Subject

Date

۱- الف)

$$H(y) = \sum_{y_i \in \{win, Loss\}} -P(y_i)\log P(y_i) = -\frac{1}{2}\log\frac{1}{2} + \left(-\frac{1}{2}\log\frac{1}{2}\right) = -\log\frac{1}{2} = \log 2 = 1$$

ب) برای x_1، x_2 و x_3 عمل split را انجام می‌دهیم و IG بدست می‌آوریم:

$$IG(S, x_1) = H(y) - \sum_{v \in x_1} \frac{|S_v|}{|S|} H_{S_v}(y)$$

$$\sum_{v \in x_1} \frac{|S_v|}{S} H_{S_v}(y) = \frac{6}{10} H_T(y) + \frac{4}{10} H_I(y)$$

$$H_T(y) = -\frac{1}{3}\log\frac{1}{3} - \frac{2}{3}\log\frac{2}{3} \simeq 0.528 + 0.389 \simeq 0.918$$

$$H_I(y) = -\frac{3}{4}\log\frac{3}{4} - \frac{1}{4}\log\frac{1}{4} \simeq 0.311 + 0.5 \simeq 0.811$$

$$\Rightarrow IG(S, x_1) = 1 - \left(\frac{6}{10} \times 0.918 + \frac{4}{10} \times 0.811\right) = 0.1248$$

برای x_2 دقیقاً مشابه حالت بالا رخ می‌دهد؛ کلاس M دقیقاً مشابه کلاس I (از لحاظ اندازه و آنتروپی) و کلاس P دقیقاً مشابه کلاس T خواهد بود. پس:

$$IG(S, x_2) = IG(S, x_1) = 0.1248$$

$$IG(S, x_3) = 1 - \left(\frac{5}{10} H_S(y) + \frac{5}{10} H_C(y)\right)$$

$$H_S(y) = -\frac{4}{5}\log\frac{4}{5} - \frac{1}{5}\log\frac{1}{5} = 0.257 + 0.464 = 0.722$$

$$H_C(y) = -\frac{4}{5}\log\frac{4}{5} - \frac{1}{5}\log\frac{1}{5} = 0.722$$

$$\Rightarrow IG(S, x_3) = 1 - \left(\frac{1}{2} \times 0.722 \times 2\right) = 0.278 \leftarrow$$

بیشترین IG، بهترین ویژگی برای split

۲_ الف) $J = \frac{1}{n}\sum (p-l)^2 = \frac{1}{n}\sum_{i=1}^{n}\left(\sum_{t=1}^{T}\beta_t G_t(x_i) - y_i\right)^2$

ب) تابع هزینه را نوشته، $\sum_{1}^{T-1}\beta_t G_t(x_i)$ را جدا می‌کنیم تا β_T به تنهایی بدست آید، سپس از آن مشتق می‌گیریم:

$$\frac{\partial}{\partial \beta_T}\frac{1}{n}\sum_{1}^{n}\left(\sum_{1}^{T}\beta_t G_t(x_i) + \beta_T G_T(x_i) - y_i\right)^2 = \frac{1}{n}\sum_{1}^{n} 2\left(\sum_{1}^{T-1}\beta_t G_t(x_i) + \beta_T G_T(x_i) - y_i\right) G_T(x_i) = 0$$

$$\Rightarrow \sum_{1}^{n}\left(\sum_{1}^{T-1}\beta_t G_t(x_i) - y_i\right) G_T(x_i) + \beta_T G_T(x_i) = 0$$

$$\Rightarrow \beta_T \sum_{1}^{n} G_T^2(x_i) = \sum_{1}^{n}\Big(y_i - \underbrace{\sum_{1}^{T-1}\beta_t G_t(x_i)}_{M_{T-1}(x_i)}\Big) G_T(x_i) \Rightarrow \beta_T = \frac{\sum_{1}^{n}\left(y_i - M_{T-1}(x_i)\right) G_T(x_i)}{\sum_{1}^{n} G_T^2(x_i)}$$

۳_ الف) احتمال اینکه یک نمونه انتخاب شود، $\frac{1}{N}$ و احتمال انتخاب نشدن آن $1-\frac{1}{N}$ است. وقتی می‌خواهیم pN نمونه‌گیری انجام دهیم، احتمال اینکه یک نمونه در هیچ نمونه‌گیری‌ای نیاید، $(1-\frac{1}{N})^{pN}$ است، یعنی:

$$\left(1-\frac{1}{N}\right)^{pN} = \left(\left(1-\frac{1}{N}\right)^{N}\right)^{p} = (e^{-1})^{p} = e^{-p} \quad \left(\lim_{n\to\infty}\left(1-\frac{1}{n}\right)^{n} = e^{-1}\right)$$

