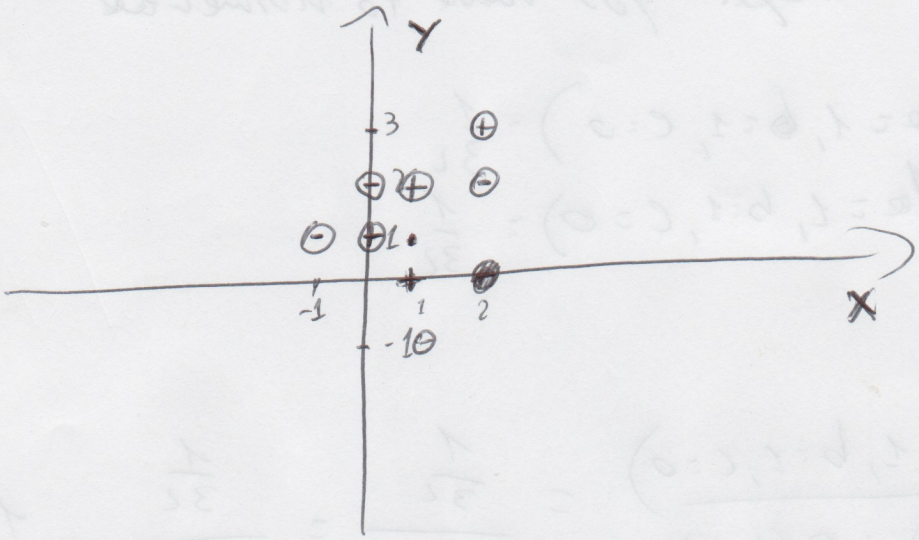
MLDM 2007 06 28

Problem 1



a) (1,1) 3- nearest- neighbor → + (3+)

b) (1,1) 5- nearest- neighbor → + (3+) (2-)

c) (1,1) 7- nearest - neighbor → - (3+) (4-)

perfect.

Problem 2

Naive Bayes classifier

$P(K=1 \mid a=1, b=1, c=0) = \frac{1}{32}$? premesse

$= P(a=1 \mid K=1)\, P(b=1 \mid K=1)\, P(c=0 \mid K=1) \cdot P(K=1)$

$= \frac{2}{4} \cdot \frac{1}{4} \cdot \frac{2}{4} \cdot \frac{4}{4} \cdot \frac{8}{64} =$

$= \frac{2}{4} \cdot \frac{1}{4} \cdot \frac{2}{4} \cdot \frac{4}{8} = \frac{1}{32}$

$P(K=1 \mid a=1, b=1, c=?) =$

Unknow

$= P(a=1 \mid K=1)\, P(b=1 \mid K=1)\, P(c=? \mid K=1)\, P(K=1)$

$= \frac{2}{4} \cdot \frac{2}{4} \cdot \frac{4}{8} = \frac{1}{16}$

Bayes yes, no and normalization.

* Law with bayes you have to normalize

$$P(K=1 \mid a=1, b=1, c=0) = \frac{1}{32}$$
$$P(K=0 \mid a=1, b=1, c=0) = \frac{1}{32}$$

normalize

$$P(K=1) \frac{P(K=1 \mid a=1, b=1, c=0)}{P(K=1) + P(K=0)} = \frac{\frac{1}{32}}{\frac{1}{32}+\frac{1}{32}} = \frac{\frac{1}{32}}{2 \cdot \frac{1}{32}} = \frac{1}{2} = 0{,}5\ \% = 50\%$$

b) $P(K=1 \mid a=1, b=1) = P(a=1 \mid K=1)\, P(b=1 \mid K=1)\, P(K=1) = \frac{1}{2} \cdot \frac{1}{4} \cdot \frac{1}{2} = \frac{1}{16}$

