Machine Learning and Data Mining An Introduction with WEKA

AHPCRC Workshop - 8/18/11 - Dr. Martin

Based on slides by Gregory Piatetsky-Shapiro from Kdnuggets
http://www.kdnuggets.com/data_mining_course/

Terminology

- Components of the input:
  - Concepts: kinds of things that can be learned
    - Aim: intelligible and operational concept description
  - Instances: the individual, independent examples of a concept
    - Note: more complicated forms of input are possible
  - Attributes (Features): measuring aspects of an instance
    - We will focus on nominal and numeric ones

witten&eibe

What's a concept?

- Data Mining Tasks (Styles of learning):
  - Classification learning:
    predicting a discrete class
  - Association learning:
    detecting associations between features
  - Clustering:
    grouping similar instances into clusters
  - Numeric prediction:
    predicting a numeric quantity
- Concept: thing to be learned
- Concept description: output of learning scheme

witten&eibe

Classification learning

- Example problems: attrition prediction, using DNA data for diagnosis, weather data to predict play/not play
- Classification learning is supervised
  - Scheme is being provided with actual outcome
- Outcome is called the *class* of the example
- Success can be measured on fresh data for which class labels are known (test data)
- In practice success is often measured subjectively

Association learning

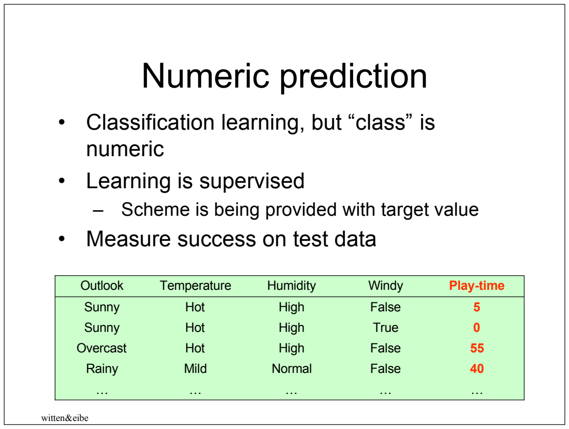
- Examples: supermarket basket analysis -what items are bought together (e.g. milk+cereal, chips+salsa)
- Can be applied if no class is specified and any kind of structure is considered "interesting"
- Difference with classification learning:
  - Can predict any attribute's value, not just the class, and more than one attribute's value at a time
  - Hence: far more association rules than classification rules
  - Thus: constraints are necessary
    - Minimum coverage and minimum accuracy

Clustering

- Examples: customer grouping
- Finding groups of items that are similar
- Clustering is *unsupervised*
  - The class of an example is not known
- Success often measured subjectively

	Sepal length	Sepal width	Petal length	Petal width	~~Type~~
1	5.1	3.5	1.4	0.2	~~Iris setosa~~
2	4.9	3.0	1.4	0.2	~~Iris setosa~~
…					
51	7.0	3.2	4.7	1.4	~~Iris versicolor~~
52	6.4	3.2	4.5	1.5	~~Iris versicolor~~
…					
101	6.3	3.3	6.0	2.5	~~Iris virginica~~
102	5.8	2.7	5.1	1.9	~~Iris virginica~~

witten&eibe

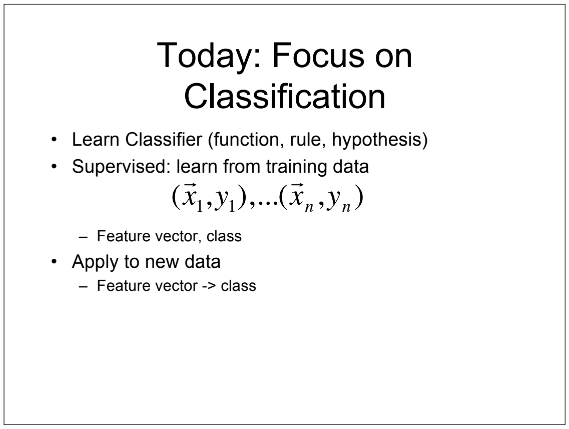


Numeric prediction

- Classification learning, but “class” is numeric
- Learning is supervised
  - Scheme is being provided with target value
- Measure success on test data

Outlook	Temperature	Humidity	Windy	Play-time
Sunny	Hot	High	False	5
Sunny	Hot	High	True	0
Overcast	Hot	High	False	55
Rainy	Mild	Normal	False	40
…	…	…	…	…

