

IZU 2023/24: Zadání 3. úkolu

Společná část zadání

Každý student má přiřazen jeden příklad, který byl zaslán e-mailem na školní účet. Pro daný příklad demonstujete princip generování rozhodovacích klasifikačních stromů algoritmem **ID3 (Iterative Dichotomiser 3)** představeným na přednášce.

Formát zadání příkladů

Zadání jednotlivých příkladů jsou uložena v textových souborech. Název souboru je ve formátu model-xloginXX.txt. Každé zadání je definováno následujícími seznamy v příslušném souboru (zde uvedené příklady jsou pouze ilustrativní):

1. Seznam atributů s výčtem hodnot.

```
attributes {  
    uver : spatna nan dobra  
    dluh : vysoky nizky  
    ruceni : zadne adekv  
    prijem : m s v  
}
```

2. Seznam tříd, do kterých jsou objekty klasifikovány.

```
classes {  
    V  
    P  
    N  
}
```

3. Seznam pozorování (příkladů/objektů), z kterých se tvoří rozhodovací strom. Každý řádek odpovídá jednomu objektu:

- 1. sloupec je identifikátor objektu
- 2. sloupec je třída, do které je objekt klasifikován
- další sloupce obsahují hodnoty atributů v tom pořadí, jak jsou definovány v seznamu atributů

```
objects {  
    1 V spatna vysoky zadne m  
    2 V nan vysoky zadne s  
    3 P nan nizky zadne s  
}
```

Řešení a formát odevzdání

Seznam objektů obsahuje vždy 15 trénovacích příkladů/objektů, podle kterých vytvořte a zapište rozhodovací strom:

- rozhodovací uzel
 - reprezentuje atribut, podle kterého se pokračuje v tvorbě podstromů
 - obsahuje název atributu a informace o vypočítaných ziscích (gains) pro všechny uvažované varianty (atributy) včetně vybraného
- listový uzel
 - reprezentuje klasifikovanou třídu
 - obsahuje pouze název třídy
- hrana mezi uzly
 - je ohodnocena konkrétní hodnotou atributu, pro který se generuje příslušný podstrom
 - obsahuje informace o objektech, kterým odpovídá zvolená cesta (seznam jejich identifikátorů)

Strukturu stromu zapište do textového souboru v následujícím formátu

- každý uzel nebo hrana bude zapsána na samostatném řádku
- pro rozhodovací uzel (podle vybraného atributu)
 - `uzel1 [label="dluh|{ruceni=0.1275|dluh=0.2917}"]`
- pro listový uzel reprezentující klasifikovanou třídu
 - `uzel2 [label="V"]`
- pro hranu
 - `uzel1 -> uzel2 [label="vysoky {2,3,4}"]`
- v části label jsou uvedeny názvy atributů a jejich hodnot podle dodaných seznamů
- čísla uvádějte na 4 desetinná místa s desetinnou tečkou
- názvy uzlů (uzel1, uzel2, ...) mohou být libovolné a slouží výhradně pro identifikaci uzlů při definici hran
- v názvech uzlů, atributů a hodnot a při uvádění čísel nepoužívejte mezery
- pořadí řádků není důležité s těmito výjimkami
 - na prvním řádku bude vždy kořenový uzel celého stromu
- nepoužívejte diakritiku, textový soubor bude ve znakové sadě latin1

Odevzdání

- Odevzdávejte zip archiv s názvem `xlogin.zip` obsahující následující soubory
 - Textový soubor s názvem `xlogin.md` – definice stromu
 - PDF soubor s názvem `xlogin.pdf` – grafické zobrazení stromu
- Poznámka: formát souboru `*.md` vychází z formátu definice grafů pro nástroj Graphviz. Po drobné úpravě souboru lze nechat vygenerovat pdf s grafickým zachycením stromu tímto nástrojem.

Součást zadání

- Textový soubor s definicí úlohy.
- Ukázkové soubory
 - Textový soubor s definicí stromu
 - Textový soubor pro Graphviz
 - PDF soubor s vygenerovaným stromem