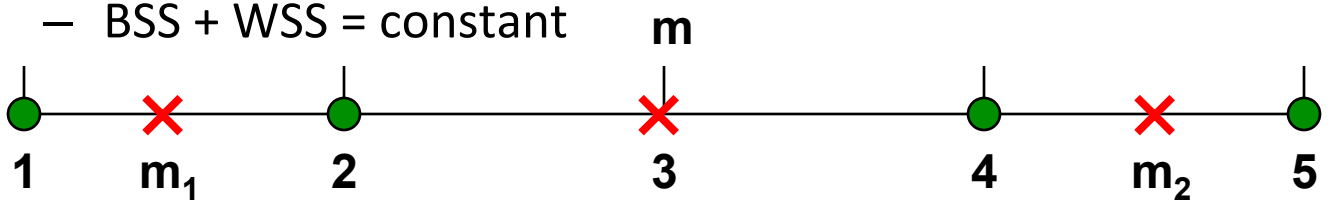
  - Separation is measured by the between cluster sum of squares

$$BSS = \sum_{i} |C_i| (m - m_i)^2$$

    - Where $|C_i|$ is the size of cluster i

Internal Measures: Cohesion and Separation

- Example: SSE
  - BSS + WSS = constant



K=1 cluster:

$$WSS = (1-3)^2 + (2-3)^2 + (4-3)^2 + (5-3)^2 = 10$$

$$BSS = 4 \times (3-3)^2 = 0$$

$$Total = 10 + 0 = 10$$

K=2 clusters:

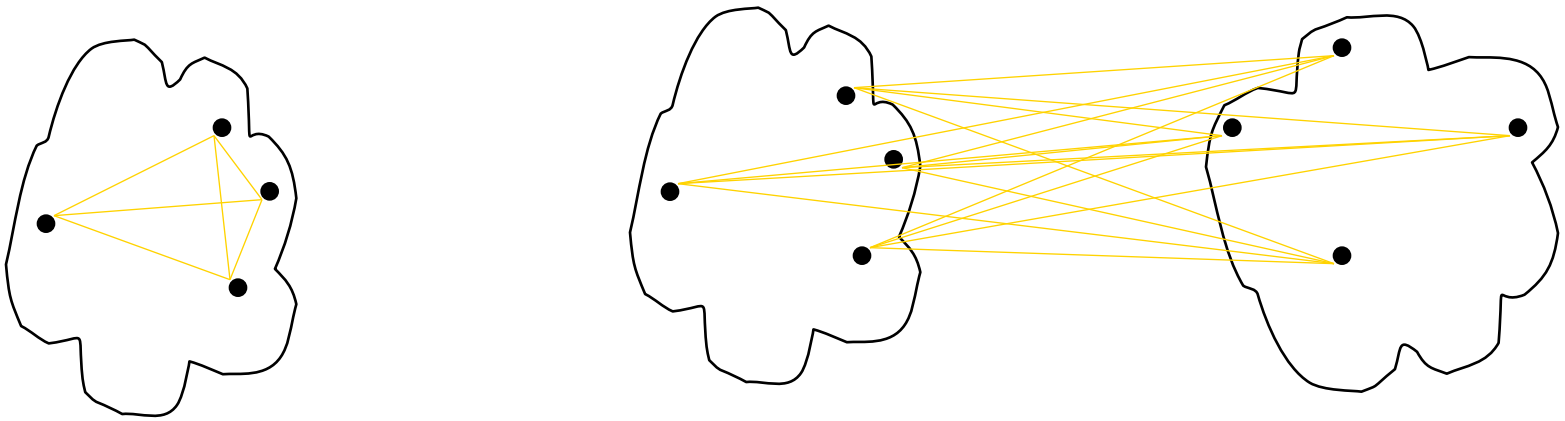
$$WSS = (1-1.5)^2 + (2-1.5)^2 + (4-4.5)^2 + (5-4.5)^2 = 1$$

$$BSS = 2 \times (3-1.5)^2 + 2 \times (4.5-3)^2 = 9$$

$$Total = 1 + 9 = 10$$

Internal Measures: Cohesion and Separation

- A proximity graph based approach can also be used for cohesion and separation.
  - Cluster cohesion is the sum of the weight of all links within a cluster.
  - Cluster separation is the sum of the weights between nodes in the cluster and nodes outside the cluster.



