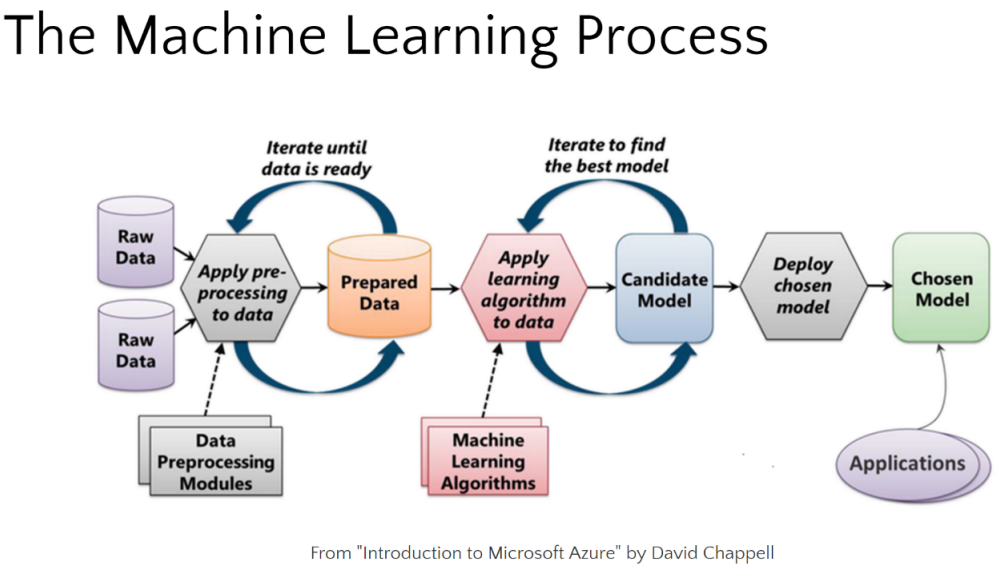
Computer Science in Medicine and Industry

Practice 1

Getting basic information about the decision systems and basics of data preprocessing (Data preprocessing and basic information about the decision systems



Short theoretical introduction

The topic of the training is to prepare data for building data mining models in the context of classification - that is, solving problems based on knowledge by assigning discrete decisions.

An example of a classification problem could be deciding whether to go to a lecture or not. Let's present below an information system (lecture information table) - see Tab. 1 and the decision system, based on which we can automatically make a decision - see Tab. 2.

The objects of the information/decision system are the individual lectures $lecture_subject_1$ .... $lecture_object_6$. The attributes of the systems are the information describing the objects, i.e. $attend\_checked$, $topic\_interesting$, $number\_of\_hours\_per\_week$.

Based on the information system from Tab. 1, each student can make an individual decision and the decision system from Tab. 2 does not have to be identical every time.

When solving a lecture attendance problem, the student makes a decision for each lecture based on its description. In this way, he creates a knowledge base (a set of solved problems) that can be useful for automatic decision-making.

To refer to an information or decision system we use descriptor notation, e.g., $attend\_checked(lecture_subject_2) = NO$, the descriptor tells us that in lecture 2 attendance is not checked.

As we write in general $(attend\_checked = NO)$ we refer to all lectures where attendance is not checked.

In our systems, the value of NO is a so-called MISSING VALUE.

In the decision-making system we can find some rules of action, for example, IF $(attend\_checked = YES) => (whether_isc? = YES)$, that is, when attendance is checked at a lecture it is definitely recommended to come to the lecture. When attendance is not compulsory, according to our system, the matter is ambiguous, because other descriptions (attributes) can affect the attendance decision.

In the exercise, we will consider two basic types of attributes: symbolic (s), i.e., $attend\_checked$, $topic\_interesting$, and numeric, e.g., $number_hour_in_tyg$..

When preparing data to create an automatic decision-making system, we mark which data are symbolic to avoid unjustified calculations on them.

Each decision-making model has a canon of data pre-processing techniques that are dedicated to it, for example, for learning artificial neural networks, the data should be normalized or standardized. For the SVM (support vector machine) technique, the data should be in numerical form. That is, symbolic needs to be transformed into numeric. Before performing k-NN classification or using NaĂŻve Bayes classifier, assuming the use of numerical metrics, the unknown values should be completed before starting work. There are many such preprocessing rules.

