Latvijas Universitāte
Datorikas fakultāte
Maģistra studiju programma
Datizraces algoritmi
Jānis Ratnieks
st. apl. nr. jr09103
3. mājas darbs

1) Uzrakstiet vismaz 1000 zīmes garu sacerējumu par Šennona entropiju.

Būtībā lekciju materiālos viss ir pateikts un lielā mērā šis būs atkārtojums ar nelieliem papildinājumiem.

Entropija ir sistēmas nesakārtotības mērs, fizikā noslēgtas sistēmas entropija pieaug līdz tā ir ieņēmusi termodinamisko līdzsvaru – tas nozīmē, ka šajā stāvoklī tai ir visvairāk iespējamo apakšlīmeņu (mikrostāvokļi), kuriem ir vienāda iestāšanās varbūtība. Apskatot kāda diskrēta gadījuma lieluma iestāšanās varbūtību (ja visas iespējamās varbūtības ir vienādas) to var aprakstīt kā $p_i=\frac{1}{n}$, bet visiem iespējamiem stāvokļiem $\sum_{i=1}^{n} p_i=1$. Šeit jāpiebilst, ka iepriekšējā izteiksmē nav obligāti, lai p_i katram stāvoklim būtu vienāda, jo stāvokļa iestāšanās varbūtība vienmēr būs 1, izņemot tad, ja vis p_i ir 0, piemēram, metot parasto metamo kauliņu iespēja uzmest 7, 8 vai 9 ir 0 katram un līdz ar to kopējā varbūtība ir 0.

Apskatot divus neatkarīgus rezultātus, kuru varbūtības ir 1/n un 1/m to kopējā varbūtība ir 1/nm, respektīvi $\frac{1}{nm}=\frac{1}{n}\frac{1}{m}$, bet ērtāk būtu, ja šīs varbūtības varētu summēt. Šādu iespēju nodrošina logaritmi, jo $\log(nm)=\log(n)+\log(m)$. Šādi tad arī tiek definēta Šenona entropija $H=-\sum_{i=1}^{n} p_i\log(p_i)$ un ar to var apreķināt kāda lieluma vidējo vērtību $E=-\sum_{i=1}^{n} p_i Y_i$.

Šeit jāpiemin, ka vislielākā entropijas vērtība būs, ja visi p_i būs vienādi ar 1/n un tas nozīmē, ka ir vienāda iespēja iestāties kādam stāvoklim, bet tas savukārt nozīmē, ka tas ir tikai un vienīgi troksnis, līdzīgi kā fizikā noslēgta sistēma ir termodinamiskā līdzsvarā. Ja pa sakaru kanālu tiek raidīta kāda informācija, tad šajā sistēmā visi p_i nebūs 1/n, jo dažiem stāvokļiem būs lielāka varbūtība un līdz ar to entropija būs mazāka. Šādi var prognozēt, cik līmeņos var sūtīt signālu, lai galā būtu skaidri zināms, kas tika sūtīts.

2) Atveriet WEKA failu *soybean.arff* uzģenerējiet ar J48 *default* parametriem lēmumu koku. Ko šī faila dati īsti nozīmē? Palasiet lekcijas slaidus un pamēģiniet atrast parametru kombināciju, kas dotu mazāka izmēra lēmumu koku, saglabājot to pašu 90% precizitāti, ko dod *default* parametri. Ja nesanāk, izmēģiniet kādu citu algoritmu. Kā parasti, WEKA vietā varat izmantot citu programmatūru.

Datu pēdējais atribūts ir slimību nosaukumi un cik daudz no pupu stādiem ir saslimuši ar attiecīgo slimību. Slimībām acīmredzot ir pazīmes (simptomi), kuri konkrētās slimības gadījumā var vai nevar izpausties. Acīmredzot uzdevums ir noteikt slimību izmantojot simptomus.
Ar noklusējuma iestatījumiem un 66% treniņdatu precizitāte sanāk 90.5%:
Number of Leaves : 61

Size of the tree : 93

Time taken to build model: 0.02 seconds

=== Evaluation on test split ===
=== Summary ===

Correctly Classified Instances	210	90.5172 %
Incorrectly Classified Instances	22	9.4828 %
Kappa statistic	0.8954	
Mean absolute error	0.0155	
Root mean squared error	0.0929	
Relative absolute error	16.1763 %	
Root relative squared error	42.4372 %	
Total Number of Instances	232	

Ja lieto *cross-validation*, tad precizitāte pieaug līdz 91.5%:

Number of Leaves : 61

Size of the tree : 93

Time taken to build model: 0.08 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances	625	91.5081 %
Incorrectly Classified Instances	58	8.4919 %
Kappa statistic	0.9068	
Mean absolute error	0.0135	
Root mean squared error	0.0842	
Relative absolute error	14.0484 %	
Root relative squared error	38.4134 %	
Total Number of Instances	683	

Ja apskatās uz datiem rūpīgāk, tad var redzēt, ka visas vērtības ir nominālas no kurām dažas ir ar vairāk kā divām vērtībām. Šajā gadījumā var iestatīt *binarySplit=True*, kas nozīmē, ka veidojot atzaru, viena vērtība var tikt nošķirta no pārējām, ja tā ir True, tad pārējās būs False. Ja šīs parametrs ir False, tad katrai no nominālajām vērtībām ir savs atzars. Koka izmēru tas samazina un dod arī vislielāko precizitāti 92.2% (lietojot 66% treniņdatu):

Number of Leaves : 45

Size of the tree : 89

Time taken to build model: 0.07 seconds

=== Evaluation on test split ===
=== Summary ===

Correctly Classified Instances	214	92.2414 %
Incorrectly Classified Instances	18	7.7586 %
Kappa statistic	0.9146	
Mean absolute error	0.0127	
Root mean squared error	0.0876	
Relative absolute error	13.1604 %	
Root relative squared error	40.0341 %	
Total Number of Instances	232	