UNIVERSITY OF LEICESTER

Theoretical Homework 1

For DATA MINING AND NEURAL NETWORKS MA4022

Gleb Vorobchuk

Leicester

March 2018

gv@student.le.ac.uk

Task №1.

Give the Bayes formula for two events and for multivalued random variables.

The probability a woman in the general population has breast cancer is 0.003. The probability that a positive mammography result occurs given that a woman has cancer is 0.50. (This is the sensitivity of the test.) The probability that a negative result will occur given that a woman does not have breast cancer is 0.97. (This is the specificity of the test.) Suppose a woman has a mammography exam and gets a positive result. What is the probability that she actually has breast cancer?

Bayes Formula:

$$P(A\,|\,B) = \frac{P(A\,|\,B) \times P(A)}{P(B)}$$

Given:

$$P(Cancer) = 0.003$$

$$P(Positive\,|\,Cancer) = 0.5$$

$$P(Negative\,|\,Cancer) = 0.97$$

$$P(Positive\,|\,Non-Cancer) = 1 - P(Positive\,|\,Cancer) = 1 - 0.97 = 0.03$$

$$P(Cancer\,|\,Pos) = \frac{P(Pos\,|\,Cancer) \times P(Cancer)}{P(Pos\,|\,Cancer) \times P(Cancer) + P(Pos\,|\,Non-Cancer) \times P(Non-Cancer))} =$$

$$= \frac{0.5 \times 0.003}{0.5 \times 0.003 + 0.03 * 0.997} = 4.77\,\%$$

UNIVERSITY OF LEICESTER

Task №2.

Give a description of classification and clustering problems. What is the difference between them? Describe KNN approach and Hart's algorithm for data reduction. Describe the K-means algorithm and prove that it terminates after a finite number of steps. Construct an example of a dataset with the minimal number of data points for which the K-means algorithm (K=2) gives different clusters for different initial positions of centres.

Classification - Predictive task of identifying a set that new observation belongs to. Based on previous observations (training set). In Machine Learning defined as an one of the subtasks of supervised learning. Classification task approximate a mapping function f from inputs variables X to discrete output variables y. The output variables y are called labels or classes. Classification requires examples of training set to be classified into at least two categories.

Clustering - Descriptive task of grouping objects in way that objects in the same group are more similar. Clustering can be formulated as a multi-objective optimization problem. Require: scalability, ability to deal with many attributes, ability to deal with outliers, high-dimensionality, unlabelled data with no predefined classes. In Machine Learning defined as an one of the subtasks of unsupervised learning.

The difference between Classification and Clustering is that the training set for classification should contain predefined classes. Also Classification predict class for the new observation when Clustering just suggest groups based on patterns in data.

KNN approach.

KNN is the algorithm used for Classification and Regression. KNN makes prediction based on the outcome of the K closest neighbours to the examine point.

In order to do that we need to have a metric of distance between a neighbour and query point. The common metric is Euclidian distance - $\sqrt{(x-p)^2}$ where x -query point, p - neigbour.

So after define the distance metric we can predict membership of the query point as averaged sum of the KNN: $y = \frac{1}{K}\sum_{i=1}^{K} y_i$

Hart's Algorithm:

Also called Condensed nearest neighbour - is an algorithm based on KNN and was developed to reduce the data set for KNN. It selects the set of prototypes U from the training data, such that 1NN with U can classify the examples almost as accurately as 1NN does with the whole data set.

Given a training set *X*, CNN works iteratively:

1 Scan all elements of *X*, looking for an element *x* whose nearest prototype from *U* has a different label than *x*.

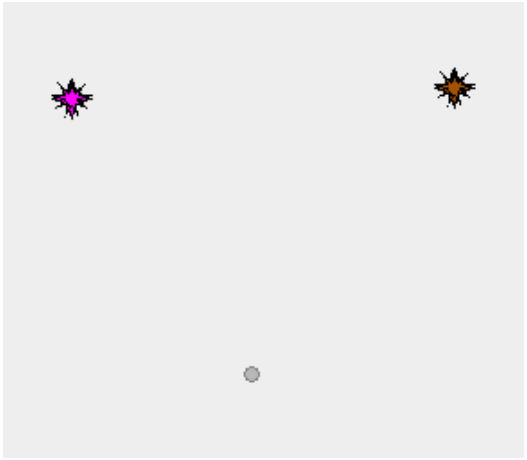
2 Remove *x* from *X* and add it to *U*

3 Repeat the scan until no more prototypes are added to *U*.

Use *U* instead of *X* for classification. The examples that are not prototypes are called "absorbed" points.[1]

