Bayesian Decision Theory

Prof. Richard Zanibbi

Bayesian Decision Theory

The Basic Idea

To minimize errors, choose the least risky class, i.e. the class for which the *expected loss* is smallest

Assumptions

Problem posed in probabilistic terms, and all relevant probabilities are known

Probability Mass vs. Probability Density Functions

Probability Mass Function, P(x)

Probability for values of discrete random variable *x*. Each value has its own associated probability

$$\chi = \{v_1, \ldots, v_m\}$$

$$P(x) \geq 0, \;\text{and}\; \sum_{x \in \chi} P(x) = 1$$

Probability Density, p(x)

$$Pr[x \in (a, b)] = \int_a^b p(x)\; dx$$

$$p(x) \geq 0 \;\; \text{and} \;\; \int_{-\infty}^{\infty} p(x)\; dx = 1$$

Probability for values of continuous random variable *x*. Probability returned is for an *interval* within which the value lies (intervals defined by some unit distance)

Prior Probability

Definition (P(w))

The likelihood of a value for a random variable representing the *state of nature (true class for the current input)*, in the absence of other information

- Informally, "what percentage of the time state X occurs"

Example

The prior probability that an instance taken from two classes is provided as input, in the absence of any features (e.g. P(cat) = 0.3, P(dog) = 0.7)

Class-Conditional Probability Density Function (for Continuous Features)

Definition (p(x | w))

The probability of a value for continuous random variable x, given a state of nature w

- For each value of x, we have a different class-conditional pdf for each class in w (example next slide)

Example: Class-Conditional Probability Densities

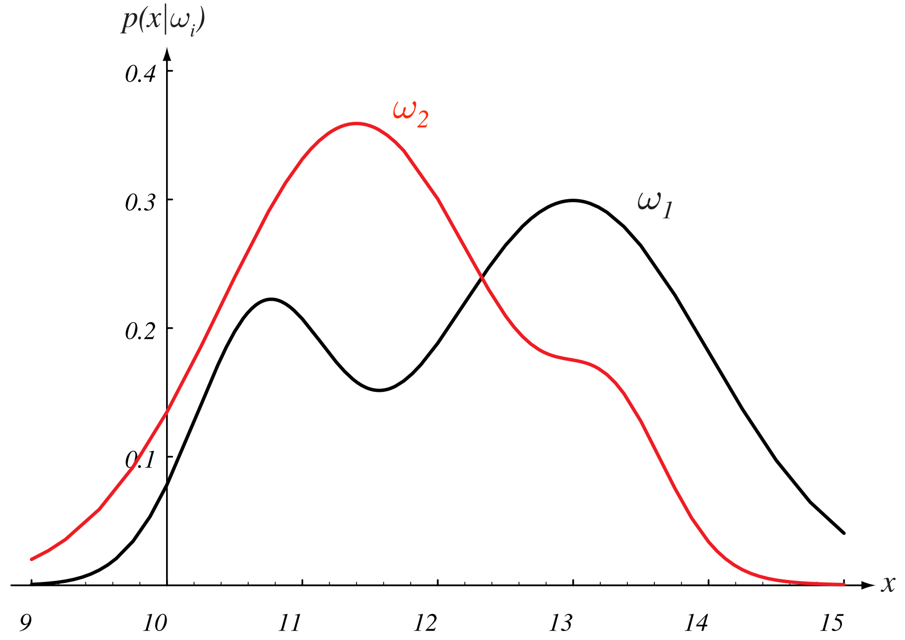

FIGURE 2.1. Hypothetical class-conditional probability density functions show the probability density of measuring a particular feature value x given the pattern is in category ω_i. If x represents the lightness of a fish, the two curves might describe the difference in lightness of populations of two 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

$$P(\omega_j|x) = \frac{p(x|\omega_j)P(w_j)}{p(x)}$$

$$\text{posterior} = \frac{\text{likelihood x prior}}{\text{evidence}}$$

$$\text{where} \;\; p(x) = \sum_{j=1}^{c} p(x|\omega_j)P(\omega_j)$$

Purpose

Convert class prior and class-conditional densities to a *posterior probability* for a class: the probability of a class given the input features ('post-observation')

