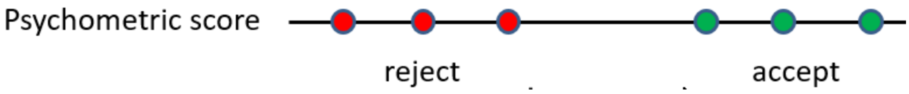
למידת מכונה הרצאה 2

למידת PAC – PAC learning. Probably Approximately Correct. "כנראה בערך נכון". בדוגמה של המטבע: בהסתברות $1-\delta$ ("כנראה"), ההטיה של המטבע היא $X \pm \epsilon$ ("בערך").

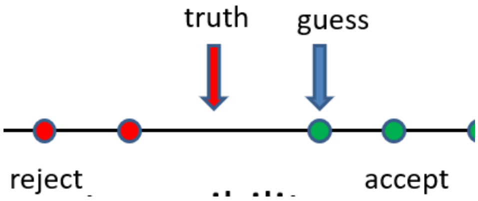
דוגמה ללמידת PAC: נתבונן ב- n תלמידים (שנדגמו באופן מקרי ואחיד) שהתקבלו או לא התקבלו לתוכנית מסוימת. ידוע לנו ציון הפסיכומטרי של כל תלמיד, וההנחה היא שזה הנתון היחיד שקובע קבלה. בהינתן תלמיד חדש וציון, האם נוכל לנחש האם הוא יתקבל או לא? הנתונים נראים כך:



החתך האמיתי נמצא איפשהו באמצע בין האדום הימני ביותר לירוק השמאלי ביותר. איפה כדאי לנו לנחש?

אם ננחש בשמאל, אין false negative. אם ננחש מימין, אין false positive. אולי באמצע?

לפני שניגש לניתוח, נזכיר את אי – השוויון: $1-x \le e^{-x}$ (או $1+x \le e^{x}$) לכל x ממשי.

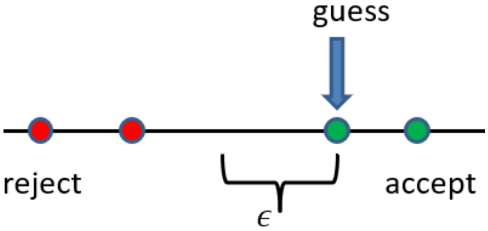


נבחר את הימין קיצון. אין false positive. אם אמרתי לתלמיד שהוא יתקבל, הוא בטוח יתקבל. מה ההסתברות ל false negative? אנחנו רוצים שבהסתברות $1-\delta$, נטעה בלכל היותר ϵ אחוז מהם. כלומר שיש רק ϵ אחוז שנמצאים באיזור הזה, הציונים הכי נמוכים שבפועל כן אמורים להתקבל. כל מה שנמצא בין **האמת** **לניחוש**, אותם אנחנו מפספסים.

נשאל את השאלה קצת אחרת. נדמיין שיש מספר תלמידים בסביבת ϵ ימנית של **האמת**. מה ההסתברות שהמדגם פספס את התלמידים בסביבה הזאת? וזאת שאלה שקל לענות עליה. זו ההסתברות שדגמנו תלמידים וכל פעם פספסנו את התלמידים בטווח הזה.

ההסתברות שדגימה מסויימת תפספס את התחום הזה היא $1-\epsilon$, וההסתברות שנפספס אותו כל פעם במשך n פעמים היא $(1-\epsilon)^n$. לפי אי השוויון זה פחות מ $e^{-\epsilon n}$. אנחנו רוצים שהשגיאה תהיה פחות מדלתא, כלומר נחשב:

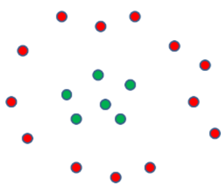
$$e^{-\epsilon n} \le \delta \Longrightarrow n \ge \frac{\ln \frac{1}{\delta}}{\epsilon}$$



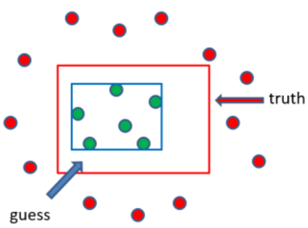
נפתח סוגריים: מה אם נשאל, מה ההסתברות שהניחוש שלנו פספס את ה-ϵ דגימות האלה? כלומר, מה ההסתברות שהמדגם פספס סביבת אפסילון שמאלית של **הניחוש**?

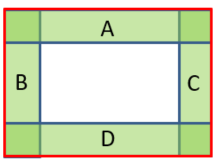
לכאורה אותו דבר. אבל הניחוש מבוצע רק אחרי הדגימה. **הניחוש הוא פונקצייה של המדגם**, אז אי אפשר לשאול איך הניחוש השפיע על המדגם.

