Chapter 6
Classification and Prediction
(3)

Outline

- Classification and Prediction
- Decision Tree
- Naïve Bayes Classifier
- Support Vector Machines (SVM)
- K-nearest Neighbors
- Accuracy and Error Measures
- Feature Selection Methods
- Ensemble Methods
- Applications
- Summary

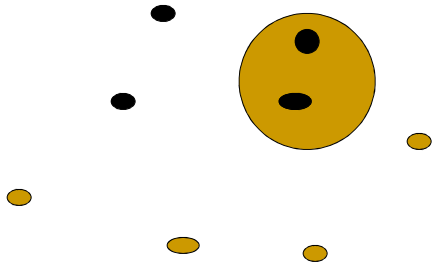
Different Learning Methods

- Eager classification Learning Method
  - Construct a classification model before test new labeled data.
    - Decision Tree, Bayesian Classification, SVM
- Instance-based Learning
  - Learning=storing all training instances
  - Classification=assigning class label to a new instance
  - Referred to as “Lazy” learning
    - K-nearest Neighbors, Case-based Reasoning

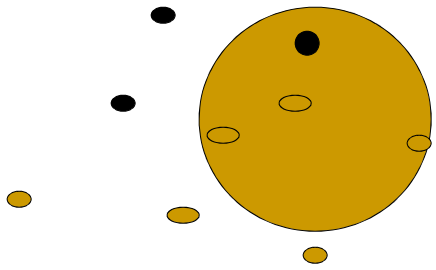
K-nearest Neighbor Algorithm

- Given a new instance x, represented by $(a_1(x), a_2(x), a_3(x),.., a_n(x))$, $a_i(x)$ denotes feature
- Find its nearest neighbor $<x', y'_k>$
- Return y'_k as the class of x

1-Nearest Neighbor



3-Nearest Neighbor



K-nearest Neighbor Algorithm

- Distance measures
  - Euclidean distance between two instances

$$dist(x_i, x_j) = \sqrt{\sum_{r=1}^{n} (a_r(x_i) - a_r(x_j))^2}$$

  - Features in different range: Normalization?!
  - Missing Data : maximum possible difference
  - Categorical data : two identical – 0, different – 1; Or some sophisticated schemes.

K-Nearest Neighbor

- Features
  - All instances correspond to points in an n-dimensional Euclidean space
  - Classification is delayed till a new instance arrives
  - Classification done by comparing feature vectors of the different points
  - Target function may be discrete or real-valued
  - Learning is very simple
  - Classification is time consuming
  - Good for incremental learning.

K-nearest Neighbor Algorithm

- Determine a good value for *K*?
- Dealing with noise?
- What's its time complexity?
- How to speed up?
  - Huge storage
  - Use representatives (a problem of instance selection) : Sampling
  - Parallel

Outline

- Classification and Prediction
- Decision Tree
- Naïve Bayes Classifier
- Support Vector Machines (SVM)
- K-nearest Neighbors
- Accuracy and Error Measures
- Feature Selection Methods
- Ensemble Methods
- Applications
- Summary

Classifier Accuracy Measures

- Accuracy of a classifier M, acc(M): percentage of test set tuples that are correctly classified by the model M
  - Error rate (misclassification rate) of M = 1 – acc(M)
  - Given *m* classes, $CM_{i,j}$, an entry in a **confusion matrix**, indicates # of tuples in class *i* that are labeled by the classifier as class *j*

Real class\Predicted class	C_1	$\neg C_1$
C_1	True positive	False negative
$\neg C_1$	False positive	True negative

Classifier Accuracy Measures

Real class \ Predicted class	C_1	$\neg C_1$	**Total**
C_1	True positive (t_pos)	False negative (f_neg)	pos
$\neg C_1$	False positive (f_pos)	True negative (t_neg)	neg

