Chapter 6. Classification and Prediction

- What is classification? What is prediction?
- Issues regarding classification and prediction
- Classification by decision tree induction
- Bayesian classification
- Rule-based classification
- Classification by back propagation
- Support Vector Machines (SVM)
- Associative classification
- Lazy learners (or learning from your neighbors)
- Other classification methods
- Prediction
- Accuracy and error measures
- Ensemble methods
- Model selection
- Summary

Lazy vs. Eager Learning

- Lazy vs. eager learning
  - Eager learning (the previously discussed methods)
    - Given a set of training set, constructs a classification model *before receiving a* new test tuple to classify
  - Lazy learning
    - Simply stores training data (or only minor processing) and *just waits* until a test tuple is given
- Lazy: much less time in training but more time in predicting

Lazy vs. Eager Learning

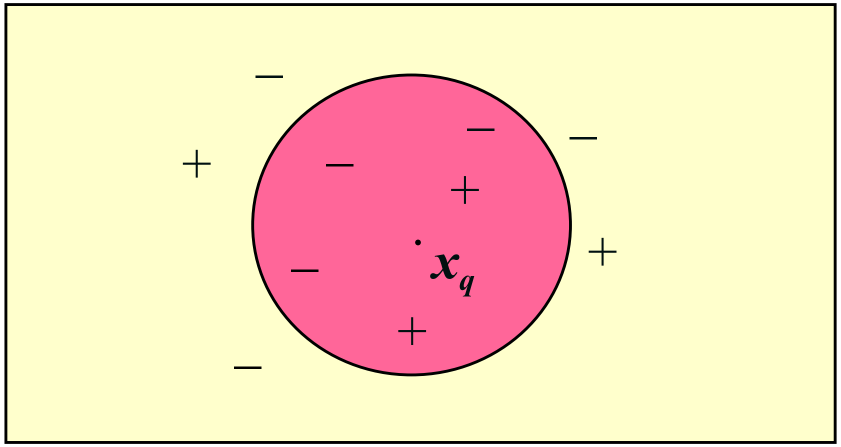
- Accuracy
  - A eager method must commit to a *single hypothesis* that covers the entire instance space
  - A lazy method effectively uses a richer hypothesis space since it uses many local functions

Lazy Learner: Instance-Based Methods

- Instance-based learning:
  - Store training examples and delay the processing (i.e., lazy evaluation)
    - Until a new instance is received to be classified
- Typical example: *k*-nearest neighbor approach

The *k*-Nearest Neighbor Algorithm

- All instances correspond to points in n-D space
  - A distance, dist($\mathbf{X_1}$, $\mathbf{X_2}$), is defined over the space
- k-nearest neighbors are retrieved in terms of the distance
- The test sample is classified by the class label of a majority of the k-nearest neighbors



Chapter 6. Classification and Prediction

- What is classification? What is prediction?
- Issues regarding classification and prediction
- Classification by decision tree induction
- Bayesian classification
- Rule-based classification
- Classification by back propagation
- Support Vector Machines (SVM)
- Associative classification
- Lazy learners (or learning from your neighbors)
- Other classification methods
- Prediction ←
- Accuracy and error measures
- Ensemble methods
- Model selection
- Summary

What Is Prediction?

- (Numerical) prediction is similar to classification
  - construct a model
  - use the model to predict a continuous value for a given input
- Prediction is different from classification
  - Classification is to predict a categorical class label
  - Prediction is to predict a value in a continuous space by using a modeled continuous-valued function

What Is Prediction?

