חיזוי שינוי חד במדד מניות   
S&P500

פרויקט בבינה מלאכותית   
236502

מגישים:

שקד מגדל 312140411

מאור ביתן {206239071}

מנחה: תומר לנגה



מבוא:

הוא מדד מניות העוקב אחרי 500 מניות גדולות בבורסות בארה"בS&P500

ואחד המדדים הנצפים ביותר בעולם, המניות המרכיבות אותו מחולקות לפי גודל כאשר 10 המניות הגדולות ביותר במדד מהוות 26% מהמדד כולו.

{https://en.wikipedia.org/wiki/S%26P\_500\_Index} הכנס כמקור

מעקב אחרי המדד מתבצע באופן קבוע בתקשורת העולמית ובפרט בתקשורת הכלכלית {הכנס כמקור רויטרס, יאהו}, כאשר שינויים חדים במדד יגיעו במקרים רבים לכותרות העיתונים, בנוסף מאחר והמדד עוקב אחרי 500 חברות אמריקאיות גדולות בשווי של מעל ל30 טריליון דולר (30,000,000,000,000) כתבות על חברות מהמדד מתפרסמות חדשות לבקרים.  
המדד כל כך משמעותי שהוא משמש בין היתר כאינדיקציה למצב הכלכלה האמריקאית.

(רעיון להוספה: לפרט על תנאים לכניסה למדד)

חלק מהמסחר המודרני בבורסה משתייך היום אל "סוחרי אלגוריתמים" שזה שם מכליל לכל הסוחרים המבצעים פעולות אוטומטיות בעזרת אלגוריתמי מסחר, בין היתר קרנות גידור וקרנות השקעה אך גם סוחרים פרטיים. האלגוריתמים יכולים להעדיף עסקאות טווח קצר (כמה שניות) או ליישם אסטרטגיית מסחר לטווח ארוך (ימים וחודשים קדימה) }https://he.wikipedia.org/wiki/%D7%9E%D7%A1%D7%97%D7%A8\_%D7%90%D7%9C%D7%92%D7%95%D7%A8%D7%99%D7%AA%D7%9E%D7%99{

על מנת ליישם אלגוריתם מסחר עליך לנסות לחזות מגמה או מחיר של מנייה או מדד בעתיד, בין אם כמה שניות קדימה או שבוע קדימה, מרגע שיש חיזוי קיימים כלי מסחר על מנת לסחור במנייה לפי החיזוי, "שורט" אם החיזוי הוא לירידה או "לונג" אם החיזוי הוא לעלייה. כמו כן יש אפשרות לקנות את המנייה עצמה.

"מדד הפחד" – הוא כינוי הניתן למדד VIX }https://finance.yahoo.com/chart/%5EVIX/#{

שנוצר ככלי למדידת תנודתיות השוק, הסיבה שהוא נקרא "מדד הפחד" קשורה לעובדה שכאשר השוק תנודתי מאוד והמדד עולה סביר יותר שמשקיע יראה תזוזות גדולות יותר בשווי המניות (לרוב לאו דווקא בכיוון אחד) מה שמוביל לפחד אצל משקיעים שאינם לוקחי סיכונים https://engineering.nyu.edu/sites/default/files/2019-03/Carr-why-is-vix-a-fear-gauge.pdf

מדד הפחד}https://www.investopedia.com/terms/v/vix.asp{

הבעיה אותה ננסה לפתור:

הבעיה אותה ננסה לפתור היא זיהוי תנודתיות חריפות במדד. הסיבה לכך היא שתנודות קטנות במדד הן משהו שקורה על בסיס כמעט יומיומי, וניסיון לזיהוי המחיר עשוי להיות מאוד בעייתי שכן לעיתים קרובות רשתות נוירונים(הכלי בו נשתמש לניבוי המגמות) מנסה להתחקות אחר מחיר היום הקודם. דבר שאולי יראה כניבוי טוב אך בפועל כאשר תהיה ירידה או עלייה חדה ביום העוקב-שימוש בכלי כזה עשוי ליצור בעיות רבות. לעומת זאת-לחזות מגמת עליה או ירידה חדה במדד ביום העוקב עשויה להיות בעלת ערך רב. נשים לב שזוהי בעיית סיווג-כלומר אנחנו רוצים לחזות ביום מסוים האם תהיה עלייה חדה, ירידה חדה או אף אחד מהם. כמו כן נרצה לקבל ניבוי אשר איננו העתקה של היום הקודם (כלומר-המסווג לא ינבא עליה או ירידה חדה בגלל שביום או יומיים הקודמים הייתה עלייה או ירידה חדה).

