

מדען נתונים, על מכונת וקטורים תומכים (SVM) כבר שמעת?

מכונת וקטורים תומכים (Support Vector Machine) הוא אלגוריתם דו-מהותי, כלומר כזה המשמש לפתרון בעיות רגרסיה ובעיות סיווג כאחד. מדובר בטכניקה ותיקה של למידה בהשגחה (Supervised Learning). יצאתי לראיין את מייסד ומנכ"ל האיגוד הישראלי למדעני נתונים מקצועיים (PDSIA), האקטואר רועי פולניצר קצת עליו וקצת על אלגוריתם ה-SVM



האקטואר רועי פולניצר הינו מדען נתונים מקצועי (Professional Data Scientist) מוסמך על ידי האיגוד הישראלי למדעני נתונים מקצועיים (PDSIA - Professional Data Scientists' Israel Association). רועי הינו מדען נתונים עצמאי מזה 10 שנים, המספק לגופים שונים פתרונות מדעיים לאתגרים הכרוכים בעבודה עם כמויות גדולות ומגוונות של נתונים, מבצע מחקרים להפקת תובנות עסקיות מנתונים עבור אותם גופים (Business Intelligence), מטייב ומסדר את מידע המשמש למחקרים ומיישם ומתקף אלגוריתמים ומודלים שונים של כריית נתונים ו-Machine Learning על המידע. יצאתי לראיין אותו בנושא מכונת וקטורים תומכים (Support Vector Machine).

את רועי אני מכיר מתוכנית ההכשרה היוקרתית, Data Science, Learning with Python Machine Learning, and Deep מכללת הי-טק (חטיבת ההדרכה של מטריקס). רועי סייע לי בהתחלה בתכנות בשפת Python באמצעות Google Colab ולאחר מכן בביצוע עבודת חקר באמצעות Jupyter notebooks, אירגון וחקר נתונים באמצעות הספריות NumPy, Scipy ו-Pandas וויזואליזציה נתונים באמצעות matplotlib ו-seaborn. ראיון זה יסוב סביב קטגוריה פופולרית של מודלים של למידה בהשגחה (Supervised Learning) המוכרת בשם מכונת וקטורים תומכים (SVM). קטגוריה זו הוצגה לראשונה על ידי קורטס וואפניק בשנת 1995. ממש כמו באלגוריתם עצי החלטה (Decision trees), באלגוריתם יער אקראי (Random forest), ובאלגוריתם השכן הקרוב ביותר (K Nearest Neighbors), שעליו רועי הרצה בתוכנית ההכשרה שלנו, גם באלגוריתם SVM ניתן להשתמש הן לסיווג והן לחיזוי/ניבוי משתנה רציף.

המכונה (ML), בניית דו"ח מחקר, יישום מודל החיזוי במערכת הייצור (Deployment) והדרכת משתמש הקצה בהטמעתו. באותה שנה קיבלתי גם הסמכה ישראלית "מנהל סיכונים מוסמך" (CRM- Certified Risk Manager) של האיגוד הישראלי למנהלי סיכונים (IARM- Israeli Association of Risk Managers). ההסמכה הזו כלה ידע בניית אשכולות (Cluster analysis) עם K-Means, ניתוח אשכולות היררכי (Hierarchical Cluster Analysis) וצמצום ממדים באמצעות ניתוח מרכיבים עיקריים (Principle Components Analysis).

בשנת 2018 קיבלת הסמכת אקטואר מלא ("Fellow") שהיא המעמד המקצועי הגבוה ביותר בלשכת מערכי השווי והאקטוארים הפיננסיים בישראל (IAVFA), איך זה קשור ללמידת מכונה?

