חלק 2 [50 נקודות]

- יש להציג את כל התוצאות עם **שלוש ספרות אחרי נקודה עשרונית** אלא אם צוין אחרת!
- יש לרשום את כל התשובות **על-גבי שאלון הבחינה בלבד**
- טיוטות החישוב ייגרסו **ללא בדיקה**

נתונות 20 תצפיות מתוך מאגר הנתונים HIGGS Data Set המכיל תוצאות סימולציה של ניסוי לזיהוי "החלקיק האלוהי" (בוזון היגס). התכונה הרציפה m_bb מסייעת לפיסיקאים להבדיל בין בוזון היגס (סיווג = 1) ליתר חלקיקים (סיווג = 0).

Record_ID	0	1	2	3	4	5	6	7	8	9
Class	1	0	1	0	0	1	1	1	0	1
m_bb	0.9	0.4	0.2	0.7	1.1	1.5	0.8	0.4	1.1	0.9

Record_ID	10	11	12	13	14	15	16	17	18	19
Class	0	0	1	0	1	0	1	1	0	0
m_bb	2	0.7	0.9	0.3	1.1	0.4	0.7	0.9	1.3	3.4

א. יש לחשב את הרווח האינפורמטיבי (Information Gain), מדד ה-Gini ומדד ה-twoing של משתנה המטרה "Class" עבור נקודת הפיצול הבאה של המשתנה m_bb : **0.7**. **30 נקודות**.

Interval	Prob. (Interval)	Prob. (0)	Prob. (1)	Entropy	Information Gain	Gini Index	Gini Drop	Twoing
<= 0.7								
> 0.7								
Total								

ב. יש לבנות רווח בר-סמך לדיוק העץ המתקבל כתוצאה מהפיצול הנ"ל ברמת-ביטחון של 95%. **10 נקודות**

פירוט החישוב (**חובה**) :

ג. יש לחשב את הערך המינימלי של מקדם הסיבוכיות α עבורו כדאי לגזום את העץ הנ"ל. **5 נקודות.**

פירוט החישוב (**חובה**) :

ד. יש לבדוק האם אי-שוויון Fano מתקיים עבור העץ שבניתם בסעיף א'. **5 נקודות.**

פירוט החישוב (**חובה**) :

בהצלחה!

דף הנוסחאות

Information Theory

- Entropy $H(X) = \sum -p(x)log_2 p(x)$ Conditional Entropy $H(Y/X) = - \Sigma\ p(x, y)*log\ p\ (y/x)$
- Mutual Information $I(X;Y) = H(Y) - H(Y/X) = \sum_{x,y} p(x,y) \bullet \log \frac{p(y/x)}{p(y)}$
- Conditional Mutual Information: $I(X;Y/Z) = H(X/Z) - H(X/Y,Z) = \sum_{x,y} p(x,y,z) \bullet \log \frac{p(x,y/z)}{p(x/z) \bullet p(y/z)}$
- Fano's Inequality: $H\ (Y/\ X_1 \ldots X_n) \leq H\ (P_e) + P_e\ log_2\ (m-1)$

Decision Trees

- Confidence Interval for an Error Rate: $Err_{Test} \pm z_\alpha \sqrt{\frac{Err_{Test}(1-Err_{Test})}{n}}$
- Confidence Interval for a difference between error rates: $\hat{d} \pm z_\alpha \sqrt{\frac{Err_{Test1}(1-Err_{Test1})}{n_1} + \frac{Err_{Test2}(1-Err_{Test2})}{n_2}}$
- Expected information needed to classify a tuple in D (before using A): $Info(D) = -\sum_{i=1}^{m} p_i \log_2(p_i)$
  - Expected information needed to classify a tuple in D (after using A): $Info_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times I(D_j)$
- Information Gain: $Gain(A) = Info(D) - Info_A(D)$
- Chi-Square Statistic: $\sum_{j=1}^{c} \sum_{i=1}^{v} \frac{(o_{ij} - e'_{ij})^2}{e'_{ij}} \Big|_{H_0} \sim \chi_\alpha^2((v-1)(c-1))$
- Apparent (pessimistic) error rate: $q = \frac{N - n_C + 0.5}{N}$
- Entropy induced by threshold T: $E(A,T;S) = \frac{|S_1|}{|S|} Ent(S_1) + \frac{|S_2|}{|S|} Ent(S_2)$
- Split Information: $SplitInfo_A(D) = -\sum_{j=1}^{v} \frac{|D_j|}{|D|} \times \log_2(\frac{|D_j|}{|D|})$
- Gini Index: $gini(T) = 1 - \sum_{j=1}^{n} p_j^2$
- Twoing Splitting Rule: $\frac{p_L p_R}{4} \left[\sum_j |p(j/t_L) - p(j/t_R)| \right]^2$
- Cost-complexity function (CART): $R_\alpha(T) = R(T) + \alpha \cdot |\tilde{T}|$