1

לימוד מבנה

למה?
- למידה מדויקת עם פחות פרמטרים
- הבנה של העולם
- פירוק טוב - ניתן לבצע חישובים?

מה עדיף?
קשתות עודפות - בעיית למידה
קשתות חסרות - לא ניתן להתגבר ע"י התאמה

הבעיה: יש $2^{O(n^2)}$ מבנים!

גישות אפשריות:

(I) אילוצים
מבחנים סטטיסטים על זוגות ושלשות...
+ אינטואיטיבי
- רגיש למבחן

(II) בייסיאני
קשה חישובית

(III) מבוסס ציון (Score)
+ מוטיבציה סטטיסטית
+ פשרות
+ לוקח בחשבון מבנה
- יקר (אבל פחות מ-II)

ציון הנראות Likelihood Score

$$L(G:D) = \max_{\theta_G} \log P(D|G, \theta_G) =$$

$$= \sum_m \log P_{\hat\theta}(X[m]) = \sum_m \sum_i \log P_{\hat\theta}(X_i[m] | Pa_i[m])$$

$$= \sum_i \sum_{pa_i} \sum_{x_i} \underbrace{M[x_i, pa_i]}_{\text{ספירה}} \cdot \log \theta_{x_i|pa_i}$$

(פ)

$$= \sum_i \sum_{pa_i} \sum_{x_i} \left[\log \hat{P}(x_i|pa_i)\right] \cdot M[x_i, pa_i]$$

$$= M \sum_i \sum_{pa_i} \sum_{x_i} \hat{P}(x_i, pa_i) \log \hat{P}(x_i|pa_i) \cdot \frac{\hat{P}(x_i)}{\hat{P}(x_i)}$$

$$= M \sum_i \sum_{pa_i} \sum_{x_i} \left[\underbrace{\hat{p}(x_i,pa_i) \log \frac{\hat{p}(x_i,pa_i)}{\hat{p}(pa_i)\hat{p}(x_i)}}_{\hat{I}(x_i;pa_i)} + \underbrace{\hat{p}(x_i,pa_i) \log \hat{p}(x_i)}_{-\hat{H}(x_i)} \right]$$

$$= M\left[\underbrace{\sum_i \hat{I}(x_i;pa_i)}_{\text{אינפורמציה בכל משפחה}} - \underbrace{\hat{H}(x_i)}_{\text{לא תלוי במבנה}} \right]$$

למה זה טוב? - הציון מתפרק (עוזר לחישוב)

- קל ללמוד עצים

מה הבעיה? $I(X;Y) \le I(X;Y,Z)$ = אינפורמציה רק גדלה.

מה המשמעות?

פתרונות אפשריים:

- להגביל מרחב חיפוש (עצים, NB, TAN)
- לשנות את ה- score

Bayesian Score

$$P(G|D) = \frac{P(D|G)\,P(G)}{P(D)}$$

נתייחס למבנה כמשתנה מקרי

$$\text{score}_B(G:D) = \log \underbrace{P(D|G)}_{\text{marginal likelihood}} + \log \underbrace{P(G)}_{\text{prior}}$$

prior (פחות משפיע)

$$\underbrace{P(D|G)}_{\text{ממוצע על פני פרמטרים}} = \int_{\Theta_G} \underbrace{P(D|G,\theta_G)}_{\text{נראות רגילה}}\, \underbrace{P(\theta_G|G)}_{\text{פריור על פרמטרים}}\, d\theta$$

ממוצע על פני פרמטרים. מה זה אומר?

צורה אחרת

$$= \prod_{m=1}^{M} \underbrace{P(X[m] \mid X[1] \dots X[m-1])}_{\text{prequential form}}$$

למה הצורה הזאת מעניינת?

ואכן ניתן להראות:

$$\frac{1}{M}\log P(D|G) \approx E_{P^*}\left[\log P(X|G,D)\right]$$

P^* – התפלגות אמיתית

נתחיל במשתנה בינארי:

$$D = \langle H, T, T, H, H \rangle$$

$$P(D) = P(X[1])\,P(X[2] \mid X[1]) \dots$$

$$= \prod_m P(X[m]=1 \mid X[1] \dots X[m-1]) = \prod_m \int_\theta P(X[m]|\theta)\,P(\theta|\cdot)\,d\theta$$

ראינו כבר בעזרת פריור דיריכלה

$$= \frac{\alpha_1}{\alpha}\cdot\frac{\alpha_0}{\alpha+1} + \frac{\alpha_0+1}{\alpha+2}\cdot\frac{\alpha_1+1}{\alpha+3}\cdot\frac{\alpha_1+2}{\alpha+4} = \frac{[\alpha_1(\alpha_1+1)(\alpha_1+2)][\alpha_0(\alpha_0+1)]}{\alpha(\alpha+1)\cdots(\alpha+4)}$$

