In The Name of Allah

Digital Media Laboratory
Sharif University of Technology

Statistical Pattern Recognition

Classification - Statistical Methods

Hamid R. Rabiee
Jafar Muhammadi, Alireza Ghassemi

Spring 2012
http://ce.sharif.edu/courses/90-91/2/ce725-1/

Agenda

- Bayesian Decision Theory
- Prior Probabilities
- Class-Conditional Probabilities
- Posterior Probabilities
- Probability of Error
- Conditional Risk
- Min-Error-Rate Classification
- Probabilistic Discriminant Functions
  - Discriminant Functions: Gaussian Density
- Minimax Classification
- Neyman-Pearson

Bayesian Decision Theory

- Bayesian Decision Theory is a fundamental statistical approach that quantifies the tradeoffs between various decisions using probabilities and costs that accompany such decisions.
  - First, we will assume that all probabilities are known.
  - Then, we will study the cases where the probabilistic structure is not completely known.

Bayesian Decision Theory

- **We are using fish sorting example to illustrate these topics.**
- **Fish sorting example revisited**
  - **State of nature is a random variable.**
  - **Define w as the type of fish we observe (state of nature, class) where**
    - $w = w_1$ for sea bass,
    - $w = w_2$ for salmon.
    - $P(w_1)$ is the a priori probability that the next fish is a sea bass.
    - $P(w_2)$ is the a priori probability that the next fish is a salmon.

Prior Probabilities

- Prior probabilities reflect our knowledge of how likely each type of fish will appear before we actually see it.
- How can we choose $P(w_1)$ and $P(w_2)$?
  - Set $P(w_1) = P(w_2)$ if they are equiprobable (uniform priors).
  - May use different values depending on the fishing area, time of the year, etc.
- Assume there are no other types of fish $P(w_1) + P(w_2) = 1$
  - (exclusivity and exhaustivity).

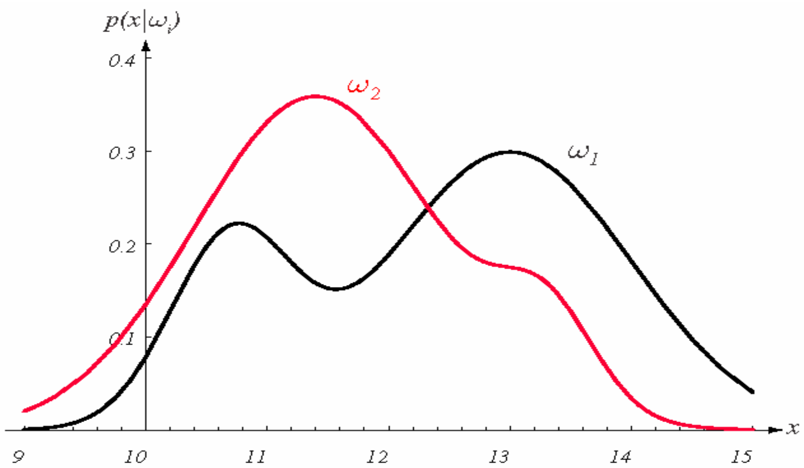
Prior Probabilities

- How can we make a decision with only the prior information?
  - Decide $\begin{cases} w_1 & \text{if } P(w_1) > P(w_2) \\ w_2 & \text{otherwise} \end{cases}$

- What is the probability of error for this decision?
  - $P(\text{error}) = \min\{P(w_1), P(w_2)\}$

Class-Conditional Probabilities

- Let's try to improve the decision using the lightness measurement x.
  - Let x be a continuous random variable.
  - Define $P(x|w_j)$ as the class-conditional probability density (probability of x given that the state of nature is w_j for j = 1, 2).
  - $P(x|w_1)$ and $P(x|w_2)$ describe the difference in lightness between populations of sea bass and salmon.
  - Hypothetical class-conditional probability density functions for two Classes.



