מקליד: עידן אריה
מרצה: ד"ר אורן גליקמן
תאריך: 13-04-2026

למידת מכונה
89-2511-02

Parameter Estimation

Bayes Decision Theory נותן את ההחלטה הכי נכונה[1], אבל היא דורשת ידע מקדים על ההסתברויות:

- $P(\omega_i)$ - priors
- $P(x|\omega_i)$ - class-conditional densities

זהו ידע שבעולם האמיתי לא תמיד יש לנו. מה שיש לנו זה:

- אינטואיציה על הpriors - למשל אם זורקים קוביה, האיטואיציה היא $\forall_{i=1}^{6} P(w_i) = \frac{1}{6}$
  - אבל אלו לא ה$P(\omega_i)$ האמיתיים.
- סט של דוגמאות - ידע אמפירי.

בפרט - אין לנו $P(x|\omega_i)$
אם המרחב סופי, אפשר פשוט לספור כמה מהדוגמאות מתאימות לכל אפשרות. אבל אם הוא לא סופי, או יותר מדי גדול, אז אי אפשר לשערך את $P(x|\omega_i)$ באמצעות ספירה, וחייבים להניח parametric form ולשערך את הפרמטרים.

למשל:
- אם יש אוסף מספרים, אפשר להניח התפלגות גאוסיאנית רב מימדית $P(x|\omega_i) \sim \mathcal{N}(\mu_i, \Sigma_i)$, ונרצה ללמוד את μ_i ואת Σ_i מתוך הdata.

יש כל מיני שיטות ללמוד את הפרמטרים.

Maximum Likelihood Estimator (MLE)

בהנחת n תצפיות $\mathcal{D} = S_n = \{x_1, \ldots, x_n\}$, רוצים למצוא פרמטר θ שמסביר אותו בצורה הכי טובה.

הרעיון: לבחור θ שממקסם את ההסתברות לראות את $\mathcal{D}$:

$$\hat{\theta}_{ML} = \arg\max_{\theta} P(S_n|\theta) = \arg\max_{\theta} \sum_{i=1}^{n} \log P(x_i|\theta)$$

עבור התפלגות בינומית - למשל, הטלת מטבע $P(H) = p, P(T) = 1 - p$, אם ראינו k פעמים עץ, הp שממקסם את הlikelihood לראות את זה הוא $\hat{p} = \frac{k}{n}$.

[1] כלומר - שממזערת loss functuin

MAP Estimation

הרעיון:
- מתחילים עם איזשהו prior $\theta : p(\theta)$
- מעדכנים ע"י חוק בייס:

$$p(\theta|\mathcal{D}) = \frac{p(\mathcal{D}|\theta)\, p(\theta)}{p(\mathcal{D})}$$

$$\hat{\theta}_{MAP} = \arg\max_{\theta} P(\theta|\mathcal{D}) \propto \arg\max_{\theta} P(\mathcal{D}|\theta)\, P(\theta)$$

הסימן $\propto$ אומר "שווה עד כדי קבוע"

הטלת מטבע

עבור הטלת מטבע, נניח $\theta \sim Beta(\alpha, \beta)$[2] ונקבל

$$\hat{\theta}_{MAP} = \frac{k + \alpha - 1}{n + \alpha + \beta - 2}$$

מתמטית, אפשר להסתכל על α ועל β בתור:

- $\alpha - 1$ פעמים לפני ראיתי עץ
- $\beta - 1$ פעמים לפני ראיתי פלי

עבור n מספיק גדול, ההשפעה של α וβ זניחה. כלומר - הprior לא פוגע בטווח הרחוק.

התפלגות אקספוננציאלית

$\mathcal{D} = S_n = \{x_1, \ldots, x_n\}$, נדגמים מתוך התפלגות אקפוננציאלית עם פרמטר $\theta > 0$:

$$P(x_i|\theta) = \theta e^{-\theta x_i}$$

הMAP יהיה:

$$\hat{\theta}_{MAP} = \frac{n}{\sum x_i + 1}$$

הרחבה של MAP - Loss-Based View

בהינתן:
- התפלגות ידועה $P(x|\theta)$, prior $P(\theta)$, וסט תצפיות $\mathcal{D}$
- פונקציית עלות $\lambda\left(\hat{\theta}, \theta^*\right)$, בהינתן $\hat{\theta}$ צפוי וθ^* אמיתי.

