# CS 559 Machine Learning
## Bayesian Decision Theory

Yue Ning

Department of Computer Science
Stevens Institute of Technology

# Review of last lecture

- Linear regression; Two special cases:
  1. Ridge regression (L2 regularization)
  2. Lasso regression (L1 regularization)
- Gradient descent algorithms
  1. BGD
  2. SGD
  3. mini-batch GD;
- Features: non-monotonicity, saturation, interactions between features;
- Maximum Likelihood estimator;
- Model selection: under-fitting, over-fitting, bias-variance;
- The curse of dimensionality

# Book: Pattern Classification

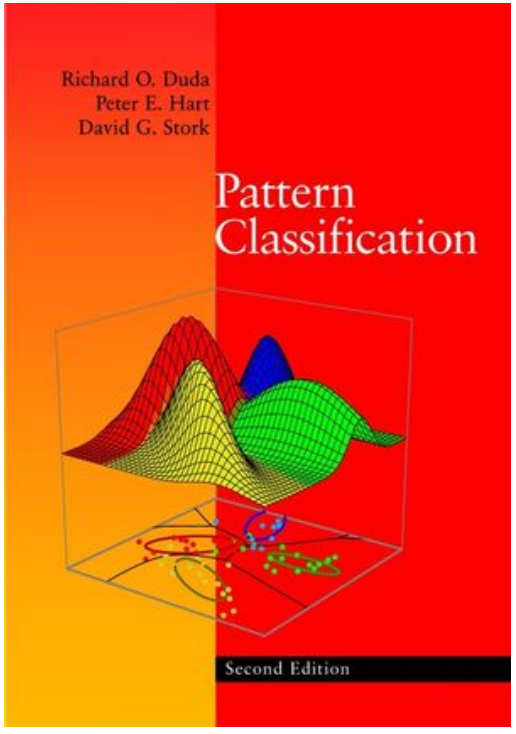


Figure: Pattern Classification by Richard O. Duda, Peter E. Hart, David G. Stork (DHS)

# Outline

# Learning Objectives

1. Understand Bayesian Decision Theory including posterior, likelihood and prior.

# Introduction

# Why and what is Bayesian decision?

- Bayesian approach: provide a mathematical rule explaining how you should change your existing beliefs in the light of new evidence.
- It allows the scientist to combine new data with their existing knowledge.
- Bayesian decision theory uses Bayes approach to analysis the problem of pattern classification.
- Quantify the trade-offs between various decisions using probability and the cost that accompany such decisions.

Assumption:

- Decision problem is posed in **probabilistic** terms.
- All of the relevant probabilities are **known**.

# Fish Example

Figure: From Internet

- Classify fish as either Salmon or Sea Bass.
- Random variable $y$ describe the fish category. (State of nature)
  - $y = y_1$: Sea Bass
  - $y = y_2$: Salmon
- Only two fish categories.

# Prior Probability

- The Prior probability reflects our prior knowledge of how likely we expect an outcome of an event **before** we actually observed such event.
- For fish example, represents how likely we are to get a sea bass or salmon before we see the next fish on the conveyor belt.
- Prior comes from prior knowledge, **NO** data have been seen yet.
- Prior might be different depending on the situation.
- If you have reliable prior knowledge, USE IT!

# Decision Rule based on ONLY Prior

- $P(y = y_1)$ or $P(y_1)$ for prior "next is sea bass".
- $P(y = y_2)$, or $P(y_2)$ for prior "next is salmon".
- $P(y_1) + P(y_2) = 1$: either $y_1$ or $y_2$ must occur.
- A decision rule prescribes what actions to take based on observed data.
- Assume **only prior** available and **equal cost** for incorrect classifications.
  - Decide $y_1$ if $P(y_1) > P(y_2)$
  - Otherwise, decide $y_2$

Limitation: Always choose the same. If the prior is uniform (e.g., $P(y_1) = P(y_2) = 0.5$), such rule behaves not well.

# Class Conditional Density

- Use class-conditional information could improve accuracy.
- A **feature** is an observable variable, e.g., lightness, length, width, etc.
- Class Conditional Density: probability density function for $x$, the feature, given the state of nature is $y$, i.e.,$p(x|y)$
- E.g., $p(x|y_1), p(x|y_2)$ describe the difference in lightness between populations of sea bass and salmon

