מבוא לבינה מלאכותית – תרגול 10

נושא:

מדדי איכות בלמידה מפוקחת

- **accuracy (and weighted accuracy)**
- **confusion matrix**
- **recall (=sensitivity), precision, specificity**
- **F1 score**
- **ROC curve & AUC**

רקע:

בתרגול הזה נעסוק במדדים הנוגעים להערכת איכות מודלים לסיווג בינארי או מולטיקלאס. כמו שיצא לנו לראות במהלך הסמסטר, כשמעריכים מודל כזה, אנחנו נחלק את הדאטא לקבוצת training ו-test, נלמד על ה-training ונקבע את איכות המודל בעזרת ה-test.

יש לציין שהרבה מהמדדים שמתאימים לסיווג הבינארי יכולים להתאים בצורות שונות לקביעת האיכות של סיווג מולטיקלאס, אבל הם יספרו לנו על תמונה חלקית (למשל, כמה דייקנו בסיווג של קבוצה אחת לעומת כל השאר).
עוד משהו ששווה לציין הוא שאין מדד אחד עדיף יחסית לשאר, כל אחד מתאים לשימוש אחר.

Accuracy:

אולי המדד הבסיסי והכי אינטואיטיבי, ולכן הוא גם מאוד נפוץ.

$$Accuracy = \frac{num.\,correct\ guesses}{total\ num.\,guesses}$$

הכי פשוט שיכול להיות. ככל שהדיוק גבוה יותר, כך המודל ייחשב טוב יותר.
קל לראות שמדד זה יכול להיות מופעל גם על מודל לסיווג בינארי וגם לסיווג מולטיקלאס.

חסרון בולט של מדד זה הוא המצב של חוסר איזון. למשל, אם יש לנו דאטא שבו התיוגים (האמיתיים) לקלאס אחד הוא 20%, לקלאס שני 70% ולקלאס שלישי 10%, נוכל בקלות להשיג דיוק של 70% על הדאטא – פשוט ננחש לכל דוגמה שהיא שייכת לקלאס השני.
פתרון אפשרי לחסרון זה: מדד ממושקל –

$$Weighted\ Accuracy = \frac{\sum_i w_i \mathbb{I}\{y_i^{true} \neq y_i^{predicted}\}}{\sum_i w_i}$$

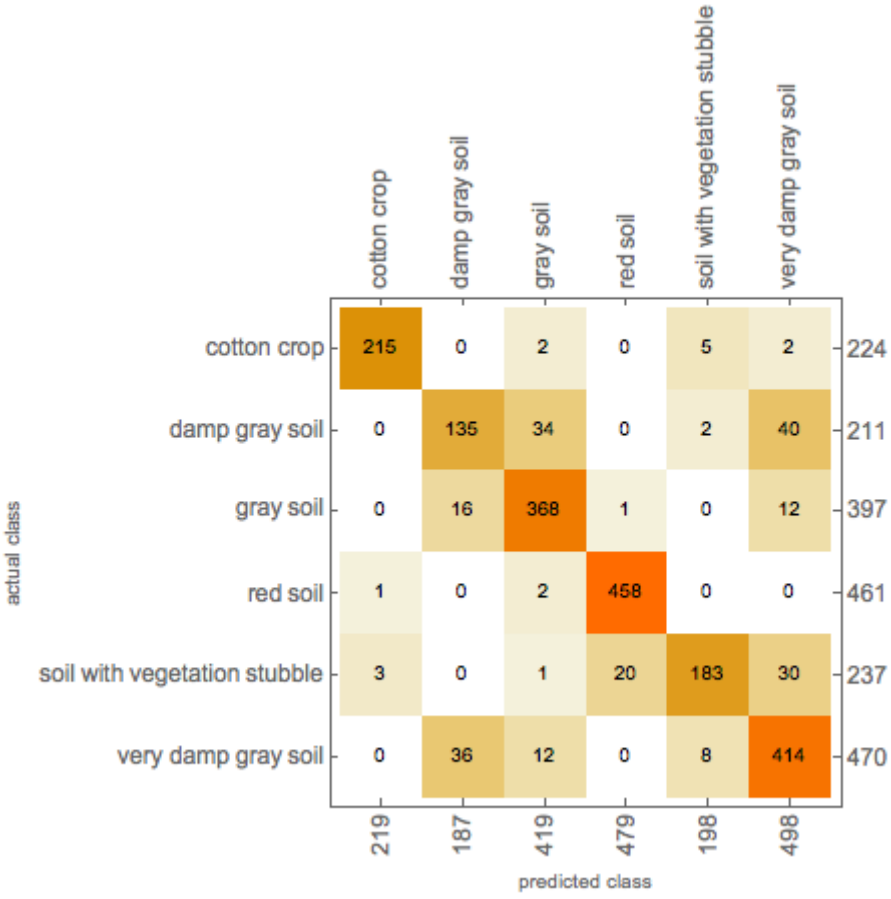
נוטים לתת משקלים גבוהים יותר לסיווג של קלאסים נדירים יותר, למשל באמצעות קביעת המשקלים להיות אחד חלקי הכמות היחסית של תיוגים מכל קלאס (למשל, אם יש לנו חצי מהדגימות כשייכות לקלאס אחד, שליש לקלאס שני ושישית לשלישי, אז המשקל לדגימות מתויגות בקלאס הראשון יהיה 2, לשני 3 ולשלישי 6).