עכשיו, דוגמה בשני מימדים:



נגיד גובה ומשקל של תינוקות. נניח שיש איזשהו מלבן שקובע מה תקין, וכל מה שבחוץ לא תקין. וצריך לשלוח את התינוק לעוד בדיקות. אנחנו רוצים לנחש מה המלבן. ניקח את המלבן הכי הדוק, כי ככה נדע בוודאות שכל מה שיש בתוך המלבן תקין. כי עדיף false positive מאשר false negative, במקרה הזה.





ושוב נשאל, כמה false positive יש לנו? נעשה את אותו הניתוח מהדוגמה הקודמת. נניח שאנחנו יודעים מה המלבן האמיתי, ונשאל מה ההסתברות שהמדגם פספס תחום מסויים מסביב למלבן האמיתי הזה. נניח שליד כל צלע (בטווח שמתפספס), יש $\epsilon/4$ אחוז מהתינוקות. ככה שבאיזורים A,B,C,D ביחד, יש לכל היותר אפסילון תינוקות. אם תהיה אפילו דגימה אחת בכל איזור, המלבן ניחוש שלנו יחרוג לתוך האיזור הירוק – כלומר יהיה קרוב יותר לאמת.

ההסתברות שהמדגם פספס איזור מסויים:

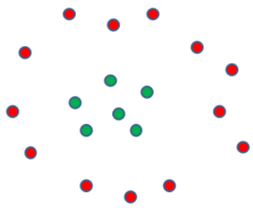
$$P(missed\ A) = \left(1-\frac{\epsilon}{4}\right)^n \le e^{-\frac{\epsilon n}{4}}$$

אז ההסתברות שהמדגם פספס אפילו אחד מהאיזורים: לפי חסם איחוד:

$$P(missed\ even\ one\ of\ A,B,C,D) \le 4\cdot P(missed\ A) \le 4\cdot e^{-\frac{\epsilon n}{4}}$$

ואנחנו רוצים ש: $e^{-\frac{\epsilon n}{4}} < \delta$, אז: $n \ge (4\ln(4/\delta))/\epsilon$.

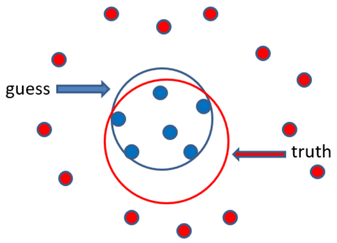
אז בהסתברות לפחות $1-\delta$, לא פספסנו אף איזור, ואנחנו טועים בלכל היותר אפסילון אחוז מהתינוקות.



אבל מה אם התחום הוא מעגל?

אז יכול להיות שנפספס את האמת, גם אם ניקח ניחוש מאד הדוק:

האנליזה שלנו נכשלת במקרה הזה. אנחנו צריכים לפתח משהו יותר כללי.



חסם הכללה – *generalization bound*:

במקרה הקודם, הנחנו שהכללים הם מלבן (או מעגל) – כלומר יש קבוצה אינסופית (ולא בת מניה) של כללים אפשריים, וצריך לבחור אחד מהם.

אם יש לנו קבוצה סופית של אובייקטים (חוקים), נוכל להוכיח:

תהי X קבוצה של נקודות, ותהי D התפלגות על X, ו-H קבוצה סופית של חוקים שממפה: $X\to\{0,1\}$. כלומר החוק מתייג את הדאטא.

תהי S קבוצה אקראית בגודל n מתוך X, לפי D (ההתפלגות).

נניח שחוק כלשהו $h\in H$ הוא **עקבי** על המדגם. כלומר הוא לא טועה לפי המדגם.

אז בהסתברות לפחות $1-\delta$, **הטעות האמיתית** של h על כל X היא:

$$e(h) \le \frac{1}{n}\left(\ln(|H|) + \ln\left(\frac{1}{\delta}\right)\right)$$

הטעות האמיתית היא ההסתברות שהחוק יטעה לגבי נקודה חדשה: $P\big(h(x) \ne labal(x)\big)$

נשים לב שיש תלות בכמות החוקים. ניגש להוכחה:

לכל חוק $h_i \in H$, אם הטעות האמיתית שלו על X גדולה מאפסילון, אז ההסתברות שהוא עקבי לכל S קטנה מ: $(1-\epsilon)^n$.

