Úkol B – Hledání podskupin v datech – kvalita bílého vína

ZS 2018

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

Tato data pocházejí ze severozápadní oblasti Portugalska zvaného Minho. Každé z vín bylo slepě hodnoceno nejméně 3 nezávislými degustátory podle stupnice 1 (velmi špatné) až 10 (excelentní). Z podstaty věci jsou však taková ohodnocení velmi subjektivní. Z toho důvodu byl každý vzorek také podroben fyzikálně-chemické analýze a zjištěny jeho základní parametry. Je tak možné zkoumat jednak vztahy mezi jednotlivými parametry, tak i mezi nimi a subjektivním hodnocením kvality.

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 s hodnocením kvality vína.

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 s hodnocením kvality vína [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* a stručně okomentovaný, přehledný zdrojový *R skript* odevzdejte pomocí *UploadSystemu*.

Název parametru	Popis parametru
Fixed acid (vázaná kyselost)	Obsah kyseliny vinné (g/l)
Volatile acid (těkavá kyselost)	Obsah kyseliny octové (g/l)
Citric acid (kyselina citronová)	Obsah kyseliny citronové (g/l)
Chlorides (chloridy)	Obsah chloridu sodného (g/l)
Free sulfur dioxide (volný SO_2)	Obsah volného oxidu siřičitého (mg/l)
Total sulfur dioxide (celkem SO_2)	Celkový obsah oxidu siřičitého (mg/l)
Density (hustota)	Hustota (g/l)
pH	Kyselost – pH
Sulphates (sulfáty)	Obsah síranu draselného (g/l)
Alcohol (alkohol)	Obsah alkoholu (% vol.)
Sweet	Víno sladké? 1 – sladké, 0 - suché
Quality (kvalita)	Výstupní veličina – subjektivní kvalita vína, stupnice 0: velmi špatné víno – 10: excelentní víno

Tabulka 1: Výčet parametrů

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.