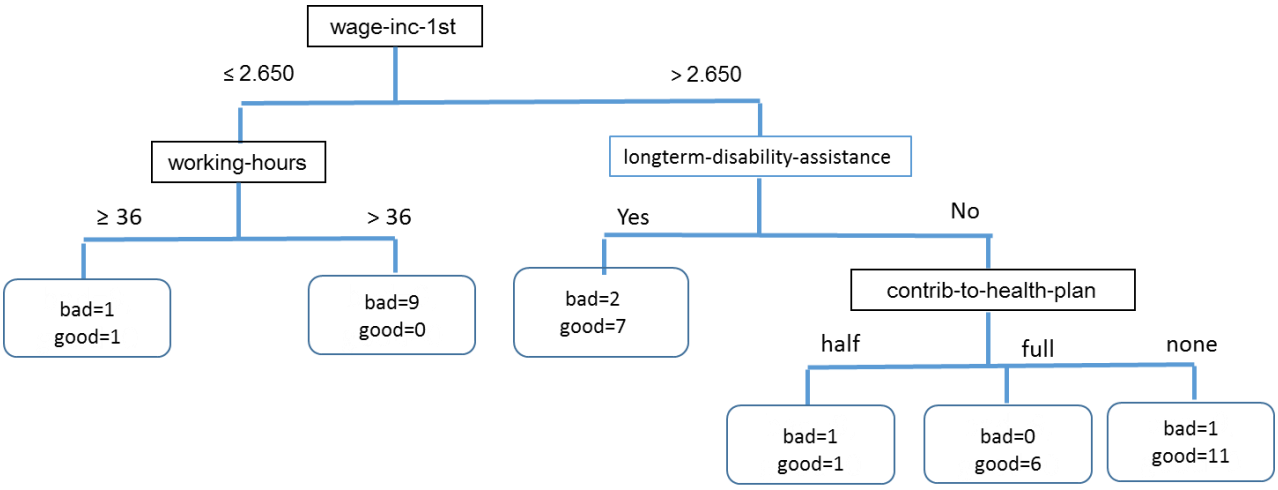
אוניברסיטת בן-גוריון בנגב
Ben-Gurion University of the Negev

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**

נתון עץ החלטה שנבנה מנתונים של 40 הסכמי שכר בקנדה. הסכם מוגדר כ"טוב" אם נתקבל ע"י ההנהלה והעובדים. הסכם מוגדר כ"רע" אם אחד הצדדים דחה אותו.



א. מהו דיוק "חוק הרוב" במסד הנתונים הזה? _____ **10 נקודות**

ב. מהו דיוק האימון של העץ הנ"ל ? _____ **10 נקודות**

ג. יש לחשב את הרווח האינפורמטיבי (Information Gain) עבור הפיצול עפ"י המשתנה contrib-to-health-plan. **15 נקודות.**

Entropy:		Conditional entropy:		Information Gain:	

ד. יש לבחון את כדאיות הפיצול עפ"י המבחן חי בריבוע. **15 נקודות**

Chi-square statistic:		Degrees of Freedom:		Conclusion:	

DF	1	2	3	4	5	6
Chi-Square	3.841	5.991	7.815	9.488	11.070	12.592

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: $q = \frac{N - n_C + 0.5}{N}$