Class-Conditional Probabilities

- How can we make a decision with only the class-conditional probabilities?
  - Decide $\begin{cases} w_1 & \text{if } P(x|w_1) > P(x|w_2) \\ w_2 & \text{otherwise} \end{cases}$
- Looks good, but prior information are not used. It may degrade decision performance
  - e.g what happens if we know a priori that 99% of fish are se basses?
- Class-conditional is known as "Maximum Likelihood", also.

Posterior Probabilities

- Suppose we know $P(w_j)$ and $P(x|w_j)$ for $j = 1, 2$, and measure the lightness of a fish as the value x.
  - **Define $P(w_j|x)$ as the a posteriori probability (probability of the state of nature being w_j given the measurement of feature value x).**
  - **We can use the Bayes formula to convert the prior probability to the posterior probability:**

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

in which

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

$P(x|w_j)$ is called the likelihood and $P(x)$ is called the evidence.

Posterior Probabilities

- How can we make a decision after observing the value of x?
  - **Decide** $\begin{cases} w_1 & \text{if } P(w_1|x) > P(w_2|x) \\ w_2 & \text{otherwise} \end{cases}$

- Rewriting the rule gives
  - **Decide** $\begin{cases} w_1 & \text{if } \dfrac{P(x|w_1)}{P(x|w_2)} > \dfrac{P(w_2)}{P(w_1)} \\ w_2 & \text{otherwise} \end{cases}$

Note that, at every x, $P(w_1|x) + P(w_2|x) = 1$.

Probability of Error

- What is the probability of error for this decision?

$$P(error\,|\,x) = \begin{cases} P(w_1|x) & \text{if we decide } w_2 \\ P(w_2|x) & \text{if we decide } w_1 \end{cases}$$

- What is the average probability of error?

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

- Bayes decision rule minimizes this error because

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

Bayesian Decision Theory

- **How can we generalize to**
  - **More than one feature? (replace the scalar x by the feature vector x)**
  - **More than two states of nature? (just a difference in notation)**
  - **Allowing actions other than just decisions? (allow the possibility of rejection)**
  - **Different risks in the decision? (define how costly each action is)**

  - **Notations for generalization**
    - Let $\{w_1, . . . ,w_c\}$ be the finite set of c states of nature (classes, categories).
    - Let $\{\alpha_1, . . . , \alpha_a\}$ be the finite set of a possible actions.
    - Let $\lambda(\alpha_i|w_j)$ be the loss incurred for taking action i when the state of nature is w_j .
    - Let x be the d-dim vector-valued random variable called the feature vector.

Conditional Risk

- Suppose we observe x and take action α_i.
  - If the true state of nature is w_j , we incur the loss $\lambda(\alpha_i|w_j)$.
  - The expected loss with taking action i is

$$R(\alpha_i \mid x) = \sum_{j=1}^{c} \lambda(\alpha_i \mid w_j) P(w_j \mid x)$$

It is also called the conditional risk.

Conditional Risk

- We want to find the decision rule that minimizes the overall risk

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

- Bayesian decision rule minimizes the overall risk by selecting the action α_i for which $R(\alpha_i|x)$ is minimum
- The resulting minimum overall risk is called the Bayesian risk and is the best performance that can be achieved.

Conditional Risk

- **Two-category classification example**
  - **Define**
    - α_1 : deciding w_1
    - α_2 : deciding w_2
    - $\lambda_{ij} : \lambda(\alpha_i \mid w_j)$
  - **Conditional risks can be written as**

$$R(\alpha_1 \mid x) = \lambda_{11} P(w_1 \mid x) + \lambda_{12} P(w_2 \mid x)$$
$$R(\alpha_2 \mid x) = \lambda_{21} P(w_1 \mid x) + \lambda_{22} 2P(w_2 \mid x)$$

Conditional Risk

- Two-category classification example
  - The minimum-risk decision rule becomes
    - Decide $\begin{cases} w_1 \text{ if} (\lambda_{21} - \lambda_{11}) P(w_1 \mid x) > (\lambda_{12} - \lambda_{22}) P(w_2 \mid x) \\ w_2 \quad \text{otherwise} \end{cases}$
  - This corresponds to deciding w1 if

$$\frac{p(x \mid w_1)}{p(x \mid w_2)} > \frac{\lambda_{12} - \lambda_{22}}{\lambda_{21} - \lambda_{11}} \frac{P(w_2)}{P(w_1)}$$

comparing the likelihood ratio to a threshold that is independent of the observation x.