$P(K=0 \mid a=1, b=1) = P(a=1 \mid K=0)\, P(b=1 \mid K=0)\, P(K=0) = \frac{1}{4} \cdot \frac{1}{4} \cdot \frac{1}{2} = \frac{1}{8}$

normalize

~~P(K~~ prob $P(K=0) = \frac{\frac{1}{8}}{\frac{1}{8}+\frac{1}{16}} = \frac{\frac{1}{8}}{\frac{2+1}{16}} = \frac{1}{8} \cdot \frac{16}{3} = \frac{16}{24} = 0{,}66$

$P(K=1) = \frac{\frac{1}{16}}{\frac{1}{8}+\frac{1}{16}} = \frac{\frac{1}{16}}{\frac{2+1}{16}} = \frac{1}{16} \cdot \frac{16}{3} = \frac{1}{3} = 0{,}33 \rightarrow 33{,}3\%$

Association rules

~~Association rules are rules base~~

~~Given a set of transactions~~, an association rule is a rule that predict the occurences of related items in a dataset

Association rules are built by inspecting the dataset.

the goal of Association rules ~~many~~ is to find all rules having (chpeck, page 8)
support > minsup threshold
Confidence > minconf threshold ect.

What is a decision tree? Is it true or false that decision tree mining can be applied to any type of data? how? if false, why?

A decision tree is a tree which is used to classify data. Decision trees have nodes, branches and leaves a node is a test on an attribute, a branch is an outcome of a test, Decision trees do not work very well with large datasets with a lot of attributes. In those cases the purity functions ~~do not work~~ do not work as expected:

Information gain: biased towards multivalued attributes
Gain Ratio: tends to prefer unbalanced splits (page 81, peck 11)

Question 5

Explain the difference between apriori and fp-growth.

Apriori follows the generate-and-test paradigm while fp-growth follows the divide-and-conquer paradigm.

Apriori is more resource consumptive because it computes the minsup for each level of the tree while fp-growth calculates the minsup while building the tree itself.

Fp-growth wins in term of computation complexity against apriori algorithm.

Question 6 (Ensembles)

What is Bagging? Is there any relation between Bagging and Bootstrap? If yes, which one? If no, why?

Bagging is an ensemble method to ~~aggregate~~ classify datasets the training procedure is based on Bootstrap (sampled with replacement).

The Bagging method gives the same weight to each classifier and classify new data based on the mayority vote of all the classifiers.

It's the easiest way to perform ensembles, the others ways are Boosting and Random Forests.

(slides 11, pack 15)

Question 7

Your company has around $2 \cdot 10^6$ customers ___ (too long to write all the question)

Decision trees should be avoided because they don't perform well on large datasets with a lot of attributes like in this case.

K-nearest-neighbor can't be used neither due to device constraints (low cpu/Ram power) because this methods needs to scan all the dataset to give a prediction

So the best algorithm to chose is the Naive Bayes classifier which performs very well on a large amount of data and is lighter than K-nearest-neighbor.

2007 07 20

Problem 1

Consider the following training data set with one real attribute X and the class attribute Y.
Using information gain, drew the first node of the decision tree for this dataset.

	S1		S2	
0	0	0	1	1	1	2	2
−	+	−	−	−	+	+	+

$info(p_1, p_2) = entropy\ p_1 \cdot entropy\ p_2$

$info(D) = -2 \cdot \frac{4}{8} \log_2 \frac{4}{8} = 1$

$info_{cs_1} = -\frac{1}{3} \log_{10} \frac{1}{3} - \frac{2}{3} \log_{10} \frac{2}{3} = 0,276$

change base $\frac{0,276}{\log_{10} 2} = 0,917$

$info_{s_1} = -\frac{1}{5} \log_{10} \frac{1}{5} - \frac{2}{5} \log_{10} \frac{2}{5}$

$\approx 0,276$

$\frac{0,276}{\log_{10} 2} = 0,97$

$info_{s_1} = \frac{3}{8} \cdot 0,917 + \frac{5}{8} \cdot 0,97 = 0,95$

$info\ gain_{s_1} = 1 - 0,95 = 0,05$

$info_{cs_2} = -\frac{2}{6} \log_{10} \frac{2}{6} - \frac{4}{6} \log_{10} \frac{4}{6} =$

