תרגיל בית 1 - חלק א'

שאלה 1

סעיף א'

מצורף מערך נתונים dataset.csv המכיל פיצ'רים שחושבו מהגהה של האותיות באנגלית, לכן הלייבל מכיל 26 ערכים שונים. (אין צורך לחקור או לעבד את הדאטה)

ערבבו את הנתונים וחלקו אותם לסט אימון וסט מבחן ביחס של 80-20.

עבור סט האימון השתמשו בשיטות הבאות לבחירת פיצ'רים:

Mutual Information (MI)

MRMR- Minimum Redundancy Maximum Relevance

ReliefF – הרחבה של Relief ל multi-class

Infinite Feature Selection (INF-FS)

Correlation Based - בחירה לפי קורלציה בין הפיצ'ר למחלקה.

Random Selection – בחירה רנדומלית של פיצ'רים.

עבור כל שיטה בדקו את אחוז הדיוק עבור [200, 100, 50, 20, 10, 5, 2] פיצ'רים ע"י אימון של RandomForest על סט המבחן המצומצם. (אין צורך לכוונן את הפרמטרים יש לבצע אימון פשוט). כמו כן, חשבו את המודל על הפיצ'רים המקוריים ללא selection ורשמו מה אחוז הדיוק.

הציגו את הנתונים בטבלה. לדוגמה:

	inf_fs	mrmr	mi	relief	random	corr
200	0.896795	0.921795	0.935256	0.918590	0.923718	0.845513
100	0.875000	0.908333	0.839744	0.891026	0.905128	0.744231
50	0.790385	0.870513	0.778205	0.847436	0.867949	0.683974
20	0.682051	0.771154	0.625641	0.560897	0.614744	0.571154
10	0.551282	0.632692	0.530769	0.411538	0.585256	0.480128
5	0.388462	0.427564	0.316667	0.263462	0.213462	0.257051
2	0.103205	0.175000	0.153205	0.096154	0.085897	0.107692

i. חוו דעתכם על טבלת התוצאות של כל אחד מהסטים מבחינת דיוק המודל מול סיבוכיות המקום (מספר הפיצ'רים שיש לשמור בפעל).

ii. בפרט, מדוע לדעתכם יש להוסיף את עמודת ה- random לטבלה?

iii. עבור בעיית multi-class משתמשים באלגוריתם ReliefF שהוא הרחבה של Relief למצב זה. הסבירו כיצד עובד ReliefF.

<u>סעיף ב'</u>

סעיף זה עוסק ב- INF-FS, בתצורה ה unsupervised של האלגוריתם.

הניחו כעת שקיימת עלות דגימה כלומר שימוש בפיצ'ר מסויים עולה כסף, נסמנו ב- $Cost_i$.

באופן כללי העלות של כל פיצ'ר שונה. נעדיף כמובן לשמר פיצ'רים עם עלות נמוכה ובעלי רלוונטיות מאשר פיצ'רים רלוונטים עם עלות גבוהה.

עברו על אלגוריתם Inf-FS (מצורף כאן) והציעו כיצד לשקלל לתוכו את רכיב העלות כך שהדירוג של פיצ'רים רלוונטים וזולים יהיה גבוה.

Algorithm 1 Unsupervised Infinite Feature Selection

Input: $F = \{\vec{f}_1, ..., \vec{f}_n\}$, α
Output: $\check{c}$ final scores for each feature
  + Building the graph
  for $i = 1 : n$ **do**
    for $j = 1 : n$ **do**
      $\sigma_{ij} = max(std(f_i), std(f_j))$
      $\overline{corr}_{ij} = 1 - |Spearman(f_i, f_j)|$
      $A(i,j) = \alpha\sigma_{ij} + (1-\alpha)\overline{corr}_{ij}$
    end for
  end for
  + Letting paths tend to infinite
  $r = \frac{0.9}{\rho(A)}$
  $\check{C} = (\mathbf{I} - rA)^{-1} - \mathbf{I}$
  $\check{c} = \check{C}\,\mathbf{e}$
  return $\check{c}$

עזרים

לנוחיותכם, מצורפים חבילות שניתן להעזר בהם. הרגישו חופשי להשתמש בכל חבילה/ספרייה או פונקציה שתרצו.

https://pypi.org/project/PyIFS/

https://epistasislab.github.io/scikit-rebate/using/#relieff

https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.mutual_info_classif.html

https://pypi.org/project/mrmr-selection

תיעוד וקריאות הקוד

- אנא הקפידו על תיעוד וקריאות הקוד. יש לציין את מטרות הפונקציות ולתת שמות משתנים בעלי משמעות.
- אין להשאיר קוד ישן כהערה, יש לכתוב קוד קצר ואלגנטי.
- הקוד צריך לרוץ ללא תקלות.

אופן הגשת תרגיל הבית

יש להגיש את התרגיל בשלשות עד לתאריך 24/04/2025 בשעה 23:59.

יש להגיש קובץ zip **יחיד** המכיל קובץ קוד .py או .ipynb וקובץ PDF בלבד.

שם הקובץ יהיה מספר הקבוצה לדוגמא group_xxx (בדקו שהינכם רשומים לקבוצות טרם הגשת התרגיל).

שימו לב! קוד שלא יצליח לרוץ יגרום להורדת ניקוד משמעותית!

הנכם מוזמנים להיעזר בפורום הקורס לשאול שאלות ולעקוב אחר תשובות שנענו עבור שאלות קודמות.

בהצלחה!