

Notițe Seminar 4

October 18, 2019

Învățare supervizată de tip clasificare (2)

Intro: În seminarul trecut v-am zis că există faza de antrenare și faza de testare, însă nu v-am zis întreaga poveste, mai există și o fază de validare...

Context: cineva vă dă un set de date despre apartamente cu atrbute de intrare și un atrbut de ieșire și vă cere să găsiți un model ML care să se comporte cât mai bine pe date noi și să furnizați un număr care să indice acest lucru.

Voi ce faceți? Vă faceți o listă cu mai mulți algoritmi (să zicem 15) (sau mai multe variante de algoritmi) pe care vreți să-i încercați (ID3, rețea neuronală, ID3 dar nu cu IG, ci cu index Gini etc) și apoi aveți cel puțin 3 opțiuni:

1. (a) antrenați 15 modele pe setul de date pe care îl aveți coresp. celor 15 alg.
 - (b) calculați acuratețea pentru fiecare model la antrenare
 - (c) selectați modelul cu acuratețea maximă și raportați acea acuratețe
 - (d) Chiar dacă modelul vostru se descurcă foarte bine pe setul de date de antrenare, acest lucru nu înseamnă că el se va descurca bine pe date noi. Cel mai probabil nu o va face. Cu alte cuvinte, ați dat peste **overfitting** (= obțineți un model care are la antrenare acuratețe mai mari decât la testare când testăm cu date noi = v-ați pliat prea mult pe datele de antrenament = ați *tocit* datele de antrenament)
 - (e) Concluzie: Nu este bine de procedat astfel.
2. (a) Să zicem că aveți 1000 de rânduri complete (input + output) despre apartamente (este locuibil?). Rândurile vor fi împărțite

în două: set de antrenare (să zicem 90% = 900 de rânduri) și set de testare (să zicem 10% = 100 de rânduri). Normal ar fi ca distribuția coloanei de ieșire să fie la fel la antrenare și la testare. De exemplu, dacă Y = (este locuibil?) este coloana de ieșire și apare cu frecvențele [700+,300-] în cele 1000 de rânduri, atunci în setul de antrenare ar fi bine să avem cam $[900 \cdot 70\% = 630+, 900 \cdot 30\% = 270-]=[630+,270-]$ drept frecvențe pentru Y , iar în setul de testare: cam $[100 \cdot 70\% = 70+, 100 \cdot 30\% = 30-]=[70+,30-]$. Dar deja vă zic prea multe...

- (b) antrenați 15 modele pe setul de date de antrenare coresp. celor 15 alg.
 - (c) calculați acuratețea pentru fiecare model pe setul de testare
 - (d) selectați modelul cu acuratețea maximă și raportați acea acuratețe
 - (e) Chiar dacă facem mai bine decât în primul caz, tot nu facem bine, pentru că acuratețea raportată este maximul celor 15 numere ce exprimă acuratețea și ne-am pliat, poate, prea mult (alegând acuratețea maximă) pe setul nostru de testare care va fi diferit de datele reale ce vor urma
 - (f) Concluzie: Nu este bine suficient de bine să procedăm astfel, pentru că numărul furnizat (acuratețea) nu reprezintă realitatea.
3. (a) Împărțiți, ca mai sus, cele 1000 de rânduri în 900 și 100.
- (b) Împărțiți cele 900 în două: să zicem 90% = 810 pentru **antrenare** și 10% = 90 pentru **validare**.
 - (c) Cele 100 de rânduri vor rămâne pentru **testare**.
 - (d) antrenați 15 modele pe setul de antrenare coresp. celor 15 alg.
 - (e) calculați acuratețea pentru fiecare model pe setul de validare
 - (f) selectați modelul cu acuratețea maximă
 - (g) antrenați cu algoritmul corespunzător modelului câștigător pe toate cele 900 de rânduri un nou model
 - (h) pentru acest model calculați acuratețea pe setul de testare și raportați această acuratețe
 - (i) Astfel scăpăm de cele două probleme de mai sus.

(j) Concluzie: Este bine să procedăm astfel.