$info_{s_2} = -\frac{2}{8} \log_{10} \frac{2}{8} =$

$info_{s_2} = \frac{6}{8} \cdot info_{cs_2} + \frac{2}{8} \cdot info_{s_2} =$

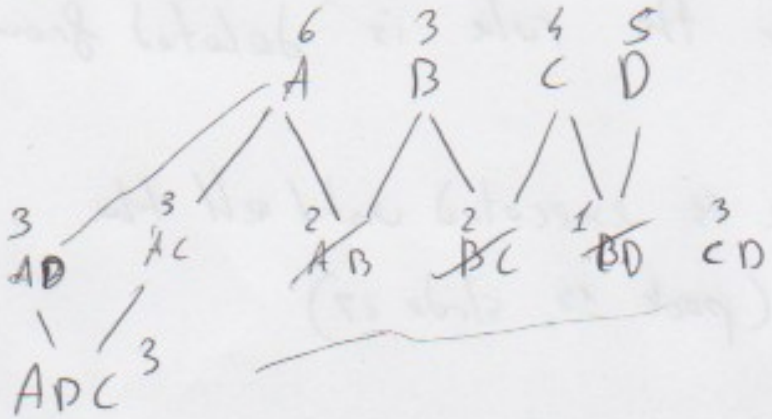
$info\ gain = 1 - info_{s_2} =$

Problem 2 (Association rules)

Consider the following dataset. Extract all the frequent itemsets with a minimum support of 3 and a minimum confidence equal to 0.8

• Confidence is defined on rules, not on itemsets

Using the Apriori algorithm



Freq itemsets A, B, C, D, AD, AC, CD, ADC

Problem 3 ~~not done~~

Illustrate the typical steps of a KDD process

KDD is a multi-step process involving data preparation, pattern searching, knowledge evaluation and refinement with iteration after modification.

the steps are: selection, preprocessing, transformation, data mining, evaluation and at the end what it's built is knowledge.

Problem 4

Briefly illustrate sequential covering algorithms

The sequential covering algorithm iteretes over a dataset trying to find a rule which classify part of the dataset, when a rule is found, all the itemsets identified by the rule is deleted from the dataset.
The rule creation stage is executed until all the dataset is covered (pack 12, slide 27)

Problem 5

Error in a 1-nearest-neighbor

		error
1	-	1
2	+	0
3	+	0
4	-	1
5	+	1
6	+	0
7	-	0
8	-	0
9	+	1
10	+	0

Cross validation error = $\frac{\text{error}}{\text{all}} = \frac{4}{10} = 0,4$ 40%

accuracy = $1 - 0,4 = 0,6$

Question 1

What is data streaming mining? What are synopses?

Data stream mining is data mining applied to streams of data and not datasets. Streams are not finite and usually contain unordered data. Data stream mining is an actual research field in data mining. Synopses are tradeoff between needed resources and performances while analyzing and researching on data streams

Question 2

Too long to copy

We have a dataset without any information about it. so our goal is to understand what the data is.

Consultant A is completely wrong. A decision tree in a big amount of data can introduce pitfalls.

Consultant B suggests a not pratical way to inspect the datasets cuz decision trees are more efficient way to classify data but in this case we just went to understand what the data is.

Consultant C says the best way to approach this problem. Hierarchical cluster will form initial groups of data, then Kmeans can show effectively the differences on each cluster.

20070503

Problem 1

Explain the similarities and the differences between naive Bayes classifiers and Bayesian networks.

Naive Bayes Classifier uses probability to compute a class label and works on all the attributes.
It assumes that attributes are independent
But it's often not so: class properties can be related.
Bayesian networks show in a quick and graphical way what are the related attributes of a class.
Bayesian networks can be used to improve performances of Bayes classifiers.

