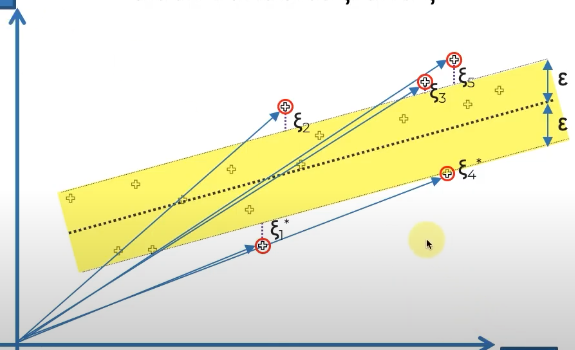
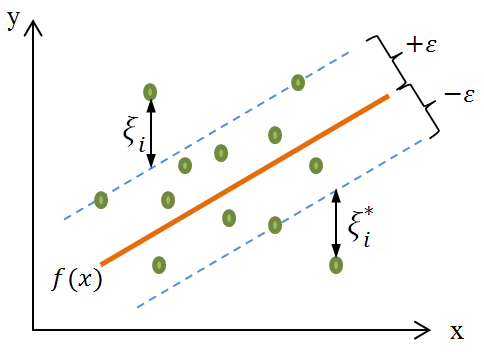
**הסבר SVR**

SVM (Support Vector Machines) הם אלגוריתמי למידת מכונה המשמשים לקלסיפיקציה, רגרסיה ומציאת רעש.

SVR (Support Vector Regressor) הוא מודל ניבוי העובד בצורה הבאה:

* מימד = פיצ'ר / עמודה



תחילה הוא ממפה את הדאטה במרחב n מימדי (בהתאם למספר הפיצ'רים) אבל כמובן לשם נוחות והבנה אנחנו נסתכל על המרחב כדו-מימדי (Y – משתנה המטרה, X – שאר הפיצ'רים).

הוא מותח קו רגרסיה כללי (הנקרא Hyperplane – הכתום – במימדים גדולים יותר מדובר במישור ולא בקו) ובהתאם לאפסילון מוגדר, הוא יוצר "צינור" שהרדיוס שלו הוא אפסילון (הנקרא -

Epsilon Insensitive Tube), השאיפה שרוב הדאטה יפול בתוך הצינור. בעצם המודל "מתעלם" מכל נקודות הדאטה שכעת נמצאות בתוך הצינור כדי למזער את השגיאה.

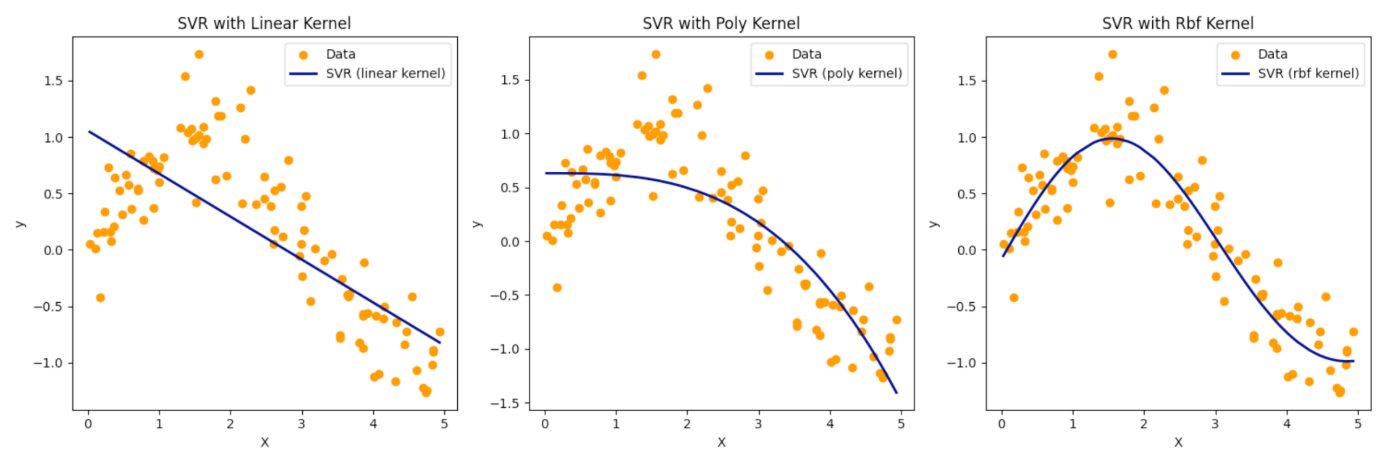
אפשר לחשוב על זה שקו הרגרסיה שלנו הוא פשוט הצינור (קו עווה מאוד) וכעת אנחנו מודדים את המרחקים של הדאטה מדפנות הצינור כשהמטרה שהמרחק (השגיאה) יהיה מינימלי.

