Test partea I

November 2020

1 Probabilități și Statistică (1p)

1. (0.1p) Fie următorul experiment aleator: aruncarea a 2 zaruri. Stabiliți care sunt spațiul de eșantionare și spațiul de evenimente pentru acest experiment. [Nu este necesar să scrieți toate elementele din spațiul de evenimente, ci doar primele și apoi scrieți "...".]

2. (0.2p) Fie evenimentele aleatoare A, B, C cu $P(A,B) \neq 0$, $P(A,C) \neq 0$, $P(A) \neq 0$. Demonstrați că:

$$P(C|A,B) = \frac{P(B|A,C) * P(C|A)}{P(B|A)}$$

3. (0.1p) Presupunem că elevul A este ascultat cu probabilitate de 0.3 iar studentul B cu probabilitate de 0.5. Dacă selectarea studenților se face în mod independent care este probabilitatea ca:
   (a) măcar unul din ei este ascultat
   (b) A este ascultat dar B nu
   (c) ambii sunt ascultați

4. (0.3p) Fie trei variabile aleatoare X, Y și Z. În tabelul de mai jos este dată distribuția probabilistă comună a acestor trei variabile.

	Z=0		Z=1	
	X=0	X=1	X=0	X=1
Y=0	0.1	0.1	0.05	0.2
Y=1	0.3	0.1	0.1	0.05

   (a) (0.05p) Indicați $P(X=0,Y=0,Z=0)$.
   (b) (0.05p) Calculați $P(Z=0)$.
   (c) (0.1p) Calculați $P(X=0,Y=0|Z=0)$.
   (d) (0.1p) Sunt variabilele X și Y independente condițional față de Z?

5. (0.3p) Pe datele de la exercițiul anterior completați tabelul următor (distribuția probabilistă comună a lui X și Y):

	X=0	X=1
Y=0		
Y=1		

   (a) (0.1p) Calculați $E[x], E[y], Var(y), Cov(X,Y)$.
   (b) (0.1p) Calculați $H(X)$ (doar înlocuire în formulă, nu trebuie dus până la capăt).
   (c) (0.1p) Calculați $H(Y|X)$ (doar înlocuire în formulă, nu trebuie dus până la capăt).

2 ID3 (4p)

1. (2p) Considerăm următorul set de date:

running nose	coughing	reddened skin	fever	ill
+	+	+	–	+
+	+	–	–	+
–	–	+	+	+
+	–	–	–	+
–	–	–	–	-
–	+	+	–	-

$H(\frac{1}{4}) = 0.81127$, $H(\frac{1}{3}) = 0.91829$, $H(\frac{1}{5}) = 0.721928$, $H(\frac{2}{5}) = 0.97095$, $H(\frac{1}{2}) = 1$

$log_2(3) = 1.5849$, $log_2(5) = 2.32192$, $log_2(7) = 2.80735$, $log_2(11) = 3.45943$, $log_2(13) = 3.7004$, $log_2(17) = 4.08746$, $log_2(19) = 4.24792$

   (a) Construiți arborele ID3.

   (b) Este arborele consistent cu datele de antrenament?

   (c) Cum este clasificat un pacient cu coughing=+, fever=+, reddened skin=-, running nose=- ?

   (d) Exprimați cu ajutorul logicii predicatelor de ordin 0, clasificarea produsă de arborele de decizie.

2. (1p) Considerăm următorul set de date cu un singur atribut de intrare continuu (înălțime) și un atribut de ieșire (gen):

Height	Gender	Counts
161	F	1
164	F	2
170	M	2
174	M	2
174	F	1
176	F	1

$H(\frac{1}{4}) = 0.81127$, $H(\frac{1}{3}) = 0.91829$, $H(\frac{1}{5}) = 0.721928$, $H(\frac{2}{5}) = 0.97095$, $H(\frac{1}{2}) = 1$

$log_2(3) = 1.5849$, $log_2(5) = 2.32192$, $log_2(7) = 2.80735$, $log_2(11) = 3.45943$, $log_2(13) = 3.7004$, $log_2(17) = 4.08746$, $log_2(19) = 4.24792$

   (a) Este setul de date consistent?

   (b) Câte praguri distincte trebuie să considerăm pentru Height atunci când căutăm atributul (optim) care trebuie pus în rădăcină?

   (c) Construiți arborele ID3.

   (d) Calculați eroarea medie la antrenament.

3. (1p) În legătură cu algoritmul ID3 fără extensia legată de lucrul cu atribute de intrare continue, argumentați dacă următoarele afirmații sunt adevărate sau false:

   (a) Algoritmul ID3 garantează obținerea arborelui de decizie optimal (ca număr de niveluri sau noduri).

   (b) Dacă nu mai sunt atribute candidat de testat într-un nod (pentru că toate atributele de intrare apar deja pe drumul de la acel nod la rădăcină), algoritmul ID3 nu mai continuă pe acel drum.