Example: Posterior Probabilities

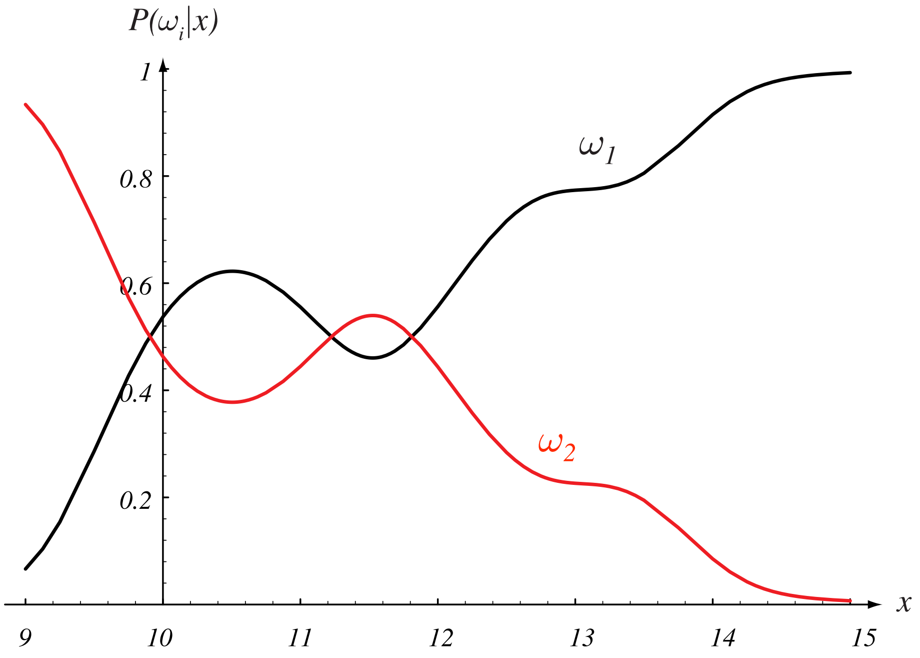

FIGURE 2.2. Posterior probabilities for the particular priors $P(\omega_1) = 2/3$ and $P(\omega_2) = 1/3$ for the class-conditional probability densities shown in Fig. 2.1. Thus in this case, given that a pattern is measured to have feature value $x = 14$, the probability it is in category ω_2 is roughly 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.

Choosing the Most Likely Class

What happens if we do the following?

Decide ω_1 if $P(\omega_1|x) > P(\omega_2|x)$; otherwise decide ω_2

A. We minimize the average probability of error. Consider the two-class case from previous slide:

$$P(error|x) = \begin{cases} P(\omega_1|x) & \text{if we choose } \omega_2 \\ P(\omega_2|x) & \text{if we choose } \omega_1 \end{cases}$$

$$P(error) = \int_{-\infty}^{\infty} P(error|x)p(x)\, dx \quad \text{(average error)}$$

Expected Loss or *Conditional Risk* of an Action

$$R(\alpha_i|\mathbf{x}) = \sum_{j=1}^{c} \lambda(\alpha_i|\omega_j)P(\omega_j|\mathbf{x})$$

Explanation

The expected (“average”) loss for taking an action (choosing a class) given an input vector, for a given *conditional loss function (lambda)*

Decision Functions and Overall Risk

$$R = \int R(\alpha(x)|x)p(x)\ dx$$

Decision Function or Decision Rule

(alpha(x)): takes on the value of exactly one action for each input vector x

Overall Risk

The expected (average) loss associated with a decision rule

Bayes Decision Rule

Idea

Minimize the overall risk, by choosing the action with the least conditional risk for input vector x

Bayes Risk (R*)

The resulting overall risk produced using this procedure. This is the best performance that can be achieved given available information.