- Major methods for prediction: regression
  - Models the relationship between one or more **predictor** variables (i.e., independent variables) and a **response** variable (i.e., dependent variable)
- Regression analysis
  - Linear and multiple regression
  - Non-linear regression
  - Other regression methods
    - Generalized linear model, Poisson regression, log-linear models, regression trees

Linear Regression

- **Linear regression**:
  - Involves a response variable y and a single predictor variable x (linear function)

    $y = w_0 + w_1 x$

    where w_0 (y-intercept) and w_1 (slope) are regression coefficients
  - Training

    To estimates the best-fitting straight line

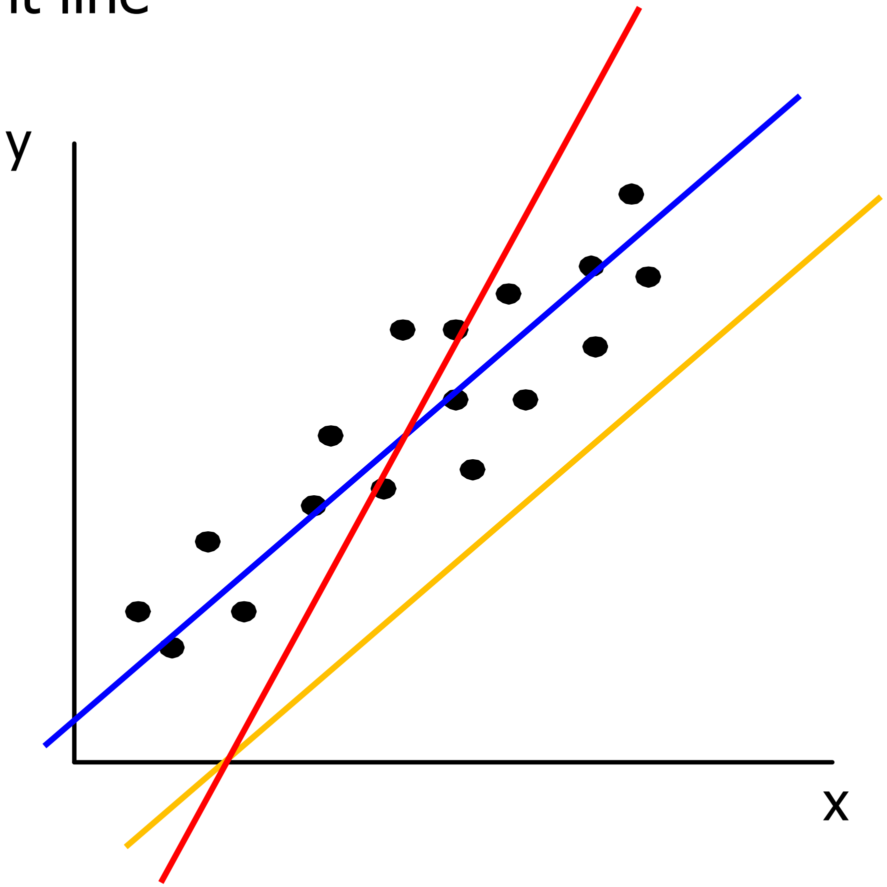
    $$y = w_0 + w_1 x$$

    w_0 and w_1 need to be estimated by using the training data

Linear Regression

- Least square method
  - To estimates the best-fitting straight line

$$y = w_0 + w_1 x$$



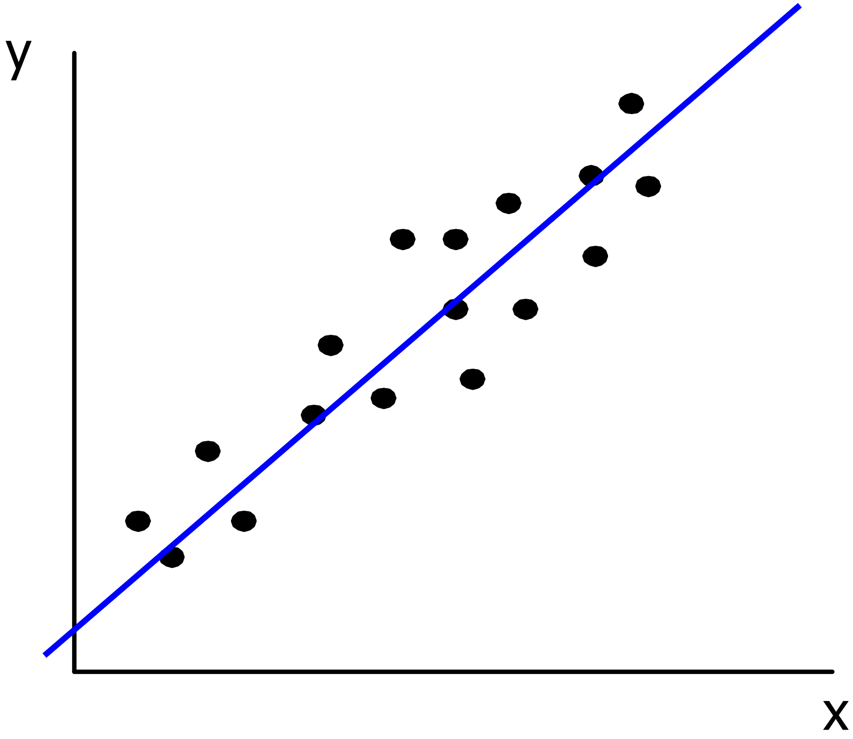
Linear Regression

- Least square method
  - To estimates the best-fitting straight line

$$y = w_0 + w_1 x$$

$$w_1 = \frac{\sum_{i=1}^{|D|} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{|D|} (x_i - \bar{x})^2}$$

$$w_0 = \bar{y} - w_1 \bar{x}$$



Linear Regression

- **Multiple linear regression**
  - Involves more than one predictor variable
    - X = <x_1, x_2, x_3, …, x_n>: predictor variables
    - y: response variable
  - Training data is composed of the form ($\mathbf{X_1}$, y_1), ($\mathbf{X_2}$, y_2),…, ($\mathbf{X_{|D|}}$, $y_{|D|}$)
  - Ex. For 2-D data, we may have: $y = w_0 + w_1 x_1 + w_2 x_2$
  - Solutions
    - By extension of the least square method
    - By using tools such as SAS and S-Plus

Nonlinear Regression

- Some nonlinear models
