

## Úkol B – Hledání podskupin v datech – bike sharing 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

Využívání služby půjčování jízdních kol má vysokou spojitost se sezónními podmínkami a parametry prostředí, např. typem počasí, srážkami, dnem v týdnu, ročním obdobím atd. Tento dataset byl pořízen během let 2011 a 2012 bikesharingovým centrem ve Washingtonu D.C., USA. Data byla agregována do hodinového měřítka a následně byly přidány údaje o počasí. *Předložený dataset obsahuje pro zjednodušení data agregovaná do celých dnů.*

### 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 celkovým počtem vypůjčených kol.

### 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 celkovým počtem vypůjčených kol [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
<b>Dteday</b>	Datum dne měření (yyyy-mm-dd)
<b>Season</b>	Roční období (1:jaro, 2:léto, 3:podzim, 4:zima)
<b>Yr</b>	Rok (0: 2011, 1: 2012)
<b>Mnth</b>	Kalendářní měsíc
<b>Holiday</b>	Prázdniny
<b>Weekday</b>	Den v týdnu (0: neděle, 1: pondělí, ..., 5: pátek, 6: sobota)
<b>Workingday</b>	Den není ani víkend ani prázdniny
<b>Weathersit</b>	Typ počasí (1: jasno až polojasno, 2: mlha, zataženo, 3: lehké sněžení, přeháňky, bouřka, 4: Silný déšť, kroupy, hustá mlha, bouřky)
<b>Temp</b>	Normalizovaná teplota v °C. Normalizace pomocí vztahu $(t-t_{min})/(t_{max}-t_{min})$ , $t_{min}=-8$ , $t_{max}=+39$
<b>Atemp</b>	Normalizovaná pocitová teplota v °C. Normalizace pomocí vztahu $(t-t_{min})/(t_{max}-t_{min})$ , $t_{min}=-16$ , $t_{max}=+50$
<b>Humidity</b>	Normalizovaná vlhkost
<b>Windspeed</b>	Normalizovaná rychlost větru
<b>Casual</b>	Počet nahodilých uživatelů
<b>Registered</b>	Počet registrovaných uživatelů
<b>Cnt</b>	Celkový počet zapůjčených kol

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.