

פרויקט בייסיאנים:

חלק א:

לפניך dataset שמתאר כמהות לקוחות שנכנסות לרשות מסעדות טבעניות הנקראת cooper branch בקנדה. נרצה לעשות תחזית כמהות הצרכים שייכנסו לחנות בעיר מסוימת ביום לפי שני משתנים מסוימים: כמהות טבעוניים שగרים בעיר ודרוג סוציאו-כלכלי של העיר.

לפניך dataset עם העמודות הבאות:

.socio_economic_profile:
מספר טבעוניים: num_vegans.
עיר: city.
מספר לקוחות שנכנסו ביום: tot_customers.
האם הנתון שיר לאיומן: is_train. אם זה true זה נתון שאיתו בונים מודלים בייסיאניים אם לא זה נתון שאיתו בודקים דיקט אחרי בניית המודל.

עליכם לטעון את dataset על ידי השורה:

```
df = pkl.load(open('../data/group_+str(id)+'regression.pkl', 'rb'))
```

משימות:

1. נא לבחור את העיר הראשונה שמופיעיה לכם בdataset ובנوت אותה מודל רגסיה ליניארית שhzוזה את מספר הלקוחות שנכנסות ליום בעיר הזאת לפי מצב סוציאו-כלכלי ומספר טבעוניים. יש לבנות מודל רגסיה עם אינטראקציה כלומר:

$$Y = b_0 + b_1 * \text{num_vegan} + b_2 * \text{socio_economic_profile} + b_3 * \text{socio_economic_profile} * \text{num_vegan}$$

נא להציג את משואת הרגסיה לפי הממצאים של `postiors`.

2. נא לעשות ניתוח שבודק האם `mcmc` התכונו להתפלגות האמיטית של `posterior` עבור אחד המשתנים לבחירתכם.

3. נא לעשות ניתוח `posterior predictive` על מנת לבדוק אם ההתפלגות שבחרתם `likelihood` ו`prior` מתאימים כאן.

4. נא לחשב את `mse` עבור `test set` של העיר שבחרתם מול התחזית שלכם. יש להציג משואה בגרף תלת מימדי את משואת הרגסיה והנתונים האמיטיים. כאשר ציר אחד זה מספר טבעוניים, ציר שני זה מצב סוציאו-כלכלי - והציר השלישי הוא מספר לקוחות לקוחות ביום.

5. נא לבנות רגרסיה היררכית כך שלכל עיר יש את אותה משווה רגרסיה מסעיף 1 אבל יש לכל הערים priors משותפים.

6. נא לחשב mse לכל עיר בנפרד על סמך set test.

7. יש לחזור על הניתוח של סעיף 1 עבור כל הערים ולחשב את mse. לאחר מכן יש להשוות mse בכל עיר בין המודל `hpooledu` והמודל היררכי. מה פרופוזיציות הערים שקיבלתם דיקט גדול יותר מודל היררכי.

8. למה מודל היררכי עשוי לעזור במקרה של הבעיה הזאת?

9. נא לבחור 2 ערים ולחשב את hd's cohen של הממוצע של מספר הלקוחות שנכנסים ביום לחנות. מה הסיכוי שהפרש תוצאות גדול מ-0? מה הסיכוי שהוא גדול מ-1?

חלק ב:

חוות שרתים של Nvidia.

כעת אתם אחראים על מערכת חוות שרתים של Nvidia.

דרישות לGPU מגיעות בקצב דרישות משתנה במהלך השבוע. המרכז חישובים שלהם יכול לספק 15 דרישות בשעה בממוצע. קצב ההגעה הוא לפי יום ושעה בשבוע. לאחר מכן יש 24 שעות ושבוע ימים סה"כ יכולים להיות לנו 168 קצבים הגעה שונים.

ברצוננו לנתח את מספר הדרישות GPU הצפוי בחוות שרתים כפונקציה של הזמן. ברשותנו יומן אירועים של דרישות המגיעים חוות שרתים. הנתונים נאספו לאורך 1000 ימים.

הנתונים מאורגנים כמיילון של 1000 ימים שונים, הממוספרים מ-0 עד 999. כל ערך במיילון הוא יומן אירועים של יום שונה.

עמודות הנתונים הן:

1. id – מזהה של הדרישה
2. Time-stamp – הזמן שבו מתרכש האירוע. יחידות הזמן הן שעות. השעה ה-i היא מספר השעות שחלפו מאז 00:00 של אותו יום.
3. Type – סוג האירוע: ביקוש (Demand) או עציבה (Release)
4. day – ערך מסווג של היום: {0: יום ראשון, 1: יום שני, 2: יום שלישי, 3: יום רביעי, 4: יום חמישי, 5: יום שישי, 6: שבת}
5. hour – השעה ביום, מספר שלם בין 0 ל-23. אם הערך הוא 0, הכוונה לשעה שבין חצות ל-00:00.
6. day_name – שם היום בשבוע.

