Oct 12th

Classification

given a training set where data is labelled w/ special attribute called class (discrete)

goal to learn the rule

Techniques

- instance based
- decision trees
- Naive Bayes
- Support vector Machines
- Neural Network

Instance Based Classifiers

- use stored training records to predict class label of unseen case

ROTE-Learners → perform classification only if the attributes of unseen record exactly match a rec in train set.

Nearest-Neighbors → use K closest records to perform classification.

Requires: training set, distance func, value for K.

How to:

- compute dist of unseen record to all training rec
- Find K nearest
- assign based on majority rules.

Note: Aggregation method

- majority rule
- weighted majority based on dist. ($w = 1/d^2$)

Scaling issues

- SHOULD SCALE TO PREVENT ATTRIBUTE DOMINATION

Note:

Limits:
- if K too small → sensitive to noise points + overfitting
- if K too big → neighborhood may include points from other classes.

Pros:
- simple to understand why a given unseen record was given a particular class
- Adapts to new attributes.

Cons:
- expensive to classify points
- KNN can be problematic in high dimensions

Decision Tree

Algorithm (HUNTS)

- specify terminating conditions

Let D_t be set of training rec that reach node t

- → if D_t contains rec that belong to same class y_t then t is a leaf node labeled as y_t
- → if D_t is empty set, then t is a leaf node labelled by default class y_d
- → if D_t contains records that belong to more than one class, use an attrib test to split the data into smaller subsets

Recursively apply procedure to each subset

Splitting Based on Nominal Attributes

- multi way split → use as many partitions as distinct values
- Binary split → categorize to binary

Discretize form ordinal categorical attribute [static → once @ beginning | Dynamic → equal interval Bucketing]

Binary: $A < V$, $A > V$

How to determine best split.

Greedy Approach - nodes w/ homogenous class distrib are preferred

Node Impurity

- gini index

$$GINI(t) = 1 - \sum_j [p(j|t)]^2$$

$p(j|t)$ rel freq of class j @ node t.

maximum $(1 - \frac{1}{n_c})$ → recs equally distributed [least interesting info]

minimum → more interesting.

$$GINI_{split} = \sum_{i=1}^{k} \frac{n_i}{n} GINI(i)$$

n_i = # of recs at child i
n = # of recs at node p.

Underfitting - when model too simple, both training and test errors are large.

overfitting - perhaps use post-pruning

Confusion Matrix

TP	FN
FP	FN

Accuracy is misleading.

(so assign costs to confusion matrix).