BITS Pilani

Pilani | Dubai | Goa | Hyderabad

S2-20_DSECFZC415
Classification and Prediction

- *The slides presented here are obtained from the authors of the books and from various other contributors. I hereby acknowledge all the contributors for their material and inputs.*
- *I have added and modified a few slides to suit the requirements of the course.*

Model Evaluation and Selection

Model Evaluation and Selection

Evaluation metrics: How can we measure accuracy? Other metrics to consider?

Use **validation test set** of class-labeled tuples instead of training set when assessing accuracy

Methods for estimating a classifier's accuracy:

- Holdout method, random subsampling
- Cross-validation
- Bootstrap

Comparing classifiers:

- Cost-benefit analysis and ROC Curves

Classifier Evaluation Metrics: Confusion Matrix

Confusion Matrix:

Given m classes, an entry, $CM_{i,j}$ in a **confusion matrix** indicates # of tuples in class i that were labeled by the classifier as class j

May have extra rows/columns to provide totals

Predicted class ->	C_1	$\neg C_1$
Actual class⇓		
C_1	**True Positives (TP)**	**False Negatives (FN)**
$\neg C_1$	**False Positives (FP)**	**True Negatives (TN)**

Classifier Evaluation Metrics: Confusion Matrix

Example of Confusion Matrix:

Predicted class ->	buy_computer = yes	buy_computer = no	Total
Actual class ⇓			
buy_computer = yes	**6954**	**46**	7000
buy_computer = no	**412**	**2588**	3000
Total	7366	2634	10000

Classifier Evaluation Metrics: Accuracy, Error Rate, Sensitivity and Specificity

Classifier Accuracy, or recognition rate: percentage of test set tuples that are correctly classified

Accuracy = (TP + TN)/All

Error rate: *1 – accuracy*, or

Error rate = (FP + FN)/All

A\P	C	¬C	
C	**TP**	**FN**	**P**
¬C	**FP**	**TN**	**N**
	P’	**N’**	**All**

- **Class Imbalance Problem**:
  - One class may be *rare*, e.g. fraud, or HIV-positive
  - Significant *majority of the negative class* and minority of the positive class
  - **Sensitivity**: True Positive recognition rate
    - **Sensitivity = TP/P**
  - **Specificity**: True Negative recognition rate
    - **Specificity = TN/N**

Classifier Evaluation Metrics: Precision and Recall, and F-measures

Precision: exactness – what % of tuples that the classifier labeled as positive are actually positive

$$precision = \frac{TP}{TP + FP}$$

Recall: completeness – what % of positive tuples did the classifier label as positive?

Perfect score is 1.0

$$recall = \frac{TP}{TP + FN}$$

Inverse relationship between precision & recall

***F* measure (*F_1* or *F*-score)**: harmonic mean of precision and recall,

$$F = \frac{2 \times precision \times recall}{precision + recall}$$

$F_ß$: weighted measure of precision and recall

– assigns ß times as much weight to recall as to precision

$$F_\beta = \frac{(1 + \beta^2) \times precision \times recall}{\beta^2 \times precision + recall}$$

Classifier Evaluation Metrics: Example

Precision = 90/230 = 39.13% *Recall* = 90/300 = 30.00%

Actual Class\Predicted class	cancer = yes	cancer = no	Total	Recognition(%)
cancer = yes	**90**	**210**	300	30.00 (*sensitivity*
cancer = no	**140**	**9560**	9700	98.56 (*specificity*)
Total	230	9770	10000	96.40 (*accuracy*)

Evaluating Classifier Accuracy: Holdout & Cross-Validation Methods

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

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

Evaluating Classifier Accuracy: Bootstrap

Bootstrap

- Works well with small data sets
- Samples the given training tuples uniformly *with replacement*
  - i.e., each time a tuple is selected, it is equally likely to be selected again and re-added to the training set

