سوال ۱)

الف)

$$I(Parent) = Entropy = -((0.4 \log_2 0.4) + (0.6 \log_2 0.6)) = 0.96$$

$$\Delta = I(Parent) - \sum_{j=1}^{k} \frac{N(v_j)}{N} I(v_j)$$

$$\Delta_A = 0.96 - ((0.7 \times 0.98) + (0.3 \times 0)) = 0.27$$

$$\Delta_B = 0.96 - ((0.4 \times 0.8) + (0.6 \times 0.63)) = 0.26$$

باید ویژگی A را انتخاب کنیم چون بهره آن بیشتر است.

ب)

$$GINI(t) = 1 - \sum_{j} [P(j|t)]^2$$

$$GINI_{split} = \sum_{i=1}^{k} \frac{n_i}{n} GINI(i)$$

$$GINI_A = ((0.7 \times 0.48) + (0.3 \times 0)) = 0.33$$

$$GINI_B = ((0.4 \times 0.37) + (0.6 \times 0.27)) = 0.31$$

در اینجا اما باید ویژگی B را انتخاب کنیم زیرا معیار GINI آن کمتر است.

ج) بله، همان‌طور که در همین سوال هم دیدیم، یکی از آن‌ها ویژگی A را برگزید و دیگری ویژگی B را انتخاب کرد. در نمودار داخل درس هم دیدیم که شیب این نمودارها با هم فرق دارند و شیب نمودار آنتروپی بیشتر است.

سوال ۲)

الف)

$$P(X|Y) = \frac{P(X,Y)}{P(Y)}$$

$$P(A|+) = \frac{0.3}{0.5} = 0.6$$

$$P(B|+) = \frac{0.1}{0.5} = 0.2$$

$$P(C|+) = \frac{0.4}{0.5} = 0.8$$

$$P(A|-) = \frac{0.2}{0.5} = 0.4$$

$$P(B|-) = \frac{0.2}{0.5} = 0.4$$

$$P(C|-) = \frac{0.5}{0.5} = 1$$

ب)

$$P(+|010) \propto P(010|+) \times P(+) = P(\bar{A}|+)P(B|+)P(\bar{C}|+)P(+) = 0.4 \times 0.2 \times 0.2 \times 0.5 = 0.008$$

$$P(-|010) \propto P(010|-) \times P(-) = P(\bar{A}|-)P(B|-)P(\bar{C}|-)P(-) = 0.6 \times 0.4 \times 0 \times 0.5 = 0$$

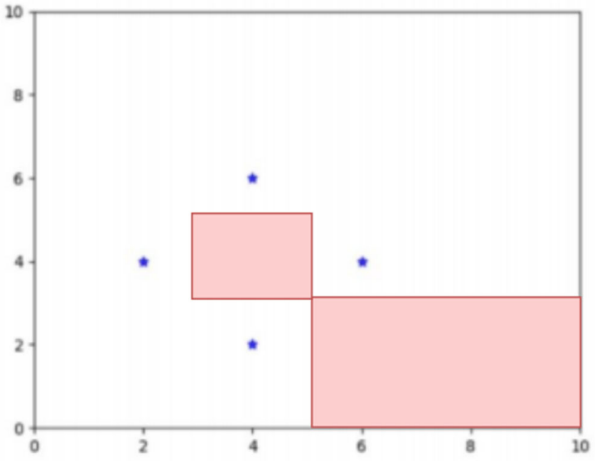
پیش‌بینی می‌شود که برچسب داده + باشد.

سوال ۳)

مثلا در زمانی که با یک مسئله‌ی حساس رو به رو هستیم علاوه بر دقتی که اندازه‌گیری می‌شود، خطا هم مهم می‌شود. در اینجا باید با کمک confusion matrix، error rate را هم در نظر بگیریم.

سوال ۴)

الف)



مرزهای تصمیم گیری در شکل مشخص شده اند.
نواحی قرمز مربوط به دایره‌های قرمز و بقیه نواحی مربوط به ستاره‌های آبی هستند زیرا هر نقطه‌ای در این نواحی ذکر شده در نظر بگیریم، نزدیک‌ترین همسایه‌شان هم‌رنگ ناحیه‌ای است که در آن قرار گرفته اند.

ب) کلاس آبی، زیرا در ناحیه آبی قرار گرفته است پس یعنی نزدیک‌ترین نقطه به آن ستاره آبی بوده است که با محاسبه‌ی فاصله نیز به همین نتیجه می‌رسیم زیرا آن‌را در قسمت الف در نظر گرفته بودیم.

ج) بله باید فضا را پیوسته در نظر بگیریم و از میانگین همسایه‌ها استفاده کنیم.

د)کاملا بستگی به داده‌ها دارد زیرا اگر مقدار کمی را انتخاب کنیم، ممکن است به خاطر نویز، دسته‌ی اشتباهی را انتخاب کنیم و در صورتی که مقدار زیادی را انتخاب کنیم نیز ممکن است یکی از همسایه‌های دسته‌بندی مناسب خودمان را انتخاب کنیم و نتیجه خراب شود پس نمی‌توان حالت کلی‌ای در نظر گرفت.

ه) به نظرم نه چون ممکن است به خاطر افزایش ابعاد داده‌ها، داده‌هایی که نزدیک هم هستند در حقیقت فاصله‌ی زیادی از هم داشته باشند که این نتیجه‌ی ما را خراب می‌کند.

و) چون Lazy است، برای آموزش پیچیدگی‌ای نداریم اما در تست شاید مجبور باشیم که فاصله‌ی آن را با تمامی داده‌ها حساب کنیم و نتیجه را بررسی کنیم؛ در نتیجه پیچیدگی آن در حالت تست بیشتر است.

ی) در اقلیدسی نرم ۲ی فاصله‌ی ویژگی‌ها را در نظر می‌گیریم اما در منهتن فاصله‌ها را با هم جمع می‌کنیم تا نتیجه را بدست آوریم.

سوال ۵)

از این روش برای مقایسه کلسیفایرهای مختلف استفاده می‌شود که در آن هر نمونه‌ی ما به میزانی برابر با سایر داده‌ها برای آموزش استفاده می‌شود و فقط یک‌بار هم برای تست از آن‌ها استفاده می‌شود.

انواع آن عبارتند از:

- Exhaustive: تمام دسته‌بندی‌های ممکن
- Non-Exhaustive: قسمتی از دسته‌های ممکن
- Nested: دسته‌بندی‌های تودرتو

در k-fold هم هر چقدر k بیشتر شود، ما شاهد افزایش واریانس و پیچیدگی زمانی و همچنین کاهش بایاس خواهیم بود.