IS422P - DATA MINING CLASSIFICATION

(PART II)

AMIRA REZK

INFORMATION SYSTEM DEPARTMENT

AGENDA

The Basics
- What is Classification?
- General Approach

Decision Tree Induction
- The Algorithm
- Attribute Selection Measures
- Tree Pruning
- Extracting Rules from Decision Trees

Bayes Classification
- Bayes' Theorem
- Naïve Bayesian Classification

Lazy Learners
- K-Nearest Neighbor Classifiers

Regression analysis
- Linear regression

Model Evaluation
- Metrics for Evaluating Classifiers Performance
- Cross-Validation
- Bootstrap

Improving Classification Accuracy
- Bagging
- Boost and AdaBoost

BAYES CLASSIFICATION METHODS

- **Naïve Bayesian classifier** → Statistical classifier that predicts the probability that a tuple belongs to a specific class
  - Based on Bayes Theorem → Bayes was an 18th century clergyman who worked on probability
  - *High accuracy*
  - *Speed*
  - *Class-conditional Independence* → Attributes' effect on class determination is **independent**

BAYES CLASSIFICATION METHODS
NAÏVE BAYESIAN CLASSIFICATION

- Given tuples with n attribute and m classes, Naïve Bayes predicts that X belong to class with highest posteriori probability
- $P(C_i\,|X) > P(C_j|X) for\ 1 \leq j\ \leq m\ , j\ \neq i$
- C_i is called the maximum posteriori hypothesis
- $P(Ci|X) = \frac{P(X|C_i)P(Ci)}{P(X)}$
- Since P(X) is constant, maximize only numerator
- If $P(C_i)$ is unknown for all i, assume uniform probability
  - Then you only have to maximize $P(X|C_i)$
  - Otherwise, $P(C_i) = \frac{|Ci_D|}{|D|}$

BAYES CLASSIFICATION METHODS
NAÏVE BAYESIAN CLASSIFICATION

- To reduce computation of $P(X|C_i)$, attributes are assumed to be independent → hence the "naïve" in the name
- $P(X|C_i) = \prod_{k=1}^{n} P(x_k|C_i) = P(x_1|C_i) \times P(x_2|C_i) \times \ldots \times P(xn|Ci)$
- If attribute is **categorical** → $P(x_k|C_i) = \frac{|Ci_{D\,xk}|}{|Ci'_{D}|}$
- If attribute is **numerical** → assume gaussian distribution →

$$P(x_k|C_i) = \frac{1}{\sqrt{2\pi\sigma ci}}\, e^{-\frac{(x_k-\mu ci)^2}{2\sigma ci^2}}$$

- Evaluate for each C_i, assign class label of class with max $P(X|C_i)$

BAYES CLASSIFICATION METHODS
NAÏVE BAYESIAN CLASSIFICATION - EXAMPLE

department	age	salary	status	count
sales	Middle aged	medium	senior	30
sales	youth	low	junior	30
sales	Middle aged	low	junior	40
systems	youth	medium	junior	20
systems	Middle aged	high	senior	20
systems	senior	high	senior	10
marketing	senior	medium	senior	10
marketing	Middle aged	medium	junior	20
secretary	senior	medium	senior	10
secretary	youth	low	junior	10

The individual has status Junior

C1 (Senior) = 80 , C2(Junior) =120
P(C1)=80/200 **P(C2)=120/200**

X: {department = "marketing"; age= "youth", salary= low}

$P(X|C1) = \frac{10}{80} \times \frac{0}{80} \times \frac{0}{80} = 0$ → P(X}C1)P(C1)=0

$P(X|C2) = \frac{20}{120} \times \frac{60}{120} \times \frac{80}{120} = 0.055$ →
P(X|C2)P(C2)=0.033

Laplacian correction or
Laplace estimator

$P(X|C1) = \frac{10}{80} \times \frac{1}{80} \times \frac{1}{80} = 0.00002$ →
P(X}C1)P(C1)=0.000008