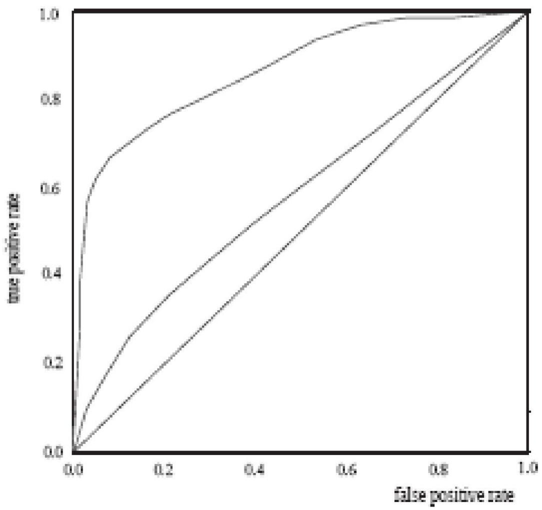
Several bootstrap methods, and a common one is **.632 boostrap**

- A data set with *d* tuples is sampled *d* times, with replacement, resulting in a training set of *d* samples. The data tuples that did not make it into the training set end up forming the test set. About 63.2% of the original data end up in the bootstrap, and the remaining 36.8% 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:

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

Model Selection: ROC Curves

- **ROC** (Receiver Operating Characteristics) curves: for visual comparison of classification models
- Originated from signal detection theory
- Shows the trade-off between the true positive rate and the false positive rate
- The area under the ROC curve is a measure of the accuracy of the model
- Rank the test tuples in decreasing order: the one that is most likely to belong to the positive class appears at the top of the list
- The closer to the diagonal line (i.e., the closer the area is to 0.5), the less accurate is the model



- Vertical axis represents the true positive rate
- Horizontal axis rep. the false positive rate
- The plot also shows a diagonal line
- A model with perfect accuracy will have an area of 1.0

Prediction

Prediction vs. Classification

- How is (Numerical) prediction similar to classification?
  - construct a model
  - use model to predict continuous or ordered value for a given input
- Difference between Prediction and classification
  - Classification refers to predict categorical class label
  - Prediction models continuous-valued functions
- Major method for prediction: regression
  - model the relationship between one or more independent or predictor variables and a dependent or response variable
- Profit, sales, mortgage rates, house values, square footage, temperature, or distance could all be predicted using regression techniques. For example, a regression model could be used to predict the value of a house based on location, number of rooms, lot size, and other factors.

Regression for Prediction

- A regression task begins with a data set in which the target values are known, e.g.
  - A regression model that predicts house values could be developed based on observed data for many houses over a period of time.
  - The data might track the age of the house, square footage, number of rooms, taxes, school district, proximity to shopping centers, and so on.
  - House value would be the target, the other attributes would be the predictors, and the data for each house would constitute a case.
- In the model build (training) process, a regression algorithm estimates the value of the target as a function of the predictors for each case in the build data.
  - These relationships between predictors and target are summarized in a model, which can then be applied to a different data set in which the target values are unknown

Prediction Techniques

- Regression analysis
  - Linear and multiple regression
  - Non-linear regression
  - Other regression methods:
    - Log-linear models,
    - Regression trees
    - etc.

Regression Analysis

- Regression analysis seeks to determine the values of parameters for a function that cause the function to best fit a set of data observations that you provide.
- The following equation expresses these relationships in symbols.

  $y = F(x,w) + e$

- Regression is the process of estimating the value of a continuous target (y) as a function (F) of one or more predictors (x1 , x2 , ..., xn), a set of parameters (w1 , w2 , ..., wn), and a measure of error (e).

Regression Analysis

- In the equation

  y = F(x,w) + e

- The predictors(x1 , x2 , ..., xn) can be understood as independent variables
- The target (y) is the dependent variable.
- The error (e), also called the residual, is the difference between the expected and predicted value of the dependent variable.
- The regression parameters are also known as regression coefficients.
- The process of training a regression model involves finding the parameter values that minimize a measure of the error, for example, the sum of squared errors.

