Naivni Bayesov klasifikator

[1]

Naivni Bayesov klasifikator predpostavlja pogojno neodvisnost vrednosti atributov pri danem razredu:

$$p(v_1, v_2, ...v_n|c) = \Pi_i p(v_i|c)$$

Naivna bayesova formula:

$$p(c|v_1, v_2, ...v_n) = p(c) * \Pi_i \frac{p(c|v_i)}{p(c)}$$

Naloga učnega algoritma je s pomočjo učne množice podatkov aproksimirati verjetnosti na desni strani enačbe.

Kako naivni Bayesov klasifikator klasificira nov primer $(v_1, v_2, ..., v_n)$**?**

Recimo, da ima ciljna spremenljivka m možnih vrednosti $(c_1, c_2, ...c_m)$. Naivni Bayesov klasifikator za vsak razred c_i po naivni Bayesovi formuli izračuna verjetnost, da primer $(v_1, v_2, ..., v_n)$ pripada razredu c_i, kar zapišemo $p(c_i|v_1, v_2, ...v_n)$. Primer klasificira v razred z največjo verjetnostjo.

Primer

Ali bo pajek ujel mravljo?

1. Barva = bela, Čas = noč

$$v_1 = \text{“}Barva = bela\text{”}$$
$$v_2 = \text{“}Cas = noc\text{”}$$
$$c_1 = DA$$
$$c_2 = NE$$

[1] Petra Kralj, Petra.Kralj@ijs.si, http://kt.ijs.si/PetraKralj/UNGKnowledgeDiscovery0607.html

Table 1: Pajek ima izkušnje z lovljenjem mravlej:

Barva	**Velikost**	**Čas**	**Ujel**
črna	velika	dan	**DA**
bela	mala	noč	**DA**
črna	mala	dan	**DA**
rdeča	velika	noč	**NE**
črna	velika	noč	**NE**
bela	velika	noč	**NE**

$$p(c_1|v_1, v_2) = \quad (1)$$

$$p(Ujel = DA|Barva = bela, Cas = noc) = \quad (2)$$

$$p(Ujel = DA) * \frac{p(Ujel = DA|Barva = bela)}{p(Ujel = DA)} * \frac{p(Ujel = DA|Cas = noc)}{p(Ujel = DA)} = \quad (3)$$

$$\frac{1}{2} * \frac{\frac{1}{2}}{\frac{1}{2}} * \frac{\frac{1}{4}}{\frac{1}{2}} = \frac{1}{4} \quad (4)$$

$$p(c_2|v_1, v_2) = \quad (5)$$

$$p(Ujel = NE|Barva = bela, Cas = noc) = \quad (6)$$

$$p(Ujel = NE) * \frac{p(Ujel = NE|Barva = bela)}{p(Ujel = NE)} * \frac{p(Ujel = NE|Cas = noc)}{p(Ujel = NE)} = \quad (7)$$

$$\frac{1}{2} * \frac{\frac{1}{2}}{\frac{1}{2}} * \frac{\frac{3}{4}}{\frac{1}{2}} = \frac{3}{4} \quad (8)$$

Bele mravlje v nočnem času ne bo ujel, ker je p(Ujel=NE| Barva = bela, Čas = noč) > p(Ujel=DA | Barva = bela, Čas = noč).

2. Barva = črna, Velikost = velika, Čas = dan

$$v_1 = \text{“}Barva = crna\text{”}$$

$$v_2 = \text{“}Velikost = velika\text{”}$$

$$v_3 = \text{“}Cas = dan\text{”}$$

$$c_1 = DA$$

$$c_2 = NE$$

$$p(c_1|v_1, v_2, v_3) = \quad (9)$$

$$p(Ujel = DA|Barva = crna, Velikost = velika, Cas = dan) = \quad (10)$$

$$p(Ujel = DA) * \frac{p(Ujel = DA|Barva = crna)}{p(Ujel = DA)} * ... \quad (11)$$

$$... * \frac{p(Ujel = DA|Velikost = velika)}{p(Ujel = DA)} * \frac{p(Ujel = DA|Cas = dan)}{p(Ujel = DA)} = \quad (12)$$

$$\frac{1}{2} * \frac{\frac{2}{3}}{\frac{1}{2}} * \frac{\frac{1}{4}}{\frac{1}{2}} * \frac{1}{\frac{1}{2}} = \frac{2}{3} \quad (13)$$

$$p(c_2|v_1, v_2, v_3) = \quad (14)$$

$$p(Ujel = NE|Barva = crna, Velikost = velika, Cas = dan) = \quad (15)$$

$$p(Ujel = NE) * \frac{p(Ujel = NE|Barva = crna)}{p(Ujel = NE)} * ... \quad (16)$$

$$... * \frac{p(Ujel = NE|Velikost = velika)}{p(Ujel = NE)} * \frac{p(Ujel = NE|Cas = dan)}{p(Ujel = NE)} * = \quad (17)$$

$$\frac{1}{2} * \frac{\frac{1}{3}}{\frac{1}{2}} * \frac{\frac{3}{4}}{\frac{1}{2}} * \frac{0}{\frac{1}{2}} = 0 \quad (18)$$

Veliko črno mravljo bo pajek v dnevnem času ujel, ker je p(Ujel=DA | Barva = črna, Velikost = velika, Čas = dan) > p(Ujel=NE | Barva = črna, Velikost = velika, Čas = dan).

V razmislek:

Ko z naivno Bayesovo formulo izračunamo verjetnosti $p(c_1|v_1, v_2)$ in $p(c_2|v_1, v_2)$ (dvorazredni problem), včasih $p(c_2|v_1, v_2) + p(c_2|v_1, v_2) \neq 1$. Zakaj?

[2]

[2] Če v dokumentu najdete napako, me prosim obvestite na Petra.Kralj@ijs.si.