$P(X|C2) = \frac{20}{120} \times \frac{60}{120} \times \frac{80}{120} = 0.055$ →
P(X|C2)P(C2)=0.033

LAZY LEARNERS
K-NEAREST NEIGHBOR CLASSIFIERS

- Delay classification until new test data is available
  - Store training data meanwhile
- Use similarity measure to compute distance between test data tuple and each of the training data tuples (Euclidian, Manhattan, …)
  - Remember to normalize if ranges vary between attributes
- *k* stands for the number of "closest" neighbors of a test data tuple according to *measured distance*
  - Majority voting of their class labels used to determine class of test tuple

LAZY LEARNERS
K-NEAREST NEIGHBOR - EXAMPLE

RID	age	Loan($)	Default
1	25	40000	No
2	35	60000	No
3	45	80000	No
4	20	20000	No
5	35	120000	No
6	52	18000	No
7	23	95000	Yes
8	40	62000	Yes
9	60	100000	Yes
10	48	220000	Yes
11	33	150000	Yes
	48	142000	?

- $D = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2}$

K –> No. of Neighbor **(Odd)**

LAZY LEARNERS
K-NEAREST NEIGHBOR - EXAMPLE

RID	age	Loan ($)	Default	Distance
1	25	40000	No	102000
2	35	60000	No	82000
3	45	80000	No	62000
4	20	20000	No	122000
5	35	120000	No	22000
6	52	18000	No	124000
7	23	95000	Yes	47000
8	40	62000	Yes	80000
9	60	100000	Yes	42000
10	48	220000	Yes	78000
11	33	150000	Yes	8000
	48	142000	?	fair

- $D = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2}$

K=1 → NN is RID 11

K=3 → NN is RID 5, 9 , 11
Default = Yes

LAZY LEARNERS
K-NEAREST NEIGHBOR - EXAMPLE

RID	age	Loan ($)	Default	Distance	Loan ($)	Distance
1	25	40000	No	102000	24.4	30.6
2	35	60000	No	82000	28.3	20.8
3	45	80000	No	62000	32.3	12.7
4	20	20000	No	122000	20.4	37.0
5	35	120000	No	22000	40.2	13.7
6	52	18000	No	124000	20.0	24.9
7	23	95000	Yes	47000	35.2	26.7
8	40	62000	Yes	80000	28.7	17.8
9	60	100000	Yes	42000	36.2	14.6
10	48	220000	Yes	78000	60.0	15.4
11	33	150000	Yes	8000	46.1	15.1
	48	142000	?	fair	44.6	

- $D = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2}$

K=1 → NN is RID 3

K=3 → NN is RID 3, 5, 9
Default = NO

REGRESSION ANALYSIS
LINEAR REGRESSION

- Linear regression is a way to model a relationship between two sets of variables. The result is a linear regression equation that can be used to make predictions about data.
- First; make a scatter plot to see if data roughly fits a line. **Why?**
- The equation has the form Y= a + bX, where Y is the dependent variable (that's the variable that goes on the Y axis), X is the independent variable (i.e. it is plotted on the X axis), b is the slope of the line and a is the y-intercept.

REGRESSION ANALYSIS
LINEAR REGRESSION- EXAMPLE

SUBJECT	AGE X	GLUCOSE LEVEL Y	XY	X2	Y2
1	43	99	4257	1849	9801
2	21	65	1365	441	4225
3	25	79	1975	625	6241
4	42	75	3150	1764	5625
5	57	87	4959	3249	7569
6	59	81	4779	3481	6561
Σ	247	486	20485	11409	40022

$$a = \frac{(\Sigma y)(\Sigma x^2) - (\Sigma x)(\Sigma xy)}{n(\Sigma x^2) - (\Sigma x)^2}$$

$$b = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{n(\Sigma x^2) - (\Sigma x)^2}$$

a = **65.1416**

b = 0**.385225**

Insert the values into the equation.

y' = a + bx

y' = 65.14 + 0.385225x

QUESTION?

NEXT …….