witten&eibe

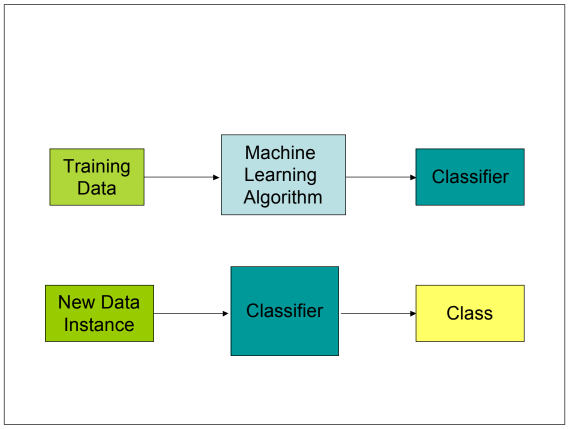


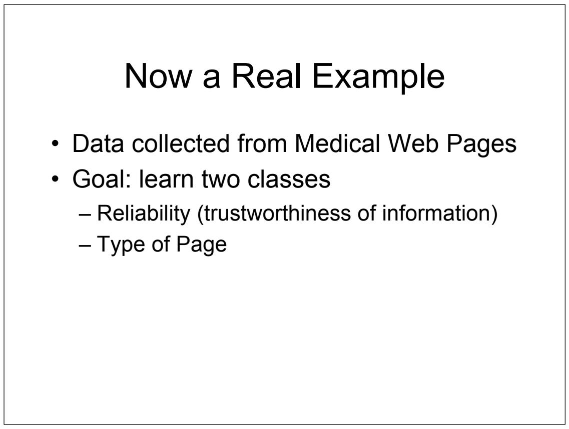
Today: Focus on Classification

- Learn Classifier (function, rule, hypothesis)
- Supervised: learn from training data

$$(\vec{x}_1, y_1), \ldots (\vec{x}_n, y_n)$$

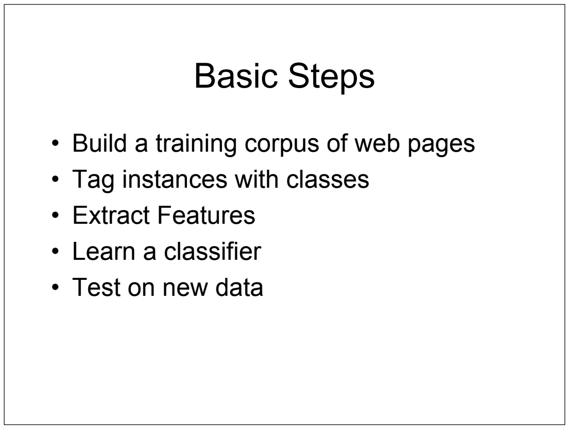
  - Feature vector, class
- Apply to new data
  - Feature vector -> class





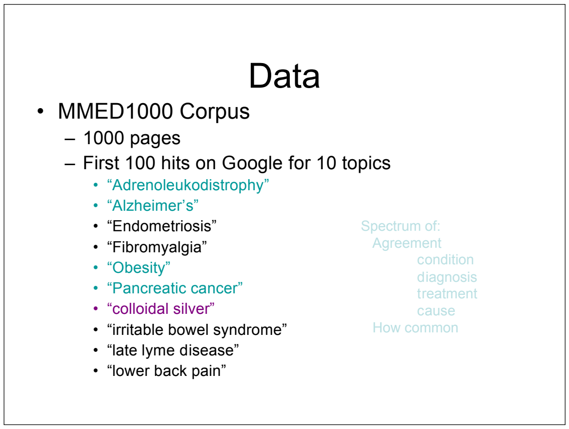
Now a Real Example

- Data collected from Medical Web Pages
- Goal: learn two classes
  - Reliability (trustworthiness of information)
  - Type of Page



Basic Steps

- Build a training corpus of web pages
- Tag instances with classes
- Extract Features
- Learn a classifier
- Test on new data



Data

- MMED1000 Corpus
  - 1000 pages
  - First 100 hits on Google for 10 topics
    - “Adrenoleukodystrophy”
    - “Alzheimer’s”
    - “Endometriosis”
    - “Fibromyalgia”
    - “Obesity”
    - “Pancreatic cancer”
    - “colloidal silver”
    - “irritable bowel syndrome”
    - “late lyme disease”
    - “lower back pain”

Spectrum of:
Agreement
condition
diagnosis
treatment
cause
How common

254 Features

- Link-Based
  - Inlinks, Outlinks, PageRank, Domain
    - Server host name, secure
- HTML markup
  - Symbols, metadata, JavaScript
  - Bold, italics, underline, font
    - Counts and frequencies

254 Features

- Text properties
  - LSA vector length, coherence, words per paragraph
  - Personal pronouns, punctuation, unique words
- Lists of words
  - Whole text, outlinks, anchor text
    - Criteria, medical, commercial, alternative
    - Disclaimer, diagnosis, shopping cart, miracle

Features Most Strongly Correlated with Reliable Pages

- “!” in anchor text - negatively
- Outlinks with same server host name and .uk
- Frequency of font changes - negatively
- “medicine” in anchor text
- Words in text
  - clinical
  - diagnosis
  - disease
  - medication
  - medicine
  - patient
  - symptom

Features Most Strongly Correlated with Unreliable Pages

- “!” in text
- Secure outlinks in the same server host name - negatively
- Number of font attributes
- Words in anchor text
  - price
  - products
  - testimonial
- Words in Text
  - doctor
  - I, me, my, them, us, we, you, your
  - miracle
  - natural
  - prevention
  - price
  - products
  - purchase
  - “to order”
  - testimonial
  - therapist

