

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/>

Datasey:

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

UČO	DATASEY
256699	Breast Cancer Wisconsin (Original) Data Set
374595	Balance Scale Data Set
395607	Blood Transfusion Service Center Data Set
410345	Census Income Data Set
374368	Dermatology Data Set
359305	Echo-cardiogram Data Set
356530	Flags Data Set
359185	Heart Disease Data Set
325154	Hepatitis Data Set
417148	Horse Colic Data Set
395986	Hill-Valley Data Set
106451	Ionosphere Data Set
359747	Internet Advertisements Data Set
396273	Japanese Credit Screening Data Set
395881	Libras Movement Data Set
396477	Lung Cancer Data Set
359789	Mammographic Mass Data Set
72665	Parkinsons Data Set
207622	Soybean (Large) Data Set
359331	SPECT Heart Data Set
256365	Tic-Tac-Toe Endgame Data Set
359226	Wall-Following Robot Navigation Data
359860	Wine Data Set
324751	Wine Quality Data Set
255651	Yeast Data Set
409930	Chess (King-Rook vs. King) Data Set
324709	Pima Indians Diabetes Data Set
113869	Car Evaluation Data Set
255838	Chess (King-Rook vs. King-Pawn) Data Set
333617	Lymphography Data Set
388433	Mushroom Data Set

Zadanie:

1. Stiahnite si pridelené datasety. Ak sa tam nachádza viac datasetov, vyberte si jeden ľubovoľný (nie príliš veľký, ani nie malý). Súbor má 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.
2. 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 potrebné 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 používať tento typ konverzie, tak aj v tomto prípade sa predpokladá, že odovzdáte súbor *.names.
3. Klasifikáciu datasetu vykonajte s defaultným nastavením parametrov klasifikátora. Na testovanie použite cross-validation (10).
4. Po ukončení výpočtu uložte celý výstup do súboru dataset_algoritmus.log, kde algoritmus $\in \{NB, IB1, IB3, JRip, DS, J48, SMO, PART, MLP\}$. Do prehľadnej tabuľky zapíšte v skratke výsledky. Tabuľka by mala mať nasledovný formát:

Algoritmu	Accuracy	Weig. Avg Precision	Weig. Avg. Recall	Čas výpočtu
J48	xx.x	xx.x	xx.x	xx

5. Vytvorte súbor unexpected.txt a zaznamenajte do neho poznámky o netypickom priebehu, ak napríklad algoritmus nedobehne, alebo o prípadných dodatočných úpravách dát (odstránenie ID atribútu..).
6. Pre najlepší algoritmus ktorý správne klasifikoval najviac inštancií (max accuracy) vyskúšajte aj beh s inými vstupnými parametrami a snažte sa ešte zvýšiť 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: cislo_dataset_algoritmus.log kde cislo označuje poradie.
7. Vypracovanú úlohu (dataset.names, dataset.data, 9x dataset_algoritmus.log, tabuľka s výsledkami, unexpected.txt, 3x cislo_dataset_algoritmus.log) odovzdajte do Odevzdávnary zazipované v jednom súbore do 10.4.2012 23:59.
8. Súbor prosím odovzdávajte v plain-texte v kódovaní UTF-8 (Windows defaultne používa cp1250/2) alebo vo formáte PDF.
9. Informácie o splnení úlohy vám zadám do poznámkového bloku.
10. V prípade nespĺnenia úlohy vám budem nútený zadať mínusové body, ktoré sa vám odpočítajú od bodov získaných v záverečnej skúške.
11. 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 (nenumrická 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,a1,c2

a1,9,a2,c1

dataset.names – popis atribútov a ich hodnôt. Prvý riadok obsahuje zoznam možných hodnôt triedy oddelených čiarkou a ukončených bodkou. Tieto musia korešpondovať s poslednými hodnotami na riadkoch v súbore dataset.data. Všetky riadky obsahujú popis atribútov v poradí, v akom sa nachádzajú v dataset.data. Chýbajúca hodnota (?) sa do výčtu nezahŕňa. Popis atribútov je nasledujúci:

meno_atribútu: [continuous | CSV zoznam hodnôt pri ordinálnom atribúte].

Napr.

c1,c2.

at1: a1,a2,a3,a4.

at2: continuous.

at3: a0,a1,a2,a3,a4.

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

dataset.data, dataset.names, dataset.info musia byť umiestnené v tom istom adresári.

Pozn. 1: Pri použití automatickej konverzie Weky by príklad vyzeral takto (uložené ako súbor CSV):

at1,at2,at3,class

a2,39,a4,c2

a4,30,a1,c2

a1,9,a2,c1

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 $k = 3$, nastaviť ako parameter KNN)

- JRip
- DS - DecisionStump
- J48
- PART
- SMO
- MLP - Multilayer Perceptron