Domáca úloha č.1 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:

- Stiahnite si pridelené datasety. Ak sa tam nachádza viac datasetov, vyberte si jeden ľubovolný (nie príliš veľký, ani nie malý). Súbory majú obvykle príponu *.names a *.data. Skontrolujte súbor *.data či je vo formáte hodnôt oddelených čiarkou a či má triedu ako poslednú hodnotu. Ak nie, preveďte súbor do tohto formátu.
- Vytvorte odpovedajúce súbory *.names, tak aby zodpovedali požadovanému formátu C4.5 popísanom nižšie. V stiahnutom súbore *.names môže byť popis jednotlivých parametrov. Ak nie, je potrbné súbor prejsť a parametre popísať. Taktiež môžete súbor

*.data premenovať na *.csv. Na prvý riadok tohto súboru napíšete mená atribútov oddelené čiarkou a importujete ho do Weky. Weka sama zabezpečí koverziu. Skontrolujte, či Weka rozpoznala správne dátové typy. Ak aj budete používat tento typ konverzie, tak aj v tomto prípade sa predpokladá, že odovzdáte súbor *.names.

- Klasifikáciu datasetu vykonajte s defaultným nastavením parametrov klasifikátora. Na testovanie použite cross-validation (10).

- Po ukončení výpočtu uložte celý výstup do súboru dataset_algoritmus.log, kde algoritmus $G \in NB, IB1, IB3, JRip, DS, J48, SMO, PART, MLP$. Do prehľadnej tabuľky zapíšte v skratke výsledky. Tabuľka by mala mať nasledovný formát:

Algoritmu	Accuracy	Weig. Avg Precision	Weig. Avg. Recall	Cas výpočtu
J48	XX.x	XX.x	XX.x	XX

- Vytvorte súbor unexpected.txt a zaznamenajte do neho poznámky o netypickom priebehu, ak napríklad algoritmus nedobehne, alebo o prípadných dodatočných úpravách dát (odstránenie ID adtribútu..).

- Pre najlepší algoritmus ktorý správne klasifikoval najviac inštancií (max accuracy) vyskúšajte aj beh s inými vstupnými parametrami a snažte sa ešte zvýšit accuracy. Všimnite si, ako jednotlivé nastavenia ovplyvňujú výsledok, prípadne dobu výpočtu. Tri najlepšie nastavenia parametrov si zapamätajte a uložte výstup aj s parametrami do súboru: číslo_dataset_algoritmus.log kde číslo označuje poradie.

- Vypracovanú úlohu (dataset.names, dataset.data, 9x dataset_algoritmus.log, tabulka s výsledkami, unexpected.txt, 3x cislo_dataset_algoritmus.log) odovzdajte do Odevzdávárny zazipované v jednom súbore do 16.5.2012 13:00.

- Súbory prosím odovzdávajte v plain-texte v kódovaní UTF-8 (Windows defaultne používa cp1250/2) alebo vo formáte PDF.

- 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.

Formát C4.5

dataset.data - čo riadok, to záznam. Hodnoty atribútov sú oddelené čiarkou, posledná hodnota je trieda. Záznam nie je ukončený bodkou. Každý atóm (nenumerická hodnota atribútu) musí byť v zozname hodnôt (popise) korešpondujúceho atribútu. Atómy nesmú byť v úvodzovkách, obsahovať medzery ani iné biele znaky. Chýbajúcu hodnotu vyjadruje otáznik. Záznamy s chýbajúcou triedou nie sú povolené. Napr. a2,39,a4,c2 a4,30,al,c2 al,9,a2,cl

dataset.names - popis atribútov a ich hodnôt. Prvý riadok obsahuje zoznam možných hodnôt triedy oddelených čiarkou a ukončených bodkou. Tieto musia korespondovat s poslednými hodnotami na riadkoch v súbore dataset.data. Všetky riadky obsahujú popis atribútov

v poradí, v akom sa nachádzajú v dataset.data. Chýbajúca hovnota (?) sa do výčtu nezahŕňa. Popis atribútov je nasledujúci: meno_atribútu: [continuous — CSV zoznam hodnôt pri ordinálnom atribúte].

Napr. cl,c2. atl: al,a2,a3,a4. at2: continuous. at3: a0,al,a2,a3,a4.

Žiadne komentáre ani prázdne riadky nie sú povolené. Poznámky môžete umiestnit do samostatného súboru dataset.info.

dataset.data, dataset.names, dataset.info musia byt umiestnené v tom istom adresári. Pozn. 1: Pri použití automatickej konverzie Weky by príklad vyzeral takto (uložené ako súbor CSV):

atl,at2,at3,class a2,39,a4,c2 a4,30,al,c2 al,9,a2,cl

Pozn. 2: Ak obsahuje váš dataset stĺpec unikátnych hodnôt (id), odstráňte ho.

Algoritmy

Klasifikačné algoritmy nájdete vo Weke na záložke classify:

- NB - Naive Bayes
- IB1
- IB3 - IBk (pre k = 3, nastavit ako parameter KNN)
- JRip
- DS - DecisionStump
- J48
- PART
- SMO
- MLP - Multilayer Perceptron