כל יום מתחילה עם 0 דרישות בשעה 00:00.

מטרת המשימה היא לאמוד את קצב הדרישות עבור כל שעה ביום.

מתוך הנתונים ניתן להשתמש בשיטות סטטיסטיות קלאסיות על מנת לאמוד את קצב ההגעה עבור כל שעה בשבוע. הבעיה היא שקיימים נתונים חסרים, אין לנו יומן אירועים עבור כל השעות.

הרעיון המרכזי הוא לאמוד את קצב הדרישת הממוצע עבור אזורי זמן שביהם קיימים נתונים, ולאחר מכן להשתמש ברגression תħallir גאוסי (GP) עבור אזורי הזמן שביהם הנתונים חסרים.

חלק זה מחולק לשלושה חלקים.

חלק 1:

משימות:

1. בצעו ניתוח בייסיאני למספר הדרישות הממוצע המגיע ביום שבת בין השעות 00:00–01:00.
2. באיזו פונקציית הסתברות (Likelihood) השתמשתם ומדוע?
3. בצעו ניתוח ניבוי פוסטרiori (Posterior Predictive Analysis) כדי לוודא שההטפלות ה-Likelihood שנבחרה אכן מתאימה.
4. על פי תוצאות הניתוח, עד כמה אתם בטוחים בהערכת ממוצע קצב הדרישות?
5. מהו קצב הדרישות המוערך? הציגו ערך יחיד המבוסס על ההטפלות הפוסטריאורית.

חלק 2:

יש לנו 24 שעות ביום ו-7 ימים בשבוע. בסך הכל מתאפשר גריד של 168 ערכים.

ברצוננו למפות את קצב ההגעה הממוצע בכל תא זמן.

במגרש הנתונים קיימים חלקים חסרים. חלק מהתאים בגריד אינם כוללים נתונים. תפקיכם הוא להשלים את הגריד.

1. חלצו את קצב הדרישת לכל תא שבו קיימים נתונים (חישוב ממוצע פשוט מספיק, אין צורך בניתוח בייסיאני נוספת).
2. השתמשו ברגression תħallir גאוסי (GP) על מנת להשלים את קצב ההגעה הממוצע עבור יתר התאים בגריד.
3. הציגו גרפַּת תלת-ממדי הכוללת:
 - a. גרפַּת פיזור (scatter plot) של הנתונים שאומדו שירות מההמפה .dataframe.
 - b. גרפַּת פיזור (scatter plot) של הנתונים שאומדו באמצעות רgression GP.
 - c. גרפַּת משטח של regression ה-GP. קלומר surface של החסויות GP.
4. מהי ההסתברות שקצב ההגעה בין השעות 10:00–11:00, ביום שבת, יהיה בין 9 ל-10 דרישות לשעה?
5. אני הציגו את המפרט המלא של ההטפלות הגאוסיינית של קצב ההגעה ביום שבת בין השעות 10:00–11:00 וביום שני בין השעות 11:00–12:00.

6. מהו המתאים (קורלציה) בין קצב ההגעה ביום ראשון לבין קצב ההגעה ביום ראשון בין השעות 00:00–08:00? 08:00–09:00?
מהו המתאים בין קצב ההגעה ביום ראשון לבין קצב ההגעה ביום ראשון בין השעות 00:00–07:00? 07:00–08:00? 08:00–09:00?
איזה ערך מתאים גבוה יותר? האם תוצאה זו הגיונית?

חישוב תוצאות של החלק התחרותי

בחלק זה נשתמש בתഴית שעשינו בחלק הקודם.
עת ברצוננו להשתמש בהם לצורך חישוי מספר הדרישות הכספיים במערכת כפונקציה של הזמן.
כל שאליכם לעשות הוא ליצור DataFrame עם תוצאות בפורמט הבא.

העלו את `true_h` ואת `df_res`.

`h` הוא מספר הדרישות הכספי האמיתית במערכת.

`df_res` הוא DataFrame הכלול כולל עמודות: `day`, `hour` ו-`rate`. נכון לעכשו עמודת `rate` מכילה אפסים; עליו למלא בה את קבועי ההגעה שחזיתם באמצעות רגסיטר GP, ולאחר מכן להריץ את הפונקציה `plot_results`.

הפונקציה תמחיש את מספר הדרישות הכספי שהעריכתם כפונקציה של הזמן, בהשוואה לעריכים האמיתיים (מתוך `h_true`).

בסוף, הפונקציה תדפיס את ערך MSE, והקובזת עם ערך MSE הנמוך ביותר תנצה.

נקודות הבונוס הן:

מקום ראשון: 3 נקודות בונוס.

מקום שני: 2 נקודות בונוס.

מקום שלישי: נקודת בונוס אחת.

שהקובזה הטובה ביותר ביותר תנצה!!

בהצלחה!