Since we got k=2 we need at least 3 data points. Screenshots made in Java Applet[2]

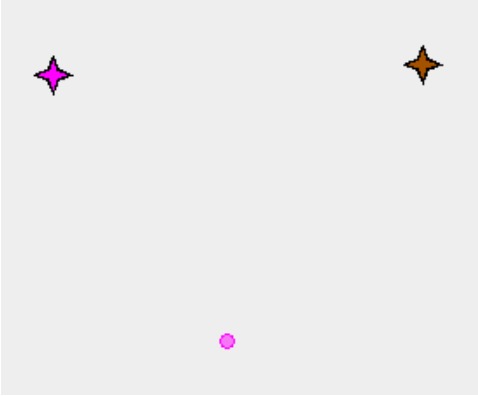
Step 1:



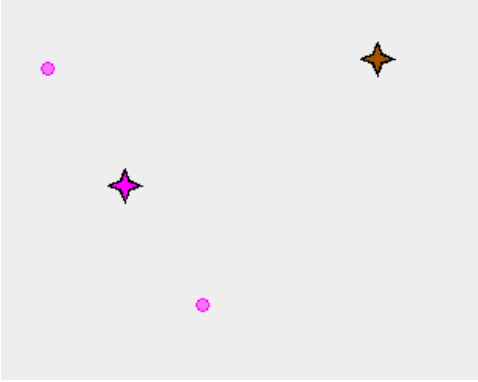
[1] https://en.wikipedia.org/wiki/K-nearest_neighbors_algorithm

[2] E.M. Mirkes, K-means and K-medoids applet. University of Leicester, 2011.

Step 2:



Step 3



As the result we observe 2 clusters.

UNIVERSITY OF LEICESTER

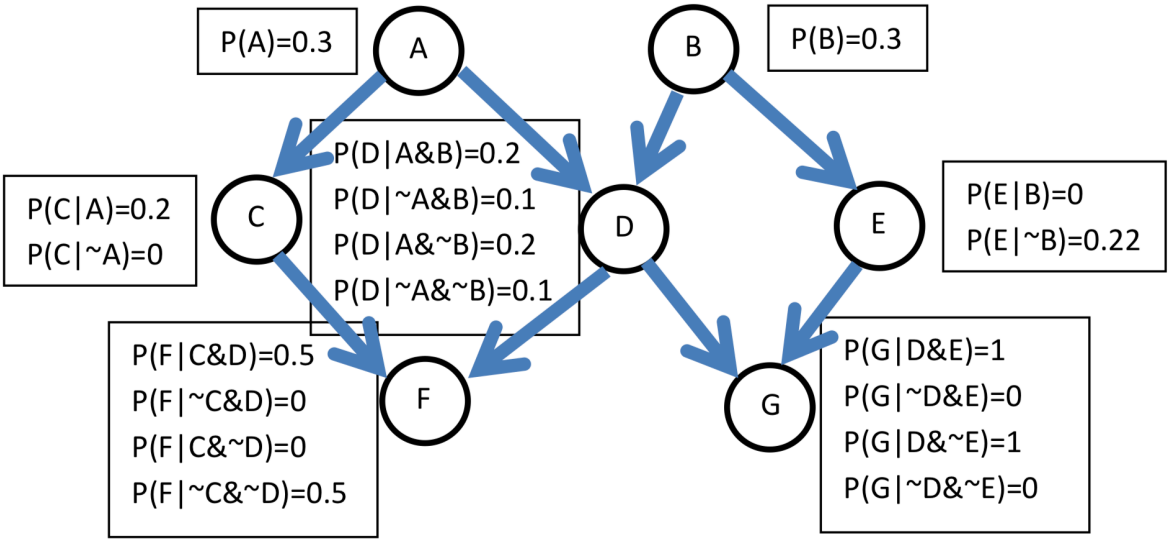
Task №3.