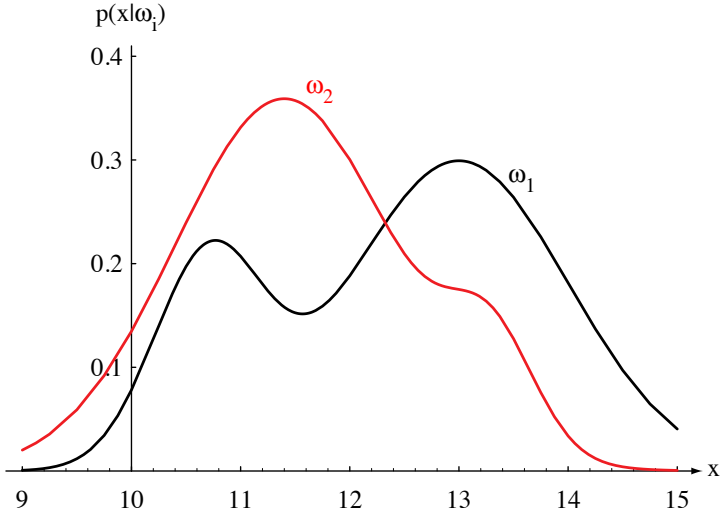


Figure: Class conditional probability [DHS book chapter 2]

# Posterior Probability

- If we know prior $P(y)$ and conditional density $p(x|y)$, as well as observed feature value $x$ (e.g., lightness of the fish), how does that affect our decision?
- Posterior probability: the probability of a certain state of nature $y$ given our observables feature $x$: $P(y|x)$
- Bayes rule:

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

$$p(x) = \sum_{i=1}^{2} p(x|y_i)P(y_i)$$

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

# Posterior Probability

- Posterior is determined by prior and likelihood.
- Example: when $P(y_1) = \frac{2}{3}$, $P(y_2) = \frac{1}{3}$

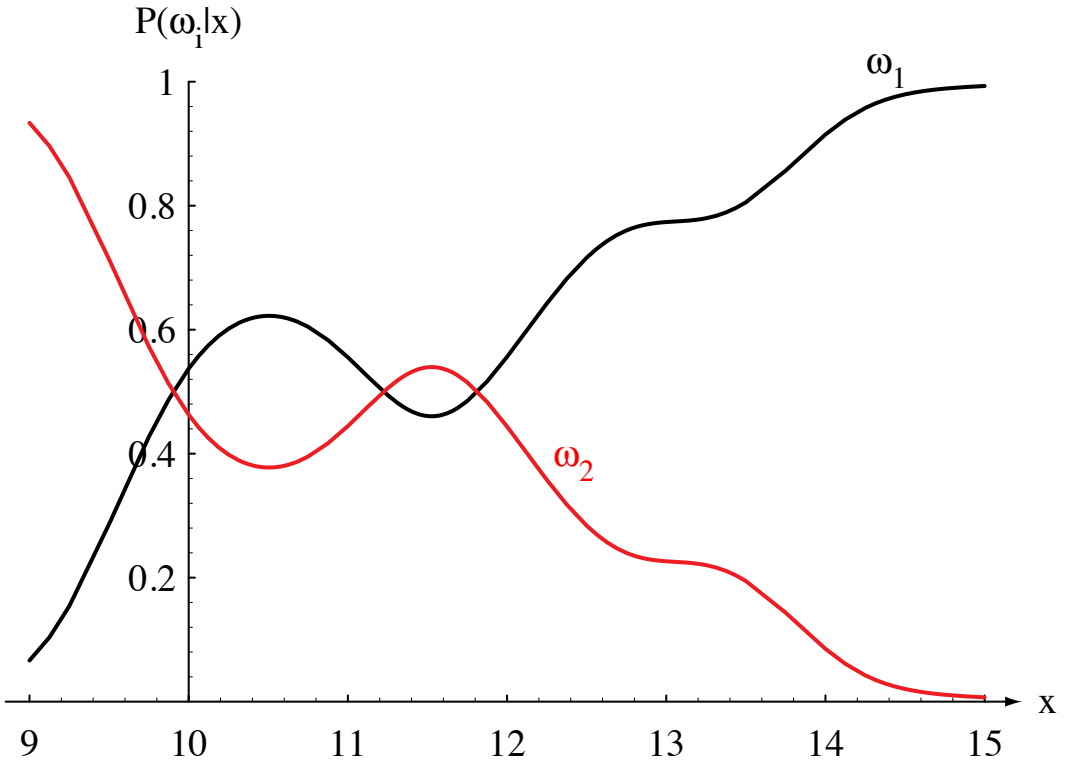


Figure: Posterior probability [DHS book chapter 2]

# Decision Rule based on Posterior

- Given observation $x$, the decision is based on posterior probability.
  - Decide $y_1$, if $P(y_1|x) > P(y_2|x)$
  - Decide $y_2$, if $P(y_2|x) > P(y_1|x)$
- Probability of error: for two class scenario, whenever we observe a particular $x$,

$$P(error|x) = \begin{cases} P(y_1|x), \text{if decide } y_2 \text{ but } y_1 \text{ is true} \\ P(y_2|x), \text{if decide } y_1 \text{ but } y_2 \text{ is true} \end{cases}$$