:תיאור פתרון הבעיה

על מנת לבצע את הניבוי הרצוי, נשתמש במערכת לומדת, ובפרט רשת נוירונים.

הבעיה הניצבת דורשת לבצע ניבוי על העתיד על סמך מידע על העבר. כלומר מדובר בבעיית רגרסיה. רשת הנוירונים בה נשתמש הינה מסוג rnn ובפרט lstm

נגדיר מספר מושגים:

רשת נוירונים-זהו מודל מתמטי חישובי אשר מבנהו דומה למבנה הרשת העצבית במוח ומשמש ללמידת מכונה. ברשת הנוירונים קיימים יחדות חישוביות רבות המקושרות זו לזו, כאשר כל אחת מקבלת פלט ומוציאה פלט כלשהוא. השימושים של רשתות נוירונים הם רבים: זיהוי פנים תמונות, דיבור וזיהוי מגמות בשוק ההון-שזו תהיה כאן המטרה שלנו.

Rnn-Reccurent neural network. זוהי קבוצה של רשתות נוירונים בעלות מבנה של גרף לא מכוון, כאשר לכל נוירון ברשת יש activation function וזיכרון, הפלט של רשת הנוירונים תלוי במצב(state) הקודם שלה והקלט שהתקבל באותו state. צורת החישוב של הפלט היא רקורסיבית ונתונה ע"י הנוסחה הבאה(כאשר היא פונקציה כלשהיא)

LSTM-זוהי רשת נוירונים מסוג rnn שבאה לפתור את בעיית ה short-term memory- מצב שבו רשת rnn מקבלת כמות גדולה של sequential data ומתקשה "לזכור" את הדאטא המוקדם ולהשתמש בו בשלבים מאוחרים יותר וכן את בעיית ה vanishing gradient – גראדיאנט הינו ערך שנועד לעדכן את משקלי הרשת, אך בבעיה זו הגראדיאנט נהיה קטן יותר ויותר וכאשר נהיה קטן מידי הוא אינו מצליח לתרום הרבה ללמידה. המיוחד ברשת זו הוא בנוירונים שלה: הנוירונים מכילים מספר פונקציות אקטיבציה(sigmoid, tanh, pointwise multiplication, pointwise addition, vector concantation) אשר מאפשרים לרשת "לשכוח" מידע קודם וכך בעצם ניתן להשתמש במידע שהתקבל גם בטווח הקצר וגם בטווח הארוך(למעשה בוררים את המידע הרלוונטי)

Softmax-פונקציית אקטיבציה שהפלטים שלה הם הסתברויות לסיווג של הקלט לתווית מסויימת. למשל אם ברצוננו לבצע סיווג בינארי לדוגמה מסוימת, אז הפלט של softmax יהיו 2 הסתברויות-אחת היא ההסתברות לכך שהסיווג הוא + וההסתברות השנייה היא שההסתברות היא –(הסיווג יכול להיות למספר תוויות). סכום ההסתברויות הוא תמיד 1.

הפונקציה בה הרשת משתמשת כדי למדוד את הביצועים שלה. היא מחושבת בזמן האימון -Loss\_function

ומטרתה למדוד את המרחק בין מה שהרשת ניסתה לנבא לבין הערך האמיתי שנדרש לנבא. הרשת מנסה למזער ערך זה.

Learning rate- הקצב שבו מתבצע האימון. ערך קטן מידי משמעו אימון איטי מידי וערך גדול מידי

עשוי לגרום לעדכון חד במשקלים ולתוצאות לא טובות.

Epoch-זהו hyperparameter שקובע את מס' הפעמים שכל הdataset יעבור דרך הרשת.

Batch size- זהו hyperparameter שקובע את החלוקה של ה dataset לקבוצות של דוגמות. Batch size x יגרום לכל שהdataset יהיה מחולק לקבוצות של x דוגמות.

**פתרון המערכת**:

הבעיה כאמור, תיפתר ע"י שימוש במערכת לומדת הפתרון יתבצע במספר שלבים:

1.איסוף הדאטא אשר ישמש ללמידה, הכולל מידע היסטורי וכתבות כלכליות.

