**Lecture 2: White Board Notes & Class discussion**

**Bayes Classifier (Probabilistic classifier):**

* Follows Bayes rule
* E = Instance/observation/record/case
* H = Class
* For example, let us assume that there are
  + Three classes: Tree, Building, Car
  + Features: Shape(X1), Color(X2), Texture(X3)
* P(Class = Tree|E) = P(Class = Tree|X1) P(Class = Tree|X2) P(Class = Tree|X3)
* P(Class = Building|E) = P(Class = Building|X1) P(Class = Building|X2) P(Class = Building|X3)
* P(Class = Car|E) = P(Class = Car|X1) P(Class = Car|X2) P(Class = Car|X3)

Naïve Bayes discussion

* Zero Frequency: If an attribute value doesn’t occur with every class value, then the probability will be zero.
* To solve this add 1 to the count of every attribute value
* Missing Values: Ignore missing values in calculation
* Numeric attributes:
  + Assumption: X1 = N(µ1, σ1), X2 = N(µ2, σ2), X3 = N(µ3, σ3)
  + P(Class = Tree|X1) =

**Decision Tree classifiers**

* Top down approach
* Splits data based on attribute values
* Example: 14 cases – 9 cases from class 1, 4 cases from class 2 with 4 attributes.
* Criterion for choosing the best attribute to split: Information Gain

Information gain

Decision trees with probabilities

< 3

< 9

Three rules in total: Sample rule and Class calculation:

IF F1 <3, F5 <9 THEN C1 = 10/13

**K-Nearest Neighbors:**

* Distance measure used for numerical data: Euclidean, Minkowski, Mahalanobis distance, etc
* Distance measure used for categorical data: Jaccard’s, Dice Coefficient
* Euclidean distance:
* Euclidian distance measure with weights for attributes:

**Clustering**

* Finding K for K-means:
* Find Scatter within and Scatter between for different values of K.
* Scatter ratio = Scatter within/Scatter between
* The K value after which the scatter ratio does not decrease much can be chosen.

**Evaluation of classifiers**

Terminology

* Resubstitution error = Training error
* Generalization error = Testing error

To evaluate classifier, we have to split the data into training and Testing and calculate Accuracy, Error rate

How to split data into training and testing

* Holdout partitioning (66% training, 34% testing)
* Stratified sampling (Selecting same percentage of cases from each class for training the classifier)
* K- fold Cross validation (N=2000, K = 10)

N= 10 (200)

N= 3 (200)

N= 9 (200)

N= 2 (200)

N= 1 (200)

Training Set Testing Set

* Leave one out (1 case used for test set)

Choose reliable Estimates

We can build reliable estimates by calculating the confidence interval for our estimates:

* Confidence interval =