8.3 Prediction

- Suppose an employee needs to predict how much rise he will get in his salary after 5 years. In this case a model is constructed based on his previous salary values that predicts a continuous-valued function or ordered value.
- Prediction is generally about the future values or the unknown events and it models continuous-valued functions.
- Most commonly used methods for prediction is regression.

8.3.1 Structure of Regression Model

- Regression Model represents reality by using the system of equations.
- Regression model explains relationship between variables and also enables quantification of these relationships.
- It determines the strength of relationship between one dependent variable with the other independent variable using some statistical measure.
- Dependent variable is usually denoted by Y.
- The two basic types of regression :

1. Linear regression
2. Multiple regressions

- The general form of regression is :

Linear regression : $Y = m + nX + u$

Multiple regression : $Y = m + n_1X_1 + n_2X_2 + n_3X_3 + ... + n_tX_t + u$

Where :

Y = The dependent variable which we are trying to predict

X = The independent variable that we are using to predict variable Y

m = The intercept

n = The slope

u = The regression residual.

- In multiple regressions each variable is differentiated with subscripted numbers.
- Regression uses a group of random variables for prediction and finds a mathematical relationship between them. This relationship is depicted in the form of a straight line (linear regression) that approximates all the points in the best way.
- Regression may be used to determine for e.g. price of a commodity, interest rates, the price movement of an asset influenced by industries or sectors.

8.3.2 Linear Regression

Regression tries to find the mathematical relationship between variables, if it is a straight line then it is a linear model and if it gives a curved line then it is a non linear model.

Simple linear regression :

- The relationship between dependent and independent variable is described by straight line and it has only one independent variable

$$Y = \alpha + \beta X$$

- Two parameters, α and β specify the (Y-intercept and slope of the) line and are to be estimated by using the data at hand.
- The value of Y increases or decreases in a linear manner as the value of X changes accordingly.
- Draw a line relating to Y and X which is well fitted to given data set.

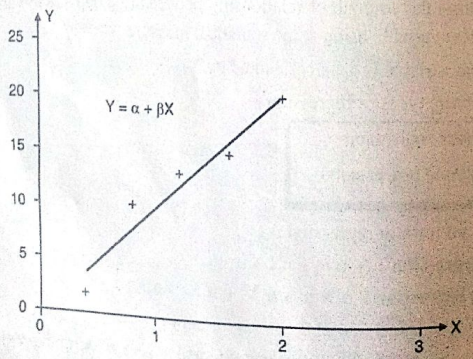

Fig. 8.3.1: Linear regression

- The ideal situation is that if the line which is well fitted for all the data points and no error for prediction.
- If there is random variation of data points, which are not fitted in a line then construct a probabilistic model related to X and Y.
- Simple linear regression model assumes that data points deviates about the line, as shown in the Fig. 8.4.1.

8.3.3 Multiple Linear Regression

- Multiple linear regression is an extension of simple linear regression analysis .
- It uses two or more independent variables to predict the outcome and a single continuous dependent variable

$$Y = a_0 + a_1 X_1 + a_2 X_2 + ... + a_k X_k + e$$

Where, Y is the dependent variable or response variable

$X_1, X_2 .....X_k$ are the independent variables or predictors

e is random error.

$a_0, a_1, a_2..... a_k$ are the regression coefficients

8.3.4 Other Regression Model

- In log linear regression a best fit between the data and a log linear model is found.
- Major assumption: A linear relationship exists between the log of the dependent and independent variables.
- Log linear models are models that postulate a linear relationship between the independent variables and the logarithm of the dependent variable, for example :

$$\log(y) = a_0 + a_1 x_1 + a_2 x_2 ... + a_N x_N$$

where y is the dependent variable; x_i, i=1,...,N are independent variables and $\{a_i$, i=0,...,N$\}$ are parameters (coefficients) of the model.

- For example, log linear models are widely used to analyze categorical data represented as a contingency table. In this case, the main reason to transform frequencies (counts) or probabilities to their log-values is that, provided the independent variables are not correlated with each other, the relationship between the new transformed dependent variable and the independent variables is a linear (additive) one.