   (c) Un atribut de intrare poate apărea în arborele produs de algoritmul ID3 de mai multe ori.

   (d) Dacă aplicăm ID3 și câștigul de informație al atributului de ieșire în raport cu un atribut de intrare X este 0, atunci X nu va apărea în arborele de decizie.

3 Bayes (3.5p)

1. (0.5p) Pe un aeroport pasagerii sunt testați foarte atent. Un terorist este arestat cu probabilitate de 0.98 iar o persoană nevinovată cu probabilitate 0.1. Știm că un pasager este terorist cu probabilitate de 0.001. Care este probabilitatea ca o persoană arestată să fie terorist?

2. (1p) Fie următorul set de date, unde A, B, Y sunt discrete:

A	B	Y
0	1	0
1	0	0
0	1	1
1	1	2
2	1	2

   (a) Estimați în sensul verosimilității maxime (MLE) $P(A = 0|Y = 1)$.

   (b) Care este decizia Bayes Naiv pentru instanța A=2, B=1. Cu ce probabilitate se ia decizia?

   (c) Ce problemă întâmpină clasificatorul Bayes Naiv? Aplicați tehnica de remediere prezentată la curs și raspundeți din nou la întrebarea precedentă.

   (d) Câți parametri trebuie să estimeze Bayes Naiv?

3. (1p) Fie următoarea distribuție comună a variabilelor aleatoare binare A, B, Y.

A	B	Y	P(a,b,c)
0	0	0	0.15
0	0	1	0.15
0	1	0	0.1
0	1	1	0.05
1	0	0	0.05
1	0	1	0.2
1	1	0	0.1
1	1	1	0.2

   (a) Cum clasifică Bayes Optimal instanța A=1, B=0? Cu ce probabilitate se ia această decizie?

   (b) Calculați rata medie a erorii pentru Bayes Optimal.

4. (1p) Se consideră variabilele aleatoare X_1, X_2, X_3 şi X_4. Aceste variabile sunt independente condiţional două câte două în raport cu variabila Y, cu excepţia perechii X_1, X_3.

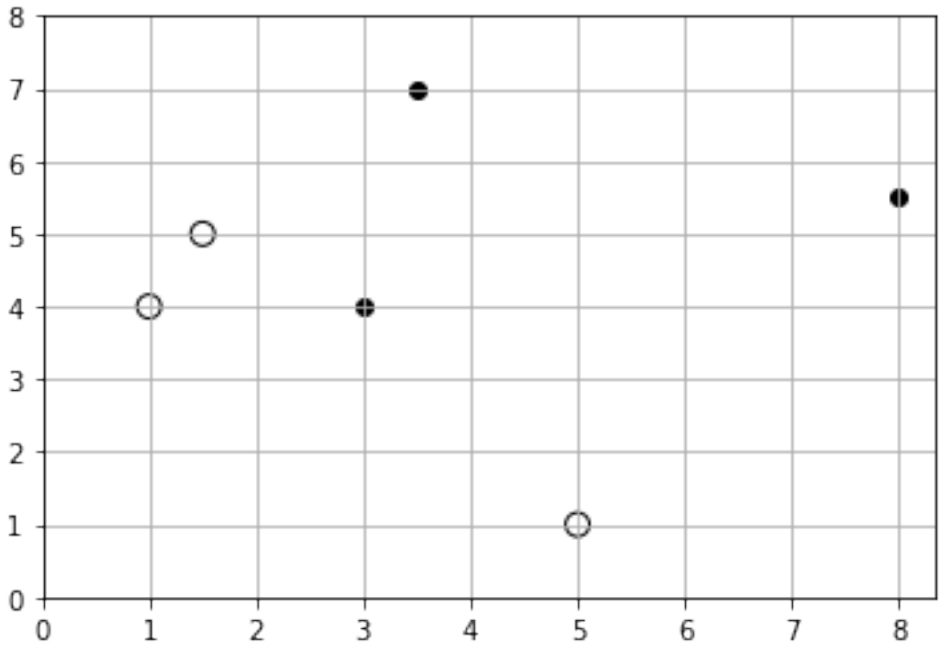
   (a) Modificați regula de decizie a algoritmului Bayes Naiv pentru a ţine cont de această particularitate a datelor.

   (b) Ce eroare ar avea noul algoritm?

   (c) Câți parametri ar trebui estimați?

4 k-NN (1.5p)

1. Fie următoarele puncte în planul 2d.



   (a) Trasați granițele de decizie pentru 1NN

   (b) Care este decizia algoritmului 3NN pentru punctul (2, 6) (vecinătățile se vor construi în manieră inclusivă). Dar cu ponderi? Dar folosind distanța Manhattan sau Cebîșev?

   (c) Calculați eroarea la CVLOO pentru 1NN, 5NN.