Úkol B – Hledání podskupin v datech – alcohol consumption

ZS 2016

Cíl

Cílem druhého úkolu je seznámit se s praktickým použitím shlukovacích metod. Shlukovací metody lze využít k odhalení vnitřní struktury dat. Poznání vnitřní struktury dat může pomoci k lepšímu porozumění datům a jevům, které zachycují, stejně tak ke konstrukci lepších klasifikátorů.

Data

Data o konzumaci alkoholu u studentů na nějaké portugalské střední škole. Data popisují studenty z hlediska jejich úspěšnosti v oborech matematiky a portugalština, spolu s parametry rodinného prostředí, času stráveného učením, cestováním do a ze školy, volným časem, konzumací alkoholu ve všední dny a o víkendech, apod. Konzumace alkoholu mezi studenty je velmi zajímavé téma a data studie mohou posloužit k objektivnímu posouzení jeho vlivu na studijní výsledky.

Hledání podskupin v datech

Pokuste se v datech nalézt vnitřní strukturu (podskupiny) na základě hodnot příznaků. Zvolte vhodnou metodu. Zamyslete se nad tím, zda je vhodné data normalizovat. Výsledky vizualizujte a pokuste se je interpretovat, např. pomocí typických (průměrných) reprezentantů. Prozkoumejte, které z příznaků se nejvíce podílí na rozdělení pozorování do jednotlivých shluků. Uvažte, jestli je možné dát nalezenou strukturu do souvislosti se studijními výsledky a konzumací alkoholu.

Požadované kroky analýzy [10 bodů]

- upravte data pro shlukování [1 b]
- zvolte vhodnou metodu shlukování [1 b]
- zvolte vhodné parametry shlukovací metody [1 b]
- prezentujte výsledky shlukování pomocí typických reprezentantů shluků [3 b]
- zjistěte v jakých příznacích a jak se jednotlivé shluky liší [3 b]
- porovnejte výsledky shlukování se studijními výsledky a konzumací alkoholu [1 b]

Výsledky upravte do formy zprávy, která bude obsahovat stručný **úvod**, popis metod, které jste použili, v sekci **metody,** výsledky jejich aplikace na data v sekci **výsledky** a závěry, které jste zjistili interpretací výsledků v sekci **závěr**. Maximální délka zprávy je 3 stránky. Zprávu ve formátu pdf odevzdejte pomocí UploadSystemu.

school	škola, kterou student navštěvuje ('GP' - Gabriel Pereira nebo 'MS' - Mousinho da Silveira)
sex	studentovo pohlaví ('F' – female/žena nebo 'M' – male/muž)
age	studentův věk (od 15 do 22)
address	typ studentova bydliště ('U' – urban/městské nebo 'R' – rural/venkovské)
famsize	velikost rodiny ('LE3' - <=3 nebo 'GT3' - > 3)
Pstatus	soužití rodičů ('T' - living together/žijí spolu nebo 'A' – apart/rozvedeni)
Medu	vzdělání matky (0 - žádné, 1 – první stupeň ZŠ, 2 druuhý stupeň ZŠ, 3 SŠ or 4 VŠ)
Fedu	vzdělání otce
Mjob	zaměstnání matky
Fjob	zaměstnání otce
reason	důvod pro studium na dané škole
guardian	opatrovník
traveltime	doba cesty do školy (1 - <15 min., 2 - 15 až 30 min., 3 - 30 min. to 1 hod, nebo 4 - >1 hod)
studytime	týdenní doba studia(1 - <2 hod, 2 - 2 to 5 hod, 3 - 5 až 10 hod, nebo 4 - >10 hod)
failures	počet předchozích nedokončení ročníku
schoolsup	podpora ze strany školy (yes nebo no)
famsup	vzdělávací podpora v rodině (yes nebo no)
paid	doučování (yes nebo no)
activities	mimoškolní aktivity (yes nebo no)
nursery	navštěvoval školku (yes nebo no)
higher	chce pokračovat na VŠ (yes nebo no)
internet	přístup k internetu doma (yes nebo no)
romantic	ve vztahu (yes neob no)
famrel	kvalita domácích vztahů (od 1 - very bad do 5 – excellent)
freetime	volný čas po škole (od 1 - very low do 5 - very high)
goout	chodí ven s kamarády (od 1 - very low do 5 - very high)
Dalc	konzumace alkoholu ve všední dny (od 1 - very low do 5 - very high)
Walc	konzumace alkoholu o víkendech (od 1 - very low do 5 - very high)
health	zdravotní stav (od 1 - very bad do 5 - very good)
absences	počet absencí (od 0 do 93)

Tabulka 1: Popis proměnných z studie konzumace alkoholu mezi studenty

Bonusová úloha – simulovaná data

Na rozdíl od hlavního úkolu je bonusová úloha zaměřená na simulovaná data. Data k bonusové úloze představují závislosti, která jsou těžká pro standardní algoritmy shlukování jako je k-means. Vaším úkolem je v tomto případě vyzkoušet na data metody shlukování prezentované v rámci DVZ a pokusit se najít takovou kombinaci parametrů metod (metriky, způsob linkování pozorování) a transformace příznaků (polynomiální příznaky, kernelová metoda), které by umožnili správně klasifikovat data v souborech *jain.csv*, *spiral.csv* a *pathbased.csv*. Všechna data obsahují správné řešení v proměnné *class*. Volbu metody zdůvodněte a vysvětlete, proč Váš přístup data správně shlukuje. Za bonusovou úlohu je možné získat až 3 body.