cohesion

separation

External Measures of Cluster Validity: Entropy and Purity

entropy For each cluster, the class distribution of the data is calculated first, i.e., for cluster j we compute p_{ij}, the 'probability' that a member of cluster j belongs to class i as follows: $p_{ij} = m_{ij}/m_j$, where m_j is the number of values in cluster j and m_{ij} is the number of values of class i in cluster j. Then using this class distribution, the entropy of each cluster j is calculated using the standard formula $e_j = \sum_{i=1}^{L} p_{ij} \log_2 p_{ij}$, where the L is the number of classes. The total entropy for a set of clusters is calculated as the sum of the entropies of each cluster weighted by the size of each cluster, i.e., $e = \sum_{i=1}^{K} \frac{m_i}{m} e_j$, where m_j is the size of cluster j, K is the number of clusters, and m is the total number of data points.

purity Using the terminology derived for entropy, the purity of cluster j, is given by $purity_j = \max p_{ij}$ and the overall purity of a clustering by $purity = \sum_{i=1}^{K} \frac{m_i}{m} purity_j$.

Final Comment on Cluster Validity

"The validation of clustering structures is the most difficult and frustrating part of cluster analysis.

Without a strong effort in this direction, cluster analysis will remain a black art accessible only to those true believers who have experience and great courage."

Algorithms for Clustering Data, Jain and Dubes

The Real World

P. Jackson and I. Moulinier. 2002. *Natural Language Processing for Online Applications*

- "There is no question concerning the commercial value of being able to classify documents automatically by content. There are myriad potential applications of such a capability for corporate intranets, government departments, and Internet publishers"

- "Understanding the data is one of the keys to successful categorization, yet this is an area in which most categorization tool vendors are extremely weak. Many of the 'one size fits all' tools on the market have not been tested on a wide range of content types."

Evaluation of PCA Analysis

- Variance of the data along each principle component provides one means for determining the minimum number of significant factors needed to explain the data. The first principle component explains the greatest percentage of the data's original total variance (that is, the variance around the global mean), which each succeeding principle component explaining less of the total variance.

```
> summary(pr.r)
Importance of components:
                          PC1     PC2      PC3       PC4
Standard deviation      1.43    0.2486   0.00788   0.000685
Proportion of Variance  0.97    0.0294   0.00003   0.000000
Cumulative Proportion   0.97    1.0000   1.00000   1.000000
   <values omitted to save space>
                          PC25      PC26      PC27     PC28
Standard deviation      7.47e-05  6.42e-05  4.9e-05  3.26e-05
Proportion of Variance  0.00e+00  0.00e+00  0.0e+00  0.00e+00
Cumulative Proportion   1.00e+00  1.00e+00  1.0e+00  1.00e+00
```

- Those principal components accounting for 99% of the total variance provide a useful subset for explaining the data. In this case, the first two principal components account for more than 99% of the total variance, suggesting that two components are all that are need to explain the data.

Measuring Classification Figures of Merit

- Not just accuracy; in the real world, there are economic measures:
  - Your choices are:
    - Do no classification
      - That has a cost (hard to compute)
    - Do it all manually
      - Has an easy-to-compute cost if doing it like that now
    - Do it all with an automatic classifier
      - Mistakes have a cost
    - Do it with a combination of automatic classification and manual review of uncertain/difficult/"new" cases
  - Commonly the last method is most cost efficient and is adopted

Bias/Variance Tradeoff

Model Loss (Error)

- Squared loss of model on test case i:

$(\text{Learn}(x_i, D) - \text{Truth}(x_i))^2$

- Expected prediction error:

$\langle \text{Learn}(x_i, D) - \text{Truth}(x_i)^2 \rangle_D$

Bias/Variance Decomposition

- $\langle L(x,D) - T(x) \rangle^2_D$

  $= \text{Noise}^2 + \text{Bias}^2 + \text{Variance}$

- Noise^2 = lower bound on performance
- Bias^2 = (expected error due to model mismatch)^2
- Variance = variation due to train sample and randomization

Bias^2

- Low bias
  - linear regression applied to linear data
  - 2nd degree polynomial applied to quadratic data
- High bias
  - constant function
  - linear regression applied to non-linear data

