Lecture 5: Bayes Classification

Richa Singh

Slides are prepared from several information sources including Duda, Hart, Stork

Probability

- Conditional probability of A given B:

$$P(A/B) = \frac{P(A,B)}{P(B)}$$

- Deriving chain rule from above:

$$P(A,B) = P(A/B)P(B) = P(B/A)P(A)$$

Bayes Theorem

$$P(A/B) = \frac{P(B/A)P(A)}{P(B)}$$

$$P(B) = P(B, A) + P(B, \bar{A}) = P(B/A)P(A) + P(B/\bar{A})P(\bar{A})$$

Bayes Theorem

$$P(A/B) = \frac{P(B/A)P(A)}{P(B)}$$

$$P(B) = P(B,A) + P(B,\bar{A}) = P(B/A)P(A) + P(B/\bar{A})P(\bar{A})$$

$$P(Disease/Symptom) = \frac{P(Symptom/Disease)P(Disease)}{P(Symptom)}$$

$$P(Symptom) = P(Symptom/Disease)P(Disease) +$$
$$P(Symptom/NoDisease)P(NoDisease)$$

Bayes Classification

An Example

- "Sorting incoming fish on a conveyor according to species using optical sensing"

Sea bass

Species

Salmon

Let us build a machine learning system that classifies between Sea Bass and Salmon

Fish Classification: Salmon vs. Sea Bass

- Set up a camera and take some sample images
- Preprocessing involves image enhancement and segmentation;
  - separate touching or occluding fishes and
  - extract fish contour

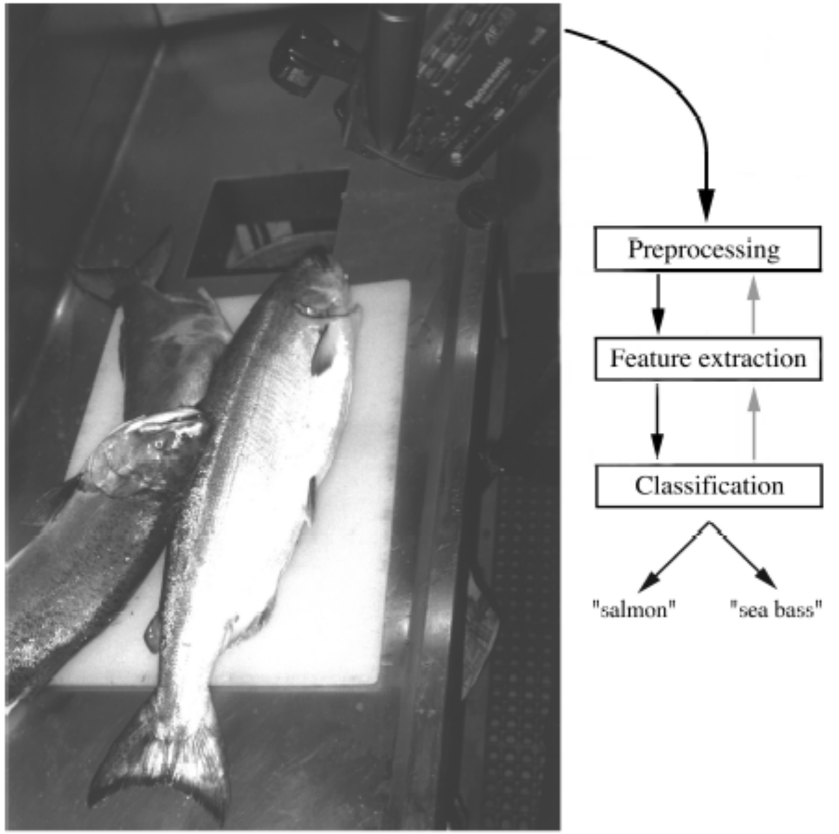


FIGURE 1.1. The objects to be classified are first sensed by a transducer (camera), whose signals are preprocessed. Next the features are extracted and finally the classification is emitted, here either "salmon" or "sea bass." Although the information flow is often chosen to be from the source to the classifier, some systems employ information flow in which earlier levels of processing can be altered based on the tentative or preliminary response in later levels (gray arrows). Yet others combine two or more stages into a unified step, such as simultaneous segmentation and feature extraction. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