Problem 2

Decision tree

$info_{10}(green_{no}) = -\frac{4}{6}\log\frac{4}{6} - \frac{2}{6}\log\frac{2}{6} = 0,117 + 0,159 = 0,276$

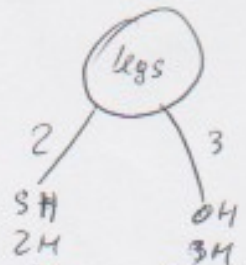
$info_2(green_{no}) = 0,9168$

$info_{10}(green_{yes}) = -\frac{1}{4}\log\frac{1}{4} - \frac{3}{4}\log\frac{3}{4} = 0,15 + 0,093 = 0,243 \rightarrow$ kg

$info_2(green_{yes}) = 0,81$

$info_{gain}(green) = \frac{4}{10} \cdot 0,81 + \frac{6}{10} \cdot 0,9168 = 0,324 + 0,55 = 0,87408$

$info_{gain}(green) = 1 - 0,87408 = 0,12582$

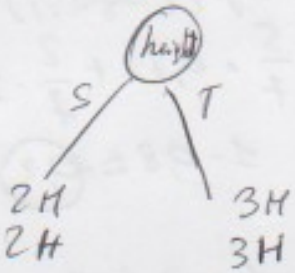


$info_{10}(legs(2)) = -\frac{5}{7}\log\frac{5}{7} - \frac{2}{7}\log\frac{2}{7} = 0,104 + 0,1554 = 0,259$

$info_2(legs(2)) : 0,8617$ $\Rightarrow info(legs) = \frac{7}{10} \cdot 0,8617 : 0,60$

$info_{10}(legs(3)) = -\frac{3}{3}\log\frac{3}{3} : -1\log 1 : \emptyset$

$infogain(legs) = 1 - 0,60 = 0,4$



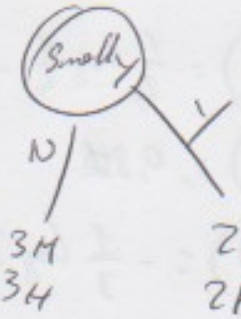
$info_{10}(height(S)) = -\frac{2}{4}\log\frac{2}{4} - \frac{2}{4}\log\frac{2}{4} = -\log\frac{2}{4} = 0,30$

$info_{10}(height(T)) = -\frac{3}{6}\log\frac{3}{6} - \frac{3}{6}\log\frac{3}{6} = -\frac{3}{3}\log\frac{3}{6} = 0,30$

$info_2(height) = \frac{0,30}{\log 2} \approx 1$

$info(height) = \frac{8}{10} \cdot 1 + \frac{5}{10} \cdot 1 = \frac{5}{10} + \frac{5}{10} = 1$

$infogain(height) = 1 - 1 = 0$



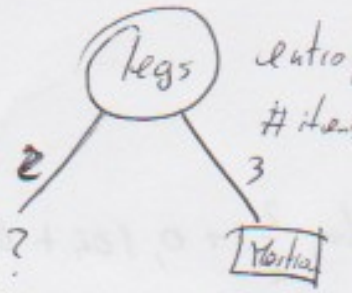
* Same as height

$info_{10}(smelly) = 0,30$

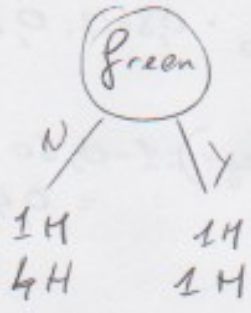
$info(smelly) = \frac{5}{10} \cdot 1 + \frac{5}{10} \cdot 1 = 1$

$infogain(smelly) = 1 - 1 = 0$

First split is legs.



entropy = 0,6
# items: 7



$info_{10}(legs(N)) = -\frac{1}{5}\log\frac{1}{5} - \frac{4}{5}\log\frac{4}{5} = 0,140 + 0,0775 = 0,21753$

$info_2(legs(N)) = 0,72262$

$info_{10}(legs(Y)) = 0,30 \Rightarrow info_2(legs(Y)) = 1$

$info(legs) = \frac{5}{7} \cdot 0,72 + \frac{2}{7} \cdot 1 = 0,8$

$infogain(legs) = 1 - 0,8 = \boxed{0,2}$

$info_{10}(height(S)) = -\frac{1}{3}\log\frac{1}{3} - \frac{2}{3}\log\frac{2}{3} = 0,159 + 0,117 = 0,2763$