המטרה: למצוא estimator $\hat{\theta}$ שממזער את תוחלת העלות

$$\hat{\theta} = \arg\min_{\hat{\theta}} \mathbb{E}_{\theta|\mathcal{D}}\left[\lambda\left(\theta, \hat{\theta}\right)\right]$$

[2] התפלגות בטא: $P(\theta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}$

עבור square loss

פונקציית השגיאה $\lambda(\theta,\theta^*)=(\theta-\theta^*)^2$. הestimator הכי טוב הוא:

$$\hat{\theta}_{SE}=\arg\min_{\hat{\theta}}\int\left(\theta-\hat{\theta}\right)^2 P(\theta|\mathcal{D})\ \mathrm{d}\theta$$

$$\vdots$$

$$\hat{\theta}_{SE}=\mathbb{E}_{\theta|\mathcal{D}}[\theta]$$

עבור 0-1 loss

$$\lambda\left(\theta^*,\hat{\theta}\right)=\begin{cases}1 & \theta^*\neq\hat{\theta}\\ 0 & \theta^*=\hat{\theta}\end{cases}$$

זה עובד רק במרחב בדיד.

$$\hat{\theta}_{0-1}=\arg\min_{\hat{\theta}}\mathbb{E}\left[\lambda\left(\theta,\hat{\theta}\right)\middle|\mathcal{D}\right]$$

$$\vdots$$

$$\hat{\theta}_{0-1}=\arg\max_{\hat{\theta}}P\left(\hat{\theta}\middle|\mathcal{D}\right)=\hat{\theta}_{MAP}$$

בגדול

לפי פונקציית הloss, נוכל לקבל posterior שונה:

- עבור loss ריבועי - נקבל ממוצע (mean)
- עבור loss ערך מוחלט - נקבל חציון (median)
- עבור loss שהוא 0 או 1 - נקבל את השכיח (mode)

כשעוברים מBayes Decision Theory לLearing from Data

בלמידת מכונה אין לנו ידע על ההתפלגויות, כמו שבייס דורש. יש לנו רק אוסף סופי של labeled data.
יש שתי שיטות:

- Generative: ללמוד $P(X|Y)$ ו$P(Y)$ ואז להפעיל בייס.
- Discriminative: ללמוד חוקיות באופן ישיר באמצעות מזעור הסיכון האמפירי.

מקליד: עידן אריה
מרצה: ד"ר אורן גליקמן
תאריך: 13-04-2026

למידת מכונה
89-2511-02

PAC Learning

המטרה של PAC זה להסיק את הsample complexity - כלומר לדעת כמה data צריך בשביל ללמוד משהו.

יש איזשהו underlying distribution שמייצר את הdata שלנו: $\mathcal{D} = P(x, y)$. יש שתי דרכים לחשוב על זה:

1. יש משהו שמייצר ביחד גם את הדוגמאות x וגם את התגיות y.
2. יש משהו שמייצר את הדוגמאות x, ומשהו אחר שמתייג אותן לפי $P(y|x)$.

נשים לב: האם אפשר להניח שהתגית היא דטרמיניסטית? לא תמיד:

- יכול להיות רעש
- לפעמים יכול להיות שלאותו x יהיו תגיות שונות בהתפלגות כזו או אחרת (למשל - בהינתן סטודנט, האם הוא יעבור את הקורס? זה לא דטרמיניסטי - אבל ההתפלגות כן תלויה בסטודנט)

המטרה: למצוא מודל h שחוזה $\hat{y}$ עבור x נתון - $h : x \to \hat{y}$ - שמסכים עם הy האמיתית, כלומר ממזער את הexpected loss $\mathbb{E}_{P(x,y)}[l(y, h(x))]$.

מניחים n התפלגויות I.I.D (בלתי תלויות ומתפלגות בצורה זהה) $S = \{(x_i, y_i) | i = 1, \dots n\}$ נדגמות מתוך התפלגות $\mathcal{D} = P(x, y)$.

