מבוא לגנומיקה חישובית ומערכתית – challenge

מטרות:

1. לבנות מסווג שבהינתן פיצ'רים מגנום של מטופל מסוים יחזיר את סוג הסרטן אשר יש לאותו מטופל עבור סרטן קשקשי בריאה (Lung Squamous Cell Carcinoma - LUSC) וסרטן קשקשי בראש ובצוואר (Head-Neck Squamous Cell Carcinoma - HNSC) בהתבסס על 100 גנים.

2. לבנות מסווג שבהינתן פיצ'רים מגנום של מטופל מסוים **ובנוסף מרמות המתילציה שלו** יחזיר את סוג הסרטן אשר יש לאותו חולה עבור אותם סוגי סרטן בסעיף הקודם בהתבסס על נתונים מאותם 100 גנים.

תיאור המשימה

מצורפים לכם הקבצים הבאים:

קובץ ראשון (train_muts_data.csv) – מכיל מוטציות עבור 80% מהמטופלים (805 מטופלים)

קובץ שני (test_muts_data.csv) – מכיל מוטציות עבור 20% מהמטופלים (201 מטופלים)

הקבצים הנ"ל מכילים את הנתונים הבאים:

1. אינדקס המטופל (case_id)
2. שם הגן בו יש את המוטציה (Gene_name)
3. מספר הכרומוזום בו יש את המוטציה ואת הגן (Chromosome)
4. פוזיציית התחלה של המוטציה (Start_Position)
5. פוזיציית סיום המוטציה (End_Position)
6. באיזה גדיל המוטציה נמצאת (Mut_Strand)
7. אופי המוטציה (Variant_Classification)
8. הרצף בגן המקורי במיקום המוטציה (Reference_Allele)
9. הרצף של המוטציה באלל 1 (Tumor_Seq_Allele1)
10. 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 קיימת גם עמודה בשם Label עם תווית (=לייבל) השווה ל-1 עבור חולי LUSC ול-2 עבור חולי HNSC

קובץ שלישי (train_feats.csv) – מכיל פיצ'רים ולייבלים עבור סט ה-TRAIN של המטופלים

קובץ רביעי (test_feats.csv) – מכיל פיצ'רים עבור סט ה-TEST של המטופלים

הקבצים האלה מכילים את הפיצ'רים של כמות המוטציות עבור כל אחד מ-100 הגנים שנבחרו.

שימו לב שזוהי רק דוגמא לפיצ'רים אפשריים שניתן לייצר. אתם צריכים לחשוב על פיצ'רים נוספים ולייצר אותם

קובץ חמישי (100_genes.csv) – מכיל את 100 הגנים שנבחרו של human עם שם הגן, הגדיל שלו, הרצף המקודד שלו (עבור הtranscript הארוך ביותר של אותו גן), והמידע על אותו חלבון\transcript (כולל מיקומי הרצף המקודד בגנום).

קובץ שישי (train_meth_data.csv) – מכיל נתוני מתילציה עבור אותם מטופלים מהקובץ הראשון

קוד פייתון אשר בהינתן מיקום של מוטציה בכרומוזום מסויים מוציא X נוקלאוטידים לפני המוטציה ו-Y נוקלאוטידים אחרי המוטציה (extract_sequences)

שימו לב שבשביל להשתמש בקוד פייתון המצורף, יש להוריד את רצף הכרומוזמים המתאים לגן (מספר הכרומוזום נמצא בקבצים המצורפים) באתר NCBI כפי שנלמד בכיתה (https://www.ncbi.nlm.nih.gov/genome/?term=human)

קובץ שביעי (meth_train_data.csv) – מכיל נתוני מתילציה עבור החולים בסט ה-TRAIN. שימו לב כי לא תוכלו לראות את כל השורות שלו אם תסתכלו עליו באקסל (אבל כשתקראו אותו בפייתון הוא ייקרא כולו).

קובץ שמיני (meth_test_data.csv) – מכיל נתוני מתילציה עבור החולים בסט ה-TEST.