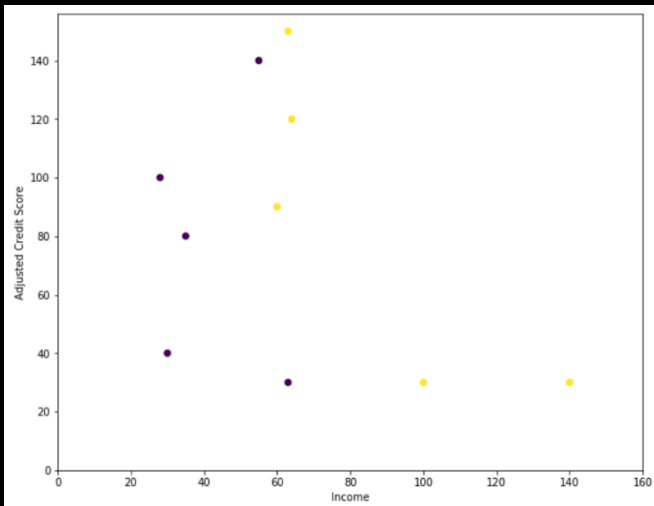
ההסמכה האמורה קשורה פחות ללמידת מכונה ויותר ללמידה עמוקה (Deep Learning) והיא כוללת ידע במה שנקרא רשתות נוירונים (Multi-layer perceptron, Recurrent neural networks, LSTM, Convolution ו-pooling) עיבוד תמונה (Object recognition and object detection) ועיבוד שפה (Application in NLP).

בשנת 2019 אתה מתחיל ללמוד Python ובסוף השנה אתה גם פותח קורס חד-יומי לתכנות בשפת Python וגם מקבל הסמכת מדען נתונים מהאיגוד הישראלי למדעני נתונים (PDSIA), איך זה קורה?

ראשית, שפת Python מאוד דומה לשפת R שאותה למדתי בשנת 2011 באוניברסיטת חיפה ובה תכנתתי מאז. כך שהיה לי בסיס טוב מאוד ל-Python. אני רק אציין שבסוף 2019 גם לימדתי קורס חד-יומי לתכנות בשפת R. לגבי הסמכת "מדען נתונים מקצועי" (PDS-Professional) (Data Scientist) שקיבלתי מהאיגוד הישראלי למדעני נתונים מקצועיים (PDSIA - Professional Data Scientists' Israel Association), ההסמכה האמורה נשענת על 4 אדנים: תכנות ב-Python, ניתוח נתונים וויזואליזציה ב-Python, למידת מכונה ולמידה עמוקה. שים לב שהרזומה שלי עד לאותו רגע כלל את 4 האדנים הללו.

ברשותך נעבור לאלגוריתם מכונת וקטורים תומכים (SVM). מתי נתקלת בו לראשונה?

אני למדתי את המודל הזה בשנת 2012, בזמנו קרו לזה מודל ולא אלגוריתם, כאשר למדתי למבחנים הבינלאומיים להסמכה בתחום של ניהול סיכונים (FRM).



האם תוכל לספר על הקשר שלך לעולם למידת המכונה ומדע נתונים?

ברצון. בשנת 2006 סיימתי תואר ראשון באוניברסיטת בן גוריון בכלכלה עם התמחות במימון, מדע נתונים ולמידת מכונה. כמובן שאז לא קרו לזה למידת מכונה אלא אקונומטריקה או כלכלה אמפירית, אבל איך שלא תהפוך את זה, בסוף התמחתי באיסוף מידע ממגוון מערכות, בעבודה עם כמויות גדולות של מידע, בעיבוד מידע לא מובנה, בניתוח סטטיסטי וביצירת חיבור בין בסיסי נתונים שונים. בשנת 2008 סיימתי תואר שני באותה אוניברסיטה במנהל עסקים עם התמחות באקטואריה, מדע נתונים ולמידת מכונה. אז כבר קרו לזה מידול סטטיסטי או ניהול סיכונים, אבל עדיין עסקתי בתואר די הרבה בניתוחים אנליטיים, בחיזוי, בכריית נתונים, באופטימיזציה, בעיבוד מידע, באנליזה לנתונים כמו גם בבניית כלי ויזואליזציה שונים להצגתם.

