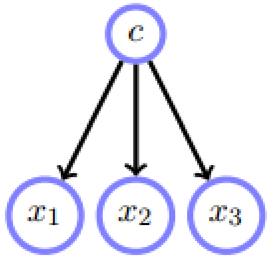
Chapter 10
Naïve Bayes

10.1 Naive Bayes and Conditional Independence

$$\mathbf{x} = (x_1, \dots, x_D)$$

$$p(\mathbf{x}, c) = p(c) \prod_{i=1}^{D} p(x_i, c)$$

$$p(c|\mathbf{x}) = \frac{p(\mathbf{x}|c)p(c)}{p(\mathbf{x})} = \frac{p(\mathbf{x}|c)p(c)}{\sum_c p(\mathbf{x}|c)p(c)}$$



$$\boxed{p(\mathbf{x}) = \sum_c p(\mathbf{x}, c) = \sum_c p(\mathbf{x}|c)p(c)}$$

- Example 10.1. Ezsurvey.org partitions radio station listeners into two groups – the 'young' and 'old'.

Age of listeners: young or old

$age \in \{\mathrm{y}, \mathrm{o}\}$ where $\mathrm{y} = \mathrm{young},\ \mathrm{o} = \mathrm{old}$

Likes or dislikes for radio stations

$r_i \in \{\mathrm{l}, \mathrm{d}\},\ i = 1,2,3,4$ where $\mathrm{l} = \mathrm{like},\ \mathrm{d} = \mathrm{dislike}$

Probability table

$p(r_1 = \mathrm{l} \mid age = \mathrm{y}\} = 0.95,\ p(r_1 = \mathrm{l} \mid age = \mathrm{o}\} = 0.03$

$p(r_2 = \mathrm{l} \mid age = \mathrm{y}\} = 0.05,\ p(r_1 = \mathrm{l} \mid age = \mathrm{o}\} = 0.82$

$p(r_3 = \mathrm{l} \mid age = \mathrm{y}\} = 0.02,\ p(r_1 = \mathrm{l} \mid age = \mathrm{o}\} = 0.34$

$p(r_4 = \mathrm{l} \mid age = \mathrm{y}\} = 0.20,\ p(r_1 = \mathrm{l} \mid age = \mathrm{o}\} = 0.92$

$p(age = \mathrm{o}) = 0.90$

Model

$$p(r_1, r_2, r_3, r_4 | age) = p(r_1|age)p(r_2|age)p(r_3|age)p(r_4|age)$$

Problem

$$p(age = \mathrm{y} \mid r_1 = \mathrm{l}, r_2 = \mathrm{d}, r_3 = \mathrm{l}, r_4 = \mathrm{d}) = ?$$

Solution

$$\frac{p(r_1 = \mathrm{l}, r_2 = \mathrm{d}, r_3 = \mathrm{l}, r_4 = \mathrm{d}|age = \mathrm{y})p(age = y)}{\sum_{age} p(r_1 = \mathrm{l}, r_2 = \mathrm{d}, r_3 = \mathrm{l}, r_4 = \mathrm{d}|age)p(age)} = 0.9161$$

$$
\begin{aligned}
p(r_1 = \mathrm{l}, r_2 = \mathrm{d}, r_3 = \mathrm{l}, r_4 = \mathrm{d} \mid age = \mathrm{y})p(age = \mathrm{y})& \\
&= p(r_1 = \mathrm{l} \mid age = \mathrm{y})p(r_1 = \mathrm{d} \mid age = \mathrm{y})p(r_1 = \mathrm{l} \mid age = \mathrm{y})p(r_1 = d \mid age = \mathrm{y})p(age = \mathrm{y}) \\
&= 0.95 \times 0.95 \times 0.02 \times 0.8 \times 0.1 \\
&= 0.0014
\end{aligned}
$$

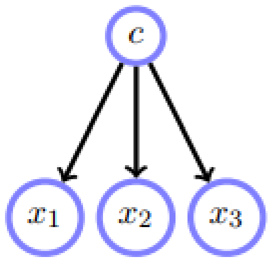
$$
\begin{aligned}
p(r_1 = \mathrm{l}, r_2 = \mathrm{d}, r_3 = \mathrm{l}, r_4 = \mathrm{d} \mid age = \mathrm{o})p(age = \mathrm{o})& \\
&= 0.03 \times 0.18 \times 0.34 \times 0.08 \times 0.9 \\
&= 1.3219 \times 10^{-4}
\end{aligned}
$$

answer: $\frac{0.0014}{0.0014+1.3219\times 10^{-4}} = 0.9161$

10.2 Estimation using Maximum Likelihood

- 10.2.1 Binary attributes

$\mathbf{x}$			c
x_1^1	$\cdots$	x_D^1	c^1
$\vdots$	$\ddots$	$\vdots$	$\vdots$
x_1^N	$\cdots$	x_D^N	c^N