حال عبارتی که از ما خواسته شده، حال امید ریاضی نمونه‌های است که انتخاب نمی‌شوند یعنی $N \times e^{-p}$

ب) خب، هنگامی دچار خطا خواهد شد که حداقل دوتا از g_i ها دچار خطا شوند. بلندترین بازه‌ای که حداقل دوتا از g_i ها خطاهایشان overlap خواهد داشت (معادل اینکه دوتایشان خطا کرده باشند)، حال ۰٫۲۵ است و حداقل طول این بازه، ۰ است (فرض کنید g_iها، متمم یکدیگر خطا کنند، آنگاه در G، هرکدام از خطای g_iها خطای دیگر را کاور می‌کنند)؛ پس محدوده خطای $E_{out}(G)$ برابر خواهد بود با: $0 \le E_{out}(G) \le 0.25$

Subject
Date

۴ـ الف) صحیح؛ هر نمونه به ازای تر N داده را به یک مدل می‌دهیم تا همزمان train شوند.

ب) غلط؛ پارامترهای مدل t ام، از مدل t-1 ام می‌آید؛ پس موازی سازی بی معنی است.

ج) غلط؛ روی بخشی از داده‌ها هر مدل train انجام می‌دهد.

د) صحیح؛ تمامی دسته بندهای، تمامی داده‌ها را learn می‌کنند.

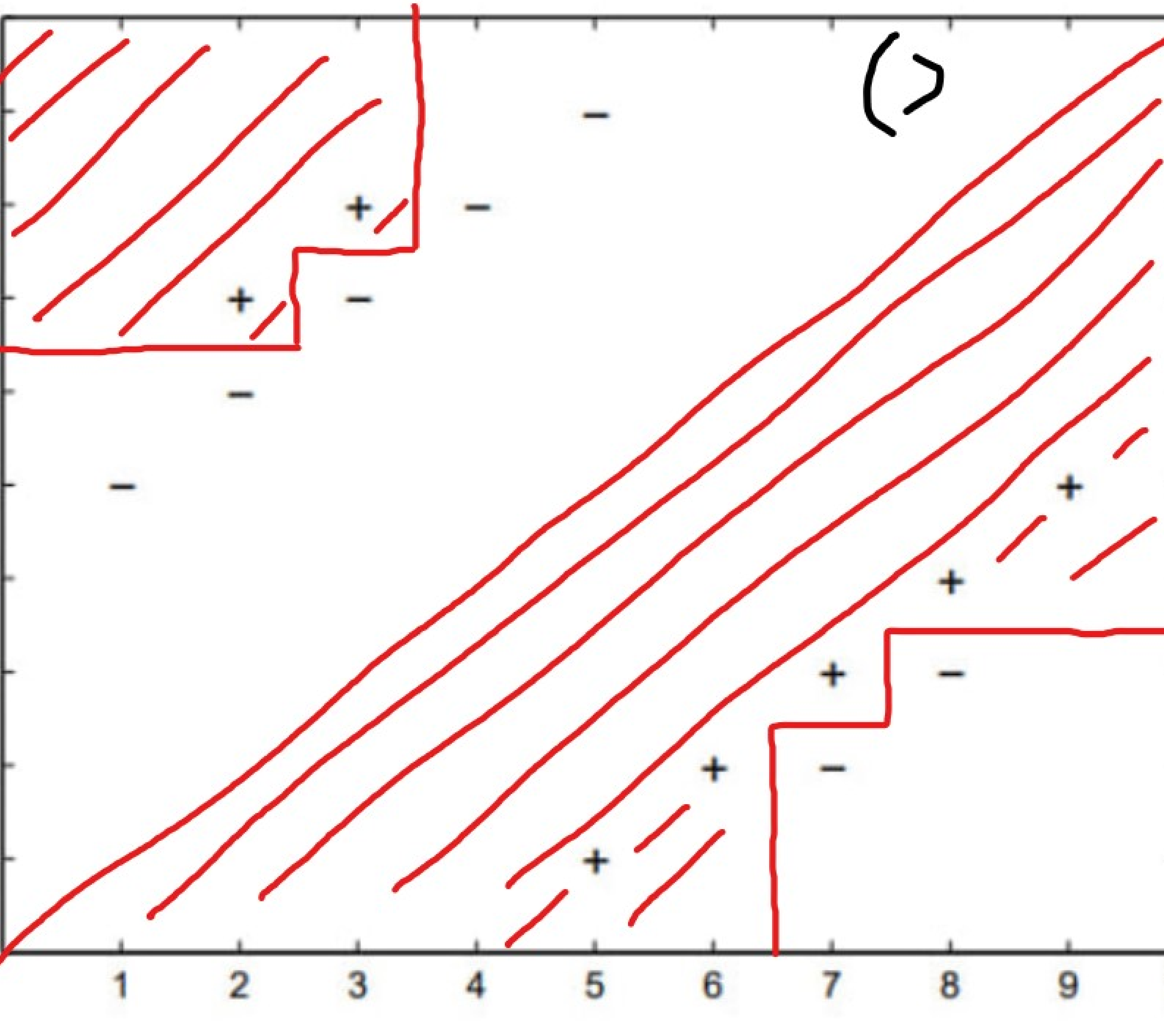
۵ـ الف) چون هر نقطه‌ای همسایه‌ی خودش (نزدیک‌ترین همسایه‌ی خودش در واقع) به شمار می‌رود، 1NN به ما خطای صفر را می‌دهد که کمینه خطای ممکن است.

