Bayesovská sieť

- Orientovaný acyklický graf
- Vrcholy: náhodné premenné $X_1 \ldots X_n$
- Nech Z_i sú predchodcovia premennej X_i
- Pre každé X_i tabuľka $\Pr(X_i|Z_i)$
- Sieť definuje celkové rozdelenie

$$\Pr(X_1 = x_1, \ldots, X_n = x_n) = \prod_{i=1}^{n} \underbrace{\Pr(X_i = x_i | Z_i = x_{Z_i})}_{\Theta_{i,x_i,x_{Z_i}}}$$

Trénovanie s úplnými dátami

$$x_1 = (x_{1,1}, x_{1,2}, x_{1,3}, \ldots, x_{1,n})$$
$$x_2 = (x_{2,1}, x_{2,2}, x_{1,3}, \ldots, x_{2,n})$$
$$\ldots$$
$$x_t = (x_{t,1}, x_{t,2}, x_{t,3}, \ldots, x_{t,n})$$

Hľadáme najvierohodnejšie parametre:

$$\Theta^* = \arg\max_{\Theta} \sum_{i=1}^{t} \log \Pr(x_{i,1}, \ldots, x_{i,n} | \Theta)$$

Nastavíme na pozorované frekvencie: $\Theta^*_{j,u,v} = \frac{\#_{j,u,v}}{n_{j,v}}$

- $\#_{j,u,v}$ je počet vzoriek, kde $X_j = u$ a $Z_j = v$
- $n_{j,v}$ je počet vzoriek, kde $Z_j = v$

Trénovanie s neúplnými dátami

- $x^{(i)}$ - pozorované dáta vo vzorke i
- $X^{(i)}$ - pozorované premenné
- $y^{(i)}$ - chýbajúce dáta vo vzorke i
- $Y^{(i)}$ - chýbajúce premenné
- Cieľ: maximalizovať vierohodnosť

$$\Theta^* = \arg\max_{\Theta} \prod_{i=1}^{t} \Pr(X^{(i)} = x^{(i)} | \Theta)$$

EM algoritmus (Expectation maximization)

- Zvoľ počiatočné $\Theta^{(0)}$
- Iteruj kým nenastane konvergencia ($k = 0, 1, \ldots$)
  - E-krok: spočítaj pre každé $i, y^{(i)}$:

$$Q_i^{(k)}(y^{(i)}) = \Pr(Y^{(i)} = y^{(i)} | X^{(i)} = x^{(i)}, \Theta^{(k)})$$

  - M-krok: spočítaj

$$\Theta^{(k+1)} = \arg\max_{\Theta} \sum_{i=1}^{t} \sum_{y^{(i)}} Q_i^{(k)}(y^{(i)}) \log \Pr(X^{(i)} = x^{(i)}, Y^{(i)} = y^{(i)} | \Theta)$$

    Maximalizujeme vierohodnosť dát doplnených vš. spôsobmi o chýbajúce hodnoty, váhované cez $Q^{(k)}$

M-krok EM algoritmu

Spočítaj

$$\Theta^{(k+1)} = \arg\max_{\Theta} \sum_{i=1}^{t} \sum_{y^{(i)}} Q_i^{(k)}(y^{(i)}) \log \Pr(X^{(i)} = x^{(i)}, Y^{(i)} = y^{(i)} | \Theta)$$

Maximalizujeme vierohodnosť dát doplnených vš. spôsobmi o chýbajúce hodnoty, váhované cez $Q^{(k)}$

Maximum pre pozorované frekvencie s váhovanými dátami:

$$\Theta_{j,u,v}^{(k+1)} = \frac{\#_{j,u,v}}{n_{j,v}}$$

- $\#_{j,u,v}$ je súčet váh doplnených vzoriek, v ktorých $X_j = u$ a $Z_j = v$
- $n_{j,v}$ je súčet váh doplnených vzoriek, v ktorých $Z_j = v$

EM algoritmus (Expectation maximization)

Označenie: $L(\Theta|X) = \prod_{i=1}^{t} Pr(X^{(i)} = x^{(i)}|\Theta)$

Lema: $L(\Theta^{(k+1)}|X) \geq L(\Theta^{(k)}|X)$, pričom rovnosť nastáva iba v lokálnom maxime alebo inflexnom bode $L(\Theta|X)$.

- Iterácie postupne zvyšujú vierohodnosť
- Môžu skončiť v lokálnom maxime