ML INFOSEC
4: Gaussian Naive Bayes

Reminder: Naive Bayes

Let $\mathcal{A}_1$, $\mathcal{A}_2$, ..., $\mathcal{A}_k$ be sets of **attributes**, C a finite set of **classes** classes and $T \subset \mathcal{A}_1 \times \cdots \times \mathcal{A}_k$ a finite set of **instances**. Moreover, let

$$F : T \to C$$

be a function, i.e. each instance $(x_1, ..., x_k)$ is classified as class $F(x_1, ..., x_k)$. How should we predict the class of a new instance $(a_1, ..., a_k) \notin T$?

Naive Bayes Classifier

For $a = (a_1, \ldots, a_k)$, the Naive Bayes classifier is given by

$$c_{NB}(a) = \operatorname{argmax}_{c \in C} P(c) \prod_{i=1}^{k} P(a_i|c).$$

Normal (Gaussian) distribution

A real (continuous) random variable X is called normally distributed with mean μ and variance σ^2 (standard deviation σ), $X \sim N(\mu, \sigma^2)$, if

$$P(a < X < b) = \int_a^b f_{\mu,\sigma}(t)dt$$

where

$$f_{\mu,\sigma}(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

Note that

$$P(X = x) = 0 \quad \forall x \in \mathbb{R},$$

but

$$P(x - \epsilon/2 < X < x + \epsilon/2) = \int_{x-\epsilon/2}^{x+\epsilon/2} f_{\mu,\sigma}(t)dt \approx \epsilon f_{\mu,\sigma}(x)$$

Simplified Play tennis 1

$\mathcal{A}_1 = \mathbb{R}$ $\quad A_1 = T$ *Temperature*
$\mathcal{A}_2 = \mathbb{R}$ $\quad A_2 = H$ *Humidity*
$C = \{yes, no\}$ $\quad$ *Play tennis*

Temperature in °C a_1	Humidity in % a_2	Class $c = F(a_1, a_2)$
30	85	no
27	90	no
28	86	yes
21	96	yes
19	80	yes
18	70	no
18	65	yes
22	95	no
19	70	yes
20	80	yes
21	70	yes
24	90	yes
27	75	yes
21	91	no

Reminder: Sample mean and variance

For data $x_1, ..., x_N \in \mathbb{R}$

$$\mu = \frac{1}{N} \sum_{j=1}^{N} x_j$$

is called their (sample) mean and

$$\sigma^2 = \frac{1}{N-1} \sum_{j=1}^{N} (x_j - \mu)^2$$

their (sample) variance (σ: (sample) standard deviation).

Simplified Play tennis 2: Summary Statistics

	Temperature T in °C		Humidity H in %	
$c =$	yes	no	yes	no
	28	30	86	85
	21	27	96	90
	19	18	80	70
	18	22	65	95
	19	21	70	91
	20		80	
	21		70	
	24		90	
	27		75	
mean μ	21.89	23.6	79.11	86.2
std. dev. σ	3.62	4.83	10.22	9.73

Gaussian Naive Bayes I

Assumption 1

For each attribute A_i with values in $\mathcal{A}_i$, $P(A_i = \cdot | c)$ is normally distributed with parameters $\mu_{A_i,c}$ and $\sigma_{A_i,c}$, where $\mu_{A_i,c}$ and $\sigma_{A_i,c}$ are the mean and the standard deviation, resp., of the sample

$$(a_i \in \mathcal{A}_i \mid F(..., a_i, ...) = c).$$

Example

Consider the attribute T and the class $c = yes$. The corresponding sample is

$$28, 21, 19, 18, 19, 20, 21, 24, 27$$

with sample mean $\mu_{T,yes} = 21.89$ and sample variance $\sigma^2_{T,yes} = 13.1$ (sample standard deviation $\sigma_{T,yes} = 3.62$).

Gaussian Naive Bayes II

Assumption 2

Fix small $\epsilon > 0$. Instead $P(a_i|c) = P(A_i = a_i|c)$ we can use

$$P(a_i - \epsilon/2 < A_i < a_i + \epsilon/2|c) \approx \epsilon f_{\mu_{A_i,c},\sigma_{A_i,c}}(a_i).$$

Gaussian Naive Bayes III: The Classifier

Gaussian Naive Bayes Classifier

For $a = (a_1, \ldots, a_k) \in \mathcal{A}_1 \times \ldots \times \mathcal{A}_k$, the Gaussian Naive Bayes classifier is given by

$$c_{GNB}(a) = \text{argmax}_{c \in C} P(c) \prod_{i=1}^{k} f_{\mu_{A_i,c}, \sigma_{A_i,c}}(a_i).$$

This follows from Assumption 2 and

$$\begin{aligned} \text{argmax}_{c \in C} P(c) \prod_{i=1}^{k} \epsilon f_{\mu_{A_i,c}, \sigma_{A_i,c}}(a_i) &= \text{argmax}_{c \in C} \epsilon^k P(c) \prod_{i=1}^{k} f_{\mu_{A_i,c}, \sigma_{A_i,c}}(a_i) \\ &= \text{argmax}_{c \in C} P(c) \prod_{i=1}^{k} f_{\mu_{A_i,c}, \sigma_{A_i,c}}(a_i). \end{aligned}$$

Simplified Play tennis 3

We wish to calculate the Gaussian Naive Bayes classifier for $a = (19, 90)$

Play tennis	$P(yes) = 9/14$	$P(no) = 5/14$
Temperature	.08203	.05258
Humidity	.02231	.03799
Product	.001167	.00071

$$c_{GNB}(19, 90) = yes$$

$$P(c = yes) = \frac{0.001167}{0.001167 + 0.00071} = 0.62174$$