Test AD si Regresie

RD (beat) >>> Dorinel

Regularizarea regresiilor liniare și logistice presupune

Select one:

a.Selectarea acelor atribute ale datelor de intrare care sunt mai informative

b.Adăugarea unui termen în funcția de cost care penalizează ponderile cu valori mari

c.Creșterea dimensiunii setului de antrenare prin adăugarea de noi exemple

d.Transformarea spațiului de intrare într-unul mai complex

Care dintre următoarele reprezintă criterii de oprire pentru construcția arborelui de decizie cu algoritmul ID3. Alegeți opțiunile corecte dintre:

1. Toate exemplele din subset fac parte din aceeași clasă
2. Câștigul informațional ajunge să fie negativ
3. Nu mai există atribute valide cu care să creăm un nod test
4. Adâncimea maximă este atinsă
5. Câștigul informațional depășește un anumit prag

☐ Select one:
☐ a.1
☐ b.4 și 5
☑ ~~e.1, 3 și 4~~
☐ d.1 și 4
☐ e.1, 2 și 3

Care dintre următoarele reprezintă criterii de oprire pentru construcția arborelui de decizie cu algoritmul ID3. Alegeți opțiunile corecte dintre:

1. Toate exemplele din subset fac parte din aceeași clasă
2. Câștigul informațional ajunge să fie negativ
3. Nu mai există atribute valide cu care să creăm un nod test
4. Adâncimea maximă este atinsă
5. Câștigul informațional depășește un anumit prag

Select one:

a.

1, 2 și 3

b.

1 și 4

c.

4 și 5

d.

1, 3 și 4

e.

1

care e aici?

Arborii de decizie construiți cu algoritmul ID3 prezintă robustețe crescută la „outliers”? Care dintre răspunsuri este cel mai plauzibil?

Select one:

a.Nu, pentru că vom obține valori negative pentru câștigul informațional în cazul lor.

b.Da, pentru că unele atribute nu sunt luate în considerare.++

c.Nu, pentru că ele (outliers) influențează sub-arborii generați ca fii ai unui nod test.

d.**Da, întrucât acest lucru este garantat prin calculul entropiei.** zic eu ++

https://datascience.stackexchange.com/questions/37394/are-decision-trees-robust-to-outliers

---

Ce tip de regresie ar trebui folosita pentru o problema in care dorim sa impartim un set de date in 2 categorii? (presupunem ca dorim sa antrenam un algoritm pentru care avem setul de date de antrenare)

Select one:

a.Regresie liniara deoarece cele 2 categorii (din setul de antrenare) pot fi etichetate cu 0 si 1. In etapa de antrenare algoritmul va avea de gasit o dreapta care trece prin cele 2 puncte (0 si 1)

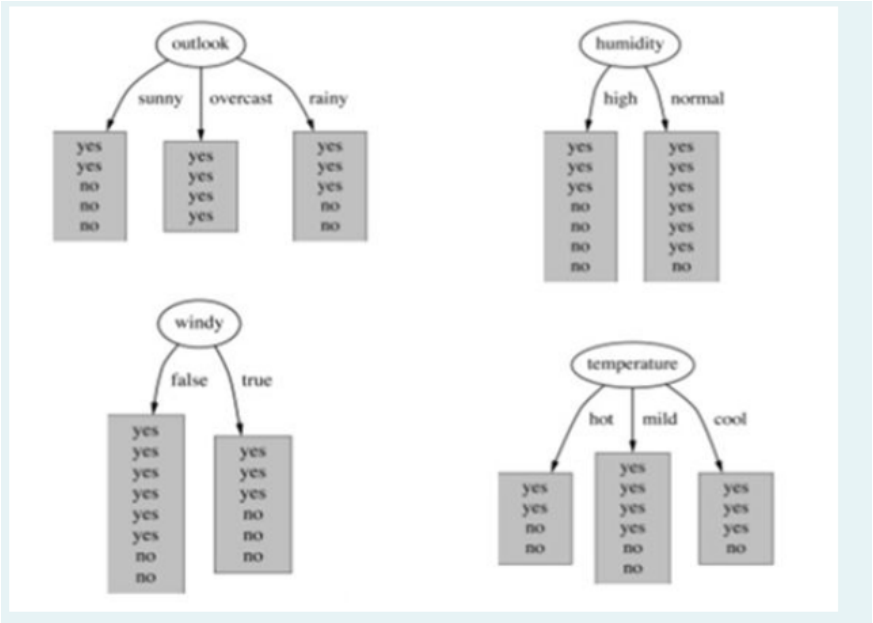
b.Regresie liniara deoarece datele pot fi separate cu ajutorul unei drepte

c.Regresie logistica deoarece nu stim cum sa codificam "categoriile" in nume reale necesare regresiei liniare.

d.Regresia logistica deoarece foloseste functia logistica pentru a modela probabilitatea ca un punct din setul de date sa se gaseasca intr-o anumita clasa.

