

## Chapter 18

# Probabilistic Methods

### 18.1 Full Bayes Classifier

Given a training dataset  $\mathcal{D}$ , with  $n$  points, in a  $d$ -dimensional space, and assuming that there are  $k$  classes, the Bayes classifier makes use of the Bayes theorem to predict the class for a new test instance,  $\mathbf{x}$ . It tries to estimate the posterior probability  $P(c_i|\mathbf{x})$  for each class  $c_i$ , and chooses the one that has the largest probability.

Recall that the Bayes theorem allows us to invert the posterior probability in terms of the likelihood and prior probability, as follows:

$$P(c_i|\mathbf{x}) = \frac{P(\mathbf{x}|c_i) \cdot P(c_i)}{P(\mathbf{x})} \quad (18.1)$$

Where  $P(\mathbf{x})$  is given as follows:

$$P(\mathbf{x}) = \sum_{j=1}^k P(\mathbf{x}|c_j) \cdot P(c_j) \quad (18.2)$$

Let  $\mathcal{D}_i$  denote the subset of points in  $\mathcal{D}$  that are labeled with class  $c_i$ , i.e.,  $\mathcal{D}_i = \{\mathbf{x}_j \in \mathcal{D} \mid \mathbf{x}_j \text{ has label } y_j = c_i\}$ . The prior probabilities can be directly estimated from the training data as follows:

$$P(c_i) = \frac{|\mathcal{D}_i|}{|\mathcal{D}|} \quad (18.3)$$

To estimate the likelihood  $P(\mathbf{x}|c_i)$ , we have to estimate the joint probability of the values in all the  $d$  dimensions  $P(x^1, x^2, \dots, x^d | c_i)$ . Assuming all dimensions are numeric, we can estimate the joint probability by assuming that each class  $c_i$  is normally distributed around some mean  $\boldsymbol{\mu}_i$ , with a corresponding covariance matrix  $\boldsymbol{\Sigma}_i$ . In a  $d$ -dimensional space  $\boldsymbol{\mu}$  is a  $d \times 1$  column vector and  $\boldsymbol{\Sigma}_i$  is a  $d \times d$  matrix. Both of these have to be directly estimated from the training data. Estimating the mean of the class  $c_i$  is straightforward:

$$\boldsymbol{\mu}_i = \frac{\sum_{\mathbf{x}_j \in \mathcal{D}_i} \mathbf{x}_j}{|\mathcal{D}_i|} \quad (18.4)$$

The covariance matrix can also be directly estimated from  $\mathcal{D}_i$ , the subset of the data with class  $c_i$ , as follows:

$$\boldsymbol{\Sigma}_i = \begin{pmatrix} \sigma_{\mathbf{x}^1 \mathbf{x}^1} & \sigma_{\mathbf{x}^1 \mathbf{x}^2} & \dots & \sigma_{\mathbf{x}^1 \mathbf{x}^d} \\ \sigma_{\mathbf{x}^2 \mathbf{x}^1} & \sigma_{\mathbf{x}^2 \mathbf{x}^2} & \dots & \sigma_{\mathbf{x}^2 \mathbf{x}^d} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{\mathbf{x}^d \mathbf{x}^1} & \sigma_{\mathbf{x}^d \mathbf{x}^2} & \dots & \sigma_{\mathbf{x}^d \mathbf{x}^d} \end{pmatrix} \quad (18.5)$$

where  $\sigma_{\mathbf{x}^a \mathbf{x}^b}$  is the covariance between dimensions  $\mathbf{x}^a$  and  $\mathbf{x}^b$  computed only from points in  $\mathcal{D}_i$ .

Once the class parameters have been estimated, we can use the multivariate normal density function to return the likelihood:

$$P(\mathbf{x}|c_i) = f(\mathbf{x}|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) = \frac{1}{\sqrt{2\pi}^d \sqrt{|\boldsymbol{\Sigma}_i|}} e^{-\frac{(\mathbf{x}-\boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}_i^{-1} (\mathbf{x}-\boldsymbol{\mu}_i)}{2}} \quad (18.6)$$

If the attributes are categorical, estimating the likelihood or the joint probability can be done by computing the fraction of times the values  $(x^1, x^2, \dots, x^d)$  co-occur, which gives us:

$$P(\mathbf{x}|c_i) = \frac{\# \text{ of times } (x^1, x^2, \dots, x^d) \text{ occurs in } \mathcal{D}_i}{|\mathcal{D}_i|} \quad (18.7)$$

For both numeric and categorical attributes it is very expensive to evaluate the joint probability. For instance for numeric attributes we have to estimate  $O(d^2)$  covariances, and as the dimensionality increases, this requires us to estimate too many parameters. Furthermore, we may not have enough data to reliably estimate these many parameters. The same problem holds for categorical attributes, since we may not have enough data to directly estimate the joint probabilities via counting.

