

## Úkol B – Hledání podskupin v datech – spambase 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

Detekce spamu je jeden z typických problémů strojového učení. Nevyžádaná pošta často slouží k reklamě, k šíření škodlivého softwaru (viry), získávání osobních informací (phishing) a nebo k jednoznačným podvodům (Nigerijský princ). Zpracováním emailů pro potřeby strojového učení se zabývá text mining, se kterým se setkáme na některém z následujících cvičení. V případě tohoto úkolu jsou už data upravena do formy datové matice, a není proto třeba provádět jakékoli zpracování skutečných emailů.

### Popis dat

Data obsahují informace o 4600 emailech. Emaily jsou popsány pomocí výskytu některých slov a znaků (money, credit, \$, ...), které jsou v proměnných ‚word\_freq\_‘ a ‚char\_freq\_‘ jako procento z celkového počtu slov/znaků. Dalšími proměnnými jsou počty velkých písmen, ‚capital\_run\_length\_average‘ je průměrná délka řetězců slov uvedených ve velkých písmenech, ‚capital\_run\_length\_longest‘ je největší délka řetězce slov uvedených ve velkých písmenech a ‚capital\_run\_length\_total‘ je celková délka řetězců slov uvedených ve velkých písmenech.

### 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 klasifikací emailu (jestli je nebo není spam).

### 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 klasifikací [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.