מחלקת ההיפוטזות $\mathcal{H}$ יכולה להיות כל מיני דברים - למשל, אם $x \in \mathbb{R}^2$, אז $\mathcal{H}$ יכול להיות "כל המלבנים המקבילים לצירים" או "כל הקווים (שחותכים את המרחב ל2)" או אפילו "כל הפונקציות האפשריות $h : (x_1, x_2) \to \{0, 1\}$". אנחנו רוצים ש$\mathcal{H}$ יהיה אקספרסיבי - אבל לא יותר מדי, כי אז נוכל פשוט "לשנן" את סט הדוגמאות במקום ללמוד משהו שיעזור לדוגמאות חדשות.

השגיאה - true error לעומת empirical error

הtrue error ($Err_{\mathcal{D}}$) מגדיר zero-one loss:

$$l_{0-1}(h(x), y) = \begin{cases} 0 & h(x) = y \\ 1 & h(x) \neq y \end{cases}$$

הempirical error (Err_S) הוא מול סט הדוגמאות:

$$Err_S(h) = \frac{1}{n} \sum_{i=1}^{n} l_{0-1}(h(x_i), y_i)$$

הנחת הrealizability

מניחים שהקשר האמיתי בין הנתונים x והתגיות y ניתן לייצוג מושלם ע"י המודל שבחרנו. כלומר: הנתונים נוצרו ע"י היפותזה שקיימת בתוך $\mathcal{H}$.

זה אומר שקיימת לפחות היפותזה "מושלמת" אחת $h^* \in \mathcal{H}$ המקיימת $Err_{\mathcal{D}}(h^*) = 0$.

מקליד: עידן אריה
מרצה: ד"ר אורן גליקמן
תאריך: 13-04-2026

למידת מכונה
89-2511-02

Empirical Rist Minimization (ERM)

בשיטת ERM בוחרים h שממזער את השגיאה האמפירית:

$$ERM\left(S\right)=\arg\min_{h\in\mathcal{H}}Err_{S}\left(h\right)$$

במקרה הrealizable, תמיד $\min_{h}Err_{S}\left(h\right)=0$.

הקשר בין Err_S ו$Err_{\mathcal{D}}$

תכונה סטטיסטית: $Err_S\left(h\right)$ הוא unbiased estimator של $Err_{\mathcal{D}}\left(h\right)$ - כלומר באינסוף, אם נדגום מספיק נקודות, Err_S יתכנס בממוצע לErr_D:

$$\mathbb{E}\left[Err_{S}\left(h\right)\right]=Err_{\mathcal{D}}\left(h\right)$$

ככל שיש יותר דוגמאות $Err_S\left(h\right)$ יהיה יותר קרוב ל$Err_{\mathcal{D}}\left(h\right)$:

$$\mathbb{E}\left[Err_{S}\left(h\right)\right]=\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}l_{0-1}\left(h\left(x_{i}\right),y_{i}\right)\right]=\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[l_{0-1}\left(h\left(x_{i}\right),y_{i}\right)\right]=Err_{\mathcal{D}}\left(h\right)$$

Version Space והיפותזות רעות

לכל $\mathcal{H}$ יש Version Space - סט של היפוטזות עבורת הטעות האמפירית היא אפס.

בעיה: זה ש$Err_S\left(h\right)=0$ לא בהכרח אומר ש$Err_{\mathcal{D}}\left(h\right)=0$

נגדיר "היפותזה רעה" כהיפותזה עבורה $Err_{\mathcal{D}}\left(h\right)>\epsilon$. תמיד יש סיכוי δ שסט דוגמאות S יגרום לERM להגיע להיפותזה רעה. לכן - לכל "הבטחה" יש רמת ביטחון $1-\delta$.