Let's move on to a set of tasks designed to demonstrate selected techniques for preparing (preprocessing) raw data for use in decision-making models.

Tabela 1: Lecture information system in the first week of semester 6

	$attend\_checked$	$topic\_interesting$	$number\_of\_hours\_per\_week$
$lecture\_subject_1$	YES	NO	2
$lecture\_subject_2$	NO	YES	1
$lecture\_subject_3$	NO	NO	2
$lecture\_subject_4$	NO	NO	1
$lecture\_subject_5$	TAK	YES	2
$lecture\_subject_6$	$DONT\_KNOW$	$DONT\_KNOW$	1

Tabela 2: Decision-making system - shows the decisions made by the concurrent student in the first week of the semester 6

	attend_checked	*topic_interesting*	*number_of_hours_per_week*	if_attend
$lecture\_subject_1$	YES	NO	2	YES
$lecture\_subject_2$	NO	YES	1	YES
$lecture\_subject_3$	NO	NO	2	NO
$lecture\_subject_4$	NO	NO	1	NO
$lecture\_subject_5$	YES	YES	2	YES
$lecture\_subject_6$	$DONT\_KNOW$	$DONT\_KNOW$	1	YES

Set of tasks to do

1) Create on the desktop folder in the format name.surname, in it put all the files related to the task.

2) Select one of decision systems available in the folder dane,
w pliku $\_info - data - discrete.txt$ in the file $\_info - data - discrete.txt$ we have description of available decision systems in the format:)

$$systemu\_name\ attr\_no.\ obj\_no$$

$$system\_name\ number\_of\_attributes\ number\_of\_objects$$

and in file $nazwa - type.txt$ we have types of attributes

$$n - atrybut\ numeryczny(numeric),$$

$$s - atrybut\ symboliczny(symbolic).$$

Remember about the restrictions for symbolic attributes.

3) Load selected decision system for instance in C++ and find the information:
Find available decision classes,
b) Find size of decision classes (number of objects in classes)),
c) Find minimal and maximal values for each attribute - apply for numerical attributes,
d) For each attribute detect the number of different available values,
e) For each attribute list the set of different, available values,
f) Compute standard deviation for each attribute in the whole system and separately for each decision class.

4) Do for selected data the following preprocessing:
a) Generate ten per cent of missing values in selected decision system, and complete the missing values with most common values or mean values (for symbolic attributes)
b) Normalize attribute values into intervals): $< -1, 1 >, < 0, 1 >, < -10, 10 >$, normalization of descriptor $a_i(ob_j)$ (i-th attribute and jth object)) into interval $< a, b >$

consists of the step:

$$a_i(ob_j) = (\frac{(a_i(ob_j) - min_{a_i}) * (b - a)}{max_{a_i} - min_{a_i}} + a$$

c) Do standarization of attributes of selectred data using the following method:

$$a_i(ob_j) = \frac{a_i(ob_j) - mean_{a_i}}{variance_{a_i}}$$

$$mean_{a_i} = \frac{\sum_{j=1}^{number\_of\_objects} a_i(ob_j)}{number\_of\_objects}$$

$$variance_{a_i} = \sqrt{\sum_{j=1}^{number\_of\_objects} (a_i(ob_j) - mean_{a_i})^2}$$

after standarization, mean value of attribute a_i is equal 0, parametr *variance* is equal 1,
d) Convert symbolic values of attribute Geography into Dummy Variables and remove one of new attributes to avoid dummy varaible trap demostration of attribute conversion a_1 na dummy varaibles (into dummy variables): for the system

a_1	a_2	a_3
$symbol_1$	1	4
$symbol_2$	2	3
$symbol_3$	1	5
$symbol_2$	1	5

First step of conversion:

$a_1.symbol_1$	$a_1.symbol_2$	$a_1.symbol_3$	a_2	a_3
1	0	0	1	4
0	1	0	2	3
0	0	1	1	5
0	1	0	1	5

We select the appeared values by 1, in the last step one of attributes should be removed to avoid self absorbtion in regression,

$a_1.symbol_2$	$a_1.symbol_3$	a_2	a_3
0	0	1	4
1	0	2	3
0	1	1	5
1	0	1	5

5) To do tasks you can use the available exemplary starter codes,