State of Nature/Prior

- Prior probabilities reflect domain expert's knowledge of *how likely it is that each type of fish will appear,* before we actually see it.
  - State of nature is a random variable: $P(\omega_1), P(\omega_2)$
  - Uniform priors: The catch of salmon and sea bass is equiprobable ($P(\omega_1) = P(\omega_2)$)
  - $P(\omega_1) + P(\omega_2) = 1$ (exclusivity and exhaustivity)

Problem Analysis

- Extract features from the images
  - Length
  - Lightness
  - Width
  - Number and shape of fins
  - Position of the mouth, etc…
- This is the set of all suggested features to explore for use in our classifier

Representation: Fish Length as Feature

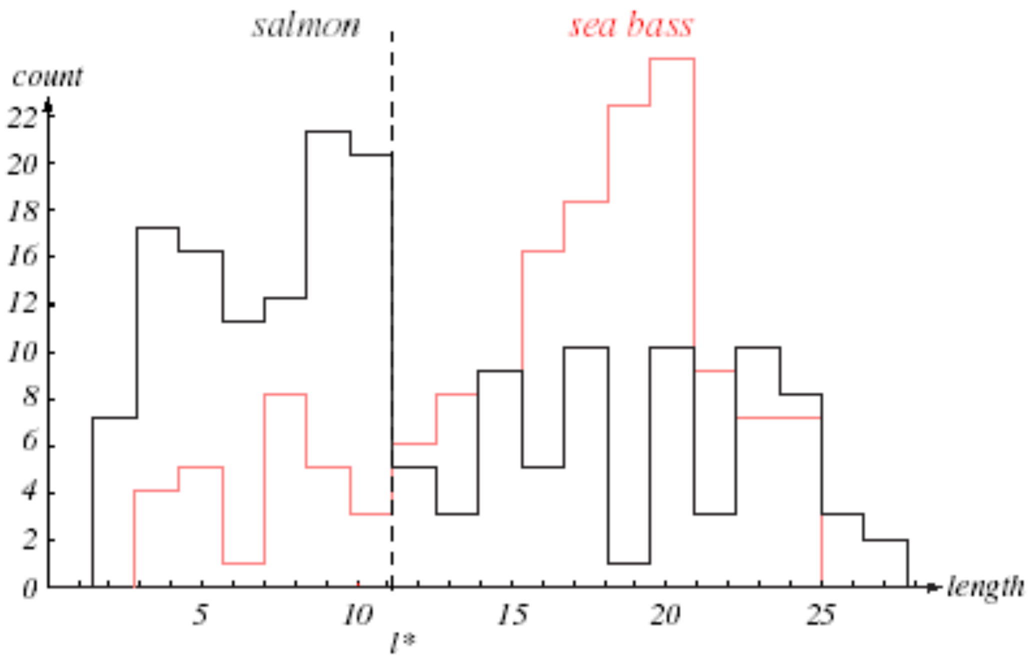


FIGURE 1.2. Histograms for the length feature for the two categories. No single threshold value of the length will serve to unambiguously discriminate between the two categories; using length alone, we will have some errors. The value marked l^* will lead to the smallest number of errors, on average. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

Training Samples

Class-conditional Probabilities

- Use of the class–conditional information
- $P(x|\omega 1)$ and $P(x|\omega 2)$ describe the difference in feature (length or lightness) between the populations of sea-bass and salmon