Se antrenează un arbore de decizie având “maximum information gain” (câștigul maxim informațional, bazat pe diferenta de entropie) drept criteriu de selecție a atributelor.

Considerând acest criteriu și următoarea situație de antrenare (cf. imaginii), care va fi atributul selectat pentru a “împărți” nodul curent:



Select one:

a.Outlook (+?) ++

b.Temperature (?)

c.Humidity

d.Windy

Dorim să prezicem prețul unui apartament din zona Pipera. Pentru a face acest lucru ce putem folosi dintre:

Select one: **ziceti repede care e**

a.Regresie liniar**ă sau arbori de de**cizie ++++

b.Arbori de deciziecare di

c.Regresie logistică

d.Regresie liniară

e.Regresie liniară sau logistică

----------over----------

In cazul algoritmului C4.5 daca avem valori necunoscute pentru un atribut in anumie exemple de invatare trebuie sa

Select one:

a.atribuim acelui atribut valoarea cu cea mai mare frecventa din exemplele de invatare

b.atribuim acelui atribut valoarea cu cea mai mica frecventa din exemplele de invatare

---

Care dintre urmatoarele sunt bune practici pentru reducerea overfitting-ului?

(a) Utilizarea unei funcții de cost cu 2 componente, care include un regulator pentru a penaliza complexitatea modelului

(b) Utilizarea unui optimizator bun pentru a reduce erorile pe datele de antrenare

(c) Construirea unei structuri de subseturi de modele imbricate, antrenarea pe fiecare subset pornind de la cel mai mic, si oprirea cand eroarea de cross-validare incepe sa creasca

(d) Eliminarea aleatoare a 50% din datele de antrenare

Select one:

a.(a) si (c)

b.(a) si (b) si (c)

c.(b) si (c)

d.(a) si (b)

e.(c)

macar a si c cred ++

Ce presupune ”lama lui Occam” in contextul arborilor de decizie?

Select one:

a.Dacă sunt mai mulți arbori de decizie corecți, se preferă cel mai simplu

b.Un arbore corect de decizie trebuie să cuprindă toate atributele din setul de date

c.Nu există mai mulți arbori corecți pentru aceeași problemă

d.Nu se poate aplica principiul în contextul arborilor de decizie

In cazul algoritmului C4.5 daca avem valori necunoscute pentru un atribut in anumie exemple de invatare trebuie sa

Select one:

a.atribuim acelui atribut valoarea cu cea mai mica frecventa din exemplele de invatare

b.atribuim acelui atribut valoarea cu cea mai mare frecventa din exemplele de invatare

---

Intr-un arbore de decizie pentru clasificarea unor exemple care contin atribute cu valori continue, are sens sa repetam pe o anumita ramura a arborelui (cale de la radacina la o frunza) un acelasi atribut.

Select one:

True

False

Avem urmatoarea matrice de confuzie pentru o problema de clasificare binara (pe coloane sunt valorile reale, iar pe linii sunt valorile prezise).

Care afirmatie este adevarata?

	Positive	Negative
Positive	23	1
Negative	12	556

Select one:

a.

Accuracy=568/592

b.

Accuracy=557/556

c.

Recall=23/24

d.

Precision=23/24

Un set de date utilizat pentru învățare supervizată are o valoare mare a entropiei informaționale dacă:

Select one:

a.

Numărul de exemple din fiecare clasă este relativ similar

b.

Foarte multe dintre exemple aparțin unui număr mic de clase, în timp ce restul de clase au un număr mic de exemple. cred ca asta? asta pare

c.

Toate exemplele fac parte din aceeași clasă

Pipera duplicat

Ce presupune ”lama lui Occam” in contextul arborilor de decizie?

Select one:

a.

Dacă sunt mai mulți arbori de decizie corecți, se preferă cel mai simplu

b.

Nu există mai mulți arbori corecți pentru aceeași problemă

c.

Nu se poate aplica principiul în contextul arborilor de decizie

d.

Un arbore corect de decizie trebuie să cuprindă toate atributele din setul de date

Avem urmatoarea matrice de confuzie pentru o problema de clasificare binara (pe coloane sunt valorile reale, iar pe linii sunt valorile prezise).

Care afirmatie este adevarata?

	Positive	Negative
Positive	23	1
Negative	12	556

Select one:

a.

Accuracy=568/592

b.

Accuracy=557/556

c.

Recall=23/24

d.

Precision=23/24