Observații:

- i. Atunci când setul de date este mic (deci, și în cazul nostru, cel cu 1000 de rânduri), vom avea cele 100 de rânduri pentru testare, însă pentru a obține la validare o acuratețe din cele 900 de rânduri rămase putem face și altfel:

A. **cross-validation cu metoda k-fold:**

- împărțim cele 900 de rânduri în, să zicem, 10 bucăți de lungimi egale, deci de 90 de rânduri fiecare: b_1, b_2, \dots, b_{10}
- b_1 va juca rol de date de validare, iar restul bucăților împreună vor reprezenta datele de antrenament \Rightarrow acuratețe₁
- b_2 va juca rol de date de validare, iar restul bucăților împreună vor reprezenta datele de antrenament \Rightarrow acuratețe₂
- ...
- b_{10} va juca rol de date de validare, iar restul bucăților împreună vor reprezenta datele de antrenament \Rightarrow acuratețe₁₀
- pentru un algoritm din cei 15, vom avea o acuratețe la validare dată de acuratețe = $\frac{\text{acuratețe}_1 + \dots + \text{acuratețe}_{10}}{10}$
- astfel, calculați acuratețea la validare pentru fiecare din cei 15 alg.
- selectați algoritmul cu acuratețea maximă
- antrenați cu algoritmul câștigător pe toate cele 900 de rânduri un nou model
- pentru acest model calculați acuratețea pe setul de testare și raportați această acuratețe

B. sau **cross-validation cu metoda leave-one-out (CVLOO):**

este cross-validation cu metoda k-fold atunci când $k =$ numărul de rânduri pentru antrenare și validare = 900, în cazul nostru

Apare așa ceva prin exercițiile din carte? Da: trebuie să știți să calculați eroarea sau acuratețea la CVLOO pentru un algoritm/model dat. Veți considera că acele rânduri furnizate în astfel de exerciții sunt pentru antrenare și validare.

vezi ex. 66/pag. 360

vezi ex. 27/pag. 458

ID3 (2)

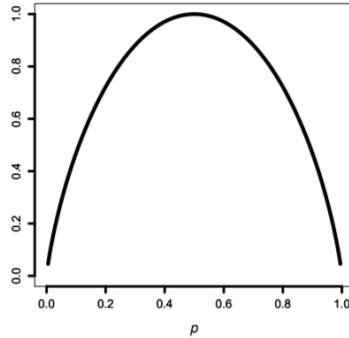
- Dacă setul de date de antrenare este consistent, atunci eroarea la antrenare produsă de modelul învățat de algoritmul ID3 este 0 (*overfitting*). Pentru a nu mai avea overfitting, putem *prosti* modelul/arboarele, prin trunchiere (*pruning*). *Nu voi intra în detaliile despre pruning pentru că momentan nu știu cât s-a insistat la curs pe acest aspect, dar ca idee vă recomand să citiți ex.20/pag.302.*
- Dacă setul de date de antrenare este inconsistent, atunci eroarea la antrenare produsă de modelul învățat de algoritmul ID3 este ... (vezi ex. 6c/pag. 274)
- Eroarea la CVLOO pentru ID3: vezi ex. 10/pag.279

ID3 (2) - calcule mai rapide

1. când vrem să selectăm un atribut de intrare pentru a-l plasa într-un nod, în loc să calculăm IG-uri, calculăm entropii condiționale medii (am discutat acest lucru seminarul trecut)
2. **putem să aplicăm formulele de la ex.33/pag.331: vezi filmuletele de pe grupul de facebook**
3. Raționamente calitative
 - (a)
 - $H[n+, n-] = 1$
 - $H[0+, n-] = H[n+, 0-] = 0$
 - (b)
 - Folosind simetria din graficul

Exemplification: Entropy of a Bernoulli Distribution

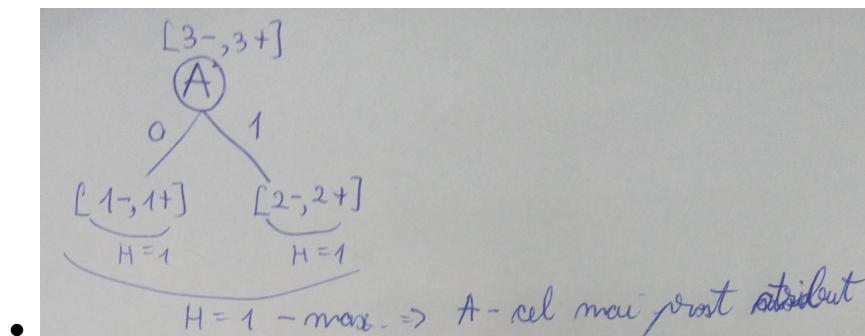
$$H(p) = -p \log_2 p - (1-p) \log_2(1-p)$$