$info_2(height(S)) = 0,9181$

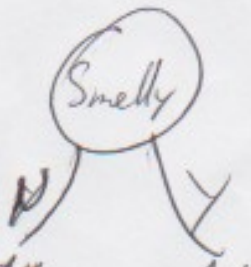
$info_{10}(height(T)) = -\frac{1}{4}\log\frac{1}{4} - \frac{3}{4}\log\frac{3}{4} = 0,15 + 0,0837 = 0,2437$

$info_2(height(T)) = 0,8095$

$info(height) = \frac{3}{7}\,0,9181 + \frac{4}{7}\,0,8056 =$

$0,3934 + 0,4628 = \boxed{0,856}$

$infogain(height) = 1 - 0,856 = \boxed{0,144}$



$info_{10}(Smelly(N)) = -\frac{1}{4}\log\frac{1}{4} - \frac{3}{4}\log\frac{3}{4} = 0,15 + 0,053$

$info_2(Smelly(N)) = 0,81$

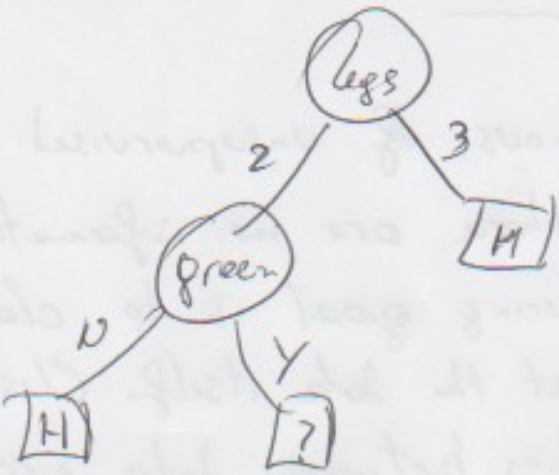
$info_{10}(Smelly(Y)) = -\frac{1}{3}\log\frac{1}{3} - \frac{2}{3}\log\frac{2}{3} = 0,160 + 0,117 = 0,277$

$info_2(Smelly(Y)) = 0,82$

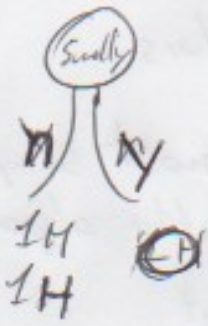
$info(Smelly) = \frac{4}{7} \cdot 0,81 + \frac{3}{7}\,0,82 = 0,4628 + 0,3542 = 0,857$