Class-conditional PDF

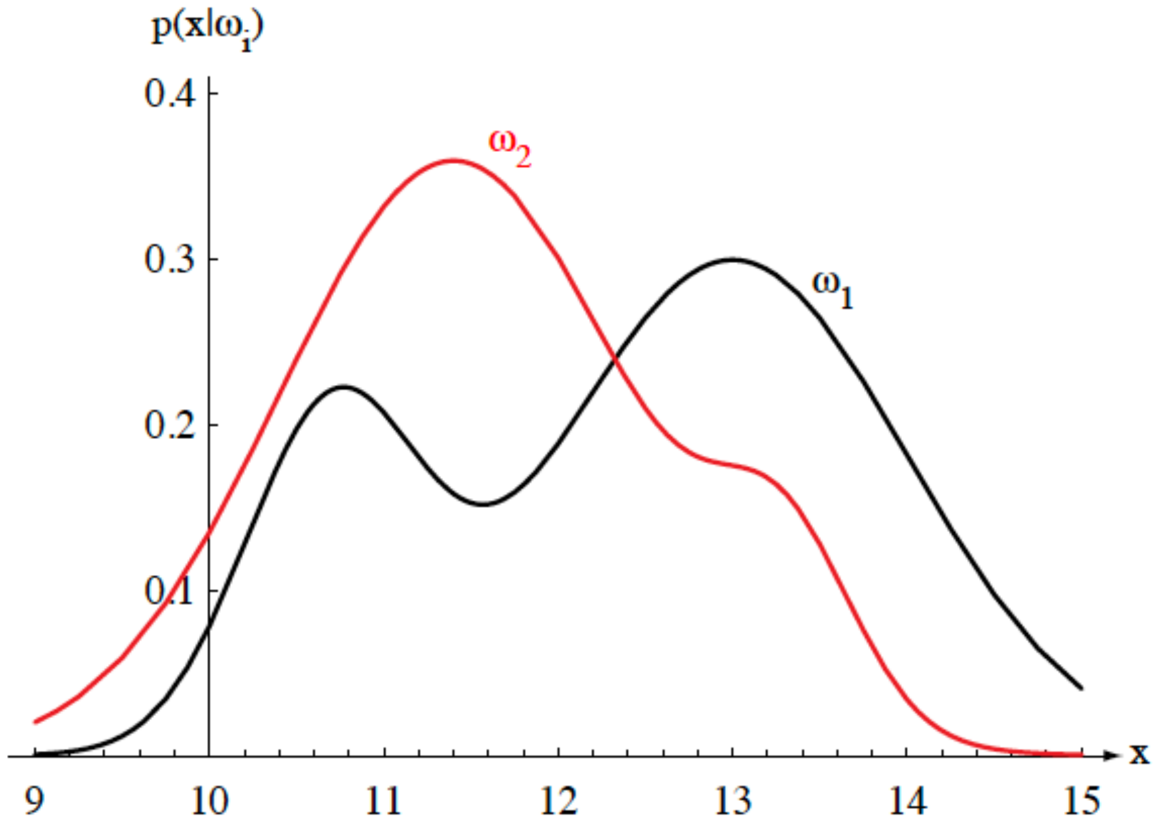


Figure 2.1: Hypothetical class-conditional probability density functions show the probability density of measuring a particular feature value x given the pattern is in category ω_i. If x represents the length of a fish, the two curves might describe the difference in length of populations of two types of fish. Density functions are normalized, and thus the area under each curve is 1.0.

The Bayes Classifier

- Use Bayes Rule!

Posterior

Likelihood

Prior

$$\underbrace{P(Y|X_1,\ldots,X_n)}_{\text{Posterior}} = \frac{\overbrace{P(X_1,\ldots,X_n|Y)}^{\text{Likelihood}}\overbrace{P(Y)}^{\text{Prior}}}{\underbrace{P(X_1,\ldots,X_n)}_{\text{Normalization Constant}}}$$