מקליד: עידן אריה
מרצה: ד"ר אורן גליקמן
תאריך: 2026-04-13

למידת מכונה
89-2511-02

PAC (Probably Approximately Correct) Learning

הגדרה: מרחב היפותזות $\mathcal{H}$ הוא PAC-learnable אם קיים אלגוריתם A שמוצא היפותזה $h = A(S_n)$ לכל מדגם S_n, כך שלכל $\epsilon, \delta > 0$ קיים $N_{\epsilon,\delta}$ (פולינומי ב$\frac{1}{\epsilon}, \frac{1}{\delta}$) כך שלכל $n > N_{\epsilon,\delta}$ ולכל התפלגות $\mathcal{D}$ מתקיים:

$$Pr\left[Err_{\mathcal{D}}\left(A\left(S_n\right)\right) < \epsilon\right] > 1 - \delta$$

כלומר: δ מייצג את רמת הבטחון (probably)
ϵ מייצג את הדיוק (approximately correct)

- עם מספיק נתונים, סביר להניח שנוכל למצוא היפותזה נכונה בקירוב.

נשים לב: זוהי תכונה של מרחב ההיפותזות $\mathcal{H}$, לא של המציאות $\mathcal{D}$.

משפט: בהינתן realizability, לכל $\epsilon, \delta > 0$, אם $n > \frac{1}{\epsilon}\ln\left(\frac{|\mathcal{H}|}{\delta}\right)$ אזי לכל התפלגות $\mathcal{D}$ ומדגם S_n מתקיים:

$$Pr\left[Err_{\mathcal{D}}\left(ERM\left(S_n\right)\right) < \epsilon\right] > 1 - \delta$$

משמעויות:

- כל מרחב $\mathcal{H}$ סופי הוא PAC-learnable באמצעות ERM.
- ϵ קטן ככל שיש יותר נתונים.
- ϵ קטן ככל שדורשים רמת ביטחון גבוהה יותר (δ קטנה יותר)
- ϵ גדל ככל ש$\mathcal{H}$ יותר גדול ומורכב.

בהינתן היפותזה רעה, ההסתברות שהיא תרמה אותנו היא:

$$P\left(Err_{S_n}\left(h\right) = 0\right) \leq \left(1 - \epsilon\right)^n$$

כי היא צריכה לתת שגיאה אפס על כל מקרה במדגם S.
כלומר - ככל שn יותר גדול, הסיכוי שנרומה בגלל מדגם רע קטן אקפוננציאלית עם n.

המקרה האגנוסטי

אגנוסטי זה אומר לא realizable. בעולם האמיתי לא מובטח לנו שיש היפותזה שנותנת אפס loss.

נגדיר: את h^* בתור ההיפותזה שממזערת את השגיאה האמיתית:

$$h^* = \arg\min_{h \in \mathcal{H}} Err_{\mathcal{D}}\left(h\right)$$

משפט: לכל $\epsilon, \delta > 0$, $n > \frac{1}{\epsilon^2}\ln\left(\frac{2|\mathcal{H}|}{\delta}\right)$, והתפלגות $\mathcal{D}$:

$$Pr\left[Err_{\mathcal{D}}\left(ERM\left(S_n\right)\right) - Err_{\mathcal{D}}\left(h^*\right) < \epsilon\right] > 1 - \delta$$

תובנות

מעבר מrealizable לאגנוסטי אומר ש:

- אנחנו לא יכולים להבטיח את השגיאה האמיתית
- החסם שלנו משתנה לרעה

חסימת הTrue Error

משפט ברמת ביטחון $1-\delta$:

$$Err_D\left(ERM\left(S_n\right)\right) < \min_{h\in\mathcal{H}} Err_S\left(h\right) + \sqrt{\frac{\ln\left(\frac{2|\mathcal{H}|}{\delta}\right)}{2n}}$$

כלומר הtrue error חסום ע"י סכום של:

1. הטייה: השגיאה עבור המדגם שקיבלנו
2. הvariance של מרחב ההיפותזות

יש כאן tradeoff. מצד אחד רוצים מרחב היפותזות גדול כדי להקטין את הbias. מצד שני - ככל ש$\mathcal{H}$ יותר גדול, מרכיב הvariance יפגע בנו (כי יהיה overfitting)
בעצם נרצה את מרחב ההיפותזות הכי קטן ופשוט שמסביר הכי הרבה data.

המקרה האינסופי

את כל הדברים האלה אפשר להרחיב למקרה האינסופי. במקרה הזה אין $|\mathcal{H}|$ - במקום זה משתמשים במדד שנקרא "VC Dimension".