Prof.dr.sc. Bojana Dalbelo Bašić

Fakultet elektrotehnike i računarstva
Zavod za elektroniku, mikroelektroniku, računalne i inteligentne sustave

www.zemris.fer.hr/~bojana
bojana.dalbelo@fer.hr

Naivan Bayesov klasifikator

v1.2

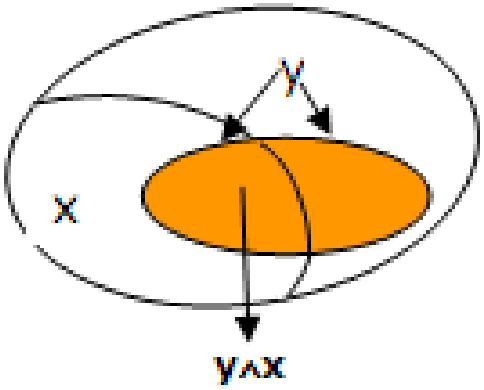
Bayesovo pravilo

$$p(y \mid x) = \frac{P(x \mid y)P(y)}{P(x)}$$

Pojašnjenj e Bayesovog pravila :

x = podaci

y = hipoteza (model)



X – prostor el. događaja

$$p(H_i \mid E_1\, E_2 \ldots E_n) =$$

$$= \frac{p(E_1 \mid H_i)\, p(E_2 \mid H_i) \ldots p(E_n \mid H_i)\, p(H_i)}{\sum_{k=1}^{m} p(E_1 \mid H_k) p(E2 \mid H_k) \ldots p(E_n \mid H_k) p(H_k)}$$

Bayesovo pravilo - primjer

- H = {h_1=(iz Skandinavije), h_2=(iz ostatka Europe)};

  $P(h_1) = 0{,}048$; $P(h_2) = 0{,}952$

- A = {osoba je plava}; $P(A) = 0{,}1$
- u Skandinaviji su gotovo svi plavi $P(A|h_1) = 0{,}85$

Kolika je vjerojatnost da je osoba plave kose iz Skandinavije?

$$P(h_1 \mid A) = \frac{P(A \mid h_1)P(h_1)}{P(A)} = \frac{0.85*0.048}{0.1} = 0.408$$

Nazivlje

- $P(h_1)$ je **a priori** vjerojatnost hipoteze h_1
- $P(h_1|A)$ je **a posteriori** vjerojatnost hipoteze h_1
- $P(A|h_1)$ je **izglednost** vjerojatnost hipoteze h_1

"likelihood" se prevodi kao vjerodostojnosti li izglednost

Specifičnosti naivnog Bayesovog klasifikatora

- primjena kod podataka koje se sastoje od više atributa
- koristi pretpostavku o nezavisnosti atributa unutar razreda zbog čega se i naziva “naivnim”
- računa **MAP-hipotezu**

Pretpostavka koju klasifikator koristi

- atributi unutar podatka su međusobno nezavisni u odnosu na razred:

$$P(x \mid h) = P(a_1, \ldots, a_T \mid h) = \prod_t P(a_t \mid h)$$

- ta pretpostavka u stvarnosti često može biti narušena
- usprkos tome, u praksi dobro funkcionira

MAP-hipoteza

- **MAP-hipoteza** (Maximum A Posteriori) je ona hipoteza za koju je P(h|D) najveći za predočene podatke D
- pišemo $\mathbf{h_{MAP}}$

MAP-hipoteza

- na temelju Bayesovog teorema:

$$h_{MAP} = \arg\max_{h_i \in H} P(h_i \mid D)$$

$$h_{MAP} = \arg\max_{h_i \in H} \frac{P(D \mid h_i) P(h_i)}{P(D)}$$

Vjerojatnost *P(D)* možemo izostaviti jer je konstantna

$$h_{MAP} = \arg\max_{h_i \in H} P(D \mid h_i) P(h_i)$$

Maksimalna izglednost (maximum likelihood ML)

U slučaju kada su sve hipoteze iz *h* jednako vjerojatne, dalje pojednostavljujemo:

$$h_{MAP} = \arg\max_{h_i \in H} P(D \mid h_i) P(h_i)$$

$$h_{ML} = \max_{h_i \in H} P(D \mid h_i)$$

Praktične primjene klasifikatora

- Medicinske dijagnoze
- Klasifikacija teksta
  - Filtracija neželjene (*spam*) pošte
  - Odabir web-stranica koje bi mogle zanimati određenog korisnika
- Proučavanje karakteristika genoma

Primjer “Dan za odbojku na pijesku”