Variance

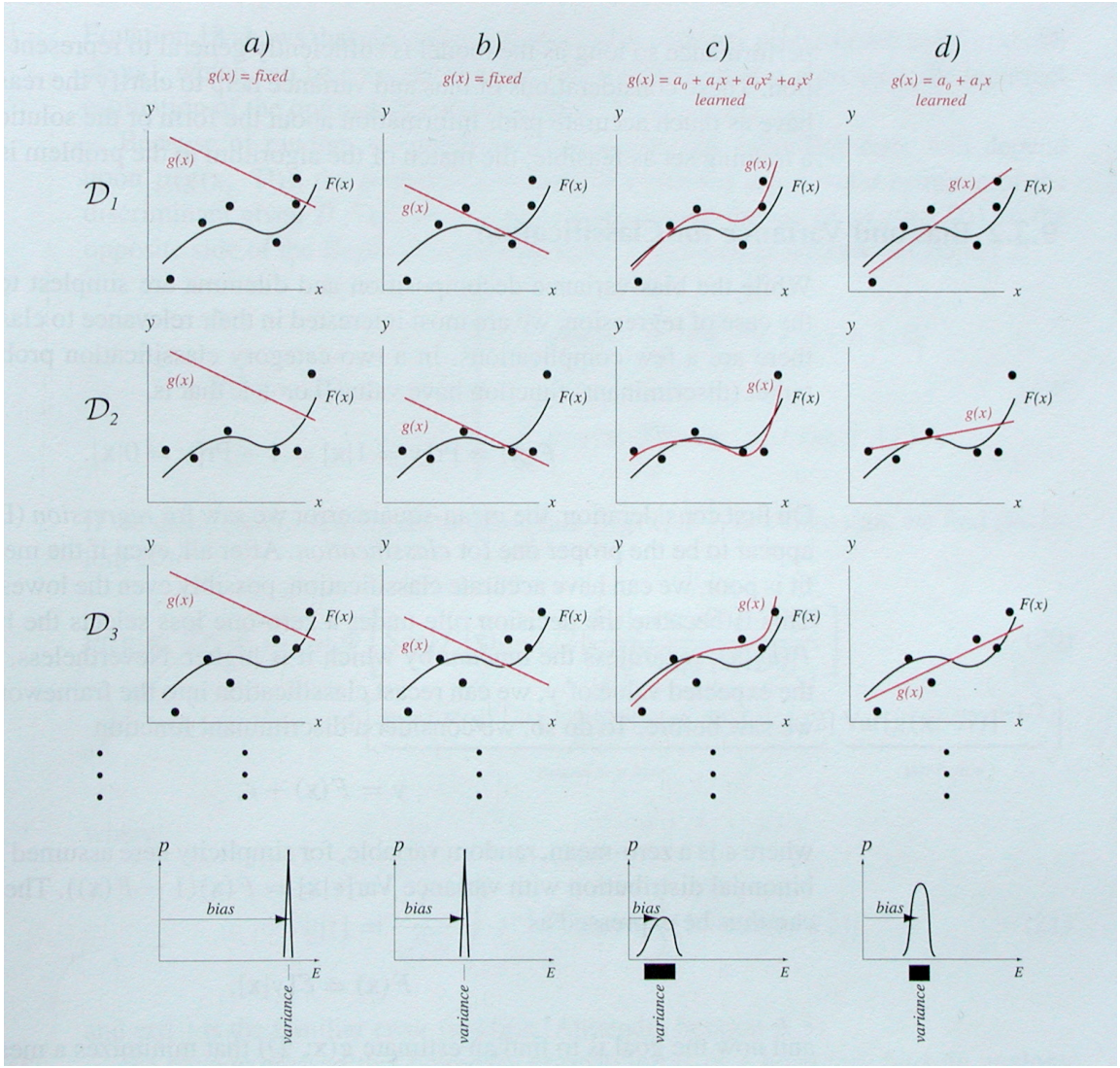
- Low variance
  - constant function
  - model independent of training data
  - model depends on stable measures of data
    - mean
    - median
- High variance
  - high degree polynomial

