

מבוא לבינה מלאכותית – תרגול 12

נושא:

מדדי איכות ללמידה מפוקחת:

- Accuracy (and weighted accuracy) -

- Confusion matrix -

- Recall-Precision, Sensitivity-Specificity -

- ROC curve & AUC -

רקע:

בתרגול הזה נעסוק במדדים הנוגעים להערכת איכות מודלים לסיווג בינארי או מולטיקלאס. כאמור במהלך הסמסטר, במודלים כאלה, מחלקים את הדאטא ל- training ו- test, לומדים על ה- training וקובעים את איכות המודל על סמך הביצועים על ה- test. הערות לפני שנתחיל:

- כמו בתרגול שעבר, גם בתרגול הזה אדלג על נושא שלימדתי בשנה שעברה (F1 score) ומוזמנים לקרוא את [התרגול שהיה אז](#).
- 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 מחלקה אחת לעומת כל השאר).
- אין מדד אחד עדיף יחסית לשאר, כל אחד מתאים לשימוש אחר, תלוי במטרה שבה מתמקדים.
- יש עוד כל מיני מדדים שהוגדרו למשימות ספציפיות. למשל, מה תהיה ההערכה שלנו למערכת Siri שבמקום להבין שאמרנו "What is the weather in Japan", חזתה שאמרנו "One theater in Japan"? זה דורש מדד שמתאים ספציפית למשימה הזאת והרבה פחות למשימות אחרות, לכן נתמקד במדדים כלליים.

:Accuracy

אולי המדד הכי בסיסי ואינטואיטיבי, לכן הוא גם נפוץ מאוד. משמש לסיווג בינארי ולמולטיקלאס.

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

הכי פשוט שיכול להיות. ככל שהדיוק גבוה יותר, כך המודל ייחשב טוב יותר.

חסרון בולט של מדד זה הוא המצב של חוסר איזון. למשל, אם יש לנו דאטא שבו 20% מהתיוגים (האמיתיים) שייכים למחלקה הראשונה, 70% לשנייה ו-10% לשלישית, 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 מחלקות נדירות יותר, למשל באמצעות קביעת המשקלים להיות אחד חלקי הכמות היחסית של תיוגים מכל מחלקה (למשל, אם התפלגות התיוגים היא $(\frac{1}{2}, \frac{1}{3}, \frac{1}{6})$ [חצי מהדגימות ממחלקה 1 וכו'], אז המשקולות לפי מחלקה יהיו $(2, 3, 6)$).

:Confusion matrix

בהרבה מקרים, הדיוק נותן רק חלק מהמידע על איכות המודל. למשל, באוסף תמונות שחצי מהן של כלבים וחצי של חתולים, דיוק של 50% יכול לומר שזיהינו נכונה את כל החתולים וטעינו על כל הכלבים, מצב כלשהו באמצע, או שזיהינו נכון את כל הכלבים וטעינו בסיווג לכל החתולים. לכן, מגדירים את מטריצת הבלבול (אני לא אחראי לתרגום), שהאיבר ה- (i, j) שלה הוא כמות הדגימות מהמחלקה ה- i שחזינו שהם שייכים למחלקה ה- j .

הערה: לפעמים התפקיד של השורות הוא התיוג האמיתי ולא של העמודות, כמו בתמונה מטה, ולפעמים מסתכלים לא על מספר התיוגים אלא מנרמלים אותו בכמות הדגימות האמיתיות מהתיוג המתאים.

דוגמה:

		cotton crop	damp gray soil	gray soil	red soil	soil with vegetation stubble	very damp gray soil	
actual class	cotton crop	215	0	2	0	5	2	224
	damp gray soil	0	135	34	0	2	40	211
	gray soil	0	16	368	1	0	12	397
	red soil	1	0	2	458	0	0	461
	soil with vegetation stubble	3	0	1	20	183	30	237
	very damp gray soil	0	36	12	0	8	414	470
		219	187	419	479	198	498	
		predicted class						

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

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

נגדיר -

		Actual Values	
		Positive (1)	Negative (0)
Predicted Values	Positive (1)	TP	FP
	Negative (0)	FN	TN

:Recall (= sensitivity), precision, specificity

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

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

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

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

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.

המסקנה מכל המקרים האלה היא שכדאי להשתמש ביותר מממד אחד מבין אלה. בדרך כלל מסתכלים יחד על recall-precision, או על sensitivity-specificity.

הערה: ישנה עקומה של precision-recall שדומה בכמה רעיונות בה לעקומת ROC שנראה מיד, גם בה משתמשים לעתים אבל לא נעבור עליה.

