Bayesian Decision Making

Introduction

- Classification different from regression in the sense that the output will be a discrete label denoting the entity of the class.

- We will study the decision making process, by relying on probabilistic inferences.

- The Bayes Theorem is very powerful, and knowledge of it helps to design the classifier, with appropriate decision surfaces.

Bayes Decision

- It is the decision making process when all underlying probability distributions are known
- Generative model
- It is optimal given the distributions are known.
- In this discussion, we assume supervised learning paradigm.

Problem

- 2 type of fish -sea bass and salmon in a conveyer belt
- Problem: Need to recognize the type of fish.
- 2 class/ category problem

Prior probability

Let us denote the 2 classes by

- ω_1 : Salmon
- ω_2 : Sea bass

- Let us say that salmon occurs more likely than sea bass. (This information is known as the prior and is generally estimated by experimentation.)

Mathematically , $P(\omega_1) > P(\omega_2)$

Estimation of prior probabilities by frequentist approach

- Assume that out of 8000 fish used for training, we observed that 6000 were salmon, 2000 sea bass

- Accordingly we have

$P(\omega_1) = 0.75$ prior probability of salmon

$P(\omega_2) = 0.25$ prior probability of sea bass

Simple Decision making based on prior knowledge

Assign unknown fish as salmon ω_1, if

$$P(\omega_1) > P(\omega_2)$$

else

assign it to sea bass ω_2.

Issues with using only prior knowledge

- Decision rule is flawed. Salmon will always be favored and assigned to a test fish.

- Such an approach will not work in practice.

- Sole dependence on prior probability for making decisions is not a good idea

Need for features

- **Solution**: Look for additional information to describe the sea bass and salmon.

- Key idea is to look for discriminative features.

- Features like length, width, color, life span, texture may describe salmon and sea bass.

Feature vector

- Assume that we have *d* features.

- Let set of features describing a fish be represented by a *d* dimensional feature vector **x**

Class Conditional Density

- For each class/ category of fish, we can associate the *d* features to come from a probability distribution function.
- This pdf is referred to as 'class conditional density'.
- The nature of the features are continuous.
- Note that, for a set of *d* features describing the fish, we work on a *d* dimensional probability distribution.

- For the time being , let us work on how to improve our decision process by incorporating a single feature x.

- Later, we extend the framework for *d* dimensional features, and also for classes greater than 2.

Class conditional Density

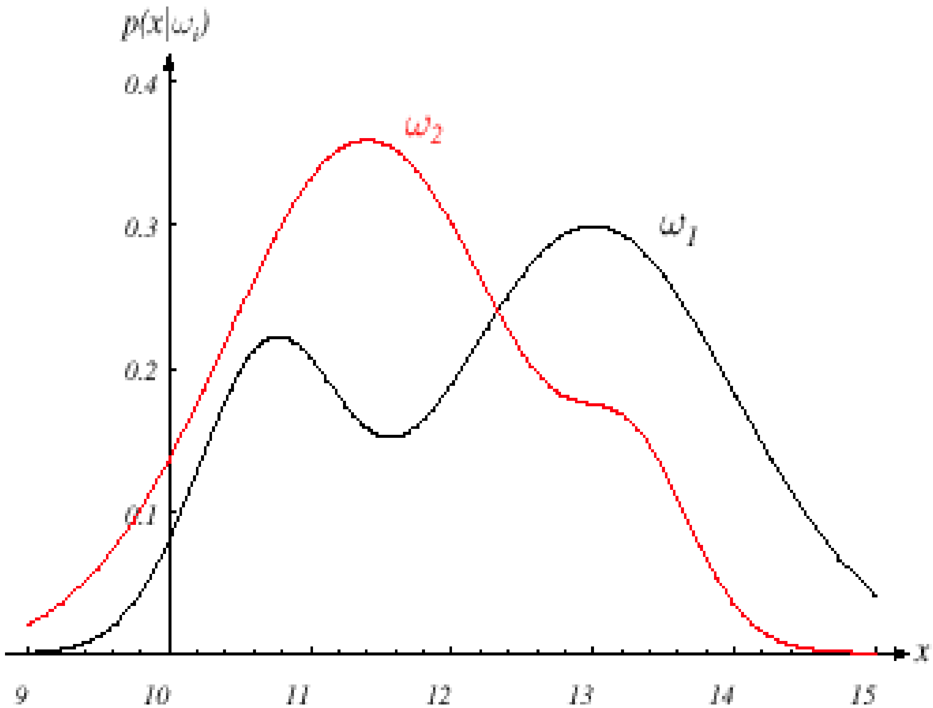

FIGURE 2.1. Hypothetical class-conditional probability density functions show the probability density of measuring a particular feature value x given the pattern is in category ω_i. If x represents the lightness of a fish, the two curves might describe the difference in lightness of populations of two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

- Bayes Theorem :