Simple Linear Regression

- <u>Simple Linear regression</u>: involves a response variable y and a single predictor variable x

  $y = w_0 + w_1 x$

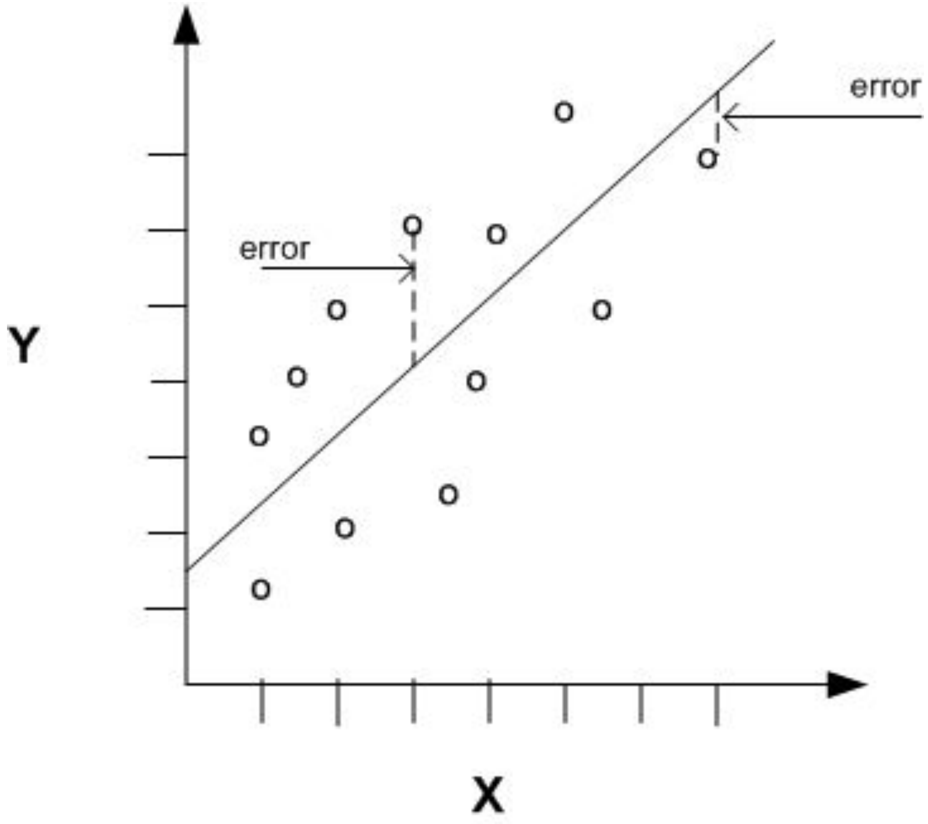
  where w_0 (y-intercept) and w_1 (slope) are regression coefficients

