

Bayesian Classifiers

Consider the following training data:

x	y
1	+
2	−
1	−
1	−

What should be the classification of a test example if $x = 1$? We can give a probabilistic argument that the best answer is “−”. Estimating probabilities from the training data we have:

$$\text{prob}(y = + \mid x = 1) = \frac{1}{3}, \quad \text{prob}(y = - \mid x = 1) = \frac{2}{3} \quad (1)$$

The classification $y = -$ is called *optimal Bayes*. The general case can be expressed as follows:

$$h = \arg \max_h \text{prob}(h \mid D)$$

The hypothesis h determined by this rule is called *the Optimal Bayes* hypothesis. In the example above D is $x = 1$, and there are two hypotheses: h_+ is $y = +$, and h_- is $y = -$. The Optimal Bayes hypothesis is h_- .

Using the Bayes theorem we have:

$$h = \arg \max_h \text{prob}(h \mid D) = \arg \max_h \frac{\text{prob}(D \mid h) \text{prob}(h)}{\text{prob}(D)} = \arg \max_h \text{prob}(D \mid h) \text{prob}(h)$$

The rule on the right hand side can be used when both $\text{prob}(D \mid h)$ and $\text{prob}(h)$ are known. The probability $\text{prob}(h)$ is called the *a-priori* probability of h . The hypothesis h determined by the above rule is called *a maximum a-posteriori*, or a MAP hypothesis. When only $\text{prob}(D \mid h)$ is known we can use the rule:

$$h = \arg \max_h \text{prob}(D \mid h)$$

The hypothesis h determined by the above rule is called *a maximum likelihood*, or a ML hypothesis. To summarize, there are three rules that use different probabilities to determine the right hypothesis:

$h_1 = \arg \max_h \text{prob}(h \mid D)$	h_1 is the Optimal Bayes hypothesis
$h_2 = \arg \max_h \text{prob}(D \mid h) \cdot \text{prob}(h)$	h_2 is the MAP (maximum a-posteriori) hypothesis
$h_3 = \arg \max_h \text{prob}(D \mid h)$	h_3 is the ML (maximum likelihood) hypothesis

In the table above both h_1 and h_2 are optimal. If there is a unique optimal hypothesis then $h_1 = h_2$. The hypothesis h_3 is not optimal.

From the probabilities in (1) we see that $y = -$ is the Optimal Bayes hypothesis. To compute the MAP hypothesis we need the following probabilities, computed with D standing for $x = 1$.

	$\text{prob}(D \mid h)$	$\text{prob}(h)$
$h_+ : y = +$	1	1/4
$h_- : y = -$	2/3	3/4

Since $1 \cdot 1/4 < 2/3 \cdot 3/4$ the MAP hypothesis is h_- , same as the Optimal Bayes hypothesis. From the above probabilities we also see that the ML hypothesis is h_+ .

Naive Bayesian

Typically the probabilities needed for MAP/ML are not available. To see this, consider the following simple extension of the example, where the feature vector has 3 values instead of 1.

x_1	x_2	x_3	y
1	1	1	+
2	2	2	-
1	1	2	-
1	2	1	-

What should be the classification if a test example is $x_1 = 1, x_2 = 2, x_3 = 2$?

Since the test example is not part of the training data we cannot compute the probabilities needed for MAP or ML. The **Naive Bayesian** technique allows us to compute an approximate MAP hypothesis by assuming conditional independence of feature values. Specifically, the assumption is as follows:

$$\text{prob}(x_1 = a_1, x_2 = a_2, x_3 = a_3 \mid h) = \text{prob}(x_1 = a_1 \mid h) \cdot \text{prob}(x_2 = a_2 \mid h) \cdot \text{prob}(x_3 = a_3 \mid h)$$

Therefore, following probabilities must be estimated for the two hypotheses:

	$\text{prob}(x_1 = 1 \mid h)$	$\text{prob}(x_2 = 2 \mid h)$	$\text{prob}(x_3 = 2 \mid h)$	$\text{prob}(h)$
$h_+ : y = +$	1	0	0	1/4
$h_- : y = -$	2/3	2/3	2/3	3/4

Since $1 \cdot 0 \cdot 0 \cdot 1/4 < 2/3 \cdot 2/3 \cdot 2/3 \cdot 3/4$ the MAP hypothesis is h_- . Since $1 \cdot 0 \cdot 0 < 2/3 \cdot 2/3 \cdot 2/3$ the ML hypothesis is also h_- .

The general Naive Bayesian rule is as follows. Given a test example $x = (x_1 = a_1 \dots, x_n = a_n)$. Suppose there are k possible labels v_1, \dots, v_k . Estimate the following probabilities from the training data:

$$\begin{aligned} &\text{prob}(x_1 = a_1 \mid h = v_1), \dots, \text{prob}(x_n = a_n \mid h = v_1), \text{prob}(h = v_1) \\ &\text{prob}(x_1 = a_1 \mid h = v_2), \dots, \text{prob}(x_n = a_n \mid h = v_2), \text{prob}(h = v_2) \\ &\vdots \\ &\text{prob}(x_1 = a_1 \mid h = v_k), \dots, \text{prob}(x_n = a_n \mid h = v_k), \text{prob}(h = v_k) \end{aligned}$$

Then compute the following k values:

$$\begin{aligned} q_1 &= \text{prob}(x_1 = a_1 \mid h = v_1) \cdot \dots \cdot \text{prob}(x_n = a_n \mid h = v_1) \cdot \text{prob}(h = v_1) \\ q_2 &= \text{prob}(x_1 = a_1 \mid h = v_2) \cdot \dots \cdot \text{prob}(x_n = a_n \mid h = v_2) \cdot \text{prob}(h = v_2) \\ &\vdots \\ q_k &= \text{prob}(x_1 = a_1 \mid h = v_k) \cdot \dots \cdot \text{prob}(x_n = a_n \mid h = v_k) \cdot \text{prob}(h = v_k) \end{aligned}$$

The Naive Bayesian decides on the label v_i if $q_i = \max(q_1, \dots, q_k)$.