Dataset

$\{(\mathbf{x}^n, c^n) | n = 1, \ldots, N\}$, $\mathbf{x}_i^n \in \{0,1\}$, $i = 1, \ldots, D$, $c^n \in \{0,1\}$

$\#(c^n = 0) = n_0$, $\#(c^n = 1) = n_1$

Parameters

$p(x_i = 1|c) = \theta_i^c$, $p(x_i = 0|c) = 1 - \theta_i^c$, $i = 1, \ldots, D$

Problem

Optimize θ_i^c

Model

$$p(\mathbf{x}|c) = \prod_{i=1}^{D} p(x_i|c) = \prod_{i=1}^{D} (\theta_i^c)^{x_i} (1 - \theta_i^c)^{1-x_i}$$

Likelihood

$$\prod_n p(\mathbf{x}^n, c^n | \theta_1^c, \ldots, \theta_D^c) = \prod_n p(\mathbf{x}^n, c^n)$$

for simplicity

Log likelihood

$$L = \log \prod_n p(\mathbf{x}^n, c^n) \; = \sum_n \log p(\mathbf{x}^n | c^n)\, p(c^n)$$

$$= \sum_{n,i} \{x_i^n \log \theta_i^c + (1 - x_i^n) \log(1 - \theta_i^c)\} + n_0 \log p(c = 0) + n_1 \log p(c = 1)$$

$$= \sum_{i,n} \{\mathbb{I}[x_i^n = 1, c^n = 0] \log \theta_i^0 + \mathbb{I}[x_i^n = 0, c^n = 0] \log\left(1 - \theta_i^0\right)$$

$$+ \mathbb{I}[x_i^n = 1, c^n = 1] \log \theta_i^1 + \mathbb{I}[x_i^n = 0, c^n = 1] \log\left(1 - \theta_i^0\right)\}$$

$$+ n_0 \log p(c = 0) + n_1 \log p(c = 1)$$

Maximize likelihood

$\frac{\partial L}{\partial \theta_i^0} = 0$, $\frac{\partial L}{\partial \theta_i^1} = 0$ implies

$$\theta_i^0 = \frac{\sum_n \mathbb{I}[x_i^n = 1, c^n = 0]}{\sum_n \left\{\mathbb{I}\left[x_i^n = 0, c^n = 0\right] + \mathbb{I}\left[x_i^n = 1, c^n = 0\right]\right\}}$$

$$= \frac{\text{number of times } x_i = 1 \text{ for class } c}{\text{number of datapoints in class } c}$$

$$\theta_i^1 = \frac{\sum_n \mathbb{I}[x_i^n = 1, c^n = 1]}{\sum_n \left\{\mathbb{I}\left[x_i^n = 0, c^n = 1\right] + \mathbb{I}\left[x_i^n = 1, c^n = 1\right]\right\}}$$

$\frac{\partial L}{\partial p(c=0)} = 0$ implies

$$p(c = 0) = \frac{\#[c^n = 0]}{N}$$

- Classification boundary

We classify a novel input $\mathbf{x}^* = (x_1^*, \dots, x_D^*)$ as class 1 if

$$p(c = 1|\mathbf{x}^*) > p(c = 0|\mathbf{x}^*)$$

Using Bayes' rule and log

$$\log p(\mathbf{x}^* \,|c = 1) + \log p(c = 1) - \log p(\mathbf{x}^*) > \log p(\mathbf{x}^* \,|c = 0) + \log p(c = 0) - \log p(\mathbf{x}^*)$$

$$\sum_i \log p(x_i^*|c = 1) + \log p(c = 1) > \sum_i \log p(x_i^*|c = 0) + \log p(c = 0)$$

$$\sum_i \left(x_i^* \theta_i^1 + (1 - x_i^*)\big(1 - \theta_i^1\big)\right) + \log p(c = 1)$$

$$> \sum_i \left(x_i^* \theta_i^0 + (1 - x_i^*)\big(1 - \theta_i^0\big)\right) + \log p(c = 0)$$

classify $\mathbf{x}^*$ as class 1 if

$$\sum_i w_i x_i^* + a > 0$$

Example 10.2 (Are they Scottish?). Consider the following vector of binary attributes:

$$(\text{shortbread, lager, whiskey, porridge, football})$$

eg) A vector $\mathbf{x} = (1, 0, 1, 1, 0)^{\mathrm{T}}$ would describe that a person likes shortbread, does not like lager, drinks whiskey, eats porridge, and has not watched England play football.