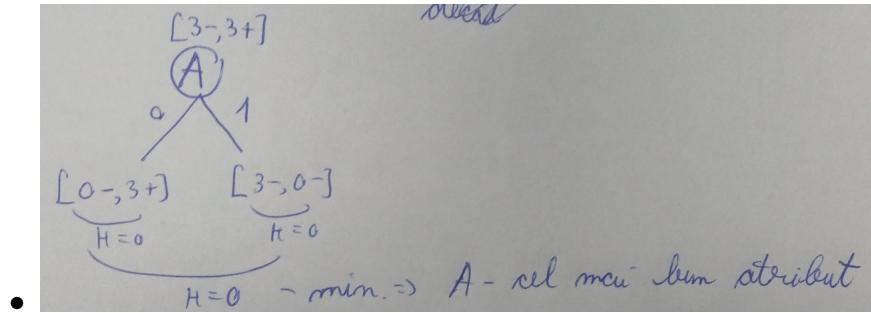


(slide preluat din <https://prof.s.info.uaic.ro/~ciortuz/SLIDES/foundations.pdf>)

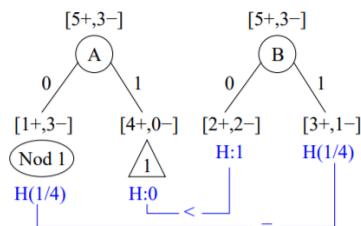
vom putea:

- spune că $H[5-, 3+] = H[5+, 3-]$
- compara $H[5-, 3+]$ cu $H[2-, 9+]$ fără a face calcule prea multe: vom compara $\frac{\min(5,3)}{5+3} = \frac{3}{8} = \frac{33}{88}$ cu $\frac{\min(2,9)}{2+9} = \frac{2}{11} = \frac{16}{88}$; cum $\frac{33}{88} > \frac{16}{88}$, vom avea: $H[5-, 3+] > H[2-, 9+]$
- la fel, putem compara $H[4-, 5+]$ cu $H[5-, 6+]$: $\frac{\min(4,5)}{4+5} = \frac{4}{9}$ cu $\frac{\min(5,6)}{5+6} = \frac{5}{11}$, adică $\frac{4}{9}$ cu $\frac{5}{11}$, adică $\frac{44}{99}$ cu $\frac{45}{99}$. Cum $\frac{44}{99} < \frac{45}{99}$, avem: $H[4-, 5+] < H[5-, 6+]$