Min-Error-Rate Classification

- Problem definition:
  - Actions are decisions on classes (α_i is deciding w_i).
  - If action α_i is taken and the true state of nature is w_j , then the decision is correct if i = j and in error if i≠j.
  - We want to find a decision rule that minimizes the probability of error.
- Define the zero-one loss function (all errors are equally costly).

$$\lambda(\alpha_i \mid w_j) = \begin{cases} 0 & \text{if } i = j \\ 1 & \text{if } i \neq j \end{cases} \quad i,j = 1,\ldots,c$$

- Conditional risk becomes

$$R(\alpha_i \mid x) = \sum_{j=1}^{c} \lambda(\alpha_i \mid wj) P(w_j \mid x) = \sum_{j \neq i} P(w_j \mid x) = 1 - P(w_i \mid x)$$

Template designed by Jafar Muhammadi

Min-Error-Rate Classification

- Minimizing the risk requires maximizing $P(w_i|x)$ and results in the minimum-error decision rule
  - **Decide w_i if $P(w_i|x) > P(w_j|x)$ for all $j \neq i$.**
- The resulting error is called the Bayesian error
  - **This is the best performance that can be achieved.**

Probabilistic Discriminant Functions

- **Discriminant functions: a useful way of representing classifiers**
  - **$g_i(x)$, i = 1, . . . , c**
    - Classifier assigns a feature vector x to class w_i if $g_i(x) > g_j(x)$ for all $j \neq i$.
  - **For the classifier that minimizes conditional risk**
    - $g_i(x) = -R(\alpha_i \mid x)$.
  - **For the classifier that minimizes error**
    - $g_i(x) = P(w_i \mid x)$.

Probabilistic Discriminant Functions

- These functions divide the feature space into c decision regions separated by decision boundaries ($R_1, \ldots, R_c$).
  - **Note that the results do not change even if we replace every $g_i(x)$ by $f(g_i(x))$ where $f(\cdot)$ is a monotonically increasing function (e.g., logarithm).**
    - This may lead to significant analytical and computational simplifications.

Discriminant Funcs: Gaussian Density

- Discriminant functions for the Gaussian density in case of min-error-rate classification, can be written as (why?):
  - $g_i(x) = \ln p(x|w_i) + \ln P(w_i)$, $p(x|w_i) = N(\mu_i, \Sigma_i)$, or

$$g_i(x) = -\frac{1}{2}(x-\mu_i)^T \Sigma_i^{-1}(x-\mu_i) - \frac{1}{2}\ln 2\pi - \frac{1}{2}|\Sigma_i| + \ln P(w_i)$$

Discriminant Funcs: Gaussian Density

- **Case 1: $\Sigma_i = \sigma^2 I$**
  - **Discriminant functions are**

$$g_i(x) = w_i^T x + w_{io}$$

Where $w_i = \frac{1}{\delta_2}\mu_i$ **and** $w_{io} = \frac{1}{\delta^2}\mu_i^T\mu_i + \ln P(w_i)$

  - (w_{i0} is the threshold or bias for the i'th category).

