(67800) שיטות הסתברותיות | תרגיל תכנותי 1

שם: רון מורן | ת"ז: 206170920

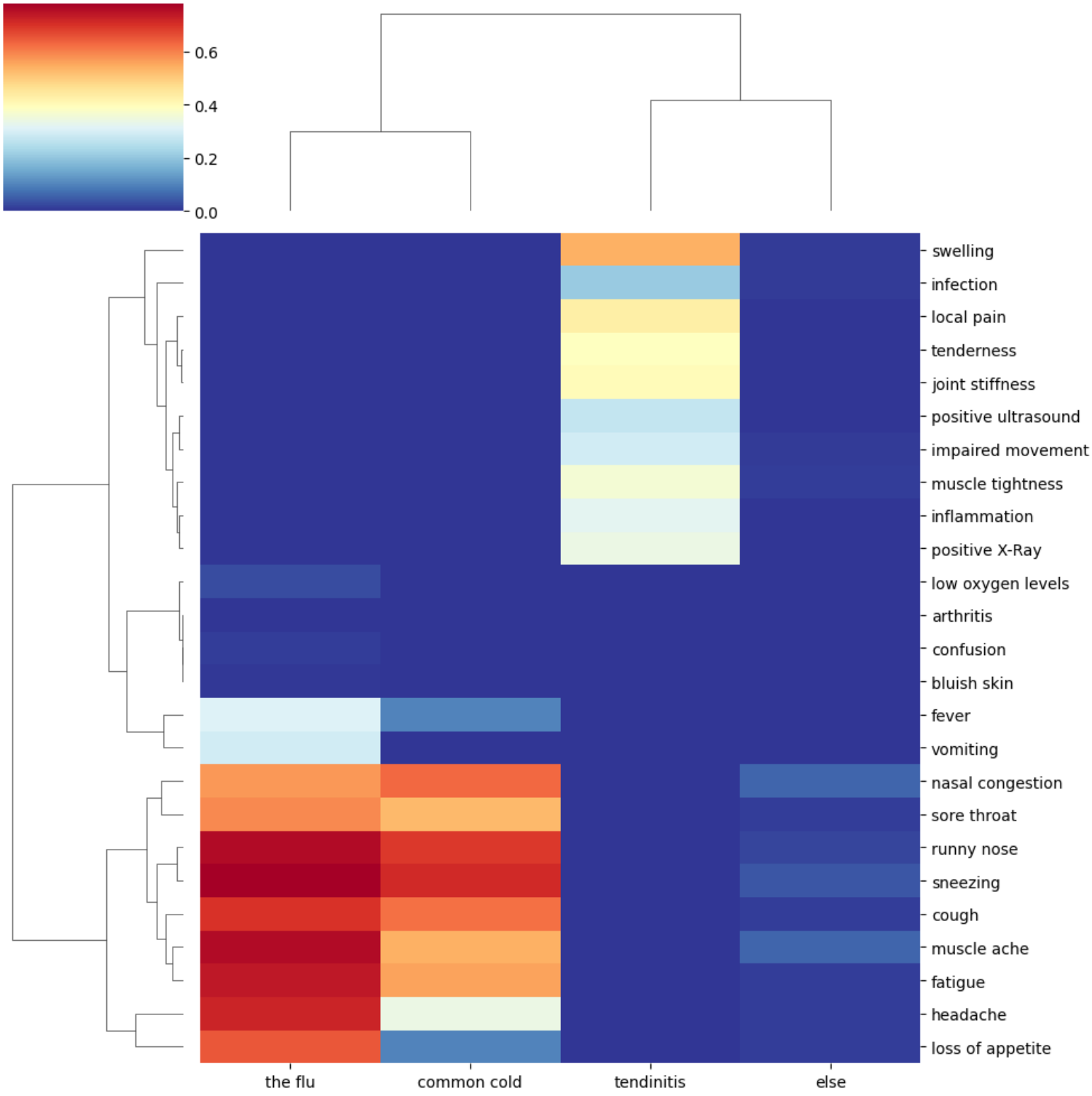
שאלה 1

$$\mathbb{P}\left(D, S_1, \ldots, S_m\right) = \mathbb{P}\left(D\right) \cdot \prod_{i=1}^{m=25} \mathbb{P}\left(S_i | D\right)$$

עבור הרשת הבייסיאנית: ישנם m משתני "סימפטום", כל CPD מוגדר ע"י 4 דרגות חופש. בנוסף, קיים prior על D שכולל 3 דרגות חופש. סה"כ ישנן $4m + 3 = 103$ דרגות חופש.