$infogain(Smelly) = 1 - 0,857 = \boxed{0,143}$

Secand split is green



infogain(height)=0

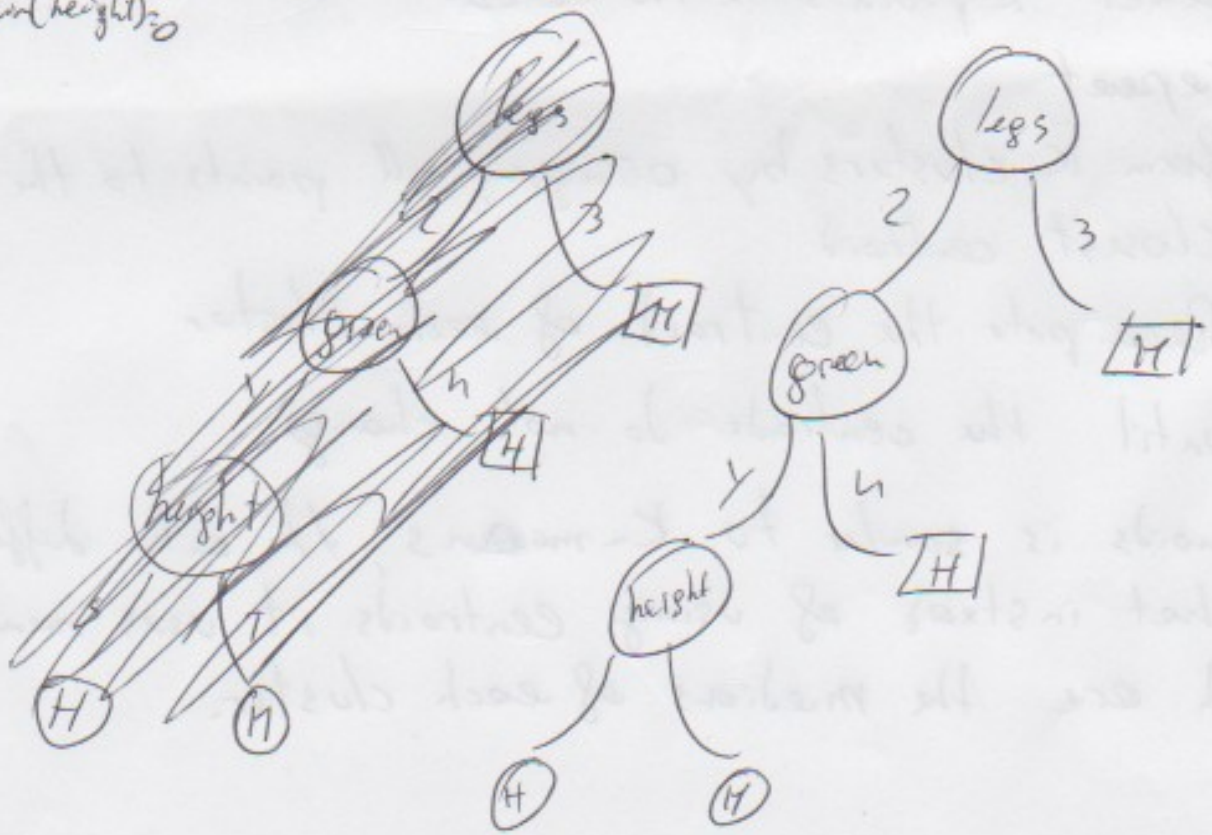


$info_{2}(smelly) = -\frac{1}{2}\log\frac{1}{2} - \frac{1}{2}\log\frac{1}{2}$

$info_2(smelly) = 1$

$info(smelly) = \frac{2}{2} \cdot 1 = 1$

$infogain(smelly) = 0$



Problem 3

Explain what is clustering ——

Clustering is a synonymous of unsupervised learning and it's applied when there are no informations about the data. Clustering goal is to classify similar data looking at the data itself. Clustering is done finding similarities between data according to the characteristics found in the data and grouping similar data objects into clusters.

The K-means algorithm is the most simple of the clustering algorithms. This is the algorithm

Select K points in the dataset

repeat

form K clusters by assigning all points to the closest centroid

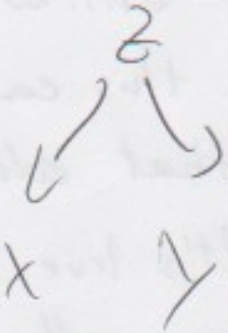
Recompute the centroid of each cluster

until the centroids do not change

Kmenoids is similar to K-means, the only difference is that instead of using centroids it uses menoids which are the medians of each cluster.

Problem 4

e) $X \in \{0,1\}$ $Y \in \{0,1\}$



b) 10 → $P(x=1|z=1)$, $P(x=1|z=0)$, $P(x=0|z=0)$
$P(x=0|z=1)$, $P(y=0|z=1)$, $P(y=1|z=1)$
$P(y=0|z=0)$, $P(y=1|z=0)$, $P(z=0)$
$P(z=1)$

Problem 5

Bootstrap, ~~cross~~ sampling with replacement to form the training set and it's ~~poss~~ suitable for small datasets coz it produces new items in the itemset.
Cross validation is sampling without replacement and in it leave one out version can reach an accuracy of 80%
with bootstrap you can achieve ~80% accuracy but the dataset must be uniformly distributed.

