

Úkol B – Hledání podskupin v datech – Auta

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 byla nasbírána v roce 1993 a pokrývají charakteristiky a cenu vybraných 93 osobních aut. U každého auta byla zjišťována cena, spotřeba, výkon, hmotnost, rozměry a další charakteristiky (viz Tabulka 1). Kompletní data jsou k dispozici v souboru cars.csv.

Manufactured	výrobce
Model	Model auta
Type	Typ auta
Min.Price	Cena za základní verzi vozu v \$1000
Max.Price	Cena za prémiovou verzi vozu v \$1000
Price	Průměrná cena (průměr Min.Price a Max.Price)
MPG.city	Údaj o spotřebě ve městě
MPG.highway	Údaj o spotřebě na dálnici
AirBags	má auto airbagy? [none, driver only, a driver & passenger]
DriveTrain	Typ náhonu [rear wheel, front wheel nebo 4WD]
Cylinders	Počet válců („Mazda RX-7 má rotační motor, proto údaj chybí)
EngineSize	Objem motoru
Horsepower	Maximální výkon
RPM	Počet otáček za minutu (při max. výkonu)
Rev.per.mile	Počet otáček na jednu míli (při nejvyšším rychlostním stupni)
Man.trans.avail	Má auto ruční převodovku?
Fuel.tank.capacity	Kapacita nádrže (v US galonech)
Passangers	Maximální počet cestujících
Length	Délka auta (v palcích)
Wheelbase	Rozvor kol (v palcích)
Width	Šířka auta (v palcích)
Turn.circle	Prostor nutný k otočení o 180 stupňů (ve stopách)
Rear.seat.room	Velikost prostoru u zadních sedadel (v palcích)
Luggage.room	Velikost zavazadlového prostoru (v kubických stopách)
Weight	Hmotnost (v librách)
Origin	Indikátor, zda je výrobce auta z USA nebo mimo USA

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 instancí do jednotlivých shluků. Uvažte jestli je možné dát nalezenou strukturu do souvislosti s klasifikací výrobce aut.

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 (které příznaky jsou relevantní pro hledání výrobce aut) [2 b]
- porovnejte výsledky shlukování s klasifikací výrobce aut [2 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.