  - **Decision boundaries are the hyperplanes $g_i(x) = g_j(x)$, and can be written as**

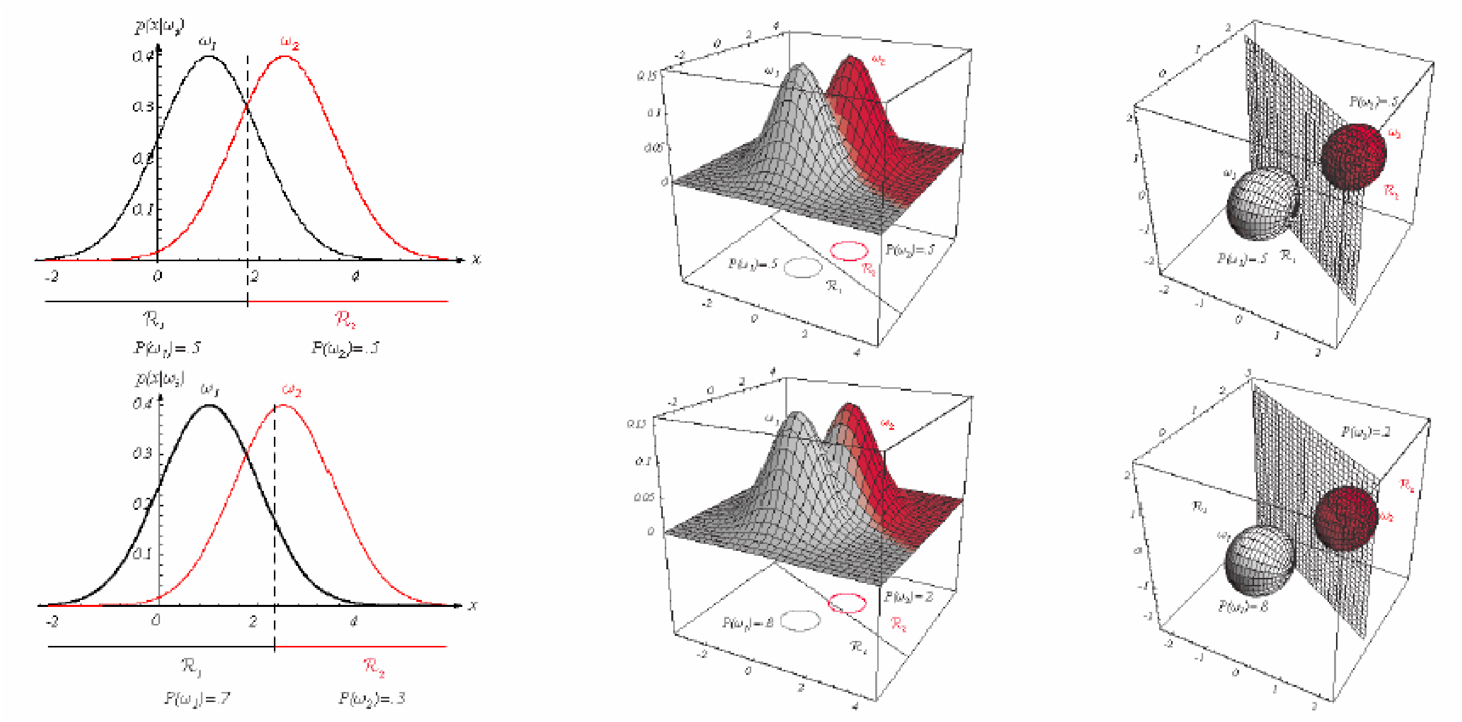
$$w_{ij}^T (x - x_0^{(ij)}) = 0$$

Where $w_{ij} = \mu_i - \mu_j$ **and** $x_o^{(ij)} = \frac{1}{2}(\mu_i + \mu_j) - \frac{\delta^2}{||\mu_i + \mu_j||^2} \ln \frac{P(w_i)}{P(w_j)}(\mu_i - \mu_j)$

Hyperplane separating R_i and R_j passes through the point $x_0^{(ij)}$ and is orthogonal to the vector w.

