1 Naív-Bayes módszer

- A_i: egy objektum i-edik paraméterének lehetséges értékeiből álló (véges) halmaz, $i = 1, \ldots, n$
- V: a lehetséges osztályokból álló halmaz (V elemeire gyakran mint *címké*kre hivatkozunk)
- $\mathcal{D}$: eloszlás $(A_1 \times A_2 \times \cdots \times A_n) \times V$ felett
- E: $\mathcal{D}$ szerint generált példahalmaz ($E \subseteq (A_1 \times A_2 \times \cdots \times A_n) \times V$)
- $(\mathbf{X}, Y) = ((X_1, \ldots, X_n), Y)$: $\mathcal{D}$ szerint generált véletlen példa (tulajdonságvektor-címke pár)

1.1 Az osztályozás

A módszer alapján abba a v osztályba sorolunk egy ismeretlen címkéjű $(a_1, \ldots, a_n)$ tulajdonságvektorral rendelkező objektumot, amely maximalizálja a $\mathbb{P}(Y = v) \prod_{i=1}^{n} \mathbb{P}(X_i = a_i | Y = v)$ értéket — tehát

$$v_{\text{Naiv}} = \operatorname*{argmax}_{v \in V} \mathbb{P}(Y = v) \prod_{i=1}^{n} \mathbb{P}(X_i = a_i | Y = v).$$

A szorzatban levő valószínűségeket az E mintahalmaz segítségével becsüljük a következőképpen. $\mathbb{P}(Y = v)$ közelítése ($v \in V$):

$$\mathbb{P}(Y = v) \approx \frac{|\{(\mathbf{x}, y) \in E : y = v\}|}{|E|},$$

$\mathbb{P}(X_i = a_i | Y = v)$ közelítése ($v \in V$, $i = 1, \ldots, n$):

$$\mathbb{P}(X_i = a_i | Y = v) \approx \frac{|\{(\mathbf{x}, y) \in E : y = v, x_i = a_i\}|}{|\{(\mathbf{x}, y) \in E : y = v\}|}.$$

1.2 m-estimate of probability

Kis mintahalmaz esetén a kisebb $\mathbb{P}(X_i = a_i | Y = v)$ értékekre adott becslések könnyen nullának adódhatnak, ami túlzott torzításhoz vezethet. Ilyen esetekben érdemes lehet a fenti becslésekbe "kívülről" beépíteni a problémával kapcsolatosan esetlegesen rendelkezésünkre álló ismereteket a következő formában. Ha tudjuk, hogy egy $\mathbb{P}(X_i = a_i | Y = v)$ értéknek megközelítőleg

mekkorának kellene lennie (jelöljük ezen értéket $p_{i,v}$-vel), akkor alkalmazhatjuk a

$$\mathbb{P}(X_i = a_i | Y = v) \approx \frac{|\{(\mathbf{x}, y) \in E : y = v, x_i = a_i\}| + p_{i,v}m}{|\{(\mathbf{x}, y) \in E : y = v\}| + m}.$$

becslést, ahol m egy megfelelően nagy konstans. (Minél biztosabbak vagyunk $p_{i,v}$ értékében, m annál nagyobb lehet). Ezen módszer neve "m-estimate of probability".

Ha ilyen információ nem áll rendelkezésünkre, akkor egy lehetőség, hogy m értékéül az adott komponens lehetséges értékeinek számát adjuk meg (azaz $m = |A_i|$), $p_{i,v}$ értékéül pedig ennek reciprokát (azaz $p_{i,c} = 1/|A_i|$). Ekkor tehát a becslésünk

$$\mathbb{P}(X_i = a_i | Y = v) \approx \frac{|\{(\mathbf{x}, y) \in E : y = v, x_i = a_i\}| + 1}{|\{(\mathbf{x}, y) \in E : y = v\}| + |A_i|}.$$