נשאל: מה ההסתברות שקיים חוק שהטעות האמיתית שלו גדולה מאפסילון, ושעדיין יצא עקבי על כל המדגם?

$$\begin{aligned}
&P(\exists h\in H : e(h) > \epsilon \cap h\ is\ consistent\ with\ S)\\
&\qquad = P\big((h_1\ is\ consistent\ with\ S\mid e(h_1)>\epsilon) \cup (h_2\ is\ consistent\ with\ S\mid e(h_2)>\epsilon) \cup \ldots\big)\\
&\qquad \le \sum_i P(h_i\ is\ consistent\ with\ S\mid e(h_i)>\epsilon) \le |H|\cdot(1-\epsilon)^n \le |H|\cdot e^{-\epsilon n}
\end{aligned}$$

ניקח: $|H|\cdot e^{-\epsilon n}\le\delta$, אז: $\epsilon \le \frac{1}{n}\left(\ln(|H|)+\ln\left(\frac{1}{\delta}\right)\right)$.

אז ההסתברות שמצאנו חוק עם טעות אמיתית שהיא לפחות $\frac{1}{n}\left(\ln(|H|)+\ln\left(\frac{1}{\delta}\right)\right)$, אבל עדיין עקבי על S, היא לכל היותר דלתא. אז אם מצאנו חוק עקבי, בהסתברות לפחות $1-\delta$ הטעות האמיתית שלו קטנה מזה.

עוד הגבלה שיש לנו במשפט הזה: אנחנו מניחים שהחוק עקבי לגמרי על כל המדגם. אם יש טעות מאוד קטנה, אולי נרצה להשתמש בו? נניח שיש חוק עם **טעות אמפירית** $\bar{e}(h)$ על המדגם. (אחוז הנקודות מהמדגם שהחוק טעה לגביהן). עדיין נוכל להוכיח:

תהי X קבוצה של נקודות, ותהי D התפלגות על X, ו-H קבוצה סופית של חוקים שממפה: $X \to \{0,1\}$. כלומר החוק מתייג את הדאטא.

תהי S קבוצה אקראית מתוך X, לפי D (ההתפלגות).

נניח שיש חוק $h \in H$ בעל טעות אמפירית $\bar{e}(h)$. אז בהסתברות לפחות $1-\delta$, הטעות האמיתית של h על כל X היא:

$$e(h) \leq \bar{e}(h) + \sqrt{\frac{\ln(2 \cdot |H|) + \ln(1/\delta)}{2n}}$$

נשים לב שהנוסחה דומה למה שראינו קודם, עד כדי קבועים. ההבדל המשמעותי הוא השורש. בגלל שאנחנו מתעסקים עם מספרים שקטנים מ1, השורש מגדיל את המספר, כלומר מחליש את החסם.

ההוכחה למשפט נובעת מאי שוויון הופדינג:

תהי B התפלגות ברנולי על $\{0,1\}$, עם פרמטר p. (ההסתברות שיצא 1). נבצע n ניסויים בת"ל, ויהי $X = \frac{1}{n}\sum t_i$ הממוצע האמפירי. אזי: $P(|X-p| > \epsilon) < 2e^{-2n\epsilon^2}$

אז ניקח חוק מסויים $h \in H$, ולכל נקודה $x \in S$ יהי $e(x) = 1_{h(x) \neq label(x)}$ (האינדיקטור למאורע שהחוק טועה בנקודה הזו). אזי $\bar{e}(h) = \sum_{x \in S} e(x)$.

אז עכשיו הנקודות עם התיוג טעות שלהן הן כולן משתנים מקריים שמתפלגים ברנולי $\{0,1\}$. אז לפי הופדינג, לכל חוק:

$$P(|\bar{e}(h) - e(h)| \geq \epsilon) \leq 2e^{-2n\epsilon^2}$$

אז בסך הכל:

$$\begin{aligned} P(\exists h \in H : |\bar{e}(h) - e(h)| \geq \epsilon) &= P\Big((|\bar{e}(h_1) - e(h_1)| \geq \epsilon) \cup (|\bar{e}(h_2) - e(h_2)| \geq \epsilon) \cup \ldots\Big) \\ &\leq \sum_i P(|\bar{e}(h_i) - e(h_i)| \geq \epsilon) \leq |H| \cdot 2e^{-2n\epsilon^2} \end{aligned}$$

נקבע $|H| \cdot 2e^{-2n\epsilon^2} \leq \delta$, ונקבל $\epsilon \geq \sqrt{\frac{\ln(2 \cdot |H|) + \ln(1/\delta)}{2n}}$. כלומר לכל חוק, בהסתברות לפחות $1-\delta$, ההפרש בין הטעות האמיתית לטעות האמפירית קטנה מהשורש הזה.

בהמשך, נגדיר מהו חוק פשוט, ונרחיב את ההגדרה לאוסף חוקים **אינסופי** (בהנחה שהם פשוטים).