8.4 Model Evaluation and Selection

Q. Justify why classification is said to be supervised learning. How is the classifier accuracy determined and also explain its various types. **MU - May 12**

- Validation test data is very useful to estimate the accuracy of model
- Various methods for estimating a classifier's accuracy are given below. All of them are based on randomly sampled partitions of data :
  - Holdout method
  - Random subsampling
  - Cross-validation
  - Bootstrap
- If we want to compare classifiers to select the best one then the following methods are used :
  - Confidence intervals
  - Cost-benefit analysis and Receiver Operating Characteristic (ROC) Curves

8.4.1 Accuracy and Error Measures

MU - May 16

Q. Explain major factors related to performance of DT based data mining techniques.

MU - Dec. 11

Accuracy of a classifier M, acc(M) is the percentage of test set tuples that are correctly classified by the model M.

Basic concepts :

1. **Partition the data randomly into three sets :** Training set, validation set and test set.
   - Training set is the subset of data used to train/build the model.
   - Test set is a set of instances that have not been used in the training process. The model's performance is evaluated on unseen data. Testing just estimates the probability of success on unknown data.
   - Validation data is used for parameter tuning but it cannot be the test data. Validation data can be the training data, or a subset of training data.
   - Generalization Error : Model error on the test data.
2. **Success :** Instance (record) class is predicted correctly.
3. **Error :** Instance class is predicted incorrectly.
4. **The confusion matrix :** It is a useful tool for analyzing how well your classifier can recognize tuples of different classes.
   - If we have only two way classification then only four classification outcomes are possible which are given below in the form of a confusion matrix :

		Predicted class		
	Class Label	C_1	C_2	**Total**
Actual class	C_1	True Positives (TP)	False Negatives (FN)	**P**
	C_2	False Positives (FP)	True Negatives (TN)	**N**
	Total	**P'**	**N'**	**All**

- **TP :** Class members which are classified as class members.
- **TN :** Class non-members which are classified as non-members.
- **FP :** Class non-members which are classified as class members.
- **FN :** Class members which are classified as class non-members.
- **P :** Number of positive tuples.
- **N :** The number of negative tuples.
- **P' :** The number of tuples that were labeled as positive.
- **N' :** The number of tuples that were labeled as negative
- **All :** Total number of tuple i.e. TP + FN + FP + TN or P + N or P' + N'

5. **Sensitivity :** True Positive recognition rate which is the proportion of positive tuples that are correctly identified

$$\text{Sensitivity} = TP/P$$

6. **Specificity :** True Negative recognition rate which is the proportion of negative tuples that are correctly identified

$$\text{Specificity} = TN/N$$

7. **Classifier accuracy or recognition rate :** Percentage of test set tuples that are correctly classified

$$\text{Accuracy} = (TP + TN)/\text{All}$$

OR

$$\text{Accuracy} = \frac{TP + TN}{P + N}$$

Accuracy is also a function of sensitivity and specificity:

$$\text{Accuracy} = \text{Sensitivity}\frac{P}{(P+N)} + \text{Specificity}\frac{N}{(P+N)}$$

8. **Error rate :** A percentage of errors made over the whole set of instances (records) used for testing.

$$\text{Error rate} = 1 - \text{accuracy, or Error rate} = (FP + FN)/\text{All}$$

Or

$$\text{Error rate} = \frac{FP + FN}{P + N}$$

9. **Precision :** Percentage of tuples which are correctly classified as positive are actual positive. It is a measure of exactness.

$$\text{Precision} = \frac{|TP|}{|TP| + |FP|}$$

10. **Recall :** Percentage of positive tuples which the classifier labelled as positive. It is a measure of completeness.

$$\text{Recall} = \frac{|TP|}{|TP| + |FN|}$$

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

$$F = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

12. **F_β :** Weighted measure of precision and recall and assigns ß times as much weight to recall as to precision

$$F_\beta = \frac{(1 + \beta^2) \times \text{Precision} \times \text{Recall}}{\beta^2 \times \text{Precision} + \text{Recall}}$$

where β is a non-negative real number.

13. **Classifiers can also be compared with respect to :**
    - Speed
    - Robustness
    - Scalability
    - Interpretability

14. **Re-substitution error rate :**
    - Re-substitution error rate is a performance measure and is equivalent to training data error rate.
    - It is difficult to get 0% error rate but it can be minimized, so low error rate is always preferable.

8.4.2 Holdout

- In holdout method, data is divided into training data set and testing data set (usually 1/3 for testing, 2/3 for training).

