Naive Bayes

10.10.2017
4.hafta
Makine öğrenmesi

Random Variables and Probability

A: Random variable

A: the next patient you examine has a cough

P(A) : fraction of possible world which A is true
↓
probability

$0 \leq P(A) \leq 1 \quad P(true) = 1$

$P(false) = 0$

S: Sample space

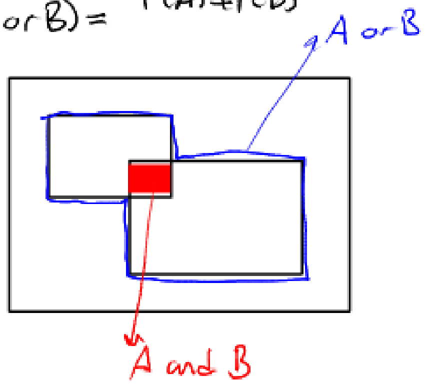
$$P(A) = \frac{\#\text{ of } A\text{'s}}{\#\text{ of elements in } S}$$

$$\sum P(A) = 1$$

Dependence of Variables

Independence of Variables

$P(A \text{ or } B) = P(A) + P(B)$



$P(A) = P(A \text{ and } B) + P(A \text{ and not } B)$

Conditional Probability

P(A|B) = Fraction of worlds in which B is TRUE that also have A TRUE

$P(H) = 1/10$
↓
headache

$P(F) = 1/40$
↓
flu

$P(H|F) = 1/2$

$P(F|H) = ?$

$$P(F/H) = \frac{P(F \text{ and } H)}{P(H)} =$$

$$P(F \text{ and } H) = P(H|F) * P(F) = \frac{1}{2} \cdot \frac{1}{40} = \frac{1}{80}$$

$$= P(F|H) * P(H)$$

$$P(F|H) = \frac{1/80}{1/10} = 1/8$$

$$\boxed{\text{Chain Rule: } P(A|B) = \frac{P(A \text{ and } B)}{P(B)}}$$

$P(B) \rightarrow \neq 0$

$$\text{Bayes Rule: } P(A|B) = \frac{P(B|A) * P(A)}{P(B)}$$

$P(B|A)$ → likelihood (Conditional probability)

$P(A)$ → prior probability

$P(B)$ → evidence

A → class

B → evidence (attribute)

$$P(\text{elma} \mid \text{ince yaprak}) = \frac{P(\text{ince y.} \mid \text{elma}) * P(\text{elma})}{P(\text{ince y.})}$$

Example

$P(\oplus|\text{cancer}) = 0{,}98$

$P(\theta|\overline{\text{cancer}}) = 0{,}97$

$P(\text{cancer}) = 0{,}08$

Test yapılıyor, hasta kanser mi?

$P(\text{cancer}|\theta) = ?$

$$\left.\begin{matrix} P(\text{cancer}|\theta) \\ P(\overline{\text{cancer}}|\oplus) \end{matrix}\right\} \left.\begin{matrix} ? \\ \text{maximum a posteriori} \end{matrix}\right\rangle \rightarrow \text{MAP}$$

$$h_{MAP} = \underset{h \in H}{\text{argmax}}\,(P(h|D))$$

$h \in H$ → cancer, $\overline{\text{cancer}}$

$$P(\text{cancer}|\oplus) = \frac{P(\oplus|\text{cancer}) * P(\text{cancer})}{P(\oplus)}$$

$$P(\overline{\text{cancer}}|\oplus) = \frac{P(\oplus|\overline{\text{cancer}}) * P(\overline{\text{cancer}})}{P(\oplus)}$$

$$h_{MAP} = \{P(\oplus|\text{cancer}) * P(\text{cancer}), P(\oplus|\overline{\text{cancer}}) * P(\overline{\text{cancer}})\}$$

$$= \{0{,}98 * 0{,}008\ ,\ 0{,}03 * 0{,}992\}$$

$$= \{0{,}07, \boxed{0{,}029}\} \rightarrow \overline{\text{cancer}}$$

$a_1, a_2, a_3, \ldots a_n$: attributes (bağımsız)

$$\underbrace{P(a_1 \cap a_2 \cap a_3 \ldots \cap a_n)}_{\text{a özellikleri independent}} = \prod_i P(a_i \mid \underbrace{v}_{\text{targetvalue (class)}})$$

$$\underset{\text{class}}{V_{MAP}} = \underset{v_J \in V}{\operatorname{argmax}} \frac{P(a_1, a_2, \ldots a_n \mid v_J) * P(v_J)}{\underbrace{P(a_1, \ldots a_n)}_{\text{constant}}} = P(a_1, \ldots a_n \mid v_J) * P(v_J) \Rightarrow$$

$$\underset{v_J \in V}{\operatorname{argmax}} P(v_J \mid a_1, \ldots a_n) = ?$$

NAIVE BAYES CLASSIFIER

$$\boxed{\Rightarrow V_{MAP} = \underset{v_J \in V}{\operatorname{argmax}} \prod_i P(a_i \mid v_J) * P(v_J)}$$

Training

① Her sınıf için sınıflara ait örnek dökümanlar ayrı ayrı tek dök. yapılır (docs$_J$)

② stop words temizlenir (a, in, the, ...)

③ $P(v_J) = \frac{|\text{docs}_J|}{|\text{All docs}|} = \frac{n_J}{N}$

④ Kelimelerden sözlük oluşturulur

	class A	class B
fantastic	0,2	0,5

n_k: w_k kelimesinin o sınıfta kaç defa geçtiği

$$P(w_k \mid v_J) = \frac{n_k \xrightarrow{} \text{k kelimesinin o sınıftaki sayısı}}{n_J \xrightarrow{} \text{j. sınıftaki kelime sayısı}}$$

$$\boxed{P(w_k \mid v_J) = \frac{n_k + \underbrace{M}_{\text{m-estimate (sabit)}}}{n_J + \underbrace{|\text{vocabulary}|}_{\text{sözlükteki toplam farklı kelime sayısı}}}}$$

Example

class A : "the cat crabs the crolls off the stairs"

class B : "It is raining cats and dogs"

Vocabulary = {cat, crab, croll, stair, rain, dog}

x = "cats eat mice and dogs bury bones"

$$V_{MAP} = \underset{\{c_a, c_b\}}{\operatorname{argmax}} \prod_i P(a_i \mid v_J) * P(v_J)$$

$$P(v_J) = 1/2 \qquad P(w_k \mid v_J) = \frac{n_{kj} + 1}{n_J + 6}$$

$$\left.\begin{aligned} &\text{a.)}\quad P(A) * P(cat|A) * P(dog|A) = \frac{1}{2} * \frac{2}{4+6} * \frac{1}{10} = 0{,}01 \\ &\text{b)}\quad P(B) * P(cat|B) * P(dog|B) = \frac{1}{2} * \frac{2}{3+6} * \frac{2}{3+6} = 0{,}024 \end{aligned}\right\} V_{MAP} \Rightarrow V_B$$

log likelihood

↓

değerler çok küçükse
değerlerin logları alınabilir