

Elyas Addawe – Saku Ihalainen – Hanna Kaimo

Hinkunuha

Tehtävä n:o 4

Metropolia Ammattikorkeakoulu
Tieto- ja viestintätekniikka
Datan käsitteily ja koneoppiminen
29.3.2021

1 Tavoitteet

Hinkunuha on fiktiivinen, mutta äärimmäisen kiusallinen sairaus. Suoraa testiä taudin toteamiseksi ei ole, mutta ihmisen sairastumisen siihen on huomattu mahdollisesti liittyvän kolmeen veressä olevaan aineeseen M1, M2 ja M3. Ei olla pystytty todistamaan kuinka paljon sairastuminen riippuu kunkin aineen ilmentymisestä veressä. Pyrimme rakentamaan testin, jolla ihmisen sairastumisen hinkunuhaan voisi todeta mittaamalla M1, M2 ja M3- aineiden määrän ihmisen veressä. Testin tuloksena tulisi arvio, onko ihmisellä hinkunuha vai ei.

Onko tämän testin rakentaminen mahdollista ja kuinka luotettava testi olisi? Tuleeko virheellisiä tuloksia ja minkälaisia ne mahdollisesti ovat?

2 Data

Data on fiktiivinen ja siinä kuvataan hinkunuha nimistä sairautta.

Attribuutit:

M1: numeerinen muuttuja, vaihteluväli 0-1

M2: numeerinen muuttuja, vaihteluväli 0-1.5

M3: numeerinen muuttuja, vaihteluväli 0-0.5

Disease: numeerinen muuttuja, vaihteluväli 0-1, vain luvut 0= henkilöllä ei ole hinkunuhaa tai 1= henkilöllä on hinkunuha

Datan koko on 723 havaintoa. Datassa ei ole muuten virheitä.

3 Datan valmistelu

Alkuperäisessä datassa oli sarakkeiden väleinä puolipisteet ja numeroiden desimaalit oli merkitty pilkuilla. Weka odottaa desimaalilukujen sisältävän pisteen eikä pilkkua. Samalla sarakkeiden erottamiseen ei saa käyttää puolipisteitä.

Avasimme gawk ohjelmiston ja teimme komennon: gawk
 'gsub(",",".")gsub(";",","){print \$0}' < hinkunuha.csv >hinkunuhaF.csv".

Muunsin pilkut pisteiksi minkä jälkeen muutin puolipisteet pilkuiksi. Siirsin syntyneen datan uuteen tiedostoon, joka on oikeassa muodossa Wekan käyttöä varten. Tämän jälkeen poistin ylimääräisen ID-sarakkeen mikä ei hyödyttänyt meitä. Wekassa muutimme disease-muuttujan nominaaliksi.

4 Mallinnus

Käytimme datan analysoinissa Wekan KNN-algoritmia. Wekassa algoritmi on nimeltään weka/classifiers/lazy/IBk. kNN- algoritmia on käytetty jo 1950- luvulta lähtien. Kaikkien attribuuttien painoarvo datassa on yhtä suuri tässä algoritmissa.

Alkutilanne hyperparametri K on 1. Käytämme Wekan "use training set"
rakentaaksemme opetusjoukon.

==== Summary ===				
Correctly Classified Instances	723	100	%	
Incorrectly Classified Instances	0	0	%	
Kappa statistic	1			
Mean absolute error	0.0018			
Root mean squared error	0.0019			
Relative absolute error	0.5511 %			
Root relative squared error	0.4775 %			
Total Number of Instances	723			

Meidän tarkkuus opetusjoukolle on 100%. Tuloksissa ei ilmaantunut yhtään virheitä.

Seuraavaksi testaamme testijoukkoa ylioppimisen varalta.

Cross validation on 10 ja k hyperparametri on 1.

== Summary ==									
Correctly Classified Instances	557								
Incorrectly Classified Instances	166								
Kappa statistic	0.54								
Mean absolute error	0.1544								
Root mean squared error	0.3903								
Relative absolute error	46.3236 %								
Root relative squared error	95.6895 %								
Total Number of Instances	723								
== Detailed Accuracy By Class ==									
	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	?	0,000	?	?	?	?	?	?	Disease
	0,782	0,242	0,778	0,782	0,780	0,540	0,762	0,722	0
	0,758	0,218	0,762	0,758	0,760	0,540	0,762	0,700	1
Weighted Avg.	0,770	0,231	0,770	0,770	0,770	0,540	0,762	0,711	

Testien tarkkuus on 77% ja opetusjoukon on 100%. Muunnamme k-parametria nostaaaksemme tarkkuutta mahdollisimman korkealle.

Laitoimme k:n arvoksi 4 ja katsomme tuloksia.

== Summary ==									
Correctly Classified Instances	626								
Incorrectly Classified Instances	97								
Kappa statistic	0.73								
Mean absolute error	0.1216								
Root mean squared error	0.2451								
Relative absolute error	36.4849 %								
Root relative squared error	60.0742 %								
Total Number of Instances	723								
== Detailed Accuracy By Class ==									
	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	?	0,000	?	?	?	?	?	?	Disease
	0,926	0,199	0,835	0,926	0,878	0,735	0,947	0,932	0
	0,801	0,074	0,908	0,801	0,851	0,735	0,947	0,926	1
Weighted Avg.	0,866	0,139	0,870	0,866	0,865	0,735	0,947	0,929	

Opetusjoukon tarkkuus laski huomattavasti ja nyt tarkkuus on noin 86.58%

Seuraavaksi testaamme testijoukkoa.

Cross validation 10 ja k:n arvo 4.