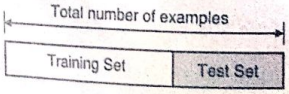

Fig. 8.4.1

- To train the classifier , training data set is used and once the classifier is constructed then use test data set to estimate the error rate of the classifier
- If the training is more than better model is constructed and if the test data is more than more accurate the error estimates.
- **Problem :** The samples might not be representative. For example, some classes might be represented with very few instances or even with no instances at all.
- **Solution :** stratification is the method which ensures that both training and testing data have equal number of samples of same class.

8.4.3 Random Subsampling

- It is a variation of the holdout method.
- The holdout method is repeated k times.
- Each split randomly selects a fixed number example without replacement.

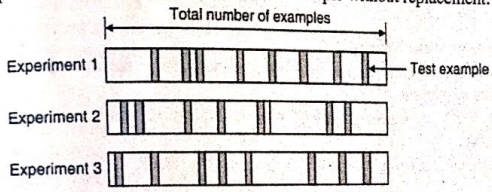

Fig. 8.4.2

- For each data split we retrain the classifier from scratch with the training examples and estimate E_i with the test examples.
- The overall accuracy is calculated by taking the average of the accuracies obtained from each iteration.

$$E = \frac{1}{K} \sum_{i=1}^{K} E_i$$

8.4.4 Cross-Validation (CV)

- Avoids overlapping test sets.
- **k-fold cross-validation :**
    - **First step :** Data is split into k subsets of equal size (usually by random sampling).

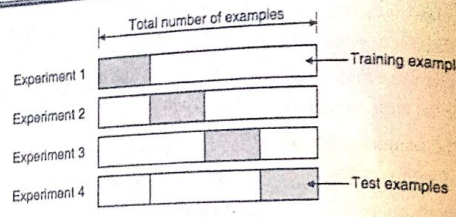

Fig. 8.4.3

- **Second step :** Each subset in turn is used for testing and the remainder for training.

- The advantage is that all the examples are used for both training and testing.
- The error estimates are averaged to yield an overall error estimate.

$$E = \frac{1}{K} \sum_{i=1}^{K} E_i$$

- **Leave-one-out cross validation :**
  - If dataset has N examples, then N experiments to be performed for Leave-one-out cross validation.
  - For every experiment, training uses N-1 examples and remaining example for testing.
- The average error rate on test examples gives the true error.

$$E = \frac{1}{N} \sum_{i=1}^{N} E_i$$

- **Stratified cross-validation:** Subsets are stratified before the cross-validation is performed.
- **Stratified ten-fold cross-validation :**
  - This gives accurate estimate of evaluation.
  - The estimate's variance get reduced due to stratification.
  - Ten-fold cross-validation is repeated ten times and finally the results are averaged based on the previous 10 results.

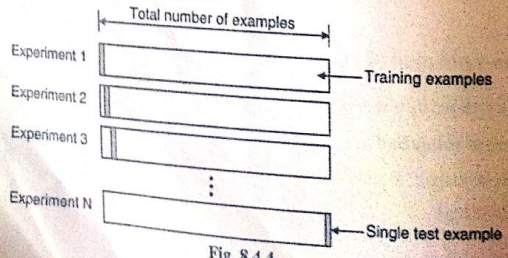

Fig. 8.4.4

8.4.5 Bootstrapping

- CV uses sampling of data set without replacement. Once the tuple or instance is selected, it cannot be selected again for training or test data.
- The bootstrap uses sampling with replacement to get the training set.
- **Training set :** A dataset of k instances is sampled with replacement k times to form the training set of k instances.
- **Test set :** This is separate dataset from the original dataset which is not the part of training dataset.
- Bootstrapping is the best error estimator for small datasets.

8.4.6 Comparing Classifier Performance using ROC Curves

- To compare two classification models, Receiver Operating Characteristic (ROC) curves are a useful visual tool. It shows the trade-off between the true positive rate and the false positive rate.
- The accuracy of the model is measured by the area under the ROC curve.
- Tuples which most likely belong to positive class should appear at the top of the list, so accordingly rank all the test tuples in decreasing order.
- The closer to the diagonal line (i.e., the closer the area is to 0.5), the less accurate is the model.
- ROC curve plots sensitivity (true positive rate) vs. 1-specificity (false positive rate).
- Always goes from (0,0) to (1,1). The more area in the upper left, the model is better.
- Random model is on the diagonal.
- "Area Under the Curve" (AUC) is a common measure of predictive performance.