- sensitivity = t_pos/(t_pos+f_neg) = t_pos / pos
- specificity = t_neg/(t_neg+f_pos) = t_neg / neg
- precision = t_pos/(t_pos+f_pos)
- accuracy = sensitivity* pos/(pos+neg) + specificity * neg/(pos+neg)
  = (t_pos + t_neg) / (pos + neg)
- Error rate = (f_pos + f_neg) / (pos + neg)

Classifier Accuracy Measures

Real class\Predicted class	buy_computer = yes	buy_computer = no	total	Recognition (%)
buy_computer = yes	6954	46	7000	99.34 (sensitivity)
buy_computer = no	412	2588	3000	86.27 (specificity)
total	7366	2634	10000	95.42 (accuracy)

- sensitivity = 6954/7000 = 99.34%
- specificity = 2588/3000 = 86.27%
- precision = 6954/7366 = 94.41%
- Accuracy = (6954+2588)/10000 = 95.42%
- Error Rate = 1 – Accuracy = 4.58%

Evaluating the Accuracy of a Classifier

- Holdout method
  - Given data is randomly partitioned into two independent sets
    - Training set (e.g., 2/3) for model construction
    - Test set (e.g., 1/3) for accuracy estimation
  - Random sampling: a variation of holdout
    - Repeat holdout k times, accuracy = avg. of the accuracies obtained

14

Evaluating the Accuracy of a Classifier

- Cross-validation (*k*-fold, where k = 10 is most popular)
  - Randomly partition the data into *k mutually exclusive* subsets, each approximately equal size
  - At *i*-th iteration, use D_i as test set and others as training set
  - Accuracy is the overall correct classifications from k iterations divided by the database size.
  - Leave-one-out: k folds where k = # of tuples, for small sized data and only one sample is left out for test each time.
  - Stratified cross-validation: folds are stratified so that class dist. in each fold is approx. the same as that in the initial data
    - ***Stratified 10-fold cross-validation*** is recommended due to its relatively low bias and variance.

Outline

- Classification and Prediction
- Decision Tree
- Naïve Bayes Classifier
- Support Vector Machines (SVM)
- K-nearest Neighbors
- Accuracy and Error Measures
- Feature Selection Methods
- Ensemble Methods
- Applications
- Summary

Feature Selection Methods

- Feature selection attempts to select the minimally sized subset of features according to the following criteria:
  - The classification accuracy is improved or does not significantly decrease;
  - The resulting class distribution, given only the values for the selected features, is as close as possible to the original class distribution, given all features
- Feature Selection methods try to remove irrelevant and redundant features from the feature space.

Three Steps of Feature Selection Methods

- Feature Evaluation
  - Goodness metrics
    - Distance: separating classes
    - Information: entropy
    - Dependency: dependence on classes
- Ranking
  - Sorts remaining inputs and assigns ranks based on importance.
- Selection
  - Identifies the subset of features

Information Gain

- It measures the information obtained for class label prediction by knowing the presence and absence of a feature.
- Features with information gain less than a certain threshold are removed.

$$IG(f) = -\sum_{i=1}^{m} p(c_i)\log p(c_i)$$
$$+ p(f)\sum_{i=1}^{m} p(c_i|f)\log p(c_i|f)$$
$$+ p(\bar{f})\sum_{i=1}^{m} p(c_i|\bar{f})\log p(c_i|\bar{f})$$

χ2 Statistic

$$\chi^2 = \sum \frac{(Observed - Expected)^2}{Expected}$$

- The x^2 measures the lack of independence between feature and class label, using the two-way contingency table of a feature and a class.
- If feature and class are independent, x^2 value is zero.
- Features with less values are removed from the feature space.

