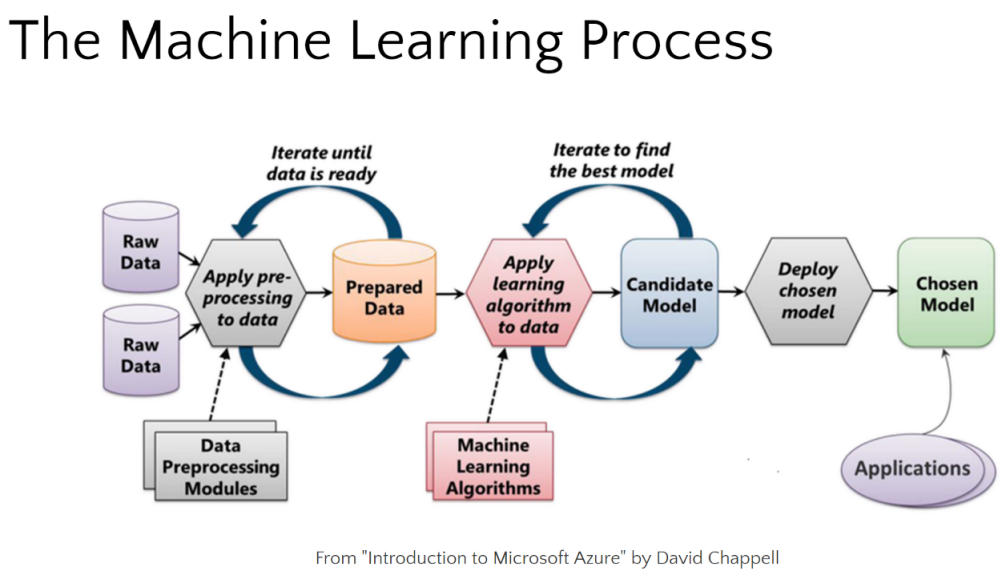
Ćwiczenie 1 - Sztuczna Inteligencja

Task 1 - Artificial Intelligence

Zdobywanie podstawowych informacji o systemach decyzyjnych i podstawy preprocesowania danych (Data preprocessing and basic information about the decision systems



Krótkie wprowadzenie teoretyczne (Short theoretical introduction)

Tematem ćwiczenia jest przygotowanie danych do budowania modeli data miningowych w kontekście klasyfikacji - czyli rozwiązywania problemów na podstawie wiedzy przez przydzielanie dyskretnych decyzji.

Przykładowym problemem klasyfikacji może być podejmowanie decyzji czy wybrać się na dany wykład czy nie. Zaprezentujmy poniżej system informacyjny (tablicę informacji o wykładach) - patrz Tab. 1 oraz system decyzyjny, na podstawie którego możemy automatycznie podejmować decyzję - patrz Tab. 2.

Tabela 1: System informacyjny o wykładach w pierwszym tygodniu semestru 6

	$czy\_obec\_spr?$	$czy\_interes\_tremat?$	$liczba\_godz\_w\_tyg.$
$wyklad\_przedmiot_1$	TAK	NIE	2
$wyklad\_przedmiot_2$	NIE	TAK	1
$wyklad\_przedmiot_3$	NIE	NIE	2
$wyklad\_przedmiot_4$	NIE	NIE	1
$wyklad\_przedmiot_5$	TAK	TAK	2
$wyklad\_przedmiot_6$	NIE_WIEM	NIE_WIEM	1

Tabela 2: System decyzyjny - pokazuje decyzje podejmowane przez konkrentego studenta w pierwszym tygodniu semestru 6

	$czy\_obec\_spr?$	$czy\_interes\_tremat?$	$liczba\_godz\_w\_tyg.$	czy_isc?
$wyklad\_przedmiot_1$	TAK	NIE	2	TAK
$wyklad\_przedmiot_2$	NIE	TAK	1	TAK
$wyklad\_przedmiot_3$	NIE	NIE	2	NIE
$wyklad\_przedmiot_4$	NIE	NIE	1	NIE
$wyklad\_przedmiot_5$	TAK	TAK	2	TAK
$wyklad\_przedmiot_6$	NIE_WIEM	NIE_WIEM	1	TAK

System informacyjny z Tab. 1, to zbiór opisanych obiektów.