Describe the structure of Bayes net: what are vertices, directed edges between vertices, and what information do we keep with vertices? Give the definition of conditional independence. Compute the probabilities P(F&G) and P(B|G) from the given Bayes net:

Vertice of Bayes Network - is the event that can occurs.

Direct edges of Bayes Network is describe connections between parent and child vertices. Parent event have affection on child event.

Conditional Independence - In probability theory, two events *R* and *B* are **conditionally independent** given a third event *Y* precisely if the occurrence of *R and* the occurrence of *B* are independent events in their conditional probability distribution given *Y*. [3]



[3] https://en.wikipedia.org/wiki/Conditional_independence

P(F&G) computed as follows:

Since some of he probabilities are equal o zero we have only 3 terms:

$$P(G \,\&\, F) = (P(G \mid D \,\&\, E) \times P(F \mid C \,\&\, D) \times P(E \mid \neg B) \times P(C \mid A) \times P(D \mid A \,\&\, \neg B) \times P(A) \times P(\neg B)) +$$
$$+ (P(G \mid D \,\&\, \neg E) \times P(F \mid C \,\&\, D) \times P(\neg E \mid B) \times P(C \mid A) \times P(D \mid A \,\&\, B) \times P(A) \times P(B)) +$$
$$+ (P(G \mid D \,\&\, \neg E) \times P(F \mid C \,\&\, D) \times P(\neg E \mid \neg B) \times P(C \mid A) \times P(D \mid A \,\&\, \neg B) \times P(A) \times P(\neg B))$$
$$= 0.000924 + 0.0018 + 0.003696 = 0.00642$$

P(B | G) computed as follows:

$$P(B \mid G) = \frac{P(B \,\&\, G)}{P(G)}$$

$$P(G) = P(B \,\&\, G) + P(\neg B \,\&\, G)$$

$$P(B \,\&\, G) = (P(G \mid D \,\&\, \neg E) \times P(\neg E \mid B) \times P(D \mid A \,\&\, B) \times P(A) \times P(B)) +$$
$$+ (P(G \mid D \,\&\, \neg E) \times P(\neg E \mid B) \times P(D \mid \neg A \,\&\, B) \times P(\neg A) \times P(B)) = 0.039$$

$$P(\neg B \,\&\, G) = (P(G \mid D \,\&\, E) \times P(E \mid \neg B) \times P(D \mid A \,\&\, \neg B) \times P(A) \times P(\neg B)) +$$
$$+ (P(G \mid D \,\&\, \neg E) \times P(\neg E \mid \neg B) \times P(D \mid A \,\&\, \neg B) \times P(A) \times P(\neg B)) +$$
$$+ (P(G \mid D \,\&\, E) \times P(E \mid \neg B) \times P(D \mid \neg A \,\&\, \neg B) \times P(\neg A) \times P(\neg B)) +$$
$$+ (P(G \mid D \,\&\, \neg E) P(\neg E \mid \neg B) P(D \mid \neg A \,\&\, \neg B) P(\neg A) P(\neg B)) = 0.091$$

So:

$$P(G) = 0.039 + 0.091 = 0.13$$

$$P(B \mid G) = \frac{0.039}{0.13} = 0.3$$

You can find table used for calculations attached to report.

Task №4.

A Paradox. A PhD student analysed data about efficiency of video lectures. He has analysed performance of 300 students in Cambridge and 300 students from another university. The data are summarised in the table below. Do the video courses increase performance? According to the table for Cambridge, the answer is 'No'. Similarly, the answer is negative for another university. However, ignoring the information about the university, and collating the data into one combined table, we find that video helps! Should we recommend the video courses? Create a Bayes net for this problem. How is it possible to avoid the 'paradox' by sampling?