אז למעשה כבר בשנת 2008 יצאת לשוק עם ידע בלמידת מכונה ומדע נתונים. אבל אני זוכר שעבדת כעוזר מחקר כמה שנים, אז האם גם שם למדת את התחום?

אכן כך. משנת 2006 ועד כמעט סוף 2010 שימשתי כעוזר מחקר של ד"ר שילה ליפשיץ ז"ל, בתחום ניהול הסיכונים במערכת הבנקאות הישראלית. ב-4 השנים הללו עסקתי בגדול שני דברים. האחד, באיסוף מידע, ביצוע מחקרים אמפיריים בעזרת שיטות מחקר וביצוע תהליכי הכנת נתונים (יבוא נתונים, תחקור, ניקוי, השלמת ערכים חסרים, נורמליזציה, הנדסת משתנים ובחירת משתנים למודל). השני, בבניית מודלי ניבוי (כגון: Decision Trees, KNN, Naïve Bayes, Logistic Regression, Ensemble Learning ו-Boosting). בחירת שיטת הערכה מתאימה (Model selection) שיפור ניבוי (כוונון הפרמטרים) וטיפול בסוגיות מיוחדות במדע נתונים (Overfitting ו-Underfitting, Data leakage).

ואז בשנת 2011 אתה מתחיל ללמוד למידת מכונה באוניברסיטת חיפה ובאוניברסיטת אריאל במקביל?

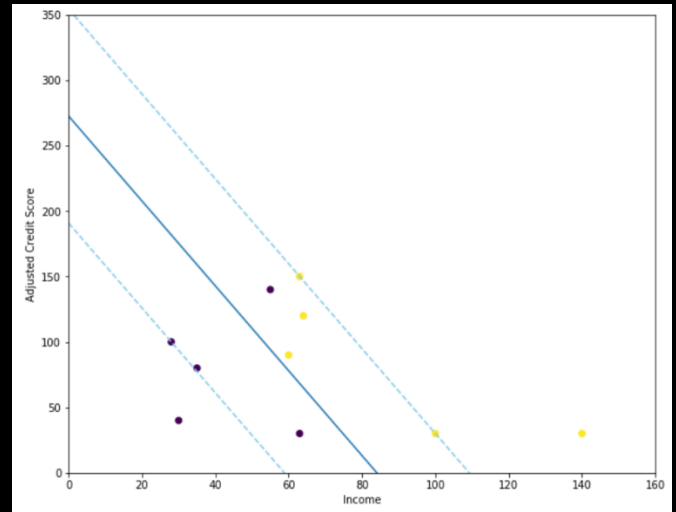
באוקטובר אותה שנה התחלתי ללמוד לימודי תעודה באקטואריה ולמידה סטטיסטית באוניברסיטת חיפה. הלימודים כללו בעיקר סטטיסטיקה תיאורית (מבוא לסטטיסטיקה, סוגי נתונים והצגתם באופן טבלאי וגרפי, מדדי מרכז ומדדי פיזור, אחוזונים, מדדי קשר והתפלגות הנתונים), הסתברות (הסתברות פשוטה במרחב הסתברותי אחיד ובמרחב הסתברותי לא אחיד, הסתברות מותנית, נוסחת בייס, משתנים מקריים בדידים, כגון: ניסוי ברנולי, התפלגות בינומית, התפלגות פואסונית, התפלגות היפרגאומטרית ומשתנים מקריים רציפים כגון: התפלגות נורמלית), הסקה סטטיסטית (אמידה נקודתית, רווחי סמך ומבחני השערות), סטטיסטיקה א-פרמטרית ותכנות בשפת R (שהיא שפת תכנות תכנות מדעי וסטטיסטי). בתחילת דצמבר 2011 התחלתי לימודי דיפלומה בניהול סיכונים באוניברסיטת אריאל בשומרון. הלימודים כללו בעיקר מודלים לינאריים בבעיות גרסיה (רגרסיה לינארית, Ridge, Lasso), מודלים לינאריים בבעיות סיווג (רגרסיה לוגיסטית ו-LDA) וסדרות עתיות (Time series).

