תרגיל בנושא סיווג רקטות – חלק ב

הבנת המסווגים השונים

אנחנו מעוניינים לטייב את איכות התוצאות ולהבין את המסווגים השונים כמיטב יכולתנו. אנו נדון כאן ב random forest (וב decision tree שהוא מקרה פרטי שלו) וב logistic regression. בשאלות הבאות, כמו בסוף החלק הקודם, נסווג את מטרות (1,4,7,10).

1. **Random Forest**
   1. קראו על random forest, למשל [פה](https://en.wikipedia.org/wiki/Random_forest).
   2. הפעילו את random forest בערכי ברירת המחדל. מה הדיוק על סט האימון? מה הדיוק על סט הבדיקה? באם קיים הבדל, האם תוכלו להבין את סיבת ההבדל? לקבלת רקע לפער בין התוצאות על סט האימון לסט הבדיקה קראו על [overfitting](https://en.wikipedia.org/wiki/Overfitting).
   3. ישנם מספר פרמטרים שמשפיעים על אופי המסווג. ביניהם מספר העצים ביער ההכרעה, עומק העצים ועוד. נסו לכוונן את העץ לקבלת תוצאות יותר טובות, במיוחד בשני הפרמטרים האמורים (מספר עצים ועומקם).
      1. נסו לפחות עשרה שילובים שונים.
      2. מה קורה לדיוק בסט האימון, בסט הבדיקה ומה קורה לפער ביניהם, כאשר מגדילים או מקטינים את הפרמטרים?
      3. מהם הפרמטרים הטובים ביותר לסט האימון? מה הפרמטרים הטובים ביותר לסט הבדיקה? מי מביניהם הוא החשוב?
   4. יער ההכרעה מוציא התפלגות של הסתברויות על פני הסיווגים השונים באמצעות הפונקציה predict\_proba. השוו בין וקטורי ההתפלגות כאשר הניבוי של המסווג היה נכון, לבין וקטורי ההתפלגות כאשר הניבוי של המסווג היה שגוי. מה ניתן להסיק?
2. **Decision Tree**

יער הכרעה בנוי מהרבה עצי הכרעה. השתמשו במחלקהDecisionTree כדי לאמן עץ הכרעה, כלומר יער הכרעה עם עץ בודד.

* 1. קבעו את עומק העץ ל 3 לכל היותר, וודאו שאין לכם יותר מדי עלים. לאיזה דיוק הגעתן?
  2. השתמשו בפונקציה plot\_tree, כדי להבין מהי הלוגיקה של עץ ההכרעה? אילו שאלות הוא שואל? האם תוכלו להסיק מהם הפיצ'רים החשובים בקלט (כלומר ב train\_data).

האם תוכלו להסביר בדיעבד מדוע אלו הפיצ'רים החשובים?

* 1. בדקו עצים עם עומקים שונים. איזה עומק נותן את הדיוק הגדול ביותר על סט האימון? ומה לגבי סט הבדיקה?
  2. היכן הפער בין סט האימון לסט הבדיקה גבוה יותר – ב decision tree או ב random forest? האם תוכלו להסביר מדוע?

1. **Logistic Regression**
   1. קיראו על logistic regression , למשל [פה.](https://en.wikipedia.org/wiki/Logistic_regression)
   2. הפעילו אתlogistic regression בערכי ברירת המחדל. מה הדיוק שלכם על סט האימון? מה הדיוק שלכם על סט הבדיקה? באם קיים הבדל, האם תוכלו להבין את סיבת ההבדל?
   3. ישנם מספר פרמטרים שמשפיעים על אופי המסווג. אחד החשובים הוא "C" שהוא גורם [הרגולריזציה](https://towardsdatascience.com/regularization-in-machine-learning-76441ddcf99a). לעיתים זהו פרמטר חשוב ביותר במניעת overfitting ולעיתים השפעתו זניחה.
      1. נסו לבדוק אותו בערכים שונים. מה קורה לדיוק בסט האימון, בסט הבדיקה ומה קורה לפער ביניהם, כאשר מגדילים או מקטינים את הפרמטר הנ"ל?
      2. מהו הפרמטר הטוב ביותר לסט האימון? מהו הפרמטר הטוב לסט הבדיקה? מי מביניהם הוא החשוב?
      3. האם פרמטר הרגולריזציה חשוב גם במקרה שלנו?
      4. המסווג מוציא התפלגות של הסתברויות על פני הסיווגים השונים באמצעות הפונקציה predict\_proba. השוו בין וקטורי ההתפלגות כאשר הניבוי של המסווג היה נכון, לבין וקטורי ההתפלגות כאשר הניבוי של המסווג היה שגוי. מה ניתן להסיק?
   4. המסווג למעשה מוצא משקולות שהם הבסיס לפונקציית המסווג. ניתן לצפות בהן בתכונות coef\_ , intercept\_.
      1. כמה משקולות בסך הכל מוצא המסווג?
      2. האם ניתן להבין מה בדיוק מבחינה טכנית המסווג עושה? מה משמעות המשקולות?