3.

למה נשתמש ב- feature selection?

כי לפעמים כמות המושגים שלנו יכולה להגיע למיליונים ומסווגים מסויימים לא יכולים להתמודד עם כמות כל כך גדולה של features.

אנו צריכים גם כדי להוריד זמן אימון של המודל, כדי להוריד זמן ריצה, כדי להיפטר מרעש כלומר ערכים מיותרים לא קשורים ואם למשל למידת מודל מיותרת שלא יהיה מסקנות לא נכונות.

למדנו בהרצאה שלוש סוגים של feature selection. Mutual Information, Chi-square, Frequency.

בשיטה Chi-square הוא שיטה המבוססת על נתונים סטטיסטיים, ויכולה לבחור מונחים בחות שימושים לפעולות סיווג. מחשבים על יד הסתברות של הופעת מילה מסויימת כיחד עם class מסוים.

בשיטה Mutual Information הוא שיטה המבוססת על מידע יותר מדויק ופרשנות מדוקת. אבל יכולה להשתמש במונחים לא שימושים נדירים. מקבלים תוצאות הכי נכונות אם מונח נמצא במסמך ומסמך שייך ל- class מסוים.

ושיטה הכי פשוטה הוא Frequency. משתמשים בה במונחים הכי נפוצים. אין בה בסיס מסוים אבל זה עובד כי עורכים עם נתונים הכי שימושיים ובכולם המקרים זו שיטה הכי יעילה.

אנו נשתמש ב-feature selection ב- NB מודל וריאנטי כי אחרת נסבול מרעש בנתונים והרבה סדרות.