Confusion matrix:

בהרבה מקרים, הדיוק נותן רק חלק מהמידע על איכות המודל. למשל, אם בזיהוי מחלה על דאטא של חצי חולים וחצי בריאים, דיוק של 50% יכול לומר שזיהינו נכונה שהחולים כולם חולים וטעינו בכל הזיהויים לבריאים, מצב כלשהו באמצע, או שזיהינו נכון שכל הבריאים בריאים וטעינו בסיווג לכל החולים.

לכן, מגדירים את מטריצת הבלבול (אני לא אחראי לתרגום):

$$Confusion\ Matrix\ (i,j) = num.samples\ predicted\ j\ and\ truely\ labeled\ i$$

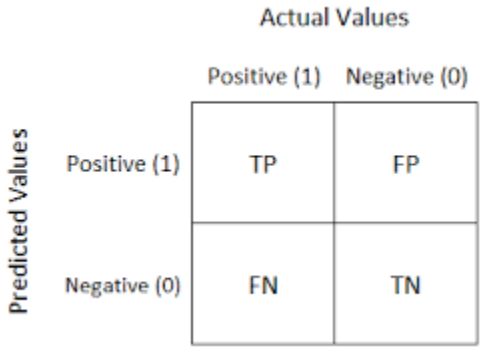


הערה: לפעמים התפקיד של השורות הוא התיוג האמיתי ולא של העמודות, כמו בתמונה, ולפעמים מסתכלים על המספר יחסית לכמות הדגימות שמתויגות (באמת) בתיוג המתאים.

המטריצה הזו נותנת את התמונה המלאה בנוגע לדיוק בסיווג (גם בינארי וגם מולטיקלאס), כי אפשר להבין ממנה כמה מדייקים בכל קלאס (על האלכסון), וכמה נוטים להתבלבל בין קלאסים מסוימים (כאן למשל: damp gray soil & very damp gray soil).

נתמקד כאן ב-confusion matrix למקרה בינארי, ממנו ייגזרו כל המדדים שנראה בהמשך:

נגדיר -



מכאן, יבוא הכל (כולל מדדים שלא נלמד).

Recall (=sensitivity), precision, specificity:

הגדרות (גם לאלה יש גרסה עברית, אבל אף אחד לא באמת משתמש בה):

$$recall = \frac{TP}{P} = \frac{TP}{TP + FN}$$

$$precision = \frac{TP}{TP + FP}$$

$$specificity = \frac{TN}{N} = \frac{TN}{TN + FP}$$

מקרי גבול מעניינים:

- סיווג כל הדוגמאות כ-positive ייתן recall מושלם (1), אבל precision נמוך ו-specificity 0.
- סיווג כל הדוגמאות כ-negative ייתן recall 0, precision לא מוגדר ו-specificity 1.
- סיווג שגוי לכל הדוגמאות ייתן אפסים בכל המדדים.
- סיווג מושלם ייתן אחדות בכל המדדים.

F1 score:

Recall ו-precision מוצגים בהרבה מקרים זה לצד זה, בעיקר במקרים בהם התיוג "positive" הוא בעל הרבה משמעות ולא ממש אכפת מ-"negative" (למשל, חולה לעומת בריא). בהמון מקרים, יש tradeoff בין ערך ה-recall לערך ה-precision, כלומר אחד מהם גבוה על חשבון שני נמוך (מתבטא במכנה, וב"בחירה" האם לטעות בתיוג כחיובי או בתיוג כשלילי).