2.עיבוד המידע- ביצוע ניסיון לבצע ניתוח סנטימנט לכתבות הכלכליות שלנו ושילוב מקורות המידע כדי לייצור דאטא סט של דוגמות time series.

3.הפעלת אלגוריתמי הלמידה וניתוח התוצאות.

נתאר את כל אחד משלבי הפתרון:

1.איסוף מידע- על מנת לחזות את מדד s&p500 אספנו כמה סוגים שונים של דאטא:

א. מהאתר של yahoo finance, אספנו מידע היסטורי על המניות הנמצאות ב s&500 מה3 שנים האחרונות. עבור כל מנייה, נתון לנו מחיר פתיחה, מחיר סגירה, high, low Volume, adj\_close. כתבנו סקריפט בפיתון על מנת להוריד באופן אוטומטי את כל המידע ההיסטורי הדרוש לנו מהאתר.

ב. מדד הפחד-השתמשנו במידע ההיסטורי על מדד הפחד מה3 שנים האחרונות. הסיבה לכך היא שמדד הפחד הינו אינדיקטור טוב לתנודות חדות במחירי מניות, ובפרט מדד s&p500 המושפע מהן באופן ישיר. את המידע על מדד הפחד הישגנו גם כן מ yahoo finance

ג. כתבות כלכליות- קיים קשר חזק בין חדשות כלכליות הנמצאות באתרי החדשות השונים לבין התנודות במחירי המניות ובפרט מחיר הסגירה של המדד אותו נרצה לנבא. הכתבות בהן השתמשנו התפרסמו באתר Reuters וגם הן מה3 שנים האחרונות.

2.עיבוד המידע-המידע אותו השגנו נלקח מ 3 מקומות שונים, ולכן היה צורך לעבד אותו כך שהמערכת תוכל ללמוד את אותו כרצף של דוגמות ממוינות לפי תאריך(שכן לסדר הדוגמות יש משמעות כי אנו מנסים לנבא משהו על העתיד לפי העבר).

עבור הכתבות הכלכליות, החלטנו כי ברצוננו לעשות להן ניתוח סנטימנט. עבור כל תאריך ב 3 השנים האחרונות, היו לנו מספר כתבות כלכליות. עבור כל כתבה כזו ביצענו ניתוח סנטימנט ע"י הסרת סימני פיסוק, הפיכת הטקסט למערך של מילים, הסרת מספרים, ביצוע lemmatization (שינוי הטיות של מילים לצורה בסיסית) וכן הסרה של stopwords שזוהי רשימה של מילים אשר עשויות להפריע לתהליך הניתוח(את הרשימה יצרנו בעצמנו). לאחר מכן ביצענו על הטקסט המעובד ניתוח סנטימנט בעזרת textblob.{אולי לפרט קצת על מה קורה בניתוח של textblob} עבור כל תאריך עשינו מיצוע על הסנטימנטים של כל הכתבות באותו תאריך, ולבסוף קיבלנו קובץ שמכיל תאריך וסנטימנט לאותו תאריך.

עבור המידע ההיסטורי, אוסף דוגמות אשר מורכבות משילוב המידע ההיסטורי של כל אחת מהמניות מ s&p500(יחד עם מדד הפחד). לשם כך יצרנו תיקייה שתכיל את כל קבצי המידע אותם נרצה לשלב, ועשינו join ביניהם לפי תאריך.

לבסוף, עשינו join בין אוסף הדוגמות המכיל את המידע ההיסטורי המשולב יחד עם קובץ הסנטימנטים לפי תאריך.

לאחר שקיבלנו את אוסף הדוגמות המשולב, נרצה להוסיף פיצ'ר נוסף של מגמה(trend) שאותו נרצה ללמוד. על מנת לקבל אותו נרצה תחילה לבצע "החלקה" למחיר הסגירה של המדד(המדד מכיל תנודות רבות ואנו מעוניינים לחזות תנודות חדות. אי ביצוע החלקה עשוי לבלבל את הרשת ולתת תוצאות פחות טובות). ה trend הוא פיצ'ר בינארי כאשר 1 משמעו עלייה ו 0 משמעו ירידה(לאו דווקא חדה) הסתברות גבוהה ל 1 או הסתברות גבוהה ל 0 נפרש כעליה או ירידה חדה וכך נבצע את ניבוי המגמות.

3.