הקבצים עם נתוני המתילציה מכילים:

1. אינדקס המטופל (case_id)
2. ה-ID של ה-probe שמדד את ערכי המתילציה (probeID)
3. ערכי הבטא המודדים את ערכי המתילציה (beta_val)
4. הכרומוזום בו יש את המתילציה (CpG_chrm)
5. פוזיציית ההתחלה של המתילציה (CpG_beg)
6. פוזיציית הסיום של המתילציה (CpG_end)
7. הגדיל בו היה ה-probe (probe_strand)
8. שם הגן בו יש את המתילציה (matching_genes)

גם כאן שימו לב כי:

- ייתכן ולמטופל מסוים יהיו כמה מקומות בהם היה לו מטליציה, כלומר, כמה שורות של נתונים
- כמו כן, עבור סט ה-TRAIN קיימת גם עמודה בשם Label עם תווית (=לייבל) השווה ל-1 עבור חולי LUSC ול-2 עבור חולי HNSC

קובץ תשיעי (Main_text_npj.pdf) – מאמר אחד שמומלץ לקרוא כדי לקבל רעיונות לפיצ'רים (תחפשו עוד)

מטלה 1 (מסווג על סמך מוטציות בלבד):

א. צרו את הפיצ'רים הבאים עבור כל מטופל:

1. כמות כל המוטציות (סכום על כל הגנים בלי להתחשב באופי שלהן)
2. כמות המוטציות מאופי מסוים (Intron, Missense וכו') לכל הסוגים האפשריים של אופי מוטציות (סכום על כל הגנים)
3. כמות המוטציות מאופי מסוים (Intron, Missense וכו') לכל הסוגים האפשריים של אופי מוטציות **לכל גן מ-100 הגנים.**

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

ג. צרו גרף בו ציר ה-Y מתאר את כמות המוטציות וציר ה-X מתאר את סוג\מיקום המוטציה עבור כל הדאטא. החלוקה של ציר X היא לפי איור 1b של המאמר המצורף.

ד. על סמך הפיצ'רים שקיבלתם וייצרתם צרו מסווג (למשל SVM) ואמנו אותו על סט ה-TRAIN כפי שנלמד בכיתה

ה. העריכו את הביצועים של המסווג שלכם לפי השגיאה הבאה על סט ה-Known Test (מתוך ה-TRAIN=Known):

Error = sum(df_known_test['Label']!=df_known_test['predict_label'])/len(df_known_test)

כאשר החיזוי של המסווג נמצא בעמודה ב-predict_label.

ו. לבסוף הציגו עבור המטופלים של סט ה-TEST מהו סוג הסרטן המתאים להם ע"י שימוש במסווג שבניתם

מטלה 2 (מסווג על סמך מוטציות ומתיליציה):

א. צרו את הפיצ'רים הבאים עבור כל מטופל:

1. ממוצע ערכי המתיליציה לכל הפרובים השייכים לגן מסוים עבור כל גן.

ב'-ו' – אותו דבר כמו במטלה 1 (סעיף ג' לא רלוונטי במטלה זו). שימו לב להשתמש גם בפיצ'רים מהמטלה הקודמת.

תיאור הגשה –

1. צרו מסמך PDF שיכיל:
   - מהם הפיצ'רים הנוספים שייצרתם וכן הסבר מדוע בחרתם להוסיף אותם
   - הגרף שייצרתם בסעיף ג' במטלה 1
   - תרשימי זרימה מפורטים של פעולת המסווגים שלכם.
   - דיון – מה מידת ההצלחה של המסווגים שלכם? כיצד מדדתם אותה? כיצד ניתן לשפר את המסווגים ?
2. צרפו את קבצי הקוד של שני המסווגים שלכם. כתבו קובץ READ_ME מסודר על אופן פעולת הקוד (מסוג PDF) לכל אחד.
3. שני קבצי csv עבור סט ה-TEST (אחד עם הסיווג של המסווג ממטלה 1 ואחד עם הסיווג של המסווג ממטלה 2) שיכילו שתי עמודות:
   - אינדקס המטופל (בעמודה case_id)
   - הסיווג המתאים לו, כלומר 1 או 2 (בעמודה predict_label)