הנקודות הכי קרובות לצינור נקראות Support Vectors כי ניתן לחשוב עליהם כוקטורים "התומכים" בצינור שלנו.

ל SVR קיימים אופציות שונות לקרנלים בהתאם לקשר הנתונים שלך.

Linear – כמו בדוגמא למעלה, כאשר יש קשר ליניארי (קורלציה) חזקה בין משתנים.

RBF – מתאים את הצינור להתנהגות הדאטה שלך ולכן מאוד מדויק אך צורך המון מהמחשב ולא מיועד לכמויות גדולות של מידע.



**פרמטרים: (משפיעים על אופי הצינור ומעניקים שליטה על שגיאות)**

Epsilon – קובע את "עובי" הצינור.

C – קובע את מידת פיתול הצינור.

ניסיונות בפועל:

תחילה ניסינו את גרסת הברירת מחדל של SVR (שהוא קרנל RBF) אך לאחר מעל 50 דקות של נסיון לאמן את המודל על הדאטה הבנו שזה לא ריאלי. לאחר מכן ניסינו להתחכם ולפצל את הדאטה לתתי קבוצות קטנות שעל כל אחד מהם מריצים את המודל ואז בסוף מריצים את המודל על כל ה Support Vectors שמצאנו בתתי קבוצות אך גם זה לא עבד. בסוף ניסינו את הקרנל הליניארי שעבד מהר מאוד אך מכיוון שהקשרים בין המשתנים כלל לא ליניאריים אז המדדי תוצאה היו נמוכים מאוד ללא מקום לשיפור משמעותי ולכן עברנו לחפש מודל שמתאים לקשרים לא ליניאריים ועל דאטה גדול.

**הסבר מפורט יותר:**

ניסיונות לשימוש באלגוריתם:

במהלך הפרויקט ניסינו שני אלגוריתמים שונים לחיזוי מחירי הטיסות: SVR עם גרעין RBF ו- LinearSVR. למרות הניסיונות לכוונן את המודלים ולשפר אותם, התוצאות שהתקבלו היו גרועות.

ניסיון ראשון - SVR עם גרעין RBF: האימון נמשך יותר מ-50 דקות, ולכן בדקנו את סיבוכיות הזמן ומצאנו שהיא בין O(N^2) ל O(N^3) – לא פרקטי לדאטה גדול.

ניסיון שני - ניסינו לשפר את הביצועים באימון ב-Batches – חלוקה למקטעים, אימון כל מקטע בנפרד ושמירת וקטורי התמיכה בלבד.

גם אימון ב-Batches לא פתר את הבעיה, ולכן הפסקנו את התהליך באמצע.

#הסיבה שזה לא עבד היא שגרעין RBF מתאים לזיהוי קשרים לא ליניאריים מורכבים, אבל הוא אינו מתאים לדאטה גדול בגלל סיבוכיות חישובית גבוהה.

ניסיון שלישי - :LinearSVR עברנו למודל LinearSVR, שהוא פשוט יותר, עם גידול ליניארי בסיבוכיות. אימנו את המודל וביצענו חיזוי, אך קיבלנו ביצועים חלשים מאוד שמראים שהמודל כמעט לא מסביר את השונות במחירי הטיסות ויש פערים גדולים בין התחזיות לערכים האמיתיים.

ניסיון רביעי - כוונון היפר-פרמטרים (Hyperparameter Tuning):

השתמשנו ב- GridSearchCV לחיפוש הערכים האופטימליים של C ו-אפסילון אך גם כאן לאחר אימון מחדש עם הפרמטרים האופטימלים, הביצועים כמעט לא השתפרו.

אמנם LinearSVR עובד טוב עבור דאטה גדול, אבל המודל הליניארי מתקשה להתמודד עם קשרים לא ליניאריים בין המשתנים. ובגלל שהתכונות שלנו כנראה לא מסבירות היטב את המחירים, המודל מתקשה לחזות.

* **באופן כללי, כל אלגוריתם שממפה את הדאטה שלו (כמו אלה) מצריך Scaling/Normalizing כדי שהטווחים יהיו שווים ובכך לא יהיו נקודות דאטה במרחק עצום מהשאר, ובכך אני מונעים הטיה.**