## 18.2 Naïve Bayesian Classifier

We saw above that the full Bayes approach is fraught with estimation related problems, especially with large number of dimensions. The naïve Bayes approach makes the “naïve” assumption that attributes are all independent. This leads to a much simpler, though surprisingly effective approach in practice. The independence assumption immediately implies that the joint probability can be decomposed into a product of dimension-wise probabilities:

$$P(\mathbf{x}|c_i) = P(x^1, x^2, \dots, x^d | c_i) = \prod_{j=1}^d P(x^j | c_i) \quad (18.8)$$

For numeric data, the naïve assumption corresponds to setting all the covariances to zero in  $\boldsymbol{\Sigma}_i$ :

$$\boldsymbol{\Sigma}_i = \begin{pmatrix} \sigma_{\mathbf{x}^1 \mathbf{x}^1} & 0 & \dots & 0 \\ 0 & \sigma_{\mathbf{x}^2 \mathbf{x}^2} & \dots & 0 \\ \vdots & \vdots & \ddots & \\ 0 & 0 & \dots & \sigma_{\mathbf{x}^d \mathbf{x}^d} \end{pmatrix} = \begin{pmatrix} \sigma_{\mathbf{x}^1}^2 & 0 & \dots & 0 \\ 0 & \sigma_{\mathbf{x}^2}^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \\ 0 & 0 & \dots & \sigma_{\mathbf{x}^d}^2 \end{pmatrix} \quad (18.9)$$

Let us plug this diagonal covariance matrix into (18.6). First note that

$$|\boldsymbol{\Sigma}_i| = \det(\boldsymbol{\Sigma}_i) = \sigma_{\mathbf{x}^1}^2 \sigma_{\mathbf{x}^2}^2 \dots \sigma_{\mathbf{x}^d}^2 = \prod_{j=1}^d \sigma_{\mathbf{x}^j}^2 \quad (18.10)$$

Also, we have

$$\boldsymbol{\Sigma}_i^{-1} = \begin{pmatrix} \frac{1}{\sigma_{\mathbf{x}^1}^2} & 0 & \dots & 0 \\ 0 & \frac{1}{\sigma_{\mathbf{x}^2}^2} & \dots & 0 \\ \vdots & \vdots & \ddots & \\ 0 & 0 & \dots & \frac{1}{\sigma_{\mathbf{x}^d}^2} \end{pmatrix} \quad (18.11)$$

and thus

$$(\mathbf{x} - \boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}_i^{-1} (\mathbf{x} - \boldsymbol{\mu}_i) = \sum_{j=1}^d \frac{(x^j - \mu_i^j)^2}{\sigma_{\mathbf{x}^j}^2} \quad (18.12)$$

Plugging these into (18.6) gives us:

$$P(\mathbf{x}|c_i) = \frac{1}{\sqrt{2\pi}^d \sqrt{\prod_{j=1}^d \sigma_{\mathbf{x}^j}^2}} e^{-\frac{\sum_{j=1}^d \left( \frac{(x^j - \mu_i^j)^2}{\sigma_{\mathbf{x}^j}^2} \right)}{2}} \quad (18.13)$$

$$= \prod_{j=1}^d \left( \frac{1}{\sqrt{2\pi} \sigma_{\mathbf{x}^j}} e^{-\frac{(x^j - \mu_i^j)^2}{2\sigma_{\mathbf{x}^j}^2}} \right) \quad (18.14)$$

$$= \prod_{j=1}^d P(x^j|c_i) \quad (18.15)$$

In other words, the joint probability has been decomposed into a product of the probability along each dimension, as required by the independence assumption. We now only have  $d$  variances to estimate, and  $d$  values to estimate for the means, for a total of only  $2d$  parameters to estimate per cluster.