System decyzyjny z Tab. 2, to zbiór rozwiazanych problemów, czyli lista wykładów, co do których mamy już podjętą decyzję.

Obiektami systemu informacyjnego/decyzyjnego są poszczególne wykłady $wyklad\_przedmiot\_1$ ... $wyklad\_przedmiot_6$.

Atrybutami systemów są informacje opisujące obiekty, czyli $czy\_obec\_spr?$, $czy\_interes\_temat?$, $liczba\_godz\_w\_tyg.$.

Na podstawie systemu informacyjnego z Tab. 1, każdy ze studentów może podjąć indywidualne decyzję i system decyzyjny z Tab. 2 nie musi być za każdym razem identyczny.

Student rozwiązując problem obecności na wykładach, podejmuje decyzję dla każdego wykładu na podstawie jego opisu. W ten sposób tworzy bazę wiedzy (zbiór rozwiązanych problemów), która może być użyteczna do automatycznego podejmowania decyzji.

Aby odwołać się do systemu informacyjnego lub decyzyjnego używamy zapisu deskryptorowego,

np. $czy\_obec\_spr?(wyklad\_przedmiot_2) = NIE$, deskryptor mówi nam, że na wykładzie 2 obecność nie jest sprawdzana.

Jak zapiszemy ogólnie ($czy\_obec\_spr? = NIE$) odwołujemy się do wszystkich wykładów, na których obecność nie jest sprawdzana.

W naszych systemach wartość NIE_WIEM jest tak zwaną wartością nieznaną (MISSING VALUE).

W systemie decyzyjnym możemy znaleźć pewne reguły działania, np. JEŻELI ($czy\_obec\_spr? = TAK$) => ($czy\_isc? = TAK$), czyli gdy na wykładzie jest sprawdzana obecność definitywnie zaleca się przychodzenie na wykład. Gdy obecność nie jest obowiązkowa, wg. naszego systemu sprawa jest niejednoznaczna, ponieważ inne opisy (atrybuty) mogą mieć wpływ na decyzję o obecności.

W ćwiczeniu rozważymy dwa podstawowe typy atrybutów: symboliczne (s), czyli $czy\_obec\_spr?$, $czy\_interes\_tremat?$, oraz numeryczne, np. $liczba\_godz\_w\_tyg.$.

Przygotowując dane do stworzenia systemu automatycznego podejmowania decyzji, zaznaczamy, które dane są symboliczne dla uniknięcia nieuzasadnionych obliczeń na nich.

Każdy model decyzyjny ma kanon technik pre-procesowania danych, które są do niego dedykowane, np. do uczenia sztucznych sieci neuronowych należy dane znormalizować lub dokonać standaryzacji. W przypadku techniki SVM (maszyny wektorów wspierających) dane powinny być w postaci numerycznej. Czyli symboliczne trzeba przekształcić w numeryczne. Przed dokonaniem klasyfikacji k-NN lub za pomocą Naiwnego klasyfikatora Bayesa, zakładając używanie metryk numerycznych, wartości nieznane powinny być uzupełnione przed rozpoczęciem pracy. Tego typu reguł preprocesowania jest bardzo wiele.

Przejdźmy do zestawu zadań, które mają na celu demonstrację wybranych technik przygotowania (preprocesowania) surowych danych do zastosowania w modelach decyzyjnych.