בשנת 2013 קיבלת שתי הסמכות במדע נתונים ולמידת מכונה?

כן. בשנת 2013 קיבלתי את ההסמכה הבינלאומית היוקרתית "מנהל סיכונים פיננסיים" (FRM- Financial Risk Manager) של האיגוד העולמי למומחי סיכונים (GARP- Global Association of Risk Professionals) שיושב בניו ג'רזי, ארה"ב. ההסמכה כללה ידע פרקטי בתחקור נתונים, ניקוי נתונים, הנדסה ובחירת משתנים, יצירת מודלי חיזוי באמצעות שיטות K Nearest Neighbors, עצי החלטה (Decision Trees), יער אקראי (Random Forest), מכונת וקטורים תומכים (SVM) והמסווג הנאיבי של הנאיבי של בייס (Naïve Bayes), והערכתם, ביצוע QA לתהליך למידת

באיזה הקשר למדת את המודל?

למדתי את מודל ה-SVM בהקשר של ניתוח ודירוג אשראי. תחשוב לרגע שיש לך תרשים פיזור של הלוואות שחלקן נפרעו במלואן וחלקן הגיעו לחדלות פירעון.



מהן המשוואות עבור הקצה העליון והקצה התחתון של הנתבי בקלסיפיקציה לינארית עם ה- m מאפיינים (קרי, משתנים מסבירים) במונחים של המשקולות (קרי, מקדמים) w_j וערכי המאפיינים x_j ?

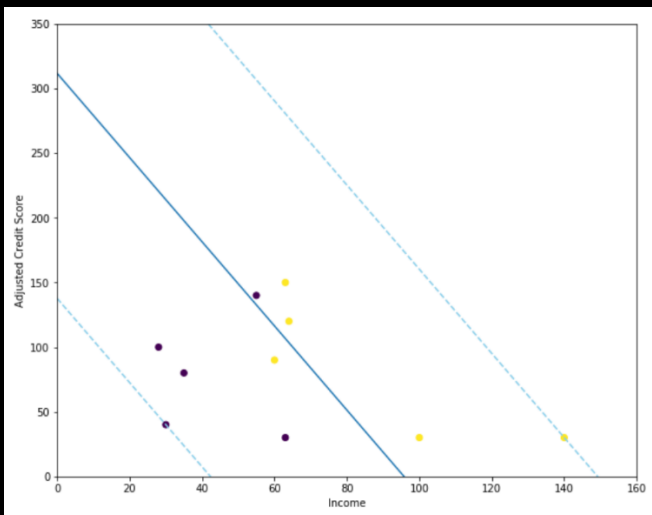
$$\sum_{j=1}^m w_j x_j = b_u$$

$$\sum_{j=1}^m w_j x_j = b_d$$

מבלי לפגוע בכלליות ניתן לכייל או לבצע קליברציה לפרמטרים כך שהמשוואות תהפוכנה ל-:

$$\sum_{j=1}^m w_j x_j = b + 1$$

$$\sum_{j=1}^m w_j x_j = b - 1$$



מודל ה-SVM, מייצר לך קווים וקטורים תומכים שיורדים משמאל לימין ומחלקים הלכה למעשה את התרשים לשתי קבוצות: משמאל למכונת הוקטורים התומכים ישנן הלוואות "רעות" ומימין למכונת הוקטורים התומכים ישנן הלוואות "טובות".

מה זה בכלל וקטור תומך?

וקטור תומך (Support vector) זהו למעשה תצפית בקצה העל-מישור (שהוא המפריד הליניארי שמפריד את המרחב לשני חצאי מרחבים שכל אחד מהם אמור להכיל בעיקר דוגמאות מסוג אחד). על מנת לא לאבד את הקוראים נכנה את העל-המישור הזה בכינוי נתיב (pathway).

הבנתי שיש SVM לבעיות רגרסיה ו-SVM לבעיות סיווג, זה נכון?