Bayes Decision Rule: Two Category Case

Bayes Decision Rule

For each input, select class with least conditional risk, i.e. choose class one if:

$$R(\alpha_1|\mathbf{x}) < R(\alpha_2|\mathbf{x})$$

where

$$\lambda_{ij} = \lambda(\alpha_i|\omega_j)$$

$$R(\alpha_1|\mathbf{x}) = \lambda_{11}P(\omega_1|\mathbf{x}) + \lambda_{12}P(\omega_2|\mathbf{x})$$

$$R(\alpha_2|\mathbf{x}) = \lambda_{21}P(\omega_1|\mathbf{x}) + \lambda_{22}P(\omega_2|\mathbf{x})$$

Alternate Equivalent Expressions of Bayes Decision Rule (“Choose Class One If…”)

Posterior Class Probabilities

$$(\lambda_{21} - \lambda_{11}) P(\omega_1|\mathbf{x}) > (\lambda_{12} - \lambda_{22}) P(\omega_2|\mathbf{x})$$

Class Priors and Conditional Densities

Produced by applying Bayes Formula to the above, multiplying both sides by p(x)

$$(\lambda_{21} - \lambda_{11}) p(\mathbf{x}|\omega_1) P(\omega_1) > (\lambda_{12} - \lambda_{22}) p(\mathbf{x}|\omega_2) P(\omega_2)$$

Likelihood Ratio

$$\frac{p(\mathbf{x}|\omega_1)}{p(\mathbf{x}|\omega_2)} > \frac{\lambda_{12} - \lambda_{22}}{\lambda_{21} - \lambda_{11}} \frac{P(\omega_2)}{P(\omega_1)}$$

The Zero-One Loss

Definition

$$\lambda(\alpha_i|\omega_j) = \begin{cases} 0 & i = j \\ 1 & i \neq j \end{cases} \quad i, j = 1, \ldots, c$$

Conditional Risk for Zero-One Loss

$$R(\alpha_i|\mathbf{x}) = \sum_{j=1}^{c} \lambda(\alpha_i|\omega_j) P(\omega_j|\mathbf{x}) = \sum_{j \neq i} P(\omega_j|\mathbf{x}) = 1 - P(\omega_i|\mathbf{x})$$

Bayes Decision Rule (min. error rate)

$$\text{Decide } \omega_i \text{ if } P(\omega_i|\mathbf{x}) > P(\omega_j|\mathbf{x}) \quad \text{for all } j \neq i$$