Cambridge	1st class	Not 1st class	% of 1st
Video	120	80	60%
Standard	75	25	75%

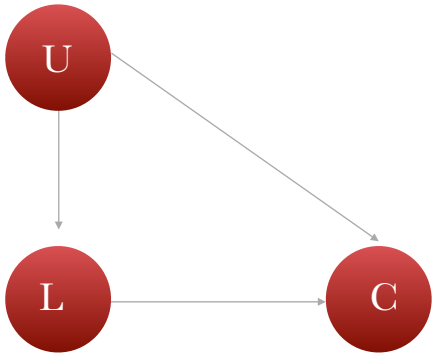
Another Uni	1st class	Not 1st class	% of 1st
Video	10	90	10%
Standard	30	170	15%

Combined	1st class	Not 1st class	% of 1st
Video	130	170	43.3%
Standard	105	195	35%

This paradox called Simpson's Paradox. It appears in several different groups of data but disappears or reverses when these groups are combined. It is sometimes given the descriptive title reversal paradox or amalgamation paradox.[4]

In order to solve this task I used example[5]

We can write distribution as: (U-Universiy, L-lecture type, C -class)



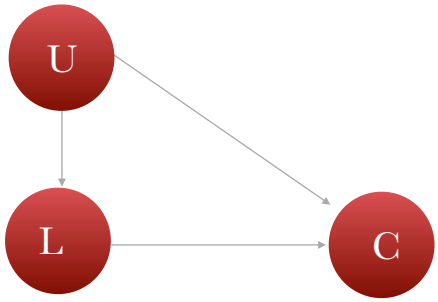
[4] https://en.wikipedia.org/wiki/Simpson%27s_paradox

[5] http://www.cimat.mx/~src/slides4.pdf

Observable Calculation computed as:

$$P(U \,\&\, L \,\&\, C) = P(C \mid U \,\&\, L) \times P(U \mid L) \times P(U)$$

In order to resolve the paradox we need to consider new distribution conditioned on lecture type, not on university. Term $P(U \mid L)$ play no role :



$$\tilde{P}(U \,\&\, C \mid L) = P(C \mid U \,\&\, L) \times P(U)$$

$$P(C \mid\mid L) = \sum_{U} \tilde{P}(U \,\&\, C \mid L) = P(C \mid U \,\&\, L) \times P(U)$$

Applying new distribution we can get away from paradox:

$$P(1_{st}Class \mid Video) = 0.5 \times 0.6 + 0.5 \times 0.1 = 0.35 = 35\,\%$$

$$P(1_{st}Class \mid Standard) = 0.5 \times 0.75 + 0.5 \times 0.15 = 0.45 = 45\,\%$$

As we can see paradox resolved. Video doesn't increase performance and we should not recommend them. Standard courses are better.

Task №5.

Give the definitions of entropy, information gain and relative information gain. Calculate the entropies H(C), H(C|A), H(C|B), information gain and relative information gain IG(C|A), RIG(C|A), IG(C|B), and RIG(C|B) for the following data table and create the decision tree for the target attribute C.

Information entropy is defined as the average amount of information produced by a stochastic source of data.[6] In other words entropy is the measure of uncertainty.

Information gain is the change in Information Entropy.

It can be written as:

$$\mathrm{H}(X) = \sum_{i=1}^{n} \mathrm{P}(x_i)\,\mathrm{I}(x_i) = -\sum_{i=1}^{n} \mathrm{P}(x_i)\log_b \mathrm{P}(x_i),$$

Where X is system with positive values $x_1, \ldots, x_n$ = , and b is the base of logarithm. Often b is equal to 2.

Here is the entropy for the system:

$$H(C) = -\frac{1}{4}\log\frac{1}{4} - \frac{1}{4}\log\frac{1}{4} - \frac{1}{2}\log\frac{1}{2} = \frac{3}{2} = 1.5$$

Then we have table containing probabilities of C given A and B:

