Test partea I

November 2020

1 Fundamente (1.5p)

1. (0.3p) Fie următorul experiment aleator: alegerea dintr-o punga dintre 3 bile: una roșie, una verde și una portocalie. Definiți două funcții de probabilitate peste spațiul de evenimente și demonstrați că respectă proprietățile unei asemenea funcții.

2. (0.2p) Fie evenimentele aleatoare A, B. Demonstrați că:

$$P(A) = P(A|B) * P(B) + P(A|\overline{B}) * P(\overline{B})$$

3. (0.5p) La un liceu 60% din elevi nu poartă nici inel nici lanț la gât. 20% poartă inel iar 30% poartă lanț. Care este probabilitatea ca un elev să poarte:

   (a) inel sau lanț

   (b) inel și lanț

4. (0.5p) Știind că X și Y sunt două variabile aleatoare discrete și independente. Demonstrați că

$$E[XY] = E[X] * E[Y]$$

5. (1p) 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=-1	0.1	0.2	0.1	0.1
Y=0	0	0.1	0	0
Y=1	0	0.1	0.2	0.1

   (a) (0.2p) Sunt variabilele X și Y independente?

   (b) (0.2p) Sunt Y și Z independente condițional față de X?

   (c) (0.25p) Calculați $E[Z], Var(Y), E[2XY]$.

   (d) (0.1p) Calculați $H(X)$ (doar înlocuire în formulă, nu trebuie dus până la capăt).

   (e) (0.1p) Calculați $H(X|Y = 0)$ (doar înlocuire în formulă, nu trebuie dus până la capăt).

   (f) (0.15p) Calculați $IG(X;Y)$ (doar înlocuire în formulă, nu trebuie dus până la capăt).

2 ID3 (3.5p)

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

X	Y	Z
0	0	0
0	1	1
1	0	1
1	1	0

   (a) (0.2p) Care este eroarea la antrenare a algoritmului ID3?

   (b) (0.7p) Care este eroarea la CVLOO?

   (c) (0.1p) În funcție de cele 2 erori ce fenomene putem observa la antrenarea ID3? Ne aflăm in vreunul din cazuri?

2. (2.5p) Considerăm următorul set de date:

running nose	coughing	fever	age	ill
+	-	+	20	-
+	–	–	20	–
–	+	+	30	+
–	+	–	30	–
+	–	+	40	+
–	-	-	40	+

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

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

   (a) (0.1p) Sunt datele consistente? Cât va fi eroarea la antrenament?

   (b) (1p) Construiți arborele ID3. Dacă la un anumit nivel există mai multe atribute cu același câștig de informație, alegeți în ordinea apariției în tabel (de la stânga la dreapta, sau crescător pentru atributul continuu)

   (c) (0.2p) Exprimați cu ajutorul logicii predicatelor de ordin 0, clasificarea produsă de arborele de decizie.

   (d) (0.1p) Cum este clasificat un pacient cu fever=+, age=32, running nose=-, coughing=+ ?

   (e) (1p) Aplicați post-pruning bottom-up bazat pe *gain ratio impurity*. Pentru un atribut de test A valoarea măsurii este:

$$\frac{H(Y) - H(Y|A)}{H(A)}$$

   unde Y este variabila de ieșire. Tăiați toate nodurile de test care au *gain ratio impurity* mai mare de 0.52.

   (f) (0.1p) Cât este eroarea medie la antrenament a noului arborelui?

3 Bayes (1p)

1. (0.5p) Fie următorul set de date, unde A, B, Y sunt discrete:

A	B	C	Y
0	0	0	0
1	0	1	1
0	1	1	1
0	1	0	1
1	1	1	1

   (a) (0.1p) Estimați în sensul verosimilității maxime (MLE) $P(A = 0|Y = 1)$.

   (b) (0.2p) Care este decizia Bayes Naiv pentru instanța A=0, B=1, C=1. Cu ce probabilitate se ia decizia?

   (c) (0.2p) Ce problemă întâmpină clasificatorul Bayes Naiv? Aplicați tehnica de remediere prezentată la curs și raspundeți din nou la întrebarea precedentă.

2. (0.5p) Fie următoarea distribuție comună a variabilelor aleatoare binare A, B, Y.

A	B	Y	P(a,b,c)
0	0	0	0.3
0	0	1	0.1
0	1	0	0.1
0	1	1	0.05
1	0	0	0.15
1	0	1	0.1
1	1	0	0.1
1	1	1	0.1

   (a) (0.2p) Cum clasifică Bayes Naiv instanța A=1, B=0? Cu ce probabilitate se ia această decizie?

   (b) (0.2p) Cum clasifică Bayes Optimal instanța A=1, B=0? Cu ce probabilitate se ia această decizie?

   (c) (0.1p) De ce exista diferențe? Demonstrați riguros.