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

$$posterior = \frac{likelihood \times prior}{evidence}$$

In the case of two categories

$$P(x) = \sum_{j=1}^{j=2} P(x \mid \omega_j)P(\omega_j)$$

Posterior probability plot for the two classes

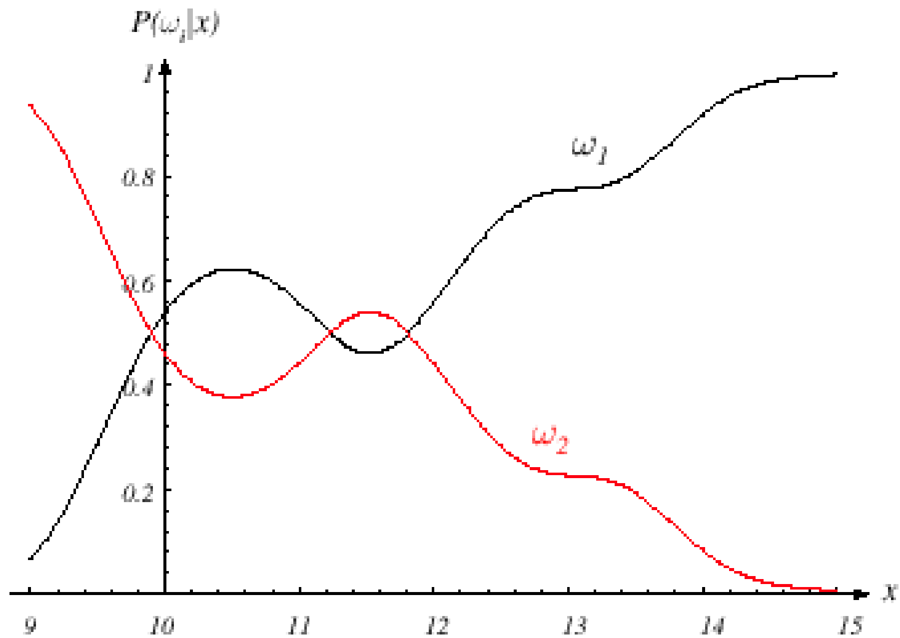

FIGURE 2.2. Posterior probabilities for the particular priors $P(\omega_1) = 2/3$ and $P(\omega_2) = 1/3$ for the class-conditional probability densities shown in Fig. 2.1. Thus in this case, given that a pattern is measured to have feature value $x = 14$, the probability it is in category ω_2 is roughly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

Decision based on posterior probabilities

- Decision given the posterior probabilities

x is an observation for which:

if $P(\omega_1 \mid x) > P(\omega_2 \mid x)$ True state of nature $= \omega_1$

if $P(\omega_1 \mid x) < P(\omega_2 \mid x)$ True state of nature $= \omega_2$

Therefore: whenever we observe a particular x, the probability of error is :

$$P(error \mid x) = P(\omega_1 \mid x) \text{ if we decide } \omega_2$$

$$P(error \mid x) = P(\omega_2 \mid x) \text{ if we decide } \omega_1$$

Decision based on posterior probabilities

- Minimizing the probability of error

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

Therefore:

$$P(error \mid x) = \min\,[P(\omega_1 \mid x), P(\omega_2 \mid x)]$$

(Bayes decision)

Probability of error

$$P(error) = \int_{-\infty}^{\infty} P(error, x)dx = \int_{-\infty}^{\infty} P(error \mid x)\, p(x)dx$$

- We want $P(error \mid x)$ to be as small as possible for every value of x

The Bayes classifier scheme strives to achieve that

Bayesian classification framework for high dimensional features and more classes

Let $\{\omega_1, \omega_2, \ldots, \omega_C\}$ be the set of "*C*" states of nature (or "categories" / "classes")

Assume , that for an unknown pattern, a *d* dimensional feature vector **x** is constructed:

From Bayes rule

$$P(\omega_j \mid \mathbf{x}) = \frac{p(\mathbf{x} \mid \omega_j)P(\omega_j)}{P(\mathbf{x})}$$

We compute the posterior probability of the pattern with respect to each of the "*c*" classes.

In the decision making step, we assign the pattern to the class for which the posterior probability is greatest.

Bayesian classification framework for high dimensional features and more classes

$$\omega_{test} = \arg\max_j P(\omega_j \mid \mathbf{x}) = \frac{p(\mathbf{x} \mid \omega_j)P(\omega_j)}{P(\mathbf{x})} \quad j = 1,2\ldots.C$$

$$P(\mathbf{x}) = \sum_{j=1}^{C} p(\mathbf{x} \mid \omega_j)P(\omega_j)$$

Evidence acts as a normalization factor ….term same for all Classes.

ω_{test} is the class for which the posterior probability is highest.

The pattern is assigned to this class.

Risk minimization framework

Let $\{\omega_1, \omega_2, \ldots, \omega_C\}$ be the set of "C" states of nature (or "categories")

Let $\{\alpha_1, \alpha_2, \ldots, \alpha_a\}$ be the set of possible "a" actions

Let $\lambda(\alpha_i / \omega_j)$ be the loss incurred for taking action α_i when the state of nature is ω_j

Risk minimization framework

The expected loss: $$R(\alpha_i) = \sum_{j=1}^{C} \lambda(\alpha_i \mid \omega_j) P(\omega_j)$$

Given an observation with vector **x**, the conditional risk is:

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

At every **x**, a decision is made: $\alpha(x)$, by minimizing the expected loss.

Our final goal is to minimize the total risk over all **x**.

$$\int R(\alpha(\mathbf{x}) \mid \mathbf{x}) p(\mathbf{x}) d\mathbf{x}$$

Reading

- Sections 2.1, 2.2 :

  Duda, Hart , Stork : Pattern Classification