:ROC (receiver operating characteristic) curve and AUC

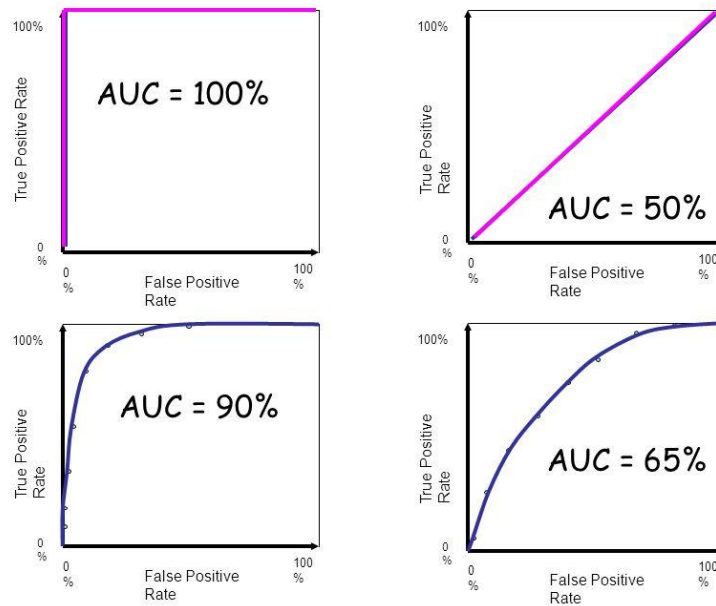
עקומת ROC היא עקומה שמתארת כמה דירוג הנקודות לפי ה-scores של המודל מתאים לתיוג הנקודות, כלומר כמה המודל מצליח לתת לנקודות המתויגות כחיוביות scores גבוהים יותר מאשר אלה שהמודל נותן לשליליות.

בשונה משאר המדדים שראינו היום, כאן מתייחסים רק לסדר של הנקודות, ללא תלות בערך סף שמעליו הדוגמות ייחשבו חיוביות ומתחתיו שליליות. כאן מחשבים לכל טווח ערכי הסף האפשריים את הביצועים.

עקומת ROC היא עקומת ה-recall (שנקרא גם ה-true positive rate – מתוך כל הדוגמות החיוביות, כמה תייגנו כחיוביים?) כתלות ב-specificity (1 – specificity) (שנקרא גם ה-false positive rate – מתוך כל מי שמתויגים כשליליים, כמה סיווגנו כחיוביים?), שנבנית באופן הבא:

- **קלט:** התיוגים של הדגימות וה-scores של כל דגימה מתוך המודל (ערכים, לאו דווקא 0 או 1, שניתנים כניחוש לגבי הדגימה).
 - יהי אוסף ערכי סף בין 0 ל-1 (כולל) שבו נשתמש.
 - נאתחל 2 מערכים באורך כמות ערכי הסף, אחד ל-true positive rate ואחד ל-false positive rate.
 - לכל ערך סף:
- נאסוף את הדגימות עם scores מעל ערך הסף. מספר הדגימות המתויגות חיוביות חלקי מספר החיוביים הכולל יהיה ה-true positive rate לערך הסף המתאים, ומספר המתויגות שליליות חלקי מספר השליליים הכולל יהיה ה-false positive rate לערך הסף המתאים.
- **פלט:** ערכי הסף ושני המערכים. מתוך שני המערכים אפשר לשרטט את העקומה.

AUC for ROC curves



שימו לב:

- הנקודה (0, 0) מתאימה לערך הסף 1, הרי מעליו אין דוגמות שחזינו שהן חיוביות. בהתאם, נקודות שקרובות לראשית הצירים מתאימות לערכי סף גבוהים (מעט מאוד חיזויים חיוביים, לכן ערכי נקודות נמוכים לשני הצירים).
- אם העקומה מתחילה לעלות בשיפוע גבוה, זה אומר שיש טווח ערכי סף גבוהים יחסית שמעליהם נתפסים הרבה מהערכים החיוביים לעומת מעט מהשליליים. לכן, אם נתמקד בלתפוס את הערכים החיוביים, נוכל להגדיר סף יחסית גבוה לתוצאות המודל שנקודות עם scores גבוהים ממנו יהיו בוודאות גבוהה חיוביות.
- הנקודה (1, 1) מתאימה לערך הסף 0, הרי לפיו תופסים כחיוביים גם את כל מה שהוא חיובי באמת וגם את כל מה ששלילי באמת. בהתאם נקודות שקרובות לפינה השמאלית העליונה מתאימות לערכי סף נמוכים (הרבה מאוד תחזיות חיוביות, לכן ערכי נקודות גבוהים לשני הצירים).
- אם העקומה מגיעה ל- (1, 1) בשיפוע נמוך, זה אומר שיש טווח ערכי סף נמוכים שמתחתיהם הרבה מהדגימות אכן שליליות ומעט מהן חיוביות (ב"התקדמות" לכיוון הפינה, ערכי הסף יורדים וזה מכניס הרבה false positives לחיזויים החיוביים, ומשאיר את החיזויים השליליים האמיתיים להיות מתחת לסף). לכן, אם נתמקד בלזהות כמה שיותר דגימות שליליות, נוכל להגדיר סף יחסית נמוך שנקודות עם scores נמוכים ממנו יהיו בוודאות גבוהה שליליות.

מתוך העקומה, מוגדר מדד ROC AUC (Area Under the Curve) – 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. ראו ציורים מטה המתארים את התפלגויות ה-scores עבור דגימות חיוביות לעומת שליליות ואת צורות עקומות ROC וערכי ה-AUC המתאימים (התעלמו מערך ה-threshold, יש לו משמעות רק בכך שמה שמעליו מקבל חיזוי חיובי ולהיפך):