Discriminant Funcs: Gaussian Density

✧ **Case 1: $\Sigma_i=\sigma^2 I$**



Discriminant Funcs: Gaussian Density

- Case 1: $\Sigma_i=\sigma^2 I$
  - Special case when $P(w_i)$ are the same for i = 1, . . . , c is the minimum-distance classifier that uses the decision rule

$$\text{assign } x \text{ to } w_{i^*} \quad \text{where } i^* = \arg\min \|x-\mu_i\|, \; i=1,\ldots,c$$

Discriminant Funcs: Gaussian Density

- Case 2: $\Sigma_i = \Sigma$
  - Discriminant functions are $g_i(x) = w_i^T x + w_{io}$

    Where $w_i = \Sigma^{-1}\mu_i$ and $w_{io} = \mu_i^T \Sigma^{-1} \mu_i + \ln P(w_i)$
  - Decision boundaries can be written as $w_{ij}^T (x - x_0^{(ij)}) = 0$

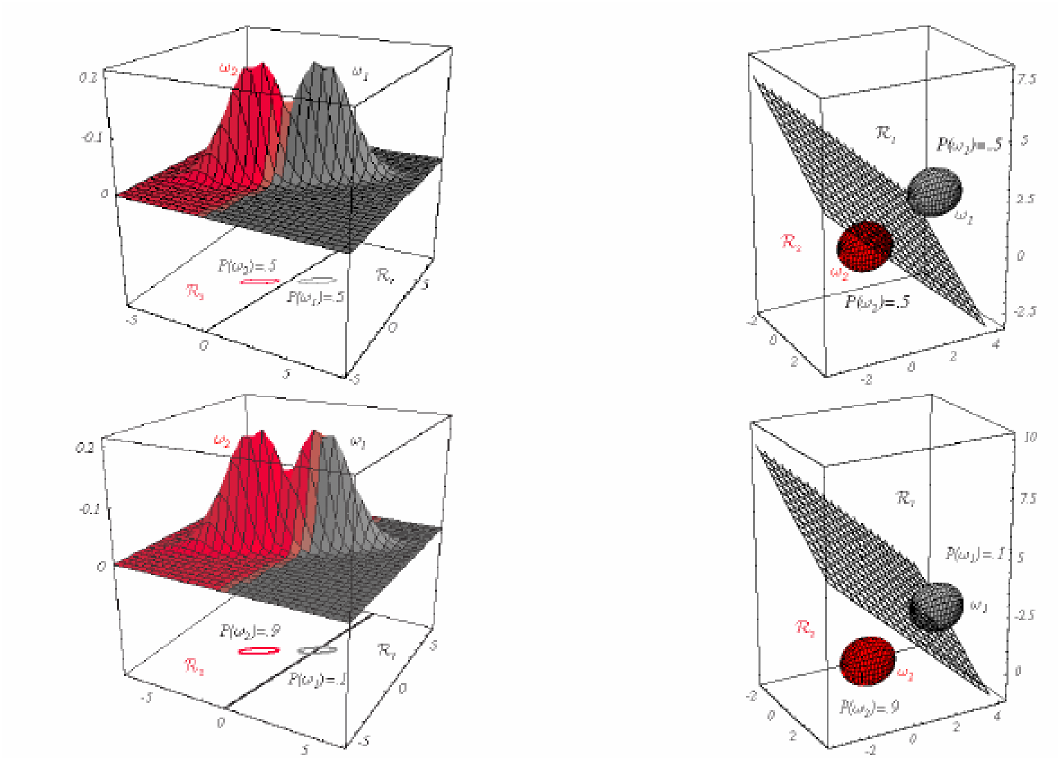
    Where $w_{ij} = \mu_i - \mu_j$ and $x_o^{(ij)} = \frac{1}{2}(\mu_i + \mu_j) - \frac{1}{(\mu_i + \mu_j)^T \Sigma^{-1} (\mu_i + \mu_j)} \ln \frac{P(w_i)}{P(w_j)} (\mu_i - \mu_j)$

    Hyperplane passes through $x_0^{(ij)}$ but is not necessarily orthogonal to the line between the means.