Label $nat \in \{\text{scottish}, \text{english}\}$

$$p(\text{sco}|\mathbf{x}) = \frac{p(\mathbf{x}|\text{sco})p(\text{sco})}{p(\mathbf{x})} = \frac{p(\mathbf{x}|\text{sco})p(\text{sco})}{p(\mathbf{x}|\text{sco})p(\text{sco}) + p(\mathbf{x}|\text{eng})p(\text{eng})}$$

Assumption

$$p(\mathbf{x}|nat) = p(x_1|nat)p(x_2|nat)p(x_3|nat)p(x_4|nat)p(x_5|nat)$$

0	1	1	1	0	0
0	0	1	1	1	0
1	1	0	0	0	0
1	1	0	0	0	1
1	0	1	0	1	0

English 6 people

1	1	1	1	1	1	1
0	1	1	1	1	0	0
0	0	1	0	0	1	1
1	0	1	1	1	1	0
1	1	0	0	1	0	0

Scottish 7 people

Prior in the database

$$p(\mathrm{sco}) = 7/13,\ p(\mathrm{eng}) = 6/13$$

ML

$$\begin{array}{ll}
p(x_1 = 1|\mathrm{eng}) = 3/6 & p(x_1 = 1|\mathrm{sco}) = 7/7 \\
p(x_2 = 1|\mathrm{eng}) = 3/6 & p(x_2 = 1|\mathrm{sco}) = 4/7 \\
p(x_3 = 1|\mathrm{eng}) = 2/6 & p(x_3 = 1|\mathrm{sco}) = 3/7 \\
p(x_4 = 1|\mathrm{eng}) = 3/6 & p(x_4 = 1|\mathrm{sco}) = 5/7 \\
p(x_5 = 1|\mathrm{eng}) = 3/6 & p(x_5 = 1|\mathrm{sco}) = 3/7
\end{array}$$

For $\mathbf{x} = (1, 0, 1, 1, 0)^{\mathrm{T}}$

$$p(\mathrm{sco}|\mathbf{x}) = \frac{p(\mathbf{x}|\mathrm{sco})p(\mathrm{sco})}{p(\mathbf{x}|\mathrm{sco})p(\mathrm{sco}) + p(\mathbf{x}|\mathrm{eng})p(\mathrm{eng})}$$

10.2.2 Multi-state variables

Data

$$\mathcal{D} = \{(\mathbf{x}^n, c^n) | n = 1, \dots, N\}$$

Goal

Determine c^* for novel input $\mathbf{x}^*$

Parameter

$$p(x_i = s|c) = \theta_s^i(c)$$

Conditional likelihood

$$\prod_{n=1}^{N} p(\mathbf{x}^n|c^n) = \prod_{n=1}^{N}\prod_{i=1}^{D} p(x_i^n|c^n)$$

$$= \prod_{n=1}^{N}\prod_{i=1}^{D}\prod_{s=1}^{S}\prod_{c=1}^{C} \theta_s^i(c)^{\mathbb{I}[x_i^n=s]\mathbb{I}[c^n=c]}$$

Conditional log-likelihood

$$L(\theta) = \sum_{i=1}^{N}\sum_{i=1}^{D}\sum_{s=1}^{S}\sum_{c=1}^{C} \mathbb{I}[x_i^n = s]\mathbb{I}[c^n = c] \log \theta_s^i(c)$$

$\mathbf{x}^n$			c^n
x_1^1	$\cdots$	x_D^1	c^1
$\vdots$	$\ddots$	$\vdots$	$\vdots$
x_1^N	$\cdots$	x_D^N	c^N

$x_i^n \in \{1, \dots, S\}$, $c^n \in \{1, \dots, C\}$

$$L(\theta) = \sum_{i=1}^{N}\sum_{i=1}^{D}\sum_{s=1}^{S}\sum_{c=1}^{C} \mathbb{I}[x_i^n = s]\mathbb{I}[c^n = c] \log \theta_s^i(c)$$

Probability condition

$$\sum_{s=1}^{S} \theta_s^i(c) = 1$$

Lagrangian

$$\mathcal{L}(\theta, \lambda) = L(\theta) + \sum_{c=1}^{C}\sum_{i=1}^{D} \lambda_i^c \left(1 - \sum_{s=1}^{S} \theta_s^i(c)\right)$$

where λ_i^c is the Lagrangian multiplier

Solution

$$\lambda_i^c = \sum_{n=1}^{N} \frac{\mathbb{I}[x_i^n = s]\mathbb{I}[c^n = c]}{\theta_s^i(c)}$$

$$\theta_s^i(c) = \frac{\sum_n \mathbb{I}[x_i^n = s]\mathbb{I}[c^n = c]}{\sum_{s'} \sum_n \mathbb{I}\left[x_i^n = s'\right]\mathbb{I}[c^n = c]}$$

$\mathbf{x}$			c
x_1^1	$\cdots$	x_D^1	c^1
$\vdots$	$\ddots$	$\vdots$	$\vdots$
x_1^N	$\cdots$	x_D^N	c^N