==== Summary ====								
Correctly Classified Instances	602							
Incorrectly Classified Instances	121							
Kappa statistic	0.6631							
Mean absolute error	0.1598							
Root mean squared error	0.3115							
Relative absolute error	47.9573 %							
Root relative squared error	76.3528 %							
Total Number of Instances	723							
==== Detailed Accuracy By Class ====								
TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
?	0,000	?	?	?	?	?	?	Disease
0,896	0,236	0,804	0,896	0,848	0,668	0,858	0,822	0
0,764	0,104	0,872	0,764	0,814	0,668	0,859	0,825	1
Weighted Avg.	0,833	0,173	0,837	0,833	0,832	0,668	0,859	0,824

Testien tarkkuus on noin 83,2% joka on huomattava ja parempi muutos edellisestä.

K parametri on nyt 8.

==== Summary ====								
Correctly Classified Instances	625							
Incorrectly Classified Instances	98							
Kappa statistic	0.7275							
Mean absolute error	0.1435							
Root mean squared error	0.2635							
Relative absolute error	43.0717 %							
Root relative squared error	64.5892 %							
Total Number of Instances	723							
==== Detailed Accuracy By Class ====								
TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
?	0,000	?	?	?	?	?	?	Disease
0,912	0,187	0,841	0,912	0,875	0,730	0,926	0,918	0
0,813	0,088	0,895	0,813	0,852	0,730	0,926	0,902	1
Weighted Avg.	0,864	0,140	0,867	0,864	0,864	0,730	0,926	0,910

Opetusjoukon tarkkuus on noin 86.4%.

Seuraavaksi testaamme testijoukkoa.

Cross validation 10 ja k:n arvo 8.

==== Summary ====								
Correctly Classified Instances	621							
Incorrectly Classified Instances	102							
Kappa statistic	0.7164							
Mean absolute error	0.1641							
Root mean squared error	0.2952							
Relative absolute error	49.2214 %							
Root relative squared error	72.3728 %							
Total Number of Instances	723							
==== Detailed Accuracy By Class ====								
TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
?	0,000	?	?	?	?	?	?	Disease
0,904	0,190	0,837	0,904	0,870	0,719	0,872	0,852	0
0,810	0,096	0,886	0,810	0,846	0,719	0,873	0,836	1
Weighted Avg.	0,859	0,145	0,861	0,859	0,858	0,719	0,873	0,844

Testien tarkkuus on nyt 85.89%.

K parametri on nyt 12.

==== Summary ====								
Correctly Classified Instances	627							
Incorrectly Classified Instances	96							
Kappa statistic	0.7331							
Mean absolute error	0.153							
Root mean squared error	0.2665							
Relative absolute error	45.9142 %							
Root relative squared error	65.3289 %							
Total Number of Instances	723							
==== Detailed Accuracy By Class ====								
TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
?	0,000	?	?	?	?	?	?	Disease
0,912	0,182	0,845	0,912	0,877	0,736	0,920	0,917	0
0,818	0,088	0,896	0,818	0,855	0,736	0,920	0,894	1
Weighted Avg.	0,867	0,137	0,869	0,867	0,867	0,736	0,920	0,906

Opetusjoukon tarkkuus noin 86.7%

Cross validation 10 ja k:n arvo 12.

==== Summary ===									
Correctly Classified Instances	616		85.2006 %						
Incorrectly Classified Instances	107		14.7994 %						
Kappa statistic	0.7026								
Mean absolute error	0.1688								
Root mean squared error	0.291								
Relative absolute error	50.634 %								
Root relative squared error	71.3268 %								
Total Number of Instances	723								
==== Detailed Accuracy By Class ===									
	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	?	0,000	?	?	?	?	?	?	Disease
	0,894	0,193	0,834	0,894	0,863	0,705	0,878	0,863	0
	0,807	0,106	0,875	0,807	0,840	0,705	0,878	0,831	1
Weighted Avg.	0,852	0,151	0,854	0,852	0,852	0,705	0,878	0,847	

Testien tarkkuus on noin 85.2%. Tarkkuus laski hieman edellisestä.

5 Arviointi

K:n arvo 1 (default), opetusjoukon tarkkuus 100%, testi 77%.

K:n arvo 4, opetusjoukon tarkkuus 86.58%, testi 83,3%

K:n arvo 8, opetusjoukon tarkkuus 86.4%, testi 85.89%

K:n arvo 12, opetusjoukon tarkkuus 86.7%, testi 85.2%

Testien tarkkuus oli 77% ja opetusjoukon tarkkuus oli 100%, kun k:n arvo oli yksi. Kun k:n arvoa aloitettiin nostamaan, testien tarkkuus nousi huomattavasti. Opetusjoukon tarkkuus kuitenkin laski.

Optimaalinen numero k:n arvoksi on 8. Testien tarkkuus oli korkeimmillaan 85.89%.

Kun menee sitä ylemmäs tarkkuus laskee hieman.

6 Täytäntöönpano

Tavoitteena oli rakentaa malli, joka analysoi merkkiaineiden M1,M2 ja M3 suhdetta hinkunuhaan. Tarkoitus on antaa arvio sairastaako henkilö hinkunuhaa vai ei. Testin suorittaminen oli mahdollista käyttäen Wekan kNN-algoritmia. Ylioppimisen varalta

aineisto pitää cross validoida. Testi osasi ennustaa hinkunuhan 85.89% tarkkuudella. Virheellisiä tuloksia oli noin 14%. Virheelliset tulokset johtuivat pääosin "FP" false positive arvosta. Henkilötä oli luokiteltu väärin sairaaksi. Väärin luokiteltujen osuus oli 14%.