ب) مقادیر زیاد: فرض کنید خطی را که دو سری داده‌ی منفی روی آن قرار دارند؛ این خط‌ها توسط ردیفی از لیبل‌های مثبت از یکدیگر جدا می‌شوند. با زیاد کردن K، لیبل‌های مثبت در نظر گرفته می‌شوند که به علت نزدیک‌تر بودن به لیبل‌های منفی، موجب افزایش خطا می‌شوند. این امر برای لیبل‌های مثبت نیز برقرار است. مقادیر کم: در دو سمتی از دیتاست که داده‌ها dense تر هستند می‌توان دید اگر داده‌هایی که فاصله اقلیدسی کمی از یکدیگر دارند، لیبل‌شان مخالف هم است پس این امر نیز در لیبل گذاری موجب خطا می‌شود.

ج) بهترین K=5 و بهترین خطا error=0.28571 (کد در اسکرین شات)

```python
import numpy as np
from sklearn.neighbors import KNeighborsClassifier
from sklearn.model_selection import LeaveOneOut

X = np.array([
    [1, 5], [2, 6], [3, 7], [4, 8], [5, 9], [7, 2], [8, 3],
    [5, 1], [6, 2], [7, 3], [8, 4], [9, 5], [2, 7], [3, 8]
])
y = np.array([
    0, 0, 0, 0, 0, 0, 0,
    1, 1, 1, 1, 1, 1, 1
])

best_k = None
best_accuracy = 0

for k in range(1, len(X)):
    loo = LeaveOneOut()
    accuracies = []

    for train_index, test_index in loo.split(X):
        X_train, X_test = X[train_index], X[test_index]
        y_train, y_test = y[train_index], y[test_index]

        knn = KNeighborsClassifier(n_neighbors=k)
        knn.fit(X_train, y_train)
        accuracy = knn.score(X_test, y_test)
        accuracies.append(accuracy)

    mean_accuracy = np.mean(accuracies)
    print(f"k = {k}, error = {1 - mean_accuracy}")
    if mean_accuracy > best_accuracy:
        best_accuracy = mean_accuracy
        best_k = k

error = 1 - best_accuracy

print(f"Best value for k: {best_k}")
print(f"Error of the best k value: {error}")
```

```
PS C:\Users\Erfan> & C:/Users/Erfan/AppData/Lo
k = 1, error = 0.7142857142857143
k = 2, error = 0.7142857142857143
k = 3, error = 0.42857142857142856
k = 4, error = 0.6428571428571428
k = 5, error = 0.2857142857142857
k = 6, error = 0.2857142857142857
k = 7, error = 0.2857142857142857
k = 8, error = 0.6428571428571428
k = 9, error = 1.0
k = 10, error = 0.7142857142857143
k = 11, error = 1.0
k = 12, error = 0.7142857142857143
k = 13, error = 1.0
Best value for k: 5
Error of the best k value: 0.2857142857142857
```


1 2 3 4 5 6 7 8 9

Subject

Date

$$H(T) = \sum_{1}^{K} -P(C_i) \log P(C_i) = \sum_{1}^{K} -\frac{1}{K} \log \frac{1}{K} = \log K$$ ۲-

$$H(T|A) = -\sum_i \sum_j \underbrace{P(T=j, A=i)} \log \underbrace{P(T=j | A=i)} = \sum_i \sum_j \left(\frac{1}{K} \frac{1}{m_A} \log K\right) = \log K$$

چون A و T مستقل هستند این عبارت را به صورت ضرب احتمال‌ها می‌توانند نوشته شود یعنی

$$\underbrace{P(T=j)}_{\frac{1}{K}} \underbrace{P(A=i)}_{\frac{1}{m_A}}$$

چون A و T مستقل هستند این عبارت برابر است با $P(T=j) = \frac{1}{K}$

منطقاً چون T از A مستقل است، دانستن A به ما اطلاعی اضافه نمی‌کند و IG باید برابر صفر باشد،

$$IG(T,A) = H(T) - H(T|A) = \log K - \log K = 0$$

که طبق محاسبات، همینطور هم هست.

PAPCO