Normalization
Constant

Bayes' Classification

- Posterior, likelihood, prior, evidence

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

  - Evidence: In case of two categories

$$\boxed{p(x) = \sum_{j=1}^{2} p(x|\omega_j)P(\omega_j)}$$

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

Posterior Probabilities

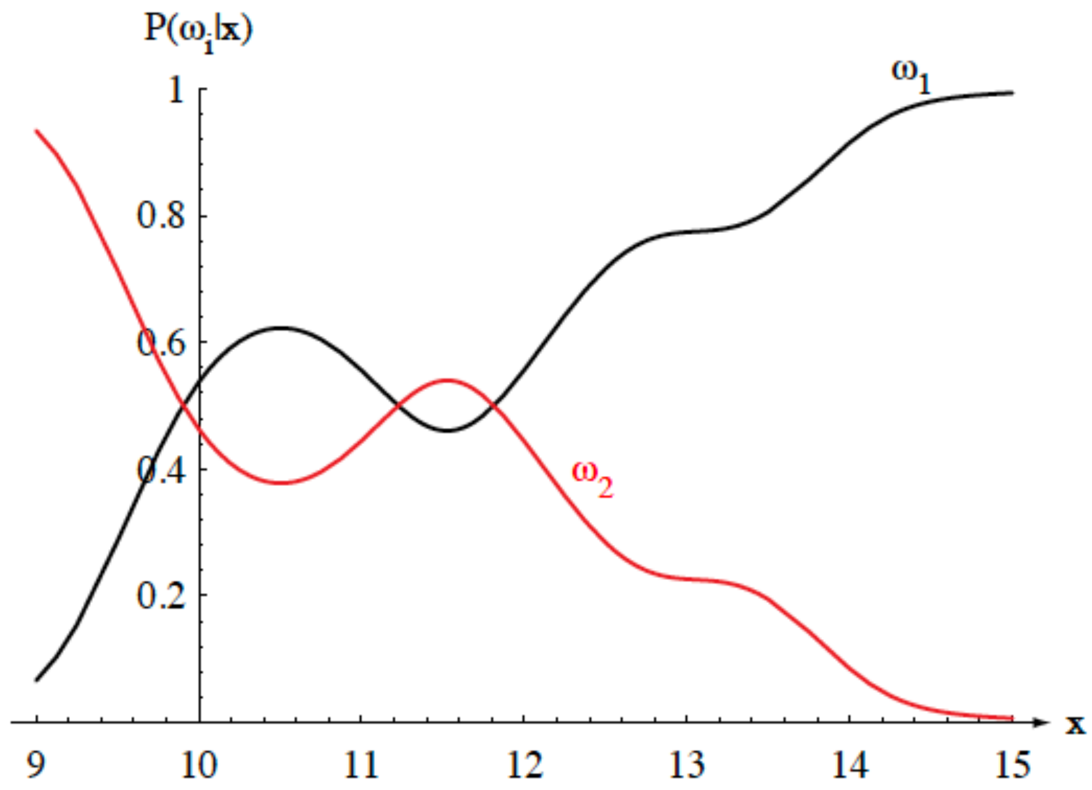


Figure 2.2: Posterior probabilities for the particular priors $P(\omega_1) = 2/3$ and $P(\omega_2) = 1/3$ for the class-conditional probability densities shown in Fig. 2.1. Thus in this case, given that a pattern is measured to have feature value $x = 14$, the probability it is in category ω_2 is roughly 0.08, and that it is in ω_1 is 0.92. At every x, the posteriors sum to 1.0.

Bayes' Decision

- Decision given the posterior probabilities

$$\text{Decide } \omega_1 \text{ if } P(\omega_1|x) > P(\omega_2|x); \text{ otherwise decide } \omega_2,$$