For categorical data, the independence assumption leads to the following direct estimate of the probability per dimension:

$$P(x^j|c_i) = \frac{\# \text{ of times value } x^j \text{ occurs in } \mathcal{D}_i}{|\mathcal{D}_i|} \quad (18.16)$$

**Naive Bayes Example:** Let us consider the dataset shown in Table 18.1. Assume that we need to classify

Id	Age	Car	Class
1	25	sports	L
2	20	vintage	H
3	25	sports	L
4	45	suv	H
5	20	sports	H
6	25	suv	H

Table 18.1: Example Dataset

the new point: (Age: 23, Car: truck). Since each attribute is independent of the other, we consider them separately. That is

$$P((23, \text{truck})|H) = P(23|H) \times P(\text{truck}|H)$$

and likewise,

$$P((23, \text{truck})|L) = P(23|L) \times P(\text{truck}|L)$$

For Age, we have  $\mathcal{D}_L = \{1, 3\}$  and  $\mathcal{D}_H = \{2, 4, 5, 6\}$ . We can estimate the mean and variance from these labeled subsets, as shown in the table below:

H	L
$\mu_H = \frac{20+45+20+25}{4} = \frac{110}{4} = 27.5$	$\mu_L = \frac{25+25}{2} = 25$
$\sigma_H = \sqrt{\frac{425}{4}} = 10.31$	$\sigma_L = \sqrt{\frac{0}{2}} = 0$

Using the univariate normal distribution, we obtain  $P(23|H) = N(23|\mu_H = 27.5, \sigma_H = 10.31) = 0.035$ , and  $P(23|L) = N(23|\mu_L = 25, \sigma_L = 0) = 0$ . Note that due to limited data we obtain  $\sigma_L = 0$ , which leads to a zero likelihood for 23 to come from class  $L$ .

For  $\text{Car}$ , which is categorical, we immediately run into a problem, since the value `truck` does not appear in the training set. We could assume that  $P(\text{truck}|H)$  and  $P(\text{truck}|L)$  are both zero. However, we desire to have some small probability of observing each values in the domain of the attribute. One simple way of obtaining non-zero probabilities is to do the *laplace correction*, i.e., to add a of count of one to the observed counts of each value for each class, as shown in the table below.

H	L
$P(\text{sports} H) = \frac{1(+1)}{4(+4)} = 2/8$	$P(\text{sports} L) = \frac{2(+1)}{2(+4)} = 3/6$
$P(\text{vintage} H) = \frac{1(+1)}{4(+4)} = 2/8$	$P(\text{vintage} L) = \frac{0(+1)}{2(+4)} = 1/6$
$P(\text{suv} H) = \frac{2(+1)}{4(+4)} = 3/8$	$P(\text{suv} L) = \frac{0(+1)}{2(+4)} = 1/6$
$P(\text{truck} H) = \frac{0(+1)}{4(+4)} = 1/8$	$P(\text{truck} L) = \frac{0(+1)}{2(+4)} = 1/6$

Note that all counts are adjusted by (+1) to obtain a non-zero probability in every case. Also, assuming that the domain of  $\text{Car}$  consists of only the four values  $\{\text{sports}, \text{vintage}, \text{suv}, \text{truck}\}$ , this means we also need to increment the denominator by  $|\text{domain}(\text{Car})| = 4$ . In other words, the probability of a given value is computed as:

$$P(v|c_i) = \frac{n_v + 1}{|\mathcal{D}_i| + |\text{domain}(c_i)|} \quad (18.17)$$

where  $n_v$  is the observed count of value  $v$  among the points in  $\mathcal{D}_i$ . Note that instead of the laplace correction, if we have some prior probability estimate for each value, we can use that too

$$P(v|c_i) = \frac{n_v + P_v}{|\mathcal{D}_i| + \sum_v P_v} \quad (18.18)$$

Using the above probabilities, we finally obtain

$$P((23, \text{truck})|H) = P(23|H) \times P(\text{truck}|H) = 0.035 \times 1/8 = 0.0044$$

$$P((23, \text{truck})|L) = P(23|L) \times P(\text{truck}|L) = 0 \times 1/6 = 0$$

We next compute

$$P(23, \text{truck}) = P((23, \text{truck})|H) \times P(H) + P((23, \text{truck})|L) \times P(L) = 0.0044 \times \frac{4}{6} + 0 \times \frac{2}{6} = 0.003$$

We then obtain the posterior probabilities as follows:

$$P(H|(23, \text{truck})) = \frac{P((23, \text{truck})|H) \times P(H)}{P(23, \text{truck})} = \frac{0.004 \times \frac{4}{6}}{0.003} = 1$$

and

$$P(L|(23, \text{truck})) = \frac{P((23, \text{truck})|L) \times P(L)}{P(23, \text{truck})} = \frac{0 \times \frac{2}{6}}{0.003} = 0$$

Thus we classify  $(23, \text{truck})$  as high risk (H).