4

הכללה:

$$P(X[1]\ldots X[m]) = \frac{[\alpha_1(\alpha_1+1)\ldots \alpha_1+M[1]-1][\alpha_0\ldots(\alpha_0+M[0]-1)]}{\alpha(\alpha+1)\ldots\ldots \alpha+M-1}$$

$$\boxed{\begin{array}{l}\Gamma(m)=(m-1)! \\ \Gamma(x+1)=x\Gamma(x) \\ \Gamma(\alpha)=\int_0^\infty t^{\alpha-1}e^{-t}dt\end{array}}$$

$$\equiv \frac{\Gamma(\alpha_1+M[1])}{\Gamma(\alpha_1)}\cdot\frac{\Gamma(\alpha_0+M[0])}{\Gamma(\alpha_0)}\cdot\frac{\Gamma(\alpha)}{\Gamma(\alpha+M)}$$

הכללה למולטינום:

$$\log P(D) = \log\Gamma(\alpha) - \log\Gamma(\alpha+M) + \sum_{i=1}^{K}\left[\log\Gamma(\alpha_i+M[x_i=i]) - \log\Gamma(\alpha_i)\right]$$

וכמובן, זה לא תלוי בסדר התצפיות.

הכללה לרשת בייסיאנית X ו-Y בת"ז:

$$P(D|G_\emptyset) = \int_{\theta_x,\theta_y} P(\theta_x,\theta_y|G_\emptyset)P(D|G_\emptyset,\theta_x,\theta_y)\,d[\theta_x,\theta_y]$$

באיזו הנחה השתמשנו פה?

$$= \left(\int_{\theta_X} P(\theta_X|G_\emptyset)\prod_m P(X[m]|\theta_X,G_\emptyset)\,d\theta_X\right)$$

$$\times\left(\int_{\theta_Y} P(\theta_Y|G_\emptyset)\prod_m P(X[m]|\theta_Y,G_\emptyset)\,d\theta_Y\right)$$

כל אחד מאלה הוא גרסת המולטינום למעלה.

<u>נכליל למקרה של תלות:</u>

$$P(D|G_{X\to Y}) =$$

$$\left(\int_{\theta_X} \text{כמו קודם} \right)$$

$$\times\left(\int_{\theta_{y|x=0}} P(\theta_{y|x=0}|G_{X\to Y}) \prod_{m:X[m]=0} P(y[m]|\underset{\theta_{y|x=0}}{x=0}) \cdot d\theta_{y|x=0}\right)$$

$$\times\left(\text{כנ"ל עבור } y|x=1 \right)$$

<u>ולהכללה הסופית:</u> נסמן $\alpha_{pa_i} = \sum_{x_i} \alpha_{x_i|pa_i}$

$$P(D|G) = \prod_i \prod_{pa_i} P(D_i|D_{pa_i}, G)$$

$$\log P(D|G) = \sum_i \sum_{pa_i} \left[\log\frac{\Gamma(\alpha_{pa_i})}{\Gamma(\alpha_{pa_i}+M[pa_i])} + \sum_i \log\frac{\Gamma(\alpha_{x_i|pa_i}+M[\cdot])}{\Gamma(\alpha_{x_i|pa_i})} \right.$$

<u>משפט (תרגיל):</u> בעזרת קרוב סטירלינג $\Gamma(x) \approx \sqrt{2\pi}\, x^{(x-\frac{1}{2})} \cdot e^{-x}$

$$\log P(D|G) \approx \underbrace{\ell(\hat{\theta}_G : D)}_{\text{ציון הנראות}} - \tfrac{1}{2}\log M \cdot \underbrace{Dim(G)}_{\text{מספר פרמטרים חופשיים.}} + O(1)$$

<u>מה המשמעות?</u>

הגדרה: נניח שהנתונים נדגמו מתוך P^* שעבורו G^* הוא p-map. נאמר שפונקציית ה-score היא קונסיסטנטית עבור $M\to\infty$ בהסתברות 1 אם:

(1) G^* הוא בעל הציון הגבוה ביותר.

(2) לכל G שאינו במחלקת שקילות ציון נמוך יותר.

משפט: BIC הוא קונסיסטנטי

(I) נניח שב-G אי-תלויות שלא קיימות ב-G^*. אזי G הוא לא I-map של P ולכן $\sum_i I_{P^*}(X_i;Pa_i^{G^*}) - \sum_i I_{P^*}(X_i;Pa_i^{G}) > 0$

(II) נניח ש-G מקודד רק חלק מהאי-תלויות שב-G^*. בפרט G יכול לייצג כל התפלגות ש-G^* מייצגת ובפרט P^*. כאשר $M\to\infty$, $\hat{P}\to P^*$ ונקבל

$$score_L(G^*:D) - score_L(G:D) \to 0$$

שלב אזי: $\Delta Bic = \frac{1}{2}[Dim(G)-Dim(G^*)]\log M > 0$

$\Leftarrow$ Bayesian Score is consistent