Zadania do wykonania (Set of tasks to do)

1) Tworzymy na pulpicie katalog w formacie Imie.Nazwisko, w nim umieszczamy wszystkie pliki dotyczące ćwiczenia (Create on the desktop folder in the format name.surname, collect all your files needed to complete the task),

2) Wybieramy jeden z systemów decyzyjnych dostępnych w katalogu dane (Select one of decision systems available in the folder dane),
w pliku $\_info - data - discrete.txt$ mamy opis poszczególnych systemów w formacie, (in the file $\_info - data - discrete.txt$ we have description of available decision systems in the format:)

$$nazwa\_systemu\ liczba\_atrybutow\ liczba\_obiektow$$

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

oraz plik $nazwa - type.txt$ w którym mamy informacje o typie atrybutów, (and in file $nazwa - type.txt$ we have types of attributes)

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

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

Pamiętamy o ograniczeniach dla atrybutów symbolicznych (remember about the restrictions for symbolic attributes).

3) Wczytujemy wybrany system np. w C++ i zdobywamy następujące informacje o systemie (load selected decision system for instance in C++ and find the information):
a) wypisujemy istniejące w systemie symbole klas decyzyjnych (find available decision classes),
b) wielkości klas decyzyjnych (liczby obiektów w klasach) (find size of decision classes (number of objects in classes)),
c) minimalne i maksymalne wartości poszczególnych atrybutów(dotyczy atrybutów numerycznych) (minimal and maximal values for each attribute - apply for numerical attributes),
d) dla każdego atrybutu wypisujemy liczbę różnych dostępnych wartości (for each attribute detect the number of different available values),
e) dla każdego atrybutu wypisujemy listę wszystkich różnych dostępnych wartości (for each attribute list the set of different, available values),
f)odchylenie standardowe dla poszczególnych atrybutów w całym systemie i w klasach decyzyjnych (dotyczy atrybutów numerycznych) (compute standard deviation for each attribute in the whole system and separately for each decision class).

4) Wykonaj dla wybranych danych następujący preprocessing (do for selected data the following preprocessing):
a) Wygeneruj 10 procent wartości nieznanych, wpisując na miejsce danych znak zapytania i napraw metodą szukania najczęściej wystepującej wartości, lub wartością średnią (dla atrybutów numerycznych), (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) Znormalizuj atrybuty numeryczne wybranego systemu na przedziały: (normalize attribute values into intervals): $< -1, 1 >$, $< 0, 1 >$, $< -10, 10 >$, normalizacja wartości (normalization of descriptor) $a_i(ob_j)$ (i-tego deskryptora obiektu j-tego (i-th attribute and jth object)) na przedział (into interval) $< a, b >$ polega na wykonaniu przeliczenia (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) Dokonaj standaryzacji wartości numerycznych wybranego systemu decyzyjnego, standaryzacja konkretnego atrybutu polega na wykonaniu operacji (Do standarization of attributes of selectred data using the following method):

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

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

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

po dokonaniu standaryzacji (after standarization), średnia wartość atrybutu (mean value of attribute) a_i jest równa (is equal) 0, parametr *variance* jest równy (is equal) 1,
d) Przeformatuj dane z pliku (Format the data from file) *Churn_Modelling.csv* do postaci (to readable form), która da się wczytać, następnie zamień atrybut symboliczny Geography na Dummy variables (convert symbolic values of attribute Geography into Dummy Variables) i skasuj jeden z trzech nowych atrybutów aby uniknąć wpadnięcia w pułapkę Dummy Variables (and remove one of new attributes to avoid dummy varaible trap). Demonstracja zamiany wartości atrybutu (demostration of attribute conversion) a_1 na dummy varaibles (into dummy variables): mając system (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

Pierwszy krok zamiany (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

jedynkami zaznaczamy występowanie konkretnej wartości, w ostatnim kroku zamiany na Dummy variables, kasujemy jeden z nowych atrybutów, aby uniknąć symetrycznego wchłaniania się wartości (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) Do wykonania zadania można wykorzystać jeden z programów demonstracyjnych(to do tasks you can use the available exemplary starter codes),