Example: Likelihood Ratio

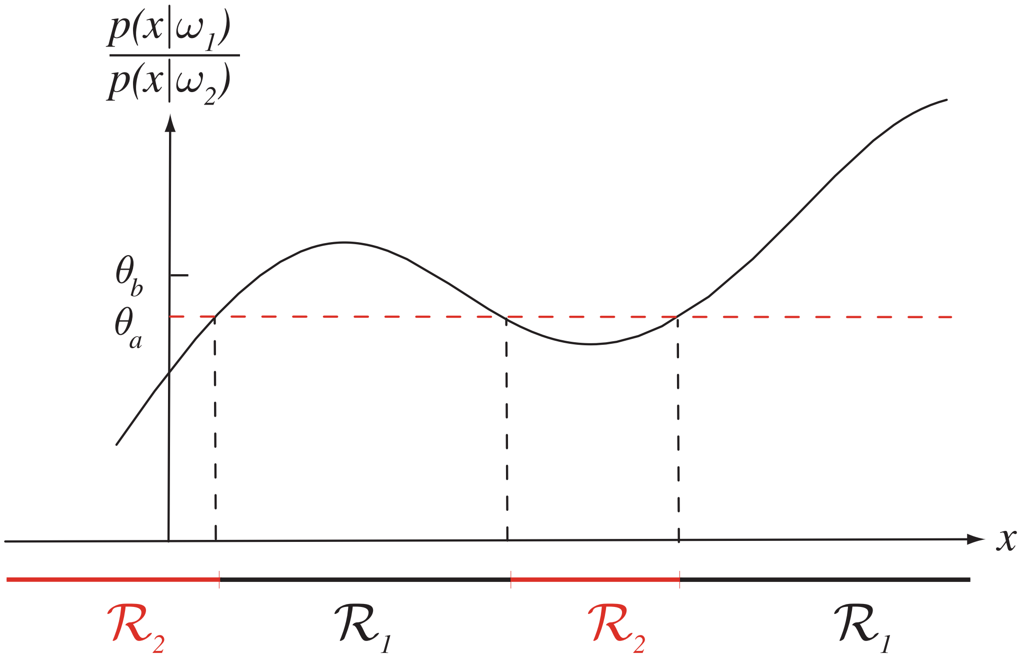

FIGURE 2.3. The likelihood ratio $p(x|\omega_1)/p(x|\omega_2)$ for the distributions shown in Fig. 2.1. If we employ a zero-one or classification loss, our decision boundaries are determined by the threshold θ_a. If our loss function penalizes miscategorizing ω_2 as ω_1 patterns more than the converse, we get the larger threshold θ_b, and hence $\mathcal{R}_1$ becomes smaller. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

Bayes Classifiers

Recall the "Canonical Model"

Decide class i if:

$$g_i(\mathbf{x}) > g_j(\mathbf{x}) \quad \text{for all } j \neq i$$

For Bayes Classifiers

Use the first discriminant def'n below for general case, second for zero-one loss

$$g_i(\mathbf{x}) = -R(\alpha_i|\mathbf{x})$$
$$g_i(\mathbf{x}) = P(\omega_i|\mathbf{x})$$

Equivalent Discriminants for Zero-One Loss (Minimum-Error-Rate)

Trade-off

Simplicity of understanding vs. computation

$$g_i(\mathbf{x}) = P(\omega_i|\mathbf{x}) = \frac{p(\mathbf{x}|\omega_i)P(\omega_i)}{\sum_{j=1}^{c} p(\mathbf{x}|\omega_j)P(\omega_j)}$$

$$g_i(\mathbf{x}) = p(\mathbf{x}|\omega_i)P(\omega_i)$$

$$g_i(\mathbf{x}) = \ln p(\mathbf{x}|\omega_i) + \ln P(\omega_i)$$

Discriminants for Two Categories

For Two Categories

We can use a single discriminant function, with decision rule: choose class one if the discriminant returns a value > 0.

Example: Zero-One Loss

$$g(\mathbf{x}) = P(\omega_1|\mathbf{x}) - P(\omega_2|\mathbf{x})$$

$$g(\mathbf{x}) = \ln \frac{p(\mathbf{x}|\omega_1)}{p(\mathbf{x}|\omega_2)} + \ln \frac{P(\omega_1)}{P(\omega_2)}$$