$x_i^n \in \{1, \dots, S\}$, $c^n \in \{1, \dots, C\}$

Goal

Determine c^* for novel input $\mathbf{x}^*$

Solution: ML

$$c^* = \underset{\mathrm{c}}{\operatorname{argmax}} \prod_i p(x_i = x_i^* | c)$$
$$= \underset{\mathrm{c}}{\operatorname{argmax}} \prod_i \theta_{x_i^*}^i(c)$$

10.3 Bayesian Naive Bayes

Data

$$\mathcal{D} = \{(\mathbf{x}^n, c^n) | n = 1, \dots, N\}$$

Goal

Find c^* for novel input $\mathbf{x}^*$

Parameters

$$p(x_i = s | c) = \theta_s^i(c)$$

Let

$$\theta^i(c) = \left(\theta_1^i(c), \dots, \theta_1^i(c)\right)$$

$$\theta(c) = \left\{\theta^i(c) \,\middle|\, i = 1, \dots, D\right\}$$

$$\theta = \left\{\theta^i(c) \,\middle|\, i = 1, \dots, D,\ c = 1, \dots, C\right\}$$

$\mathbf{x}$			c
x_1^1	$\cdots$	x_D^1	c^1
$\vdots$	$\ddots$	$\vdots$	$\vdots$
x_1^N	$\cdots$	x_D^N	c^N

$x_i^n \in \{1, \dots, S\}$, $c^n \in \{1, \dots, C\}$

$\mathbf{x}$			c
x_1^1	$\cdots$	x_D^1	c^1
$\vdots$	$\ddots$	$\vdots$	$\vdots$
x_1^N	$\cdots$	x_D^N	c^N

$x_i^n \in \{1, \dots, S\}$, $c^n \in \{1, \dots, C\}$

The posterior

$$p(\theta(c) \mid \mathcal{D}) = \prod_i p\left(\theta^i(c) \middle| \mathcal{D}\right)$$

$$\begin{aligned} p\left(\theta^i(c) \middle| \mathcal{D}\right) &\propto p\left(\theta^i(c), \mathcal{D}\right) \\ &= p\left(\mathcal{D} \middle| \theta^i(c)\right) p\left(\theta^i(c)\right) \\ &= p\left(\theta^i(c)\right) \prod_{n:c^n=c} p\left(x_i^n \middle| \theta^i(c)\right) \end{aligned}$$

The prior

$$p\left(\theta^i(c)\right) = \mathrm{Dirichlet}\left(\theta^i(c) \middle| u^i(c)\right)$$

where $u^i(c)$ is hyperparameter

Definition 8.27 (Dirichlet Distribution)**.** The Dirichlet distribution is a distribution on probability distributions, $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_Q)$, $\alpha_i \geq 0$, $\sum_i \alpha_i = 1$:

$$p(\boldsymbol{\alpha}) = \frac{1}{Z(\mathbf{u})}\delta\left(\sum_{i=1}^{Q} \alpha_i - 1\right) \prod_{q=1}^{Q} \alpha_q^{u_q - 1} \mathbb{I}\left[\alpha_q \geq 0\right] \tag{8.3.29}$$

where

$$Z(\mathbf{u}) = \frac{\prod_{q=1}^{Q} \Gamma(u_q)}{\Gamma\left(\sum_{q=1}^{Q} u_q\right)} \tag{8.3.30}$$

It is conventional to denote the distribution as

$$\text{Dirichlet}\left(\boldsymbol{\alpha}|\mathbf{u}\right) \tag{8.3.31}$$

The parameter $\mathbf{u}$ controls how strongly the mass of the distribution is pushed to the corners of the simplex. Setting $u_q = 1$ for all q corresponds to a uniform distribution, fig(8.6). In the binary case $Q = 2$, this is equivalent to a Beta distribution.

Goal

Determine c^* for novel input $\mathbf{x}^*$

Solution: MAP

$$c^* = \underset{c}{\operatorname{argmax}}\, p(c^* |\, \mathbf{x}^*, \mathcal{D})$$

$$\begin{aligned} p(c^*|\mathbf{x}^*, \mathcal{D}) &\propto p(c^*, \mathbf{x}^*, \mathcal{D}) \\ &\propto p(\mathbf{x}^*|\mathcal{D}, c^*)p(c^*|\mathcal{D}) \\ &= p(c^*|\mathcal{D}) \prod_i p(x_i^*|\mathcal{D}, c^*) \end{aligned}$$

$\mathbf{x}$			c
x_1^1	$\cdots$	x_D^1	c^1
$\vdots$	$\ddots$	$\vdots$	$\vdots$
x_1^N	$\cdots$	x_D^N	c^N

$x_i^n \in \{1, \ldots, S\}$, $c^n \in \{1, \ldots, C\}$