  - Can be formulated by a polynomial function
  - Ex: $y = w_0 + w_1 x + w_2 x^2 + w_3 x^3$
- A polynomial regression model can be transformed into linear regression model
- For example, the above formula is convertible to a linear one with new variables: $x_2 = x^2$, $x_3 = x^3$

$$y = w_0 + w_1 x + w_2 x_2 + w_3 x_3$$

Chapter 6. Classification and Prediction

- What is classification? What is prediction?
- Issues regarding classification and prediction
- Classification by decision tree induction
- Bayesian classification
- Rule-based classification
- Classification by back propagation
- Support Vector Machines (SVM)
- Associative classification
- Lazy learners (or learning from your neighbors)
- Other classification methods
- Prediction
- Accuracy and error measures
- Ensemble methods
- Model selection
- Summary

Classifier Accuracy Measures

Confusion matrix

$CM_{i,j}$, an entry, indicates # of tuples in class *i* that are labeled by the classifier as class *j*

		Classified	
		C_1	C_2
Ground truth	C_1	True positive	False negative
	C_2	False positive	True negative

classes	buy_computer = yes	buy_computer = no	total	recognition(%)
buy_computer = yes	6954	46	7000	99.34
buy_computer = no	412	2588	3000	86.27
total	7366	2634	10000	95.52

- Accuracy of a classifier M, acc(M): percentage of tuples (in a test set) that are correctly classified by the model M (=(6954+2588)/10,000)
  - Error rate (misclassification rate) of M = 1 – acc(M)

Classifier Accuracy Measures

		Classified	
		C_1	C_2
Ground truth	C_1	True positive	False negative
	C_2	False positive	True negative

classes	buy_computer = yes	buy_computer = no	total	recognition(%)
buy_computer = yes	6954	46	7000	99.34
buy_computer = no	412	2588	3000	86.27
total	7366	2634	10000	95.52