אכן. קלסיפיקציית SVM פירושה יצירת נתיב מסוים לסיווג תצפיות לאחת משתי קטגוריות, בעוד שרגרסיית SVM פירושה שימוש בנתיב מסוים לחיזוי/ניבוי ערכו של משתנה יעד (קרי, משתנה מוסבר) רציף.

מהי המטרה בקלסיפיקציית SVM?

המטרה בקלסיפיקציית SVM היא כאמור למצוא את הנתבי שמשווה את התצפיות נכון ככל האפשר כך שהתצפיות שבקבוצה אחת שוכבות מצידו האחד של הנתבי והתצפיות שבקבוצה האחרת שוכבות מצידו האחר של הנתבי.

מה ההבדל בין קלסיפיקציה מסוג הפרדה קשיחה לבין קלסיפיקציה מסוג הפרדה רכה?

בקלסיפיקציה מסוג הפרדה קשיחה (hard margin), המטרה היא למצוא את רוחב הנתבי כך שלא יתרחש סיווג שגוי של תצפיות (בהנחה שנתבי כזה אכן קיים). מאידך, בקלסיפיקציה מסוג הפרדה רכה (soft margin) פונקציית המטרה כוללת בחובה תחלופה (trade-off) בין רוחב הנתבי לבין מידת הסיווג השגוי.

מה קורה לרוחב של הנתבי ככל שהמשקולות w_j עולות?

הרוחב של הנתבי יורד ככל שהמשקולות עולות. זוהי הסיבה לכך שאנו ממזערים או מביאים למינימום פונקציה שכוללת את סכום ריבועי המשקולות.

מה קורה לרוחב של הנתבי ככל שהעלות המיוחסת להפרות (Violations Cost) עולה?

הרוחב של הנתבי יורד ככל שאנו נותנים משקל רב יותר להפרות.

כיצד נמדדת מידת ההפרה בקלסיפיקציה לינארית מסוג הפרדה רכה?

מידת ההפרה מוגדרת כמרחק הקצר ביותר שעל התצפית לזוז במרחב המאפיינים, על מנת שתסווג נכונה.

האם תוכל לקראת סיום הראיון לשפוך אור על מה שאתה עושה כמדען נתונים עצמאי?

בגדול כמדען נתונים עצמאי, במסגרת פרוייקטים של מדע נתונים שאני מבצע ללקוחות, אני תמיד מתחיל בכתיבת פרוטוקול מחקר לפני תחילת כל פרוייקט. לאחר מכן אני מכין את הנתונים שאני מקבל מהלקוחות על מנת שהנתונים יהיו נקיים ומוכנים לאנליזה. לאחר מכן, מתחיל את השלב האהוב עליי שהוא פיתוח מודלים מנבאים תוך שימוש בכלים מתאימים, עריכת ניתוחים וביצוע בקרת איכות של הפרוייקט. כמובן שכל פרוייקט אני חותם בכתיבת דו"ח סיכום המסביר את המתודולוגיה ששימשה אותי בפרוייקט כמו גם את המודלים שפיתחתי ואת נוהלי התיקוף שביצעתי. לרוב, אני מתבקש על ידי הלקוחות לבוא ולהטמיע אצלם את המודלים שפיתחתי ואף להדריך את משתמשי הקצה כיצד לעשות בהם שימוש נכון ומושלם. עד לפני שנתיים מרבית המודלים שבניתי היו מודלים של בעיות רגרסיה (Regression), מודלים של בעיות סיווג (Classification) ומודלים של בעיות ניתוח אשכולות (Clustering). בשנים האחרונות התחלתי לבנות מנועי המלצה (Recommendation systems) ומודלים של בעיות זיהוי אנומליות (Anomaly detection) ובשנה האחרונה אני בונה גם מודלים של ניתוח טקסט ועיבוד שפה טבעית (Text Analytics & NLP) ומודלים של זיהוי תמונות (Image processing).



כיצד אם כך היית מגדיר בשניים שלושה משפטים את מה שאתה עושה כמדען נתונים?

