# Domáca úloha č.2 k predmetu PV056

## Prerekvizity:

Nainštalovaný program Weka 3, ktorý si môžete stiahnuť na adrese: `http://www.cs.waikato.ac.nz/ml/weka/`

## Datasety:

Datasety si môžete stiahnuť na adrese: `http://archive.ics.uci.edu/ml/datasets.html`. Každý z vás má svoju vlastnú sadu datasetov. Pridelené datasety nájdete v tabuľke:

| | |
|---|---|
| 207622 | Internet Advertisements Data Set |
| 211069 | Mammographie Mass Data Set |
| 324426 | Tic-Tac-Toe Endgame Data Set |
| 324573 | Wine Data Set |
| 324751 | Parkinsons Data Set |
| 333279 | Hepatitis Data Set |
| 356530 | Flags Data Set |
| 357886 | Ionosphere Data Set |
| 359305 | Echo-cardiogram Data Set |
| 359441 | SPECT Heart Data Set |
| 359721 | Libras Movement Data Set |
| 359789 | Hill-Valley Data Set |
| 359940 | Heart Disease Data Set |
| 373924 | Lung Cancer Data Set |
| 374346 | Balance Scale Data Set |
| 374454 | Chess (King-Rook vs. King) Data Set |
| 374513 | Wall-Following Robot Navigation Data |
| 374595 | Breast Cancer Wisconsin (Original) Data Set |
| 396136 | Census Income Data Set |
| 409040 | Yeast Data Set |
| 409717 | Soybean (Large) Data Set |
| 410034 | Dermatology Data Set |
| 410345 | Blood Transfusion Service Center Data Set |
| 410446 | Japanese Credit Screening Data Set |
| 418142 | Horse Colic Data Set |
| 432053 | Wine Quality Data Set |

## Zadanie:

### Data Sety

- Na analýzu použite dátové sady, ktoré ste si vytvorili v prvej úlohe, prípadne si stiahnite dataset a predspracujte ho tak, ako bolo popísané v prvej úlohe.
- V tejto úlohe by ste si mali precvičit analýzu za pomoci Zmiešaných metód strojového učenia.

- Konkrétne sa jedná o metódy: Bagging a Vote ktoré nájdete na záložke classify, medzi meta klasifikátormi.
- Vašou úlohou bude za pomoci týchto klasifikátorov dosiahnut lepšie výsledky klasifikácie ako pri prvej úlohe, alebo minimálne porovnatelné.
- Pri oboch algoritmoch nastavujete ďalšie algoritmy, ktoré vykonajú samotnú analýzu. Použite tieto algoritmy: J48, RandomForest, NaiveBayes, SMO a ďalšie ľubovolné 3.
- Pri Baggingu nastavujete len jeden klasiŕikátor, preto vykonajte analýzu na každom z vašich 7 algoritmov, ale pokúste sa nastavit parametre baggingu (a vybraného algoritmu) tak, aby ste dosiahli ešte lepších výsledkov.
- Pri Vote môžete nastavit viacero klasiŕikátorov. Môžete ich lubovolné miešat, takže pri riešení je prípustná akákolvek podmnožina vašich 7 klasifikátorov. Dôležité je, aby ste si všimli parameter "combinationRule" a pohrali sa sním tak, aby ste dostali čo najlepší výsledok. Vašou úlohou bude dosiahnut v tomto prípade lepší výsledok (alebo aspoň porovnatelný) ako bol Váš najlepší dosiahnutý výsledok na tejto dátovej sade z minulej úlohy spomedzi všetkých použitých algoritmov. V tomto prípade mi odovzdáte prvých 7 najlepších výsledkov. Vo väčšej polovici prípadov vyžadujem, aby ste použili kombináciu aspoň 3 algoritmov.
- Dáta predspracujte klasicky, tak ako v prvej úlohe. Samozrejme, môžete sa s nimi pohrať aj viac ak to uznáte za vhodné.
- V prípade Baggingu mi výsledky zapíšte do tabuľky v nasledovnom formáte:

| Alg. | Acc. l.úl. | Acc. 2.úl. | Param, baggingu | Param, algoritmu | Zlepšenie o |
|---|---|---|---|---|---|
| J48 | XX.x | XX.x | ... | ... | + /-XX.X |

- V prípade Vote mi výsledky zapíšte do tabuľky v nasledovnom formáte (zoradené od najlepšieho po najhorší):

| Algoritmy | Alg. param. | Accuracy | Vote params | Best 1.úl. | Zlepšenie o |
|---|---|---|---|---|---|
| Alg A Alg B AlgC | params A params B params C | XX.x | ... | ... | + /-XX.X |
| Alg B Alg E Alg A | params B params E params A | XX.x | ... | ... | + /-XX.X |

- Vytvorte súbory unexpected-bagging.txt a unexpected-vote.txt a zaznamenajte do nich poznámky o netypickom priebehu, ak napríklad algoritmus nedobehne, alebo o prípadných dodatočných úpravách dát.
- Ak niekto chce, môže si vyskúšat túto úlohu aj naprogramovať. Weka má relatívne dobrú dokumentáciu na webe aj s praktickými ukážkami a programovanie je velmi intuitívne a jednoduché. Osobne si myslím, že naprogramovanie tejto úlohy bude pre vás rýchlejšie ako keby ste to mali vyklikávat v GUI.
- Vypracovanú úlohu (2x tabuľka s výsledkami (bagging.pdf. vote.pdf), arff súbor s vašou dátovou sadou, 2x unexpected.txt (unexpected-bagging.txt, unexpected-vote.txt)) odovzdajte do Odevzdávárny zazipované v jednom súbore do 16.05.2014 13:00.
- Ak sa to rozhodnete úlohu naprogramovať, zašlite mi aj zdrojové kódy vášho riešenia.

- Súbory s riešením prosím nevkladajte do žiadneho podadresára! (Povolené sú len súbory s naprogramovaným riešením aby boli v zvlášt adresári)
- Informácie o splnení úlohy vám zadám do poznámkového bloku.
- V prípade nesplnenia úlohy vám budem nútený zadat mínusové body, ktoré sa vám odpočítajú od bodov získaných v záverečnej skúške.
- Ak by ste mali nejaké nejasnosti, alebo by ste si nevedeli rady, napíšte mi stručný e-mail na 173001@mail.muni.cz a do predmetu mailu zadajte aspoň kód predmetu. Všeobecné otázky prosím riešte cez diskusné fórum.

gl