Sources of Variation in Unsupervised Learning

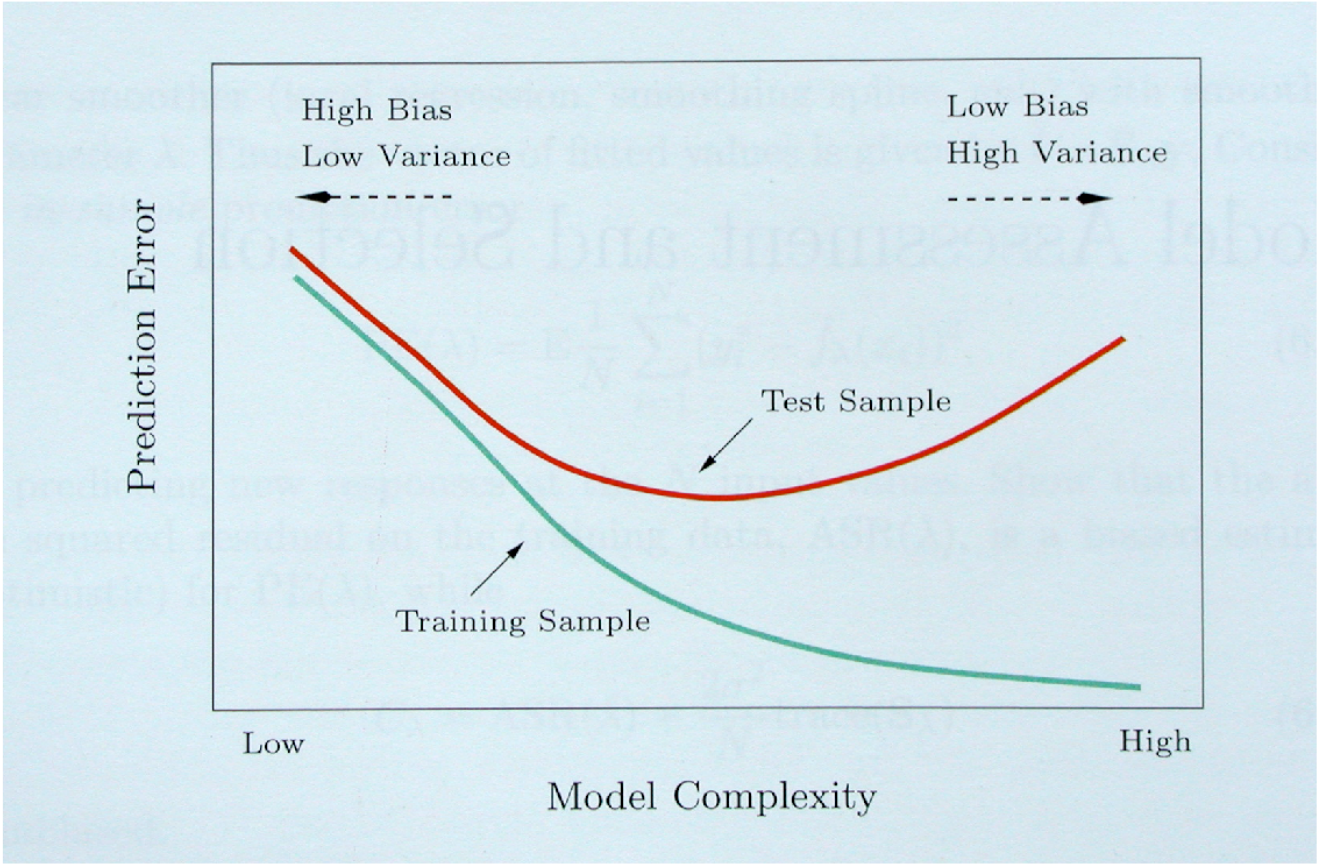
- noise in targets or input attributes
- bias (model mismatch)
- training sample
- randomness in learning algorithm
- randomized subsetting of train set:
  - cross validation, train and early stopping set

Bias/Variance Tradeoff

- Often:
  - low bias => high variance
  - low variance => high bias
- Tradeoff:
  - bias^2 vs. variance


a)
g(x) = fixed
b)
g(x) = fixed
c)
g(x) = a0 + a1x + a0x2 +a3x3
learned
d)
g(x) = a0 + a1x
learned
D1
D2
D3
y
x
g(x)
F(x)
p
bias
variance
E

Bias/Variance Tradeoff



Hastie, Tibshirani, Friedman "Elements of Statistical Learning" 2001

Reduce Variance Without Increasing Bias

- Averaging reduces variance:

$$Var(\overline{X}) = \frac{Var(X)}{N}$$

- Average models to reduce model variance
- One problem:
  - only one train set
  - where do multiple models come from?

How do you know that you have a good classifier?

- Is a feature contributing to overall performance?
- Is classifier A better than classifier B?
- Internal Evaluation:
  - Measure the performance of the classifier.
- External Evaluation:
  - Measure the performance on a downstream task

Summary

- Regression evaluation
- Overall model/classifier evaluation
  - Metrics: Precision, Recall, F-statistic
  - Meta-figures: Precision-recall, ROC
  - Trade-offs
- Cluster Evaluation

In Class Exercise:

- Perform a cross-validation on a simulated data set

a) Generate a simulated data set as follows:

```
set.seed(1)
y=rnorm(100)
x=rnorm(100)
y=x-2*x^2+rnorm(100)
```

In this data, what is *n* and what is *p*? Write out the model used to generate the data in equation form.

In Class Exercise:

- Create a scatterplot of X against Y. What do you find?
- Set a random seed, and then compute the LOOCV errors that result from fitting the following models:

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

$$Y = \beta_0 + \beta_1 X + + \beta_2 X^2 + \varepsilon$$

In Class Exercise:

- Repeat the above computation using different random seed, how do results differ?
- Which model has the lowest LOOCV error? Why?
- How does the above relate to the statistical significance of the coefficient estimates that result from each model fitting?