- Alternative accuracy measures (e.g., for cancer diagnosis)

  sensitivity = t-pos/pos (recall) /* true positive recognition rate */

  specificity = t-neg/neg /* true negative recognition rate */

  precision = t-pos/(t-pos + f-pos)

  accuracy = sensitivity * pos/(pos + neg) + specificity * neg/(pos + neg)

  = t-pos/(pos+neg) + t-neg/(pos+neg) = (t-pos+t-neg)/(pos+neg)

Predictor Error Measures

- Measure predictor accuracy
  - Measures how far off the predicted value is from the actual known value (i.e., ground truth)
- **Loss function**: measures the error between y_i and the predicted value y_i'
  - Absolute error: $| y_i – y_i'|$
  - Squared error: $(y_i – y_i')^2$

Predictor Error Measures

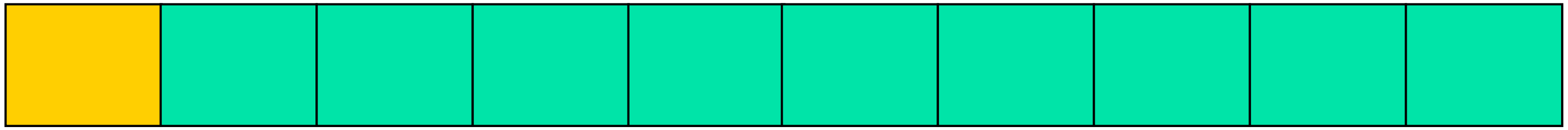
- Test error (generalization error): the average loss over the test set
  - Mean absolute error: $\frac{\sum_{i=1}^{d} | y_i - y_i' |}{d}$ Mean squared error: $\frac{\sum_{i=1}^{d} (y_i - y_i')^2}{d}$
  - Relative absolute error: $\frac{\sum_{i=1}^{d} | y_i - y_i' |}{\sum_{i=1}^{d} | y_i - \bar{y} |}$ Relative squared error: $\frac{\sum_{i=1}^{d} (y_i - y_i')^2}{\sum_{i=1}^{d} (y_i - \bar{y})^2}$
  - The mean squared-error exaggerates the presence of outliers
  - Square-root mean squared error, similarly, square-root relative squared error are popularly used
    - To get the same magnitude as the predicted quantity

Evaluating the Accuracy of a Classifier or Predictor

- 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 k accuracies obtained

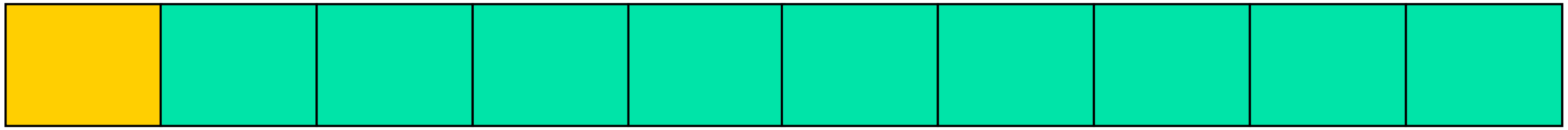
Evaluating the Accuracy of a Classifier or Predictor

- Cross-validation (*k*-fold, where k = 10 is most popular)
  - Randomly partition the data into *k mutually exclusive* subsets, each having approximately equal size
  - At *i*-th iteration, use D_i as a test set and others as a training set



Evaluating the Accuracy of a Classifier or Predictor

- Leave-one-out:
  - Special case of cross-validation, for *small* sized data
  - k folds where k = # of tuples
- *Stratified* cross-validation
  - Special case of cross-validation
  - Folds are stratified so that class distribution in each fold is approximately the same as that in the initial data



Evaluating the Accuracy of a Classifier or Predictor

- Bootstrap
  - Works well with a *small data set* (i.e., insufficient training samples)
  - Samples the given training tuples uniformly *with replacement*
    - i.e., each time a tuple is selected, but it is *re-added* to the training set for being equally likely to be *selected again*