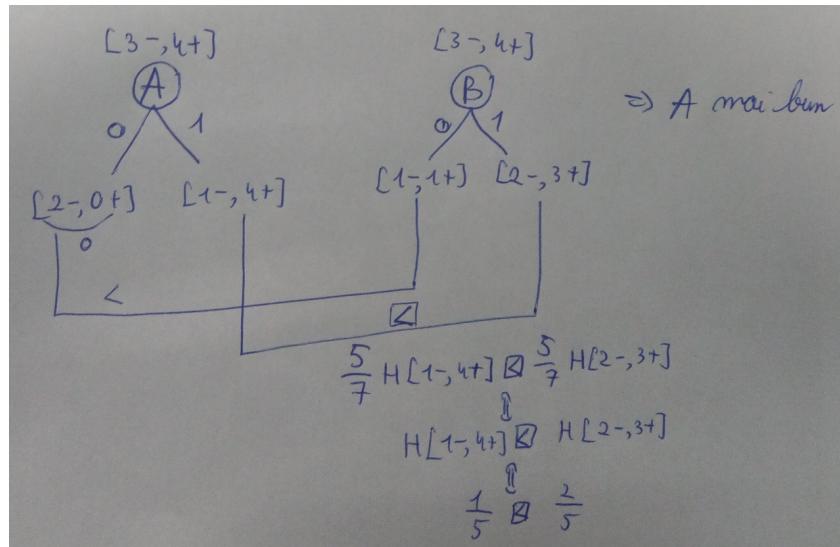


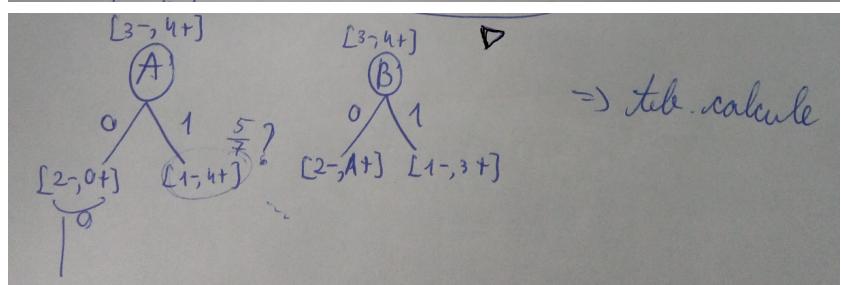
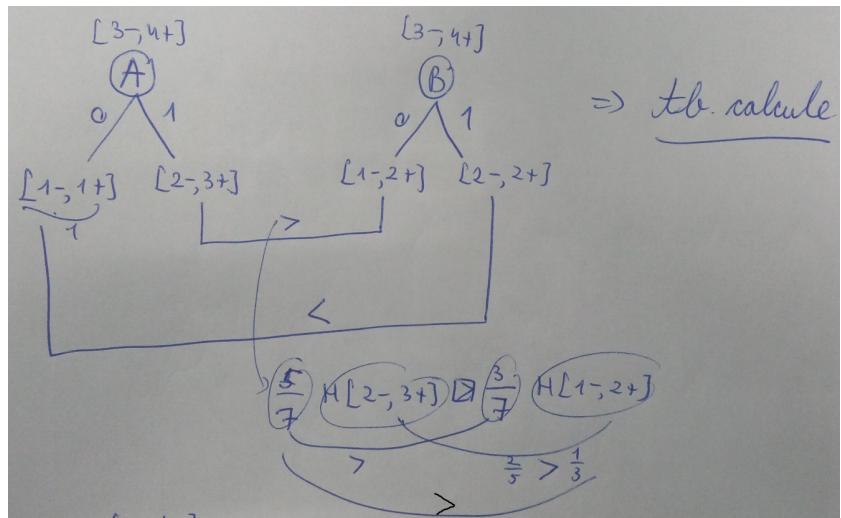
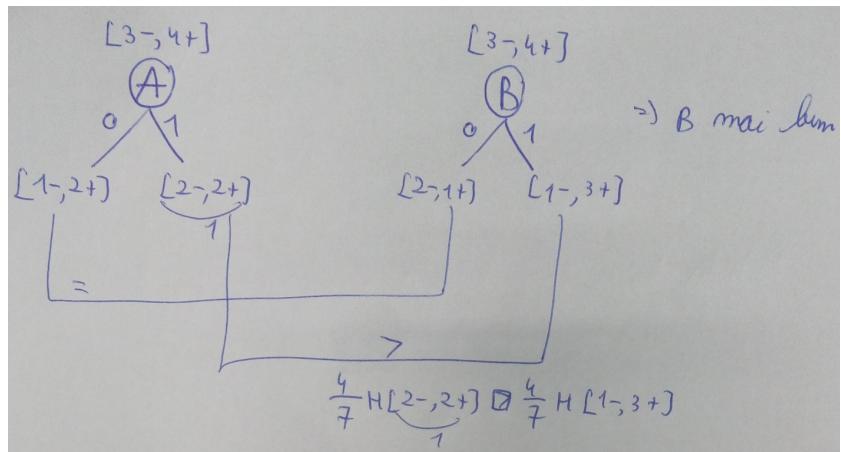
- La pagina 269 din culegere aveți undeavă:



Mai facem precizarea că semnele < și = din figura alăturată se referă de fapt nu [doar] la entropiile condiționale specifice, ci [și] la produsul acestora cu ponderile asociate în mod corespunzător: $\frac{4}{8}H[1+, 3-] = \frac{4}{8}H[3+, 1-]$ și respectiv $\frac{4}{8}H[4+, 0-] < \frac{4}{8}H[2+, 2-]$.

Veți putea compara compașii de decizie fără calcule în anumite situații, dar nu uitați că semnele de >, <, = se referă **nu doar la entropii, cât și la ponderi**. Vedeti și următoarele exemple:





Observație: rezolvări de tipul *mi se pare că nodul acesta va da IG-ul cel mai mare (pentru că are un nod frunză deja etc.)* nu se puntează. Ori faceți calcule, ori faceți un raționament calitativ valid (ca mai sus sau altele nemenționate aici: simetria pentru o formula booleană etc.).

Extensie ID3 - lucru cu atributे continue

vezi ex.10/pag.279

vezi ex.12/pag.282

Schemă de final

1. Învățare supervizată de tip clasificare (2)
 - (a) antrenare
 - (b) testare
 - (c) validare
 - (d) cross-validare cu metoda k-fold
 - (e) cross-validare cu metoda leave-one-out
2. ID3 (2)
 - (a) eroarea la antrenare pe set de date consistent/inconsistent
 - (b) eroarea la CVLOO
3. ID3 (2) - calcule mai rapide
 - (a) formule ex.33/pag.331 - vezi filmulete fb
 - (b) raționamente calitative
4. Extensie ID3 - lucru cu atrbute continue