EARIN, example exam questions in the area of data mining

1. From the training set shown below in the table with the help of top-down decision tree induction algorithm create a decision binary tree (use the smallest entropy cost for attribute value tests resulting in node with two branches - one for true achieved attribute value condition, one for false for instances without the chosen value: $E_{a_{i,j}}(T) = -\frac{|T^{small}_{a_{i,j}}|}{|T_{a_{i,j}}|}\log_2(\frac{|T^{small}_{a_{i,j}}|}{|T_{a_{i,j}}|}) - \frac{|T^{big}_{a_{i,j}}|}{|T_{a_{i,j}}|}\log_2(\frac{|T^{big}_{a_{i,j}}|}{|T_{a_{i,j}}|})$, where $a_{i,j}$ denotes the chosen attribute value). The attribute age should be discretized using two thresholds 30 and 65 years. The attribute risk will be the label class. Stop criteria: every leaf node has a one target class, so if you have no attribute values with entropies equal to zero than you choose the attribute value with the smallest entropy and continue to choose the next conditions based on not used attribute values for all two branches both true and false (of this chosen condition value with smallest entropy).

x	age	car	risk
1	18	mini	big
2	35	mini	small
3	50	racer	big
4	66	van	big
5	18	racer	big
6	35	van	small
7	60	mini	small
8	70	racer	big
9	25	van	small

2. Having two points: the first one with a positive class (3,3) and the second with a negative class (1,1) find Support Vector Machine and its support vectors.

3. Explain the training of Naive Bayes classifier and its use in predicting class values of unlabelled data using the given conditional probabilities table (probabilities of attribute values the given class).

	influenza C_1	cold C_2	pneumonia C_3	allergy C_4
headache $P(A_1) = 0.3$	$P(C_1\|A_1) = 0.2$	$P(C_2\|A_1) = 0.2$	$P(C_3\|A_1) = 0.3$	$P(C_4\|A_1) = 0.3$
cough $P(A_2) = 0.2$	$P(C_1\|A_2) = 0.3$	$P(C_2\|A_2) = 0.4$	$P(C_3\|A_2) = 0.1$	$P(C_4\|A_2) = 0.2$
sneeze $P(A_3) = 0.2$	$P(C_1\|A_3) = 0.3$	$P(C_2\|A_3) = 0.2$	$P(C_3\|A_3) = 0.2$	$P(C_4\|A_3) = 0.3$
temperature $P(A_4) = 0.3$	$P(C_1\|A_4) = 0.3$	$P(C_2\|A_4) = 0.1$	$P(C_3\|A_4) = 0.5$	$P(C_4\|A_4) = 0.1$

Find the Naive Bayes classifier hypothesis $h(C|A_1 \cap A_2 \cap A_3 \cap A_4)$.

4. Describe the process of creation the ROC curve using the given positive class output of the classifier with its probabilities and real training set classes.

Real classes	1	1	0	0	1	0	0	1	0	0
Predicted classes	1	1	1	1	1	0	0	0	0	0
Probs of positives	0.9	0.9	0.6	0.6	0.6	0.3	0.3	0.1	0.1	0.1