Example: Decision Regions for Binary Classifier

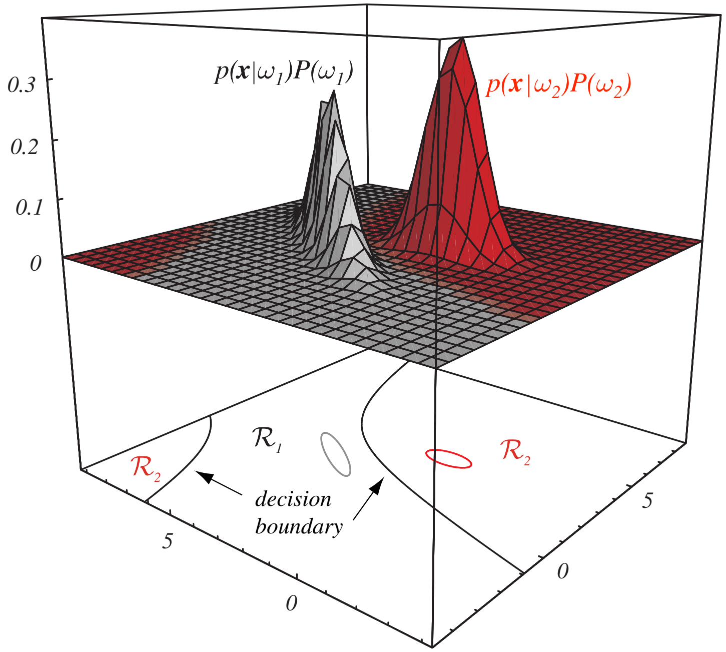

FIGURE 2.6. In this two-dimensional two-category classifier, the probability densities are Gaussian, the decision boundary consists of two hyperbolas, and thus the decision region $\mathcal{R}_2$ is not simply connected. The ellipses mark where the density is $1/e$ times that at the peak of the distribution. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

The (Univariate) Normal Distribution

Why are Gaussians so Useful?

They represent many probability distributions in nature quite accurately. In our case, when patterns can be represented as random variations of an ideal prototype (represented by the mean feature vector)

- Everyday examples: height, weight of a population

Univariate Normal Distribution

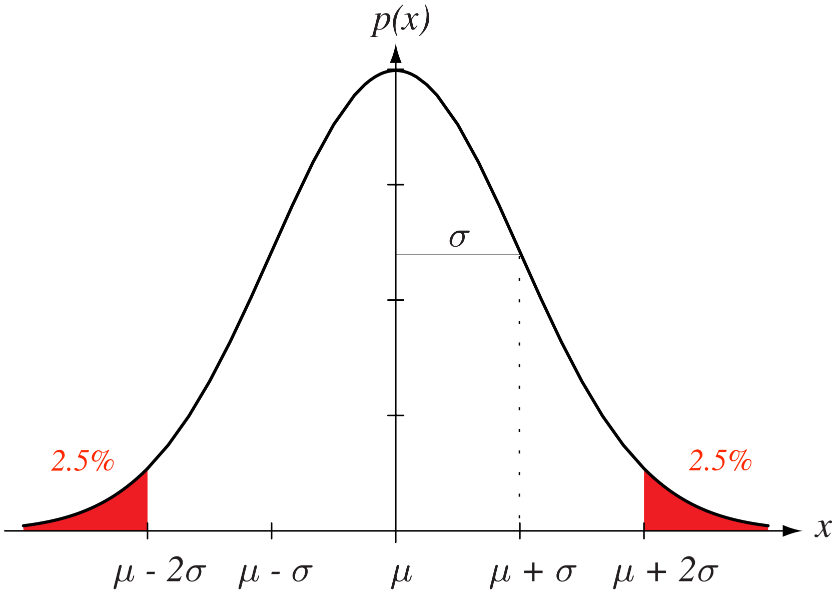

FIGURE 2.7. A univariate normal distribution has roughly 95% of its area in the range $|x - \mu| \leq 2\sigma$, as shown. The peak of the distribution has value $p(\mu) = 1/\sqrt{2\pi}\sigma$. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

Formal Definition

Peak of the Distribution (the mean)

Has value: $\frac{1}{\sqrt{2\pi}\sigma}$

Definition for Univariate Normal

$$P(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]$$

Def. for mean, variance

$$\mu = \int_{-\infty}^{\infty} x \; p(x) \; dx$$

$$\sigma^2 = \int_{-\infty}^{\infty} (x-\mu)^2 p(x) \; dx$$

Multivariate Normal Density

Informal Definition

A normal distribution over two or more variables (*d* variables/dimensions)

Formal Definition

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{d/2}|\mathbf{\Sigma}|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x}-\mu)^t\mathbf{\Sigma}^{-1}(\mathbf{x}-\mu)\right]$$

$$\mu = \int_{-\infty}^{\infty} \mathbf{x}\ p(\mathbf{x})\ d\mathbf{x}$$

$$\mathbf{\Sigma} = \int (\mathbf{x}-\mu)(\mathbf{x}-\mu)^t p(\mathbf{x}) d\mathbf{x}$$