Evaluating the Accuracy of a Classifier or Predictor

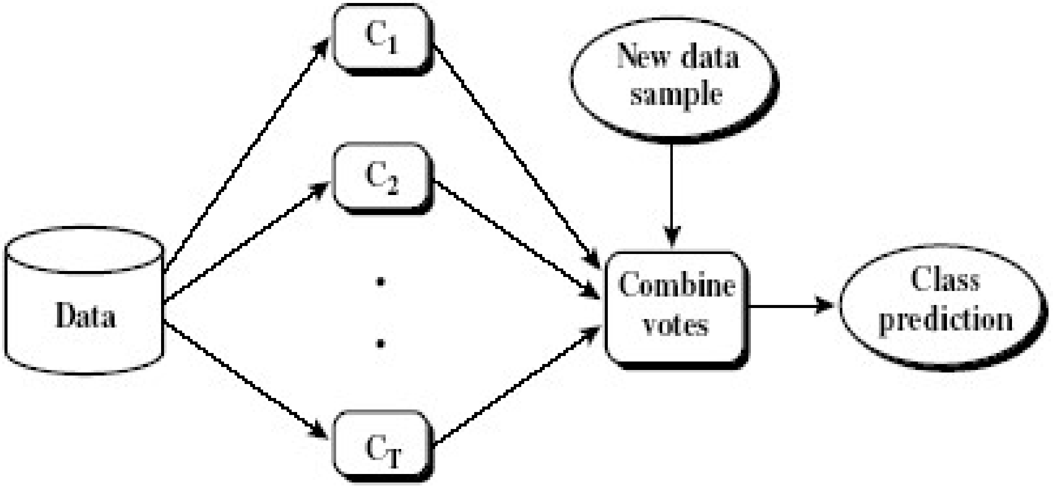
- .632 bootstrap: common one among bootstrap methods
  - Given a whole data set of *d* samples, it is sampled *d* times, with replacement, resulting in a **training set** of *d* samples
  - The data samples that did not make it into the training set end up forming the **test set**
  - About 63.2% of the original data will end up in the bootstrap (i.e., training set), and the remaining 36.8% will form the test set (since $(1 – 1/d)^d \approx e^{-1} = 0.368$)
  - Repeat the sampling procedure k times, overall accuracy of the model:

$$acc(M) = \sum_{i=1}^{k} (0.632 \times acc(M_i)_{test\_set} + 0.368 \times acc(M_i)_{train\_set})$$

Chapter 6. Classification and Prediction

- What is classification? What is prediction?
- Issues regarding classification and prediction
- Classification by decision tree induction
- Bayesian classification
- Rule-based classification
- Classification by back propagation
- Support Vector Machines (SVM)
- Associative classification
- Lazy learners (or learning from your neighbors)
- Other classification methods
- Prediction
- Accuracy and error measures
- Ensemble methods
- Model selection
- 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
  - Boosting: weighted vote with a collection of classifiers
  - Ensemble: combining a set of heterogeneous 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**

Bagging: Bootstrap Aggregation

- Prediction
  - Can be applied to the prediction of continuous values by taking the average value of each prediction for a given test tuple
- Accuracy
  - Often significant better than a single classifier derived from D
  - For noise data: not considerably worse, more robust
  - Improved accuracy in prediction

Boosting

- Analogy
  - Consult several doctors, based on a combination of weighted diagnoses — weight based on the previous diagnosis accuracy
- How boosting works?
  - Weights are assigned to each training tuple
    - Weights are used in sampling for building a training set (bootstrap)
  - 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 (in sampling) 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

Summary

- What is classification / prediction?
- Decision tree induction
- Bayesian classification
- Rule-based classification
- Associative classification
- Lazy learners (k-NN classifiers)
- Prediction (regression)
- Accuracy and error measures
- Ensemble methods