אני מבצע מחקרי מידע מעמיקים לשם הפקת תובנות עסקיות, אני מנקה, מטייב ומסדר את מידע המשמש למחקרי מידע, אני מנתח את המידע באמצעות כלים סטטיסטיים ומייצר ויזואליזציות שלו ולבסוף אני מפעיל אלגוריתמים שונים של מידול, כריית מידע ו-Machine Learning על הנתונים.

מידת ההפרה עבור תצפית בעלת תוצאה חיובית (positive outcome) הינה:

$$\max \left(b + 1 - \sum_{j=1}^m w_j x_j, 0 \right)$$

מידת ההפרה עבור תצפית בעלת תוצאה שלילית הינה:

$$\max \left(\sum_{j=1}^m w_j x_j - b + 1, 0 \right)$$

כיצד ניתן להרחיב את המתודולוגיה לקלסיפיקציה לינארית כך שתתאים גם לקלסיפיקציה א-לינארית?

פשוט מתמירים את המאפיינים הקיימים ויוצרים מאפיינים חדשים כך שניתן יהיה להשתמש בהם בקלסיפיקציה לינארית.

מהו ציון דרך (Landmark) ומהי פונקציית בסיס רדיאלי גאוסייני (RBF- Radial Basis Function)?

ציון דרך הוא נקודה במרחב המאפיינים, אשר עשויה להתאים או שלא להתאים לתצפית מסוימת, ומשמשת ליצירת מאפיין חדש. פונקציית בסיס רדיאלי גאוסייני היא פונקציה של המרחק של תצפית מסוימת מציון דרך מסוים והיא משמשת ליצירת מאפיין חדש לצורך קביעת נתיב א-לינארי ב-SVM. למעשה פונקציית בסיס רדיאלי גאוסייני הינה מאפיין סינטטי שערכו עבור תצפית מסוימת יורד ככל שהתצפית מתרחקת מציון הדרך.

האם תוכל להסביר מהי המטרה של רגרסיית SVM?

ברגרסיית SVM המטרה היא למצוא נתיב מסוים, בעל רוחב שנקבע מראש, שחוצה את המרחב שהוגדר על ידי היעד והמאפיינים. הנתיב נועד לכלול תצפיות רבות ככל האפשר, כאשר תצפיות שמחוץ לנתיב יוצרות הפרות. פונקציית המטרה ממזערת את מידת ההפרות ומקפלת בתוכה מידה מסוימת של רגולריזציה. אציין כי בדומה לרגרסיה מסוג Ridge או Lasso, הרגולריזציה נועדה לאפס (ב-Lasso) או להשאיף לאפס (ב-Ridge) את המשקולות של המאפיינים במודל ובכך ומקטינה את מידת הסיבוכיות של המודל, שני דברים שביחד מקטינים את ה-Overfitting של המודל.

מהם ההבדלים העיקריים בין רגרסיית SVM לבין רגרסיה לינארית פשוטה?

אני רואה 4 הבדלים עיקריים ברורים. ההבדל הראשון הוא שברגרסיית SVM הקשר שבין היעד לבין המאפיינים מיוצג על ידי נתיב ולא על ידי קו בודד. ההבדל השני הוא שברגרסיית SVM שגיאת הניבוי (error predictor) נוספת כאפס כאשר תצפית מסוימת שוכבת בתוך הנתיב. ההבדל השלישי הוא שברגרסיית SVM השיאויות עבור התצפיות שמחוץ לנתיב מחושבות כהפרש שבין ערך היעד לבין הנקודה הקרובה ביותר בנתיב, שעקבית עם ערכי המאפיינים. ההבדל הרביעי שברגרסיית SVM קיימת רגולריזציה מובנית (built-in) בתוך פונקציית המטרה עצמה.

מהי רגולריזציה?

שאלה טובה. רגולריזציה היא פישוט של מודל מסוים במטרה להימנע מ-Overfitting על ידי איפוס או השאפה לאפס של משקולות המאפיינים במודל.