לכן, מסתכלים על מדד F1, המוגדר כממוצע ההרמוני של ה-recall וה-precision:

$$\mathrm{F1} = \frac{1}{\frac{1}{recall} + \frac{1}{precision}} = 2 \cdot \frac{precision \cdot recall}{precision + recall}$$

הכי טוב – 1 (כאשר $\mathrm{precision} = \mathrm{recall} = 1$), הכי גרוע – 0.

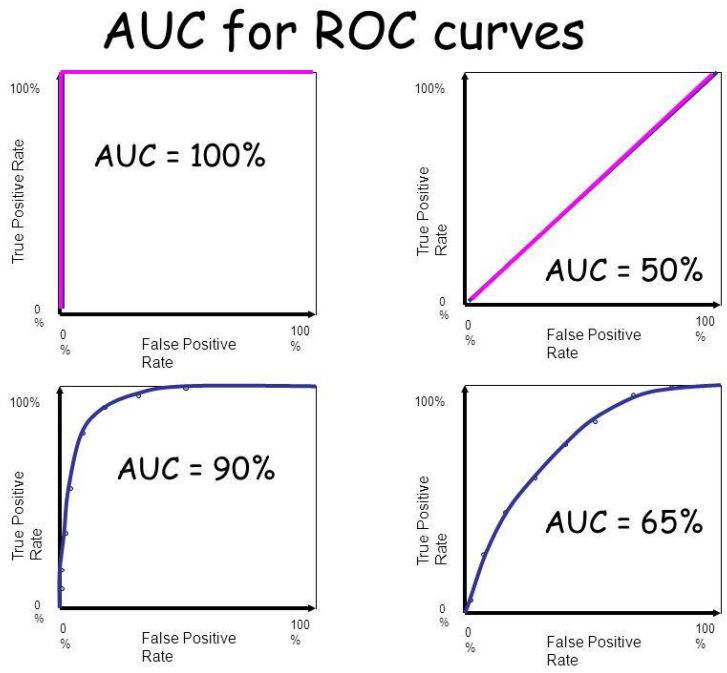
ROC (receiver operating characteristic) curve and AUC:

עקומת ROC היא עקומה שמתארת את איכות הניחוש.

זו עקומת ה-recall (שנקרא גם ה-true positive rate) כתלות ב-$1 - \text{specificity}$ (שנקרא גם ה-false positive rate). היא נבנית באופן הבא:

- בהינתן התיוגים של הדגימות וה-scores של כל דגימה (ערכים, לאו דווקא 0 או 1, שניתנים כניחוש לגבי הדגימה).
- יהי אוסף ערכי סף בין 0 ל-1 (כולל) שבו נשתמש.
- ניקח 2 מערכים באורך כמות ערכי הסף, אחד ל-true positive rate ואחד ל-false positive rate.
- לכל ערך סף:
  - נאתחל TP=0, FP=0.
  - לכל דגימה:

    אם ה-score של הדגימה גדול מערך הסף, נוסיף 1 ל-TP אם הדגימה חיובית ואחרת נוסיף 1 ל-FP.
  - נחלק את ה-TP וה-FP במספר הדגימות הכולל לקבלת false positive rate ו-true positive rate, אותם נוסיף למערך.
- לאחר קבלת הנקודות, אפשר לשרטט את העקומה.



מתוך העקומה, מוגדר מדד ROC AUC (Area Under the Curve), שהוא השטח שמתחת לעקומת ROC. המשמעות שלו – זה הסיכוי שהמודל שלנו ייתן פרדיקציה גבוהה יותר לערך חיובי אקראי מאשר לערך שלילי אקראי (כשאנחנו מניחים שהתיוג האמיתי המתאים לכל דגימה חיובית גבוה מהתיוג המתאים לכל דגימה שלילית).

ככל שהשטח הזה קרוב יותר ל-1, כך המדד נחשב טוב יותר. לעומת זאת, AUC של 0.5 הוא ערך לניחוש אקראי, ולכן כל ערך מתחת אליו נחשב גרוע במיוחד, וערך AUC מעליו אומר שהמודל מסוגל ללמוד משהו על הדאטא.