Algorithms

- Selecting features
  - Information Gain
  - Statistics
    - correlation, regression
- Classification
  - Naïve Bayes
  - Decision Tree (C4.5)
  - SVM
  - N-Closest
- Using SPSS and Weka

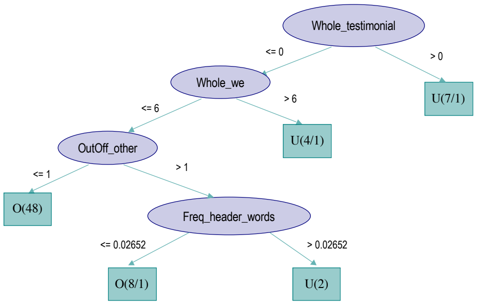
"Data Mining: Practical machine learning tools with Java implementations," by Ian H. Witten and Eibe Frank, Morgan Kaufmann, San Francisco, 2000.

http://www.cs.waikato.ac.nz/ml/weka/

Decision Trees

- Nodes are the features
- Maximize information gain
  - Expected reduction in entropy by partitioning examples according to given attribute $-\sum p_i \log_2 p_i$
  - Entropy =
    - 0 when all in same class
    - 1 when equal number in each class

Decision Tree: U-O

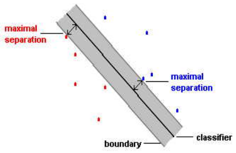


Naïve Bayes

- Want to calculate P(C|f)
  - C = class, f = set of features observed
  - Use Bayes Rule: P(C|f) = (P(C)p(f|C))/p(f)
  - Reduces to computing p(f|C)
    - Assume features conditionally independent given the class
    - p(f|C) = product(p(x_i |C))

Support Vector Machines

- Maximum Margin Hyperplane
  - Optimal separation between classes
- Kernel method
  - Nonlinear transformation between dot product spaces
    - Nonlinear space transformed to linear space
- http://www.ucl.ac.uk/oncology/MicroCore/HTML_resource/images/svm_1.jpg



Latent Semantic Analysis

- Provides measures of the semantic relatedness, quality, and quantity of information contained in discourse
- Implementation: Four Basic Steps
  - Term by document (context) matrix
  - Convert matrix entries to weights
  - Singular Value Decomposition (SVD) performed on matrix
  - Reduce Rank of matrix
    - all but the k highest singular values are set to 0
    - produces k-dimensional approximation of the original matrix (in least-squares sense)
    - this is the "semantic space"

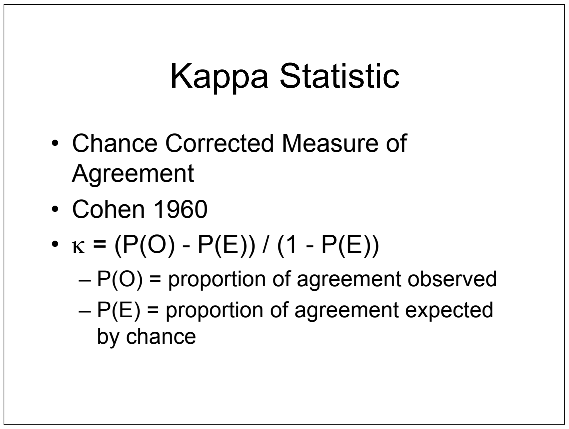
N-Closest

- Version of K-Nearest Neighbor algorithm
- Finds closest pages in LSA semantic space to page P
  - Checks their classes
  - Sums cosines for each class
  - Predicts Ps class based on largest sum

Standard Performance Measures

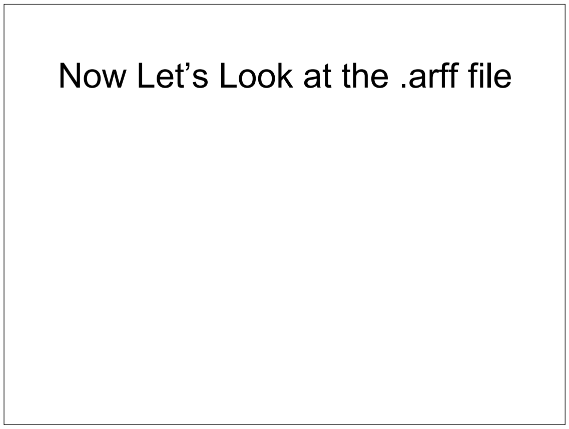
- Accuracy: percent correct
  - For classifier: (a+d)/(a+b+c+d)
- Precision: portion of selected items that the system got right
  - For class R: a/(a+b)
- Recall: portion of the target items that the system selected
  - For class R: a/(a+c)
- F-Measure: (2*precision*recall)/(precision+recall)

	R is correct	O is correct
R predicted	a	b
O predicted	c	d



Kappa Statistic

- Chance Corrected Measure of Agreement
- Cohen 1960
- $\kappa = (P(O) - P(E)) / (1 - P(E))$
  - P(O) = proportion of agreement observed
  - P(E) = proportion of agreement expected by chance



Now Let’s Look at the .arff file