מבוא למערכות לומדות – תרגיל 2

מגישים:....

# הגדרת הבעיה

בתרגיל זה החלתנו לבחון האם נוכל לנבא ציון שכותב ביקורת נותן למוצר על סמך ניתוח הטקסט בביקורת. למעשה ברצוננו לבצע ניתוח של סנתימנט התגובה, כלומר, האם הכותב חיובי, שלילי או נטרלי ביחס למוצר. מאחר וביקורת על מוצר תלויה באופי המוצר (לדוגמה מחשב שמתחמם מהר זה דבר רע ותנור כזה דוקא מעוד רצוי), החלתנו לבנות את הפטרון שלנו מסביב למצלמות ואביזרים למצלמות.

## איסוף הדוגמאות

לצורך איסוף הדוגמאות ביצענו זחילה ב Amzone.com והורדנו מוצרים שעונים על השאילתה "camera". לאחר מכן לקחנו כל מוצר ובעזרת הסקריפט מ <http://www.esuli.it>, הוצאנו את התגובות למוצר בתוספת הדירוג שהמשתמש נתן למוצר, הקותרת והתוחן. את הדוגמאות חילקנו לדוגמאות אימון ודוגמאות מבחן ביחש של 3:7.

# תיאור הפטרון

## הגדרת Features

מאחר ומדובר בטקסט השתמשנו בפטרון המומלץ והידוע של Bag of Words. בחרנו מילים מהכותרת ומהתוחן של הביקורת וספרנו כמה פעמים הן הופיעו. לאחר מכן הסרנו מילים שהופיעו בכותרת של אותו מוצר שעבורו נכתבה הביקורת. לבסוף הוספנו את כל צימדי המילים (Bi-Gram) שהופי בטקסט. סך הכל כיבלנו 2000 דוגמאות אימון שעליהם חושבו 12050 featurs שונים.

## בחירת מאפיינים

כאשר יש בידינו למעלה מ12000 מאפיינים, ברור כי בידינו יותר מידי מאפיינים, לכן, בהתאם לדרישות התרגיל בחנו מספר דרכים לניפוי המאפיינים המיותרים. המימדים הגדולים של הבעייה הקשו עליינו בהינתן מגבלות חומרה וזמן הגשה של התרגיל, לכן, מספר פטרונות קלאסיםת כגון חיפוש יוריסטי (חמדן וגנטי) או PCA, שרצינו לבחון נפלו בשלב זה. לבסוף החלתנו לבחון שיתות מבוססות ציון המחושב מהאנטרופיה של המאפיין, שיטת חיפוש סתוכסטית ושילוב בין דירוג תחונות להורדת מימד בעזרת PCA.

בנוסף לאלו היה ניסיון לדרגת תחונות בשיטה נוספת, בנינו עץ סיווג C4.5 (ללא שלב הגיזום) ונתנו ציון לתחונות על סמך העומק בו הן הופיעו. מאחר ובעץ ששואף להיות מאוזן בעל 2000 עלים יש סדר גודל של של 200 צמתים שיטה זו לה איפשרה לנו לבחור מספיק מאפיינים על מנת לבצע סיווג מדוייק בעזרץ המסווגים שבחרנו.

## דירוג תכונות

שיטה זו נותנת ציון לכל אחד מהמאפיינים, וכאשר בוחרים n מאפיינים מקבלים את n אלו שכיבלו ציון הכי גבוה. אנו השתמשנו בשתי שיטות לקביעת ציונים אלו:

כאשר H מסמן את האנטרופייה.

שתי השיטות דומות במהותן אך שונונות בציון הסופי שהן נותנות לכל תחונה. מדד IG מחשבת את ההפרש בין האנטרופיה לאנטרופיה המותנית (מה מוסיפה תחונה זו לאנטרופייה). לעומתה SU, מחשבת את הפרש האנטרופייה מהאנטרופיה המותנית המנורמלת ומוסיפה לה את האנטרופיה היחסית של המאפיין. לשני החישובים השתמשנו בספריית הקוד הפתוח weka.

<תוצאות +מבחן סטתיסתי>

## חיפוש סטוכסטי

זוהי שיטת חיפוש חמדנית במרחב המאפיינים. השיטה מורכבת ממספר שלבים, תחילה אנו נותנים ציון לכל המאפיינים בעזרת IG, ציון זה מכתיב את ההסתברות להיבחר, זאת ביחס הפוך לציון שקיבלה. לאחר מכן בכל שלב אנו בוחרים כ 50 תתי קבוצות בגודל 1% מכמות התחונות ההתחלתית. את התחונות שבחרנו אנו מנפים מהקבוצה שנותרה בסיום השלב הקודם ומבצעים הערכה לאיכות קבוצות התחונות החדשות. את השיערוך הזה אנו מבצעם באמצעות מסווג מהיר (Naïve Bayes). לבסוף בוחרים את הקבוצה בעלת הציון המקסימלי. אנו חוזרים על התהליך עד להגעה למספר התחונות הרצוי. המימוש בשלב זה הוא מימוש שלנו פרט לדירוג התחונות.

<תוצאות +מבחן סטתיסתי>

## שילוב IG עם PCA

כפי שהוזכל לעיל בעיות ביצועים מנעו מאיתנו לבצע PCA על כלל המאפיינים. לכן, החלתנו לבצע אלגורים זה על מספר התחונות שביכולתינו להריץ (2000). לקחנו את דירוג התחונות IG, כפי שתואר לעיל, חילקנו אותו לקבוצות של 2000 מאפיינים. את שיטת הורדת המימד PCA, הפעלנו רק על הקבוצה האחרונה (בעלת הציונים הנמוכים). כאשר לקבל מספרים שונים בתחום של הקבוצה האחרונה הגדלנו והקטנו את ערך הואריאנס. ערך הוואריאנס הוא הסכום המנורמל של הערכים העצמיים של הוקטורים המתאימים, והגדלתו מביא כמות אקספוננציאלית של נאפיינים (מתוך ה 2000). את שני החלקים מימשנו בעזר ספריית weka.

<תוצאות +מבחן סטתיסתי>

## המסווגים

...

השוואה בין שלושת המסווגים + הסבר על למה הוא טוב ומה החסרונות

# תוצאות

## ודאות הסיווג

לסרטת בעזרתו גרף Precision-Recall של מחלקה אחת נבחרת

## הרצת train test

הרצה על מידע אחר