DAN	VRIJEME	TEMPERATURA	VLAGA	VJETAR	IGRATI ODBOJKU?
Dan 1	Sunčano	Visoka	Visoka	Slab	**Ne**
Dan 2	Sunčano	Visoka	Visoka	Jak	**Ne**
Dan 3	Oblačno	Visoka	Visoka	Slab	**Da**
Dan 4	Kišno	Srednja	Visoka	Slab	**Da**
Dan 5	Kišno	Niska	Normalna	Slab	**Da**
Dan 6	Kišno	Niska	Normalna	Jak	**Ne**
Dan 7	Oblačno	Niska	Normalna	Jak	**Da**
Dan 8	Sunčano	Srednja	Visoka	Slab	**Ne**
Dan 9	Sunčano	Niska	Normalna	Slab	**Da**
Dan 10	Kišno	Srednja	Normalna	Slab	**Da**
Dan 11	Sunčano	Srednja	Normalna	Jak	**Da**
Dan 12	Oblačno	Srednja	Visoka	Jak	**Da**
Dan 13	Oblačno	Visoka	Normalna	Slab	**Da**
Dan 14	Kišno	Srednja	Visoka	Jak	**Ne**

Rješenje upotrebom naivnog Bayesovog klasifikatora

- Klasificiraj svako novo pojavljivanje podatka $x=(a_1,\ldots a_n)$ kao:

$$h_{Naive\ Bayes} = \arg\max_h P(h)P(\mathbf{x}\mid h) = \arg\max_h P(h)\prod_{i=1}^{n} P(a_i \mid h)$$

- Da bi to izračunali na temelju primjera za učenje, moramo procijeniti parametre iz primjera za učenje
  - Za svaku hipotezu h

$$\hat{P}(h) := \text{procijeni}\, P(h)$$

  - Za svaku vrijednost atributa a_t za svaku pojavu podatka

$$\hat{P}(a_t \mid h) := \text{procijeni}\, P(a_t \mid h)$$

Pomoću primjera iz tablice, klasificiraj zadani **novi** podatak ***x***
(vektor koji se ne nalazi u postojećoj tablici):

x = (Vrijeme = sunčano, Temp = niska, Vlaga = visoka, Vjetar = jak)

Igrati odbojku ili ne?

$$h_{NB} = \underset{h \in [da, ne]}{\operatorname{argmax}} P(h)P(x|h) = \underset{h \in [da, ne]}{\operatorname{argmax}} P(h)\prod_{t} P(a_t|h)$$

$$h_{NB} = \operatorname{argmax} P(h)\, P(\text{Vrijeme} = \text{sunčano}|h)\, P(\text{Temp} = \text{niska}|h)$$
$$P(\text{Vlaga} = \text{visoka}|h) P(\text{Vjetar} = \text{jak}|h)$$

Pomoću primjera iz tablice, klasificiraj zadani **novi** podatak ***x***
(vektor koji se ne nalazi u postojećoj tablici):
X = (Vrijeme = sunčano, Temp = niska, Vlaga = visoka, Vjetar = jak)

$$P(IgratiOdbojku = da) = 9/14 = 0.64$$
$$P(IgratiOdbojku = ne) = 5/14 = 0.36$$
$$P(Vrijeme = suncano \mid IgratiOdbojku = da) = 2/9 = 0.22$$
$$P(Vrijeme = suncano \mid IgratiOdbojku = ne) = 3/5 = 0.60$$
$$P(Temperatura = niska \mid IgratiOdbojku = da) = 3/9 = 0.33$$
$$P(Temperatura = niska \mid IgratiOdbojku = ne) = 1/5 = 0.20$$
$$P(Vlaga = visoka \mid IgratiOdbojku = da) = 3/9 = 0.33$$
$$P(Vlaga = visoka \mid IgratiOdbojku = ne) = 4/5 = 0.80$$
$$P(Vjetar = jak \mid IgratiOdbojku = da) = 3/9 = 0.33$$
$$P(Vjetar = jak \mid IgratiOdbojku = ne) = 3/5 = 0.60$$

$$P(da)P(\text{suncan} \mid da)P(niska \mid da)P(visoka \mid da)P(jak \mid da) =$$
$$0.64 * 0.22 * 0.33 * 0.33 * 0.33 = 0.0053$$
$$P(ne)P(\text{suncan} \mid ne)P(niska \mid ne)P(visoka \mid ne)P(jak \mid ne) =$$
$$0.36 * 0.6 * 0.2 * 0.8 * 0.6 = \mathbf{0.020}7$$
$$\Rightarrow odgovor: IgratiOdbojku(D) = ne$$

$$h_{MAP} = "ne"$$

Sažetak

- Bayesovo pravilo se može pretvoriti u klasifikator
- Naivan Bayesov klasifikator je jednostavan, ali učinkovit klasifikator za vektorske podatke (podatke s više atributa)
- Pretpostavlja da su atributi nezavisni u odnosu na razred (klasu)

- Neki drugi klasifikatori:
  - stroj s potpornim vektorima (SVM)
  - k-NN
  - stabla odluke
  - neuronske mreže
  - …