Homework Assignment 3 (CS 145)

Due: **Monday, Nov 09 at 12:00 (Noon)**
Only **hardcopies** are accepted.

1. Bi-Clustering (30 points)

1.1 If we are asked to do clustering in dataset with high dimensionalities, can we still use Euclidean distance to measure the distance between objects? (2 pts) Justify your answer. (8 pts)

Solution: No. (2 pts)

Because of the curse of dimensionality, high dimensional data is extremely sparse. The distance measure becomes meaningless. (8 pts)

1.2 Read pages 515 and 516 of the textbook, then prove that I x J is a bicluster with coherent values **if and only if**, for any i_1, i_2 in I and j_1, j_2 in J, $e_{i1j1} – e_{i2j1} = e_{i1j2} – e_{i2j2}$ (20 pts)

Solution:

(Direction only-if) In a bi-cluster I x J, every entry can be represented by $e_{ij} = c + a_i + b_j$, where a_i and b_j are the adjustments for row i and column j, respectively.

We have $e_{i1j1} – e_{i2j1} = c + a_{i1} + b_{j1} –c – a_{i2} –b_{j1} = a_{i1}-a_{i2}$. Similarly, $e_{i1j2} –e_{i2j2} = a_{i1} – a_{i2}$. Thus, $e_{i1j1} – e_{i2j1} = e_{i1j2} – e_{i2j2}$ (10pts)

(Direction if) Let $c = -e_{11}$, $a_i = e_{i1}$ and $b_j = e_{1j}$. Since for any i_1, i_2 in I, j_1, j_2 in J, $e_{i1j1} – e_{i2j1} = e_{i1j2} – e_{i2j2}$, we can construct $e_{ij} – e_{1j} = e_{i1} – e_{11}$. That is, $e_{ij} = -e_{11} + e_{i1} + e_{1j} = c + a_i + b_j$.

I x J is a bi-cluster with coherent values. (10pts)

2. Classification (70 points)

2.1 Decision Tree (40 points)

Color	Size	Age	Inflated
Yellow	Small	Adult	T
Yellow	Small	Child	T
Yellow	Small	Adult	T
Yellow	Small	Child	T
Yellow	Large	Adult	T
Purple	Small	Child	F
Purple	Small	Adult	F
Purple	Small	Child	F
Purple	Large	Adult	T

(1) Construct a decision tree to predict the variable "inflated". Please use information gain to measure the goodness of a feature. Use 2 as the base of the logarithm and build the decision tree. (20 pts)

Solution:

Information = I(6,3)

E(Color) = 5/9 I(5,0) + 4/9 I(3,1)

E(Size) = 7/9 I(4,3) + 2/9 I(2,0)

E(Age) = 5/9 I(4,1) + 4/9 I(2,2)

Information gain of color is the largest.

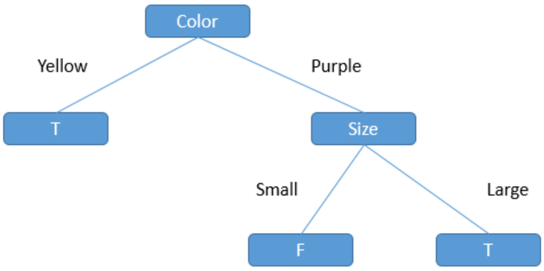
Size	Age	Inflated
Small	Child	F
Small	Adult	F
Small	Child	F
Large	Adult	T

Information = I(3,1)

E(Size) = 3/4 I(3,0) + 1/4 I(1,0)

E(Age) = 2/4 I(2,0) + 2/4 I(1,1)

Information gain of size is the largest



(20 pts)

(2) Please check the correctness of your decision tree first and then make the following prediction. Given a data entry "Purple, Large, Child", what is the prediction of "Inflated"? (5 pts)

Solution: the prediction is T

(3) Without using calculators, compare the following information I(5,5), I(5,6), I(6,5), I(5,7), I(10,10), I(10,12), I(10,13) (15 pts)

I(5,7)<I(10,13)<I(5,6)=I(6,5)=I(10,12)<I(5,5) = I(10,10)

Hint: Information measures the purity of data.

2.2 Naïve Bayesian Classifier (30 pts)

(1) Given a data entry "Purple, Large, Child", what is the prediction of "Inflated" if we use naive Bayesian classifier? (5 pts) Please state the theorem you use (5 pts), the assumptions you make (5 pts), and show calculations needed to make the prediction (15 pts)

Solution:

(1) The prediction of “inflated” is T

(2) We Use Bayesian theorem.

(3) Attributes are conditionally independent given an object.

(4) P(Color=Purple| Inflated=T) = 1/6

P(Color=Purple| Inflated=F) = 3/3

P(Size=Large| Inflated=T) = 2/6

P(Size=Large| Inflated=F) = 0/3

P(Age=Child | Inflated=T) = 2/6

P(Age=Child | Inflated=F) = 2/3

P(Inflated=T) = 6/9

P(Inflated=F) = 3/9

X=(Purple, Large, Child)

Attributes are conditionally independent

P(X| Inflated=T) = P(Color=Purple| Inflated=T) P(Size=Large| Inflated=T) P(Age=Child | Inflated=T)

=1/6 * 2/6 * 2/6

P(X| Inflated=T)P(Inflated=T)= 1/6 * 2/6 * 2/6*6/9

Similarly

P(X| Inflated=F)= P(Color=Purple| Inflated=F) P(Size=Large| Inflated=F) P(Age=Child | Inflated=F)

=3/3 * 0/3 * 2/3

P(X| Inflated=F)P(Inflated=F)= 3/3 * 0/3 * 2/3*3/9

P(X| Inflated=T)P(Inflated=T) > P(X| Inflated=F)P(Inflated=F) (15pts)