# Minimize the Probability of Error

- Minimize the probability of error.
- Decide $y_1$ if $P(y_1|x) > P(y_2|x)$; otherwise decide $y_2$.
- $P(error|x) = \min[P(y_1|x), P(y_2|x)]$
- Also minimize the average probability of error:

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

# Bayesian Decision Rule

- Decide $y_1$ if $P(y_1|x) > P(y_2|x)$; otherwise decide $y_2$.
- (Equivalent):
  Decide $y_1$, if $p(x|y_1)P(y_1) > p(x|y_2)P(y_2)$
  - *evidence* $p(x)$: unimportant for making a decision.
  - If for some $x$, we have $p(x|y_1) = p(x|y_2) \rightarrow$ decision rely on prior.
  - If have uniform prior$\rightarrow$ decision rely on likelihood.
- Assumption: equal cost for each decision.
- Summary: Given both prior and likelihoods, Bayesian decision rule combines them (through posterior probability) for decision making which achieves minimum probability of error.

# Bayesian Decision Theory

# Bayesian Decision Theory-Continuous Feature

Generalize the previous fish example in several ways:

- allow the use of more than one feature. (length, weight etc)
- allow more than two states of nature. (tilapia, sardine etc)
- allow actions other than deciding the state of nature. (Not make a decision)
- introduce loss function more general than the probability of error. (some classification mistakes are more costly than others)

# Notation

- feature vector $\mathbf{x} = (x_1, x_2, ..., x_d) \in R^d$: allow use of more than one feature.
- $y_1, y_2, ..., y_c$: finite set of $c$ states of nature, i.e., categories.
- $\alpha_1, \alpha_2, ..., \alpha_a$: a finite set of possible actions.
- $\lambda(\alpha_i|y_i)$: loss function, describes the loss incurred for taking action $\alpha_i$ when state of nature is $y_i$.
- $P(y_i)$: prior probability that state of nature is $y_i$.
- $p(\mathbf{x}|y_i)$: state conditional probability for $\mathbf{x}$.

# Posterior Probability

Bayes formula:

$$P(y_i|\mathbf{x}) = \frac{p(\mathbf{x}|y_i)P(y_i)}{p(\mathbf{x})}$$

The evidence $p(\mathbf{x})$:

$$p(\mathbf{x}) = \sum_{i=1}^{c} p(\mathbf{x}|y_i)P(y_i)$$

# Conditional Risk

- Observe **x**, take action $\alpha_i$, if true state of nature is $y_j \rightarrow$ loss $\lambda(\alpha_i|y_j)$.
- The expected loss, or conditional risk, of taking action $\alpha_i$ is:

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

- For given observation **x**, selecting the action that minimizes the conditional risk.

# Overall Risk

- **Decision rule**: function $\alpha(\mathbf{x})$: $R^d \rightarrow \{\alpha_1, ..., \alpha_a\}$, indicate which action to take for every possible observation $\mathbf{x}$.
- The overall risk: expected loss associated with a given decision rule $\alpha(\mathbf{x})$ considering all possible observations:

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

- Choose $\alpha(\mathbf{x})$ that minimizes the overall risk.
- The minimum overall risk $R^\star$ is called **Bayes risk**, best performance we can get.

# Bayesian Decision Rule

To minimize the overall risk, we compute conditional risk for $i = 1..., a$:

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

Select action $\alpha_i$ that has minimum conditional risk:

$$\alpha^\star = \arg\min_{\alpha_i} R(\alpha_i|\mathbf{x})$$

- Bayesian decision rule minimizes the overall risk. (**Minimum Risk Decision**)

# Three Approaches for Decision Problem

# Three Approaches

In general, we have three distinct approaches for decision problem which are (in increasing order of complexity):

- **Discriminant function**: find a function $g(\mathbf{x})$ which maps each input $\mathbf{x}$ directly onto a class label.
- **Discriminative models**: approaches that model the posterior probabilities directly (i.e., $p(y_k|\mathbf{x})$).
- **Generative models**: approaches that model the joint distribution $p(y_k, \mathbf{x})$.
  - Specifically, determining the class-conditional densities $p(\mathbf{x}|y_k)$ and prior class probabilities $p(y_k)$ for each class $y_k$ individually. Then use Bayes' theorem to find posterior $p(y_k|\mathbf{x})$.

# Acknowledgements

Slides adapted from Prof. Tian Han's lecture.