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

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

Questions

Review

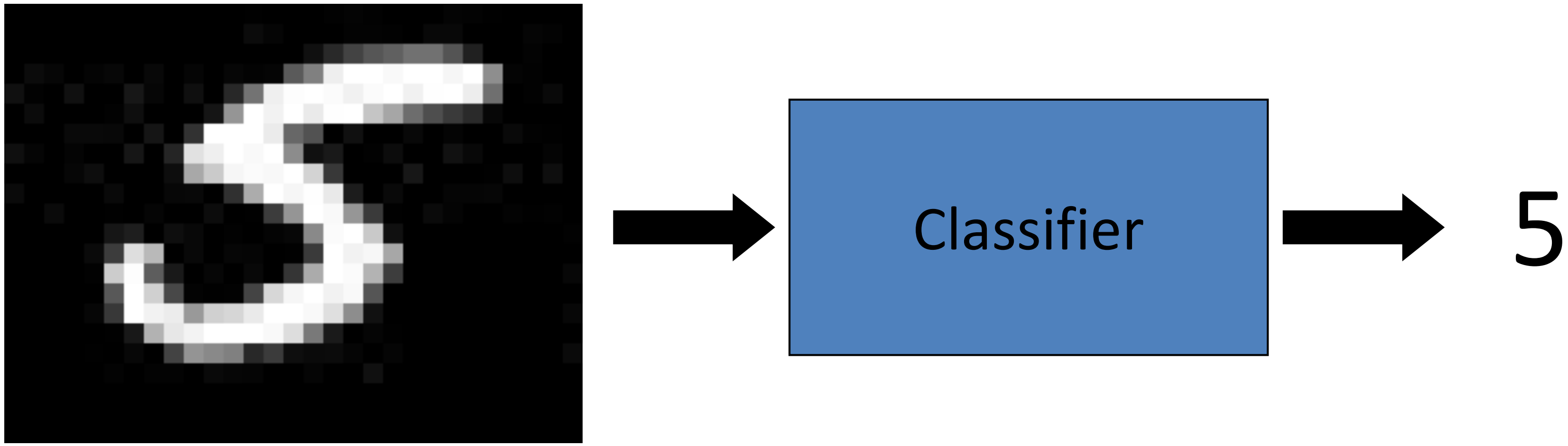
- Classification based on a single feature
- Two class classification
- Sample is assigned to one of the two classes
- The cost of making a false accept or a false reject is same

Bayesian Decision Theory

- Generalization of the preceding ideas
  - Use of more than one feature
  - Use more than two states of nature
  - Allowing actions other than decide on the state of nature
    - Allowing actions other than classification primarily allows the possibility of rejection
    - Refusing to make a decision in close or bad cases!
  - Introduce a loss function which is more general than the probability of error
    - The loss function states how costly each action taken is

Another Application

- **Digit Recognition**



- $X_1,\ldots,X_n \in \{0,1\}$ (Black vs. White pixels)
- $Y \in \{5,6\}$ (predict whether a digit is a 5 or a 6)

The Bayes Classifier

- Let's expand this for our digit recognition task:

$$
\begin{aligned}
P(Y=5|X_1,\ldots,X_n) &= \frac{P(X_1,\ldots,X_n|Y=5)P(Y=5)}{P(X_1,\ldots,X_n|Y=5)P(Y=5)+P(X_1,\ldots,X_n|Y=6)P(Y=6)} \\
P(Y=6|X_1,\ldots,X_n) &= \frac{P(X_1,\ldots,X_n|Y=6)P(Y=6)}{P(X_1,\ldots,X_n|Y=5)P(Y=5)+P(X_1,\ldots,X_n|Y=6)P(Y=6)}
\end{aligned}
$$

- To classify, we will simply compute these two probabilities and predict based on which one is greater

Model Parameters

- For the Bayes classifier, we need to "learn" two functions, the likelihood and the prior

- How many parameters are required to specify the prior for our digit recognition example?

- How many parameters are required to specify the likelihood?
  - (Supposing that each image is 30x30 pixels)

- # of parameters for modeling $P(X_1,\ldots,X_n|Y)$:
  - $2(2^n-1)$

Just 1

Model Parameters

- The problem with explicitly modeling $P(X_1,\ldots,X_n|Y)$ is that there are usually way too many parameters:
  - We' ll run out of space
  - We' ll run out of time
  - And we' ll need tons of training data (which is usually not available)

The Naïve Bayes Model

- The *Naïve Bayes Assumption*: Assume that all features are independent **given the class label Y**

- Equationally speaking:

$$P(X_1, \ldots, X_n|Y) = \prod_{i=1}^{n} P(X_i|Y)$$