Template designed by Jafar Muhammadi

Discriminant Funcs: Gaussian Density

- Case 2: $\Sigma_i = \Sigma$



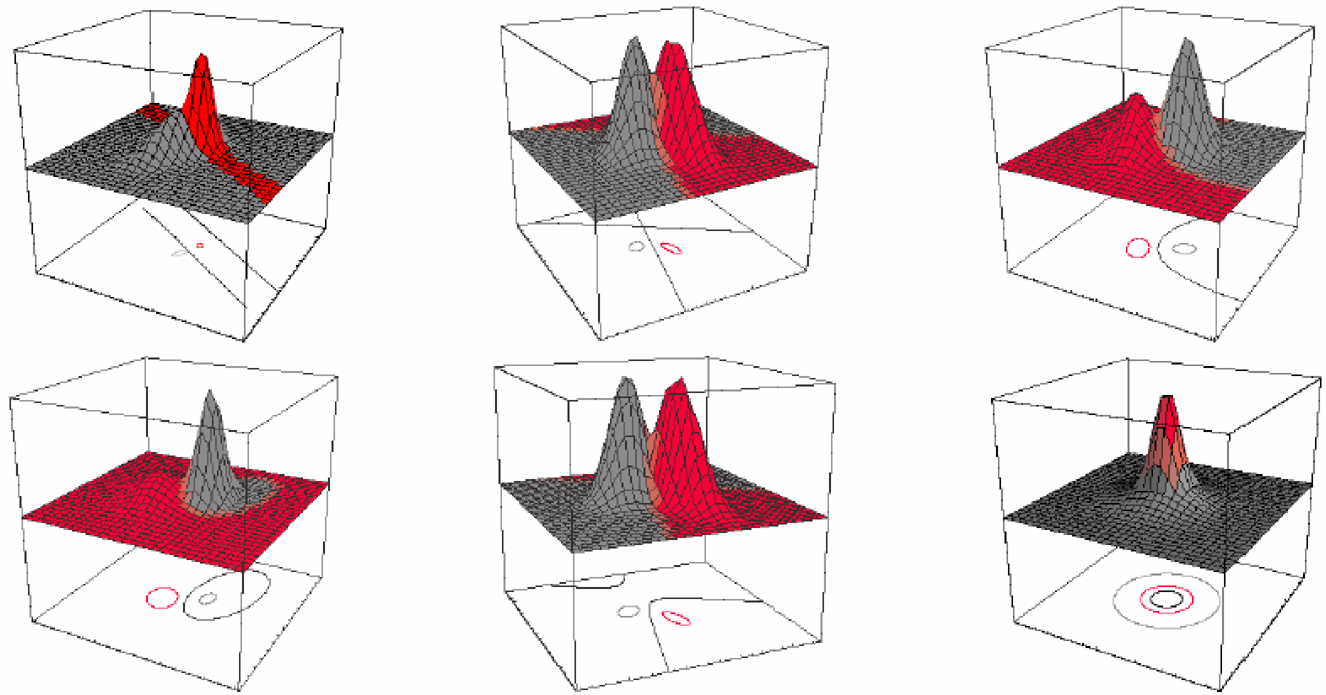
Discriminant Funcs: Gaussian Density

- Case 3: Σ_i= Arbitrary
  - Discriminant functions are $g_i(x) = x^T W_i x + w_i^T x + w_{io}$

    Where $W_i = -\frac{1}{2}\Sigma_i^{-1}$, $w_i = \Sigma_i^{-1}\mu_i$ and $w_{io} = -\frac{1}{2}\mu_i^T \Sigma_i^{-1}\mu_i - \frac{1}{2}\ln|\Sigma_i| + \ln P(w_i)$
  - Decision boundaries are hyperquadrics

