**פרק 4 : מודלים מבוססי עצים**

חלק תיאורטי:

1. Decision Trees:

Read chapter 9 for ESL, pages 305-313.

Explain to yourself the following, then discuss it with your supervisor.

1. How might one use a classification tree to return class probabilities?
2. What are the three measures used for building classification trees? What are their relative merits? What’s used for regression trees?
3. How can the Gini index be interpreted?
4. What are the disadvantages of decision trees?
5. What does pruning aim to solve? How is it done?
6. What might be a problem with classification decision trees when a categorical feature with many possible values? Suggest a solution to the problem.
7. Random Forset:

קרא מתוך ISLR עמודים 315-321.

1. What is the idea of bagging?

2. What is the trade off in choosing the number m of features to sample in every node?

3. What’s Out Of Bag Sampling? What is the size of the Out Of Bag set when the dataset is large?

4. How can one measure feature importance in random forest?

Now read about random forest in chapter 15 of ESL, pages 587-596. Skip 15.3.3 (Proximity Plots).

5. Solve exercise 15.1. What is its implication on random forest?

6. Demonstrate in code that when the number of variables is large, but the fraction of relevant variables is small, random forests are likely to perform poorly with small m.

1. Isolation Forest:

* לקרוא את ההסבר האינטואיטיבי בבלוגפוסט הבא:

<https://quantdare.com/isolation-forest-algorithm/>

* ואז קריאה יותר מעמיקה כאן: <https://towardsdatascience.com/outlier-detection-with-isolation-forest-3d190448d45e>

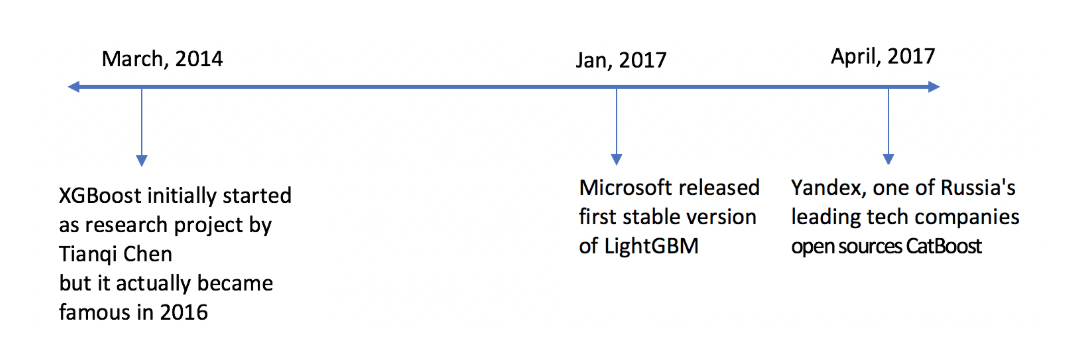
1. Boosting methods:

לקרוא את שני הבלוגפוסטים הבאים:

* בעיקר מבוא להבדלים בין bagging ל-boosting והסבר לאינטואיציה על boosting עם גזעי החלטה (עצים בגוגה אחד):<https://medium.com/mlreview/gradient-boosting-from-scratch-1e317ae4587d>
* כניסה לעומק לאחת השיטות הראשונות הטובות לboosting בשם adaboost והשוואה שלה לgradient boosting סטנדרטי: <https://towardsdatascience.com/boosting-algorithms-explained-d38f56ef3f30>

1. Advanced Boosting methods:

ראשית שים לב לסדר יציאתם של שלושת המאמרים:



וכעת נעבור לקריאה! (מצפים לשמוע ווהו מהעמדה שלך...)

* קרא את הבלוגפוסט הבא על מבוא מתמטי לGradient Boosting <https://towardsdatascience.com/the-good-old-gradient-boosting-f4614b0e62b0>
* קרא את הבלוגפוסט הבא על XGBOOST (זאת על מנת לקבל overview של האלגוריתם, לאחר מכן גם תקרא/י את המאמר עצמו) <https://towardsdatascience.com/xgboost-b736c2ab10ce>
* קרא את המאמר של XGBOOST (מצורף בתיקייה) – להתמקד בחלק הסיווגי ולא החישובי (נכנסים הרבה ליכולת מקבול ושימוש יותר טוב בביזור המודל על מספר מכונות, אין צורך להיכנס שם לעומק הנושא - אבל כן להבין את הקונספטים הכלליים שהשתמשו בהם ואיפה אפשר לשפר ביצועים).
* קרא את הבלוגפוסט על lightGBM <https://towardsdatascience.com/lightgbm-800340f21415>

בחלקים בהם את/ה מרגיש/ה כי לא מסבירים את כל הפרטים (למשל בחלק על EFB) לך למאמר (מצורף בתיקיה) ונסה להבין משם. אם עדיין הנושא לא ברור נסה חיפוש בפוסטים אחרים באינטרנט (גם במאמר עצמו הם לא סגרו כמה פינות שמן הסתם היו חייבים לסגור בקוד עצמו של הספרייה).

* קרא את הבלוגפוסט של טל פרץ (מי שהקים את הצוות הזה) על catboost [https://towardsdatascience.com/https-mediumtm-com-talperetz24-mastering-the-new-generation-of-gradient-boosting-db04062a7ea2](https://towardsdatascience.com/https-medium-com-talperetz24-mastering-the-new-generation-of-gradient-boosting-db04062a7ea2)

גם כאן אנו חושבים שהנקודה העיקרית של המאמר - ordered boosting, יחסית קשה להבנה (פשוט באמת לא מצאנו בלוגפוסט אחר יותר טוב), ושווה להבין אותה לעומק מהפסאודו-קוד שבמאמר עצמו שנמצא בתיקייה (אחרי הפסאודו-קוד יש גם הסבר שעושה יותר שכל). שוב, אם עדיין הנקודה לא ברורה נסה את מזלך אצל הרב גוגל :).

חלק מעשי:

בתרגיל זה נצלול לתרגיל הקאגל הראשון שלך. אומנם עד כה קראת וראית את מודלי העצים השונים כאלגוריתם מסווג, ניתן להפעיל אותם גם בתור רגרסור, והם אפילו די נפוצים.

הקאגל הבא: <https://www.kaggle.com/c/new-york-city-taxi-fare-prediction/overview>

מתאר מחירי מוניות בניו יורק, מכיל עמודות המתארות את הנסיעה, ומבקש שתייצר חוזה תעריפים.

עליך לבנות מסווג בעל דיוק מירבי לחיזוי הנ"ל.

בפתרונך עליך לכלול השלבים הבאים:

* אקספלורציה של המידע.
* ייצור פיצ'רים חדשים היכולים לעזור בעבודת המודל.
* בחינת מודלים (עצים ושיטות נוספות שלמדת)
* השוואה בין המודלים שנבחנו.
* מסקנות.