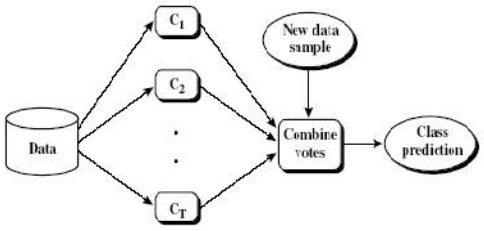
	c	$\neg c$
f	A	B
$\neg f$	C	D

$$\chi^2(f,c) = \frac{N(AD-CB)^2}{(A+C)(B+D)(A+B)(C+D)}$$

Outline

- Classification and Prediction
- Decision Tree
- Naïve Bayes Classifier
- Support Vector Machines (SVM)
- K-nearest Neighbors
- Accuracy and Error Measures
- Feature Selection Methods
- Ensemble Methods
- Applications
- Summary

Ensemble Methods: Increasing the Accuracy



- Ensemble methods
  - Use a combination of models to increase accuracy
  - Combine a series of k learned models, M_1, M_2, …, M_k, with the aim of creating an improved model M*
- Popular ensemble methods
  - Bagging: averaging the prediction over a collection of classifiers or majority vote
  - Boosting: weighted vote with a collection of classifiers

Bagging: Bootstrap Aggregation

- Analogy: Diagnosis based on multiple doctors' majority vote
- Training
  - Given a set D of *d* tuples, at each iteration *i*, a training set D_i of *d* tuples is sampled with replacement from D (i.e., bootstrap)
  - A classifier model M_i is learned for each training set D_i
- Classification: classify an unknown sample **X**
  - Each classifier M_i returns its class prediction
  - The bagged classifier M* counts the votes and assigns the class with the most votes to **X**
- Accuracy
  - Often significant better than a single classifier derived from D
  - For noise data: not considerably worse, more robust

23

Boosting

- Analogy: Consult several doctors, based on a combination of weighted diagnoses—weight assigned based on the previous diagnosis accuracy
- How boosting works?
  - Weights are assigned to each training tuple.
  - A series of k classifiers is iteratively learned.
  - After a classifier M_i is learned, the weights are updated to allow the subsequent classifier, M_{i+1}, to *pay more attention* to the training tuples that were misclassified by M_i
  - The final M* combines the votes of each individual classifier, where the weight of each classifier's vote is a function of its accuracy
- Comparing with bagging: boosting tends to achieve greater accuracy, but it also risks overfitting the model to misclassified data

24

Outline

- Classification and Prediction
- Decision Tree
- Naïve Bayes Classifier
- Support Vector Machines (SVM)
- K-nearest Neighbors
- Accuracy and Error Measures
- Ensemble Methods
- Applications
- Summary

Spelling Correction

- Google use naïve bayes classifiers to correct misspellings that users type in. Suggestions are based on information not only on the frequencies of similar spelled word typed by millions of other users, but also on the other words in your phrase.

Predicting Fraudulent Reporting

- Fraudulent Financial Reporting
  - Auditing firm needs to detect whether a company submitted a fraudulent financial report.
  - Based on the company's history data that whether the company has legal charges.

Outline

- Classification and Prediction
- Decision Tree
- Naïve Bayes Classifier
- Support Vector Machines (SVM)
- K-nearest Neighbors
- Accuracy and Error Measures
- Ensemble Methods
- Applications
- Summary

Summary (I)

- Classification/Prediction is a form of data analysis that can be used to extract models describing important data classes or to predict future data class labels.
- Effective and scalable methods have been developed for decision trees induction, Naive Bayesian classification, Support Vector Machine (SVM), nearest neighbor classifiers, and etc.
- Stratified k-fold cross-validation is a recommended method for accuracy estimation. Bagging and boosting can be used to increase overall accuracy by learning and combining a series of individual models.

29

Summary (II)

- There have been numerous comparisons of the different classification and prediction methods, and the matter remains a research topic
- No single method has been found to be superior over all others for all data sets
- Issues such as accuracy, training time, robustness, interpretability, and scalability must be considered and can involve trade-offs, further complicating the quest for an overall superior method

30