Machine Learning: Bayesian decision theory

by Elena Battini Sönmez
İstanbul Bilgi University

based on ‘Pattern Classification’
by Duda, Hart, Stork

Bayesian decision theory :

1. Fundamental statistical approach to the problem of pattern classification
2. It assumes that:
   - the decision problem (classification) is posed in probabilistic terms (find out the most probable class), and
   - all relevant probabilities valeus are known

Prior probability:

1. The ‘state of nature’ (class) is a random variable, w:
   - $P(w_i)$ = probability of class_i
   - Having ‘c’ classes, $P(w_1)+\ldots + P(w_c)=1$

2. Decision rule based on the prior probability (in case of 2 classes):

   if $P(w_1)>P(w_2)$ then w_1 else w_2

Generally, we know something more than the prior: after some observations of samples belonging to different classes we may learn some features dominant in some classes.

Class conditional probability:

1. It is the likelihood of every class, $p(x|w_i)$
2. It is the probability to have feature 'x' in a sample of $class_i$

Ex: w1=sea bass, w2=salmon
After some observations of sea bass and salmon, we learn their likelihoods (next slide)

Class conditional probability:

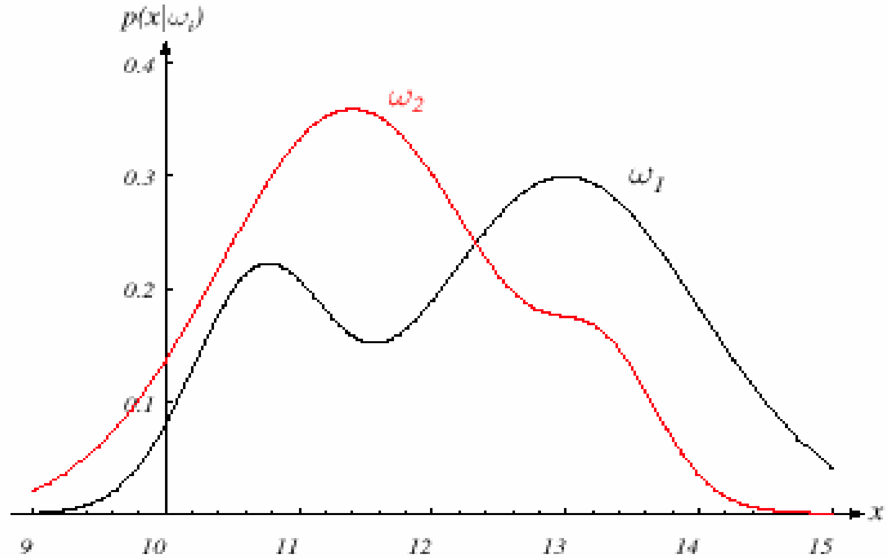

FIGURE 2.1. Hypothetical class-conditional probability density functions show the probability density of measuring a particular feature value x given the pattern is in category ω_i. If x represents the lightness of a fish, the two curves might describe the difference in lightness of populations of two types of fish. Density functions are normalized, and thus the area under each curve is 1.0. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

Bayes formula:

It defines the posterior probability, $P(w_j \mid x)$, by combining prior, $P(w_j)$, and likelihood, $p(x \mid w_j)$:

$$P(wi|x) = \frac{p(x|w_i)P(wi)}{p(x)}$$

where $p(x)$=evidence

$$\mathrm{p(x)} = \sum_{j=0}^{c} p(x|w_i)P(wi)$$

Obs: *P(w)* is a *probability mass function*, because w is a *discrete* random variable; $p(x \mid w)$ is a *probability density function*, because feature x is a *continuos* random var

Bayes formula and decision rule:

Informally: 'posterior prob = likelihood*prior'
Because the evidence is simply a scalar factor

Bayes decision rule: it is based on the posterior probability (in case of 2 classes):

$$\text{if } P(w_1 | x) > P(w_2 | x) \text{ then } w_1 \text{ else } w_2$$

Likelihood, prior and posterior probabilities:

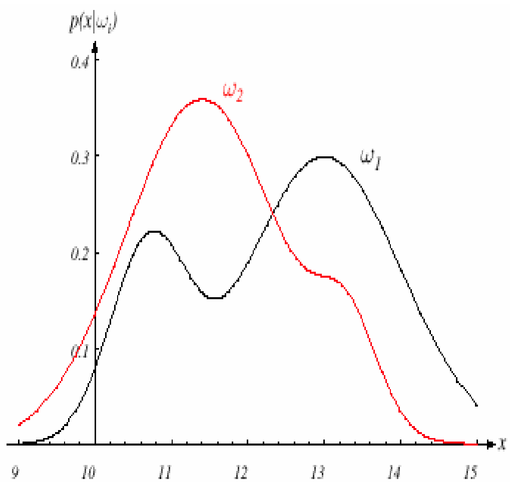

FIGURE 2.1. Hypothetical class-conditional probability density functions show the probability density of measuring a particular feature value x given the pattern is in category ω_i. If x represents the lightness of a fish, the two curves might describe the difference in lightness of populations of two types of fish. Density functions are normalized, and thus the area under each curve is 1.0. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

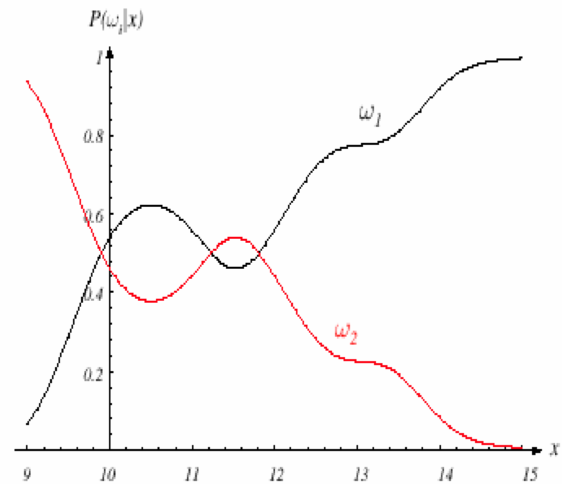

FIGURE 2.2. Posterior probabilities for the particular priors $P(\omega_1) = 2/3$ and $P(\omega_2) = 1/3$ for the class-conditional probability densities shown in Fig. 2.1. Thus in this case, given that a pattern is measured to have feature value $x = 14$, the probability it is in category ω_2 is roughly 0.08, and that it is in ω_1 is 0.92. At every x, the posteriors sum to 1.0. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

Probability of error:

$$
P(error|x) = \begin{cases} P(w1|x) \text{ if we decided } w_2 \\ P(w2|x) \text{ if we decided } w_1 \end{cases}
$$

Bayesian decision theory minimizes probability of error:

'decides w1 if $P(w_1|x)>P(w_2|x)$ otherwise decide w2'

Therefore:

$$
P(error|x)=\min \{P(w_1|x),P(w_2|x)\}
$$