עבור ההתפלגות המשותפת: כמה השמות אפשריות ישנן ל m משתנים בינאריים ומשתנה D כך ש $|Val\left(D\right)| = 4$? ישנן $2^m \cdot 4 = 2^{27}$ השמות כאלה.

שאלה 2



ניתן לראות שהתסמינים של שפעת ושל צינון מאוד דומים - בהינתן חולי מסוג שפעת, ההסתברות לאף סתום, גרון כואב וכו' גבוהה למדי וכך גם עבור צינון. למעט, אולי, כאב ראש ואבדן תיאבון - ההסתברויות דומות.

עם זאת, עבור דלקת גידים הסימפטומים מובהקים יותר - אמנם היא לא גורמת בהסתברות גבוהה כל כך למפרקים קשיחים, כאב מקומי וכו' אך ההסתברות לסימפטומים המתאימים לצינון ושפעת אפסית בהינתן דלקת גידים.

שאלה 3

לחישוב log ההתפלגות השולית, באמצעים מעט יותר יציבים נומרית:

$$\log\left(\mathbb{P}\left(S_1,\ldots,S_m\right)\right)=\log\left(\sum_{d\in Val(D)}\mathbb{P}\left(D,S_1,\ldots,S_m\right)\right)=\log\left(\sum_{d\in Val(D)}\mathbb{P}\left(D\right)\cdot\prod_{i=1}^{m=25}\mathbb{P}\left(S_i|D\right)\right)$$

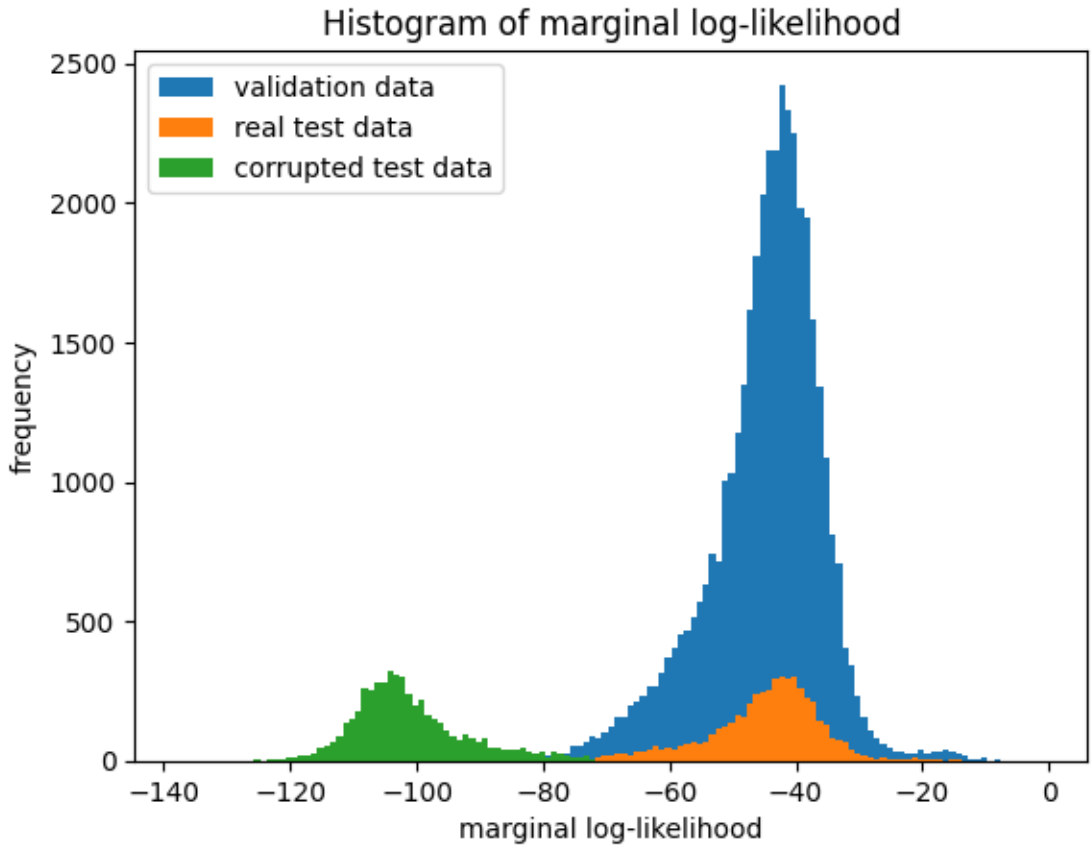
נחשב את הביטוי בתוך הסכום:

$$\mathbb{P}\left(D\right)\cdot\prod_{i=1}^{m=25}\mathbb{P}\left(S_i|D\right)=\exp\left(\log\left(\mathbb{P}\left(D\right)\cdot\prod_{i=1}^{m=25}\mathbb{P}\left(S_i|D\right)\right)\right)=\exp\left(\log\left(\mathbb{P}\left(D\right)\right)+\log\left(\sum_{i=1}^{25}\mathbb{P}\left(S_i|D\right)\right)\right)$$

וסה"כ:

$$\log\left(\mathbb{P}\left(S_1,\ldots,S_m\right)\right)=\log\left(\sum_{d\in Val(D)}\exp\left(\log\left(\mathbb{P}\left(D\right)\right)+\log\left(\sum_{i=1}^{25}\mathbb{P}\left(S_i|D\right)\right)\right)\right)$$

ההיסטוגרמה המצורפת, שכוללת חלוקה למידע מושחת (רחוק יותר משלוש סטיות תקן מהממוצע של הולידציה):



ניתן לראות את ההתכנסות החזקה יחסית בסט הולידציה של המודל סביב הסתברות של e^{-40} לקיומם ולאי קיומם של סימפטומים מסויימים. בסט הבדיקה ניתן לראות שתי סדרות שונות מאוד: אחת דומה בהתפלגותה לסט הולידציה, זה המידע האמיתי (כתום). שאר המידע (ירוק) הוא מידע שגוי עם הסתברויות נמוכות בכמה סדרי גודל ממוצע סט הולידציה.

שאלה 4

חלק א:

$$\mathbb{P}\left(D|S_1,\ldots,S_m\right)=\frac{\mathbb{P}\left(D,S_1,\ldots,S_m\right)}{\sum_{d\in Val(D)}\mathbb{P}\left(D,S_1,\ldots,S_m\right)}=\frac{\mathbb{P}\left(D\right)\cdot\prod_{i=1}^{m=25}\mathbb{P}\left(S_i|D\right)}{\sum_{d\in Val(D)}\mathbb{P}\left(D\right)\cdot\prod_{i=1}^{m=25}\mathbb{P}\left(S_i|D\right)}$$

חלק ב:

נשים לב כי המכנה קבוע לכל קלט $d\in Val\left(D\right)$. לכן, כדי לבחור את d שנותן את הערך הגדול ביותר, נחשב רק את המונה. עשינו זאת כבר בשאלה 1.

חלק ג: המסווג צודק עבור 95.9% מהקלסיפיקציות.

שאלה 5

הדיוק עבור תיוג מלא של הנתונים הוא 82.85%. אחוז הדיוק נמוך יותר שכן המודל שלנו מסוגל לתת קלסיפיקציה בודדת - הסבירה ביותר בהינתן הסימפטומים והשכיחות של כל מחלה (MAP). על כן, בכל מקום שיש בו יותר מתיוג יחיד, המודל יכשל בוודאות. ככל שישנם יותר תיוגים מרובים של מחלות לחולה, כך המודל ישגה יותר.

שאלה 6

ניתן לאפשר מספר סיווגים ע"י קביעת רף מעבר, cutoff. מעליו נקבל כל מחלה עם ערך 1 ומתחתיו נדחה עם ערך 0. נוכל לוודא כי ישנו לפחות ערך אחד עם 1 באמצעות קביעה שתמיד נבחר ב d שממקסם את הביטוי בסעיף א, ובשאר רק אם הם חוצים את ה cutoff האמור.