- <u>Method of least squares</u>: estimates the best-fitting straight line

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

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

Linear Regression With a Single Predictor



Multiple Linear Regression

<u>Multiple linear regression</u>: involves more than one predictor variable

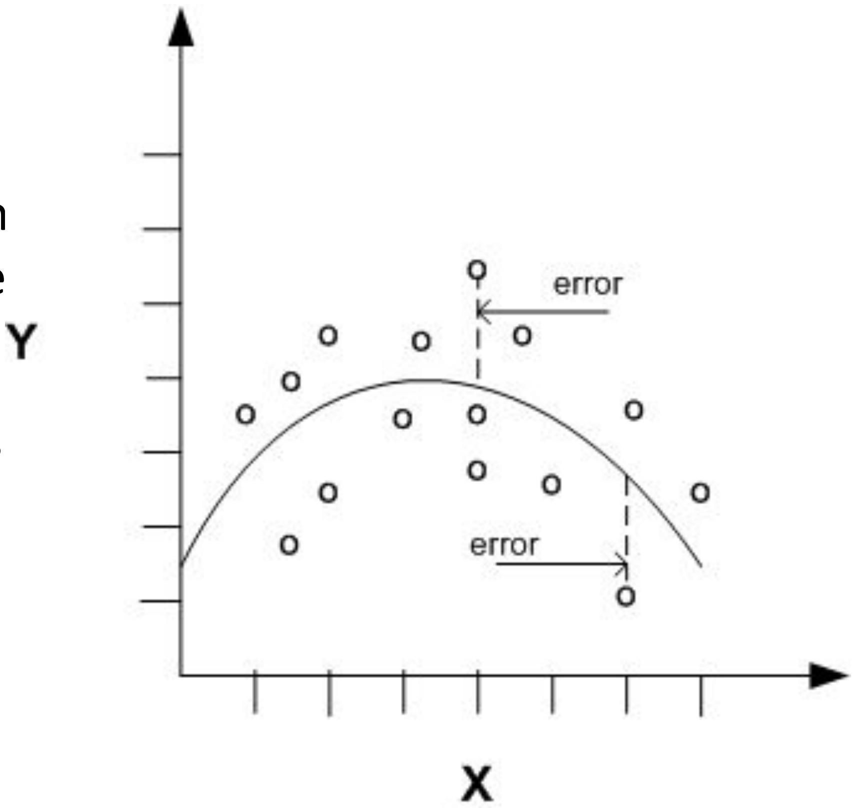
- Training data is of the form ($\mathbf{X_1}$, y_1), ($\mathbf{X_2}$, y_2),…, ($\mathbf{X_{|D|}}$, $y_{|D|}$)
- e.g. For 2-D data, we may have:

$$y = w_0 + w_1 x_1 + w_2 x_2$$

- Solvable by extension of least square method or using SAS, S-Plus
- Many nonlinear functions can be transformed into the above

Nonlinear Regression

- Often the relationship between x and y cannot be approximated with a straight line. In this case, a nonlinear regression technique may be used. Alternatively, th data could be preprocessed to make the relationship linear.
- Nonlinear regression models define y as function of x using an equation that is more complicated than the linear regression equation



Nonlinear Regression

- Some nonlinear models can be modeled by a polynomial function
- A polynomial regression model can be transformed into linear regression model. For example,
    - $y = w0 + w1\ x + w2\ x^2 + w3\ x^3$
  - convertible to linear with new variables: $x2 = x^2$, $x3 = x^3$
    - $y = w0 + w1\ x + w2\ x2 + w3\ x3$
- Other functions, such as power function, can also be transformed to linear model
- Some models are intractable nonlinear (e.g., sum of exponential terms)
  - possible to obtain least square estimates through extensive calculation on more complex formulae

Regression Trees and Model Trees

- Regression tree: proposed in CART system (Breiman et al. 1984)
  - CART: Classification And Regression Trees
  - Each leaf stores a continuous-valued prediction
  - It is the average value of the predicted attribute for the training tuples that reach the leaf
- Model tree: proposed by Quinlan (1992)
  - Each leaf holds a regression model—a multivariate linear equation for the predicted attribute
  - A more general case than regression tree
- Regression and model trees tend to be more accurate than linear regression when the data are not represented well by a simple linear model

Prescribed Text Books

	Author(s), Title, Edition, Publishing House
T1	Tan P. N., Steinbach M & Kumar V. “Introduction to Data Mining” Pearson Education
T2	Data Mining: Concepts and Techniques, Third Edition by Jiawei Han, Micheline Kamber and Jian Pei Morgan Kaufmann Publishers
R2	Principles of Data Mining, Second Edition by Max Bramer Springer © 2013
R1	Predictive Analytics and Data Mining: Concepts and Practice with RapidMiner by Vijay Kotu and Bala Deshpande Morgan Kaufmann Publishers

Simple Linear Regression Example

x	y
Area (in sq. m)	Rent (in 000s of Rupees)
172	42
150	35
181	46
174	40
194	50

Can we predict rent for a house of 160 sq. m. in the locality?