A\C	-1	0	1
0	2/4	2/4	0
1	0	2/4	2/4

B\C	-1	0	1
0	2/4	1/4	1/4
1	2/4	1/4	1/4

[6] https://en.wikipedia.org/wiki/Entropy_(information_theory)

$$H(C\,|\,A) = \frac{1}{2}(-\frac{2}{4}\log\frac{2}{4} - \frac{2}{4}\log\frac{2}{4}) + \frac{1}{2}(-\frac{2}{4}\log\frac{2}{4} - \frac{2}{4}\log\frac{2}{4}) = 1$$

$$H(C\,|\,B) = \frac{1}{2}(-\frac{2}{4}\log\frac{2}{4} - \frac{1}{4}\log\frac{1}{4} - \frac{1}{4}\log\frac{1}{4}) + \frac{1}{2}(-\frac{2}{4}\log\frac{2}{4} - \frac{1}{4}\log\frac{1}{4} - \frac{1}{4}\log\frac{1}{4}) = \frac{3}{2}$$

Information Gains:

$$IG(C\,|\,A) = H(C) - H(C\,|\,A) = 1.5 - 1 = 1$$

$$IG(C\,|\,B) = H(C) - H(C\,|\,B) = 1.5 - 1.5 = 0$$

Relative Information Gains:

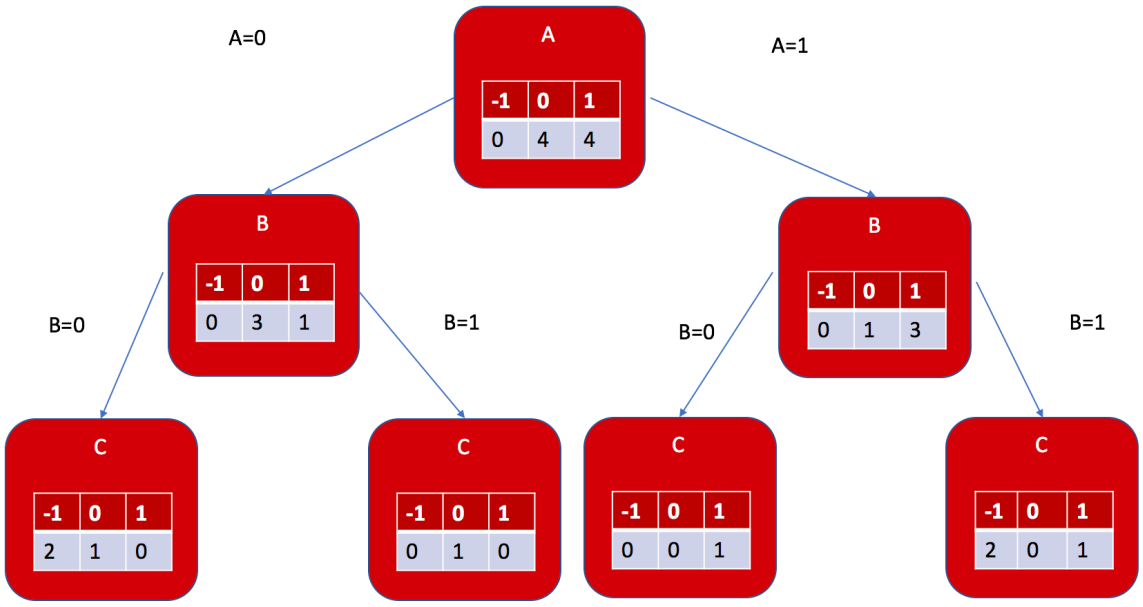
$$RIG(C\,|\,B) = \frac{H(C) - H(C\,|\,A)}{H(C)} = \frac{1.5 - 1}{1.5} = \frac{1}{3}$$

$$RIG(C\,|\,B) = \frac{H(C) - H(C\,|\,B)}{H(C)} = \frac{1.5 - 1.5}{1.5} = 0$$

Since Information Gain for B equal to 0 we will split node at A.

Now we have only one feature to work with.

Here is the final tree. Second row of the tables is the number of samples.



References:

1. https://en.wikipedia.org/wiki/K-nearest_neighbors_algorithm
2. E.M. Mirkes, K-means and K-medoids applet. University of Leicester, 2011
3. https://en.wikipedia.org/wiki/Conditional_independence
4. https://en.wikipedia.org/wiki/Simpson%27s_paradox
5. http://www.cimat.mx/~src/slides4.pdf
6. https://en.wikipedia.org/wiki/Entropy_(information_theory)