מבוא לגנומיקה חישובית ומערכתית - challenge:

מטרה - לבנות ממיין שבהינתן פיצ'רים מגנום של מטופל מסוים יחזיר את סוג הסרטן אשר יש לאותו מטופל עבור סרטן מוח (LGG) וסרטן הרחם (UCEC) בהתבסס על 100 גנים.

תיאור המשימה -

מצורפים לכם הקבצים הבאים:

קובץ ראשון (TRAIN SET DATA) -- מכיל נתונים עבור 80% מהמטופלים (844 מטופלים).

קובץ שני (TEST SET DATA) -- מכיל הנתונים עבור 20% מהמטופלים (211 מטופלים).

הקבצים מכילים את הנתונים הבאים:

1. אינדקס מטופל (Case_id).
2. שם הגן בו יש את המוטציה (Gene name).
3. מספר הכרומוזום בו יש את המוטציה והגן שייך אליו (Chromosome).
4. באיזה גדיל הגן נמצא (Strand).
5. פוזיציית התחלה של המוטציה (Start position).
6. פוזיציית סיום המוטציה (End position).
7. הרצף בגן המקורי במיקום במוטציה (Reference_Allele).
8. הרצף של המוטציה באלל 1 (Tumor_Seq_Allele1).
9. הרצף של המוטציה באלל 2 (Tumor_Seq_Allele2).

- שימו לב כי ייתכן ולמטופל מסוים יהיו כמה מוטציות, כלומר, כמה שורות של נתונים.
- כמו כן, עבור סט ה-TRAIN ישנה עמודה נוספת של Lable שמייצגת את סוג הסרטן (LGG-1, UCEC-2).

קובץ שלישי (TRAIN SET FEATURES) -- מכיל פיצ'רים ולייבלים עבור סט ה- TRAIN של המטופלים.

קובץ רביעי (TEST SET FEATURES) -- מכיל פיצ'רים עבור סט ה- TEST של המטופלים.

הקבצים מכילים את הפיצ'רים הבאים:

1. כמות מוטציות בכל גן עבור כל מטופל.
2. כמות מוטציות באזור מסוים בכל גן (אינטרונים, UTR...) עבור כל מטופל.
3. כמות מוטציות כללית לכל מטופל.

שימו לב שמדובר פה רק בדוגמאות של פיצ'רים שניתן לייצר. אתם צריכים לחשוב על פיצ'רים נוספים ולייצר אותם.

קובץ mat - מכיל את הגנום של human עם שם הגן, אינדקס הגן והרצף עצמו.

קוד מטלב אשר בהינתן מיקום של מוטציה בכרומוזום מסוים מוציא X נוקליאוטידים לפני המוטציה ו-Y נוקליאוטידים אחרי המוטציה (extract_sequences).

שימו לב שכדי להשתמש בקוד מטלב זה, יש להוריד את רצף הכרומוזום המתאים לגן (מספר הכרומוזום נמצא בקבצי האקסל המצורפים) באתר NCBI כפי שנלמד בכיתה (https://www.ncbi.nlm.nih.gov/genome/?term=human)

א. סעיף חימום: צרו את הפיצ'רים הבאים בעצמכם: כמות מוטציות כללית לכל מטופל (פיצ'ר 3) וכמות מוטציות באזור מסוים בכל גן בכל המטופלים (פיצ'ר 2). ודאו כי הפיצ'רים שיצרתם זהים לפיצ'רים אשר נתונים לכם.

ב. צרו פיצ'רים נוספים כראות עיניכם (כפי שהוסבר והודגם בכיתה).

ג. צרו גרף בו ציר וואי מתאר את כמות המוטציות וציר איקס מתאר את סוג/מיקום המוטציה עבור כל הדאטא שיש לכם. החלוקה של ציר איקס היא לפי איור 1b במאמר המצורף.

ד. על סמך הדאטא הנתון לכם, צרו ממיין באמצעות שימוש אלגוריתם SVM ואמנו אותו על סט ה-TRAIN כפי שנלמד בכיתה.

ה. העריכו את הביצועים של הממיין שלכם לפי השגיאה הבאה:

Error=(sum(predict_label~=true_lable)/number of test patients)*100

ו. לבסוף הציגו עבור המטופלים של סט ה-TEST מהו סוג הסרטן המתאים להם על ידי שימוש בממיין שבניתם.

תיאור הגשה –
1. צרו מסמך PDF אשר יכיל:
* מהם הפיצ'רים הנוספים שיצרתם וכן הסבר מדוע בחרתם להוסיף אותם.
* את הגרף שיצרתם בסעיף ג.
* תרשים זרימה מפורט של אופן פעולת הממיין שלכם.
* דיון – מה מידת ההצלחה של הממיין שלכם? כיצד מדדתם אותה? כיצד ניתן לשפר את הממיין?

2. צרפו את קבצי הקוד של הממיין שלכם. כתבו קובץ READE ME, מסודר על אופן הפעלת הקוד.

3. קובץ מטלב (פורמט של cell array) שיכיל שתי עמודות: שם המטופל והקלסיפיקציה המתאימה לו.