28/2/12

Εξόρυξη Δεδομένων, Φροντιστήριο 15/03/2012, Αντικείμενο: Κατηγοριοποίηση

1° Ερώτημα:

The following table presents the various values of the attributes for the last ten games when the winning captain decided to bat first and the final outcome of the game.

Independent Variables			Dependent Variable
Outlook	**Humidity**	**Number of batsmen in team > 6**	**Final Outcome**
Sunny	High	Yes	Won
Overcast	High	No	Lost
Sunny	Low	No	Lost
Sunny	High	No	Won
Overcast	Low	Yes	Lost
Sunny	Low	Yes	Won
Sunny	Low	No	Lost
Sunny	High	No	Won
Sunny	Low	Yes	Won
Sunny	Low	Yes	Won

Να κατασκευαστεί το δένδρο απόφασης στο οποίο οι υψηλότεροι κόμβοι αντιστοιχούν σε γνωρίσματα που αποδίδουν μεγαλύτερο κέρδος πληροφορίας.

2° Ερώτημα:

We have data from an survey and objective testing with two attributes (acid durability and strength) to classify whether a special paper tissue is good or not. Here is four training samples.

X1 = Acid Durability (seconds)	X2 = Strength (kg/square meter)	Y = Classification result
7	7	Bad
7	4	Bad
3	4	Good
1	4	Good

Now the factory produces a new paper tissue that pass laboratory test with X1 = 3 and X2 = 7. Without another expensive survey, can we guess what the classification of this new tissue is? Good probably :P
Apply the k-nn classification algorithm (suppose k=3 and in order to calculate the distance use the Euclidian distance) Good (Good(2) - Bad(1))