Problem 6 (1LY)

Overfitting is when our model of the class to recognize is full of unuseful attributes. ~~It's said that the class is~~ this can reduce the power of mining algorithms that collapse in a simple search into a database. It's true the statement because in small datasets some attributes could look ~~related~~ related when they aren't, thus this can lead to an overfitted model.

Problem 7

No, it's not a good offer. Since nominal attributes can have many values an automatic tool, if not well designed, can lead to overfitted rules. I would like to really thest that software and see how it handles cases with many values for nominal attributes.

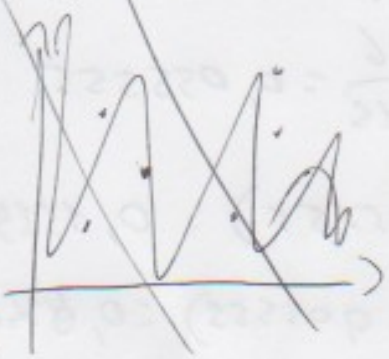
2007 09 17

Problem 1

K-means

~~pick K=2 random points~~

a)



(page 32, pack Bayes)

b)

$P(play? = yes \mid X = \langle Sunny, hot, high, N \rangle)$, $P(play = yes \mid X = \langle sunny, hot, high, N \rangle)$.

$P(play = yes \mid x = \langle sunny, hot, high, N \rangle)$

$= P(x = \langle sunny, hot, high, N \rangle \mid play = yes)\, P(play = yes)$

$= \frac{2}{9} \cdot \frac{2}{9} \cdot \frac{3}{9} \cdot \frac{6}{9} \cdot \frac{9}{15} = 0,00658$

$P(play = no \mid x = \langle sunny, hot, high, N \rangle)$

$= P(x = \langle sunny, hot, high, N \rangle \mid play = no)\, P(play = no) = \frac{3}{6} \cdot \frac{2}{6} \cdot \frac{5}{6} \cdot \frac{2}{6} \cdot \frac{6}{15}$

$= 0,01851$

Normalize

$P(yes) = 0,00658 / (0,00658 + 0,01851) = 0,263 \Rightarrow 26\%$

$P(no) = 0,01851 / (0,00658 + 0,01851) = 0,7404 \Rightarrow 74\%$

So the result is No

b)
Like before

$P(\text{sunny}, \text{hot}, \text{high} \mid \text{play}=\text{yes}) = \frac{2}{9} \cdot \frac{2}{9} \cdot \frac{3}{9} \cdot \frac{9}{15} = 0,008876$

$P(\text{sunny}, \text{hot}, \text{high} \mid \text{play}=\text{no}) = \frac{3}{6} \cdot \frac{2}{6} \cdot \frac{5}{6} \cdot \frac{6}{15} = 0,055555$

normalisation

$P(\text{yes}) = 0,008876 / (0,008876 + 0,055555) = 0,1519$

$P(\text{no}) = 0,055555 / (0,008876 + 0,055555) = 0,8476$

$P(\text{yes}) = 15\,\%$

$P(\text{no}) = 85\,\%$

c)
as seen in point a $P(\text{no}) = 74\,\%$ so the classifier classifies $X = (\text{sunny}, \text{hot}, \text{high}, N)$ as play=no!

Problem 2

Plant

$\text{info}_{10}(\text{Plant}(\text{usa})) = -\frac{2}{4} \log \frac{2}{4} - \frac{2}{4} \log \frac{2}{4} = 0,30$

~~info~~ ~~info(Plant~~

$\text{info}_{10}(\text{Plant}(\text{ca})) = -1 \log 1$

$\text{info}_{10}(\text{Plant}(\text{cin})) = -1 \log 1$

$\text{info}(\text{Plant}) = 1 \rightarrow \text{info gain} = 0$

Type

$\text{info}_{10}(\text{type}(A)) = -\frac{2}{5} \log \frac{2}{5} - \frac{3}{5} \log \frac{3}{5} \rightarrow \text{info}_2(\text{type}(A)) =$

$\text{info}_{10}(\text{type}(B)) = -\frac{5}{5} \log \frac{5}{5} = 0$

$\text{info}(\text{type}) = \frac{5}{10} \cdot$