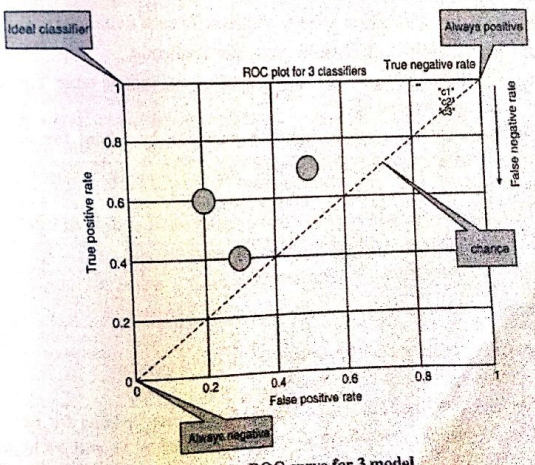

Fig. 8.4.5 : ROC curve for 3 model

8.5 Combining Classifiers

- Combining classifier is an ensemble methods which increases the Accuracy
- To get new improved model M* , combine a series of n learned models, $M_1, M_2, \ldots, M_n$
- Popular ensemble methods are :
  1. Bagging
  2. Boosting

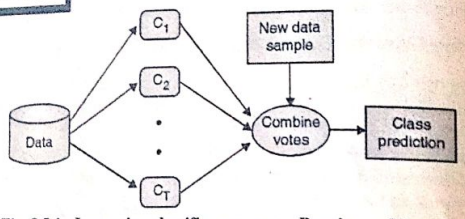

Fig. 8.5.1 : Increasing classifier accuracy : Bagging and boosting

8.5.1 Bagging or Bootstrap Aggregation

- It averages the prediction from the collection of various classifiers used.
- A bootstrap method is used, for data set D of n tuples, for each iteration n tuples are sampled with replacement from D
- A classifier model M from each iteration is learned for each training set.
- For unknown sample Y, each classifier gives class prediction.
- The bagged classifier M* uses the voting method i.e. the sample tuple Y is assigned the class with the most votes to tuple Y.
- For continuous values, it can be used for prediction by taking the average of all predictions for given sample.
- Accuracy :
  - Uses voting method so it is better than a single classifier derived from D.
  - For noise data not considerably worse, more robust.
  - Proved improved accuracy in prediction.

8.5.2 Boosting

- In boosting, each training tuple has weight
- n number of classifiers are learned iteratively.
- After learning of M_i classifier, every time the weights are updated for next classifier learning i.e. M_{i+1}. So if the tuples which were misclassified by M_i will get higher weight for next classifier.

- Use voting method, check the votes of each classifier to get the final M* which helps to get the accuracy.
- The extended boosting algorithm works for the prediction of continuous values.
- Boosting tends to accomplish greater accuracy as compared to bagging, there is a risk of overfitting the model.

8.5.3 Random Forest

- Each classifier in the group is a **decision tree** classifier.
- Decide the split based on the random selection of attributes at each node.
- Voting method is used during classification, each tree votes and most common class is selected.
- Two Methods to construct Random Forest :
  1. **Forest-RI (random input selection) :**
     - It randomly selects N attributes for the split at the node.
     - It uses the CART methodology which allows the tree to grow to maximum size.
  2. **Forest-RC (random linear combinations) :**
     - Based on the existing attributes, it creates new attributes.
     - It reduces the correlation between individual classifiers.
- It is faster than bagging or boosting.

8.6 University Questions and Answers

May 2010

Q. 1 What is Classification ? What are the issues in classification ? Apply statistical based algorithm to obtain the actual probabilities of each event to classify the new tuple as a tall. Use the following data : *(Ans. : Refer Sections 8.1, 8.1.5 and Ex. 8.2.6)* **(10 Marks)**

Person ID	Name	Gender	Height	Class
1	Kristina	Female	1.6 m	Short
2	Jim	Male	2 m	Tall
3	Maggi	Female	1.9 m	Medium
4	Marya	Female	2.1 m	Tall
5	Stephanie	Female	1.7 m	Short
6	Bob	Male	1.85 m	Medium
7	Catherine	Female	1.6 m	Short
8	Dave	Male	1.7 m	Short
9	Wilson	Male	2.2 m	Tall