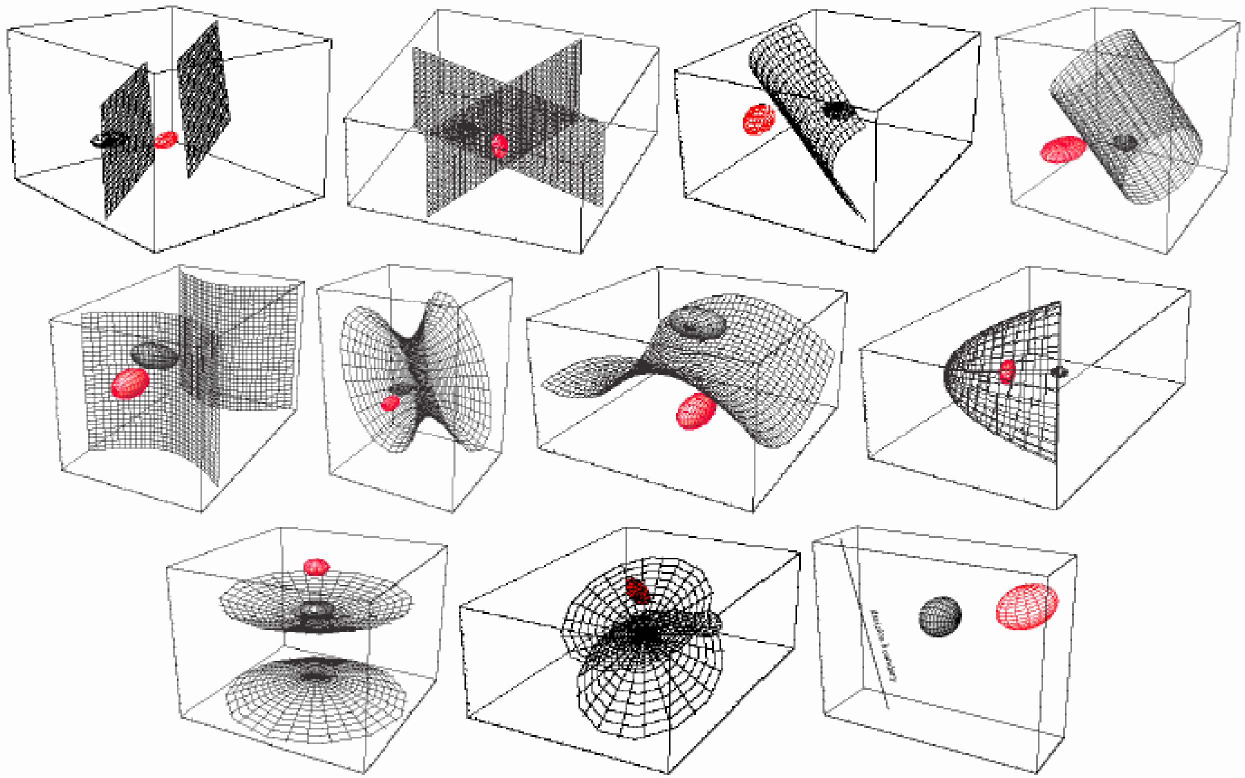
Discriminant Funcs: Gaussian Density

✧ **Case 3: Σ_i= Arbitrary**



Discriminant Funcs: Gaussian Density

✧ Case 3: Σ_i= Arbitrary



Minimax Classification

- In many real life applications, prior probabilities may be unknown, or time-varying, so we can not have a Bayesian optimal classification.
- However, one may wish to minimize the max possible overall risk.
  - The overall risk is,

$$R = \int_{R_1} [\lambda_{11} P(w_1) P(x|w_1) + \lambda_{12} P(w_2) P(x|w_2)] dx$$
$$+ \int_{R2} [\lambda_{21} P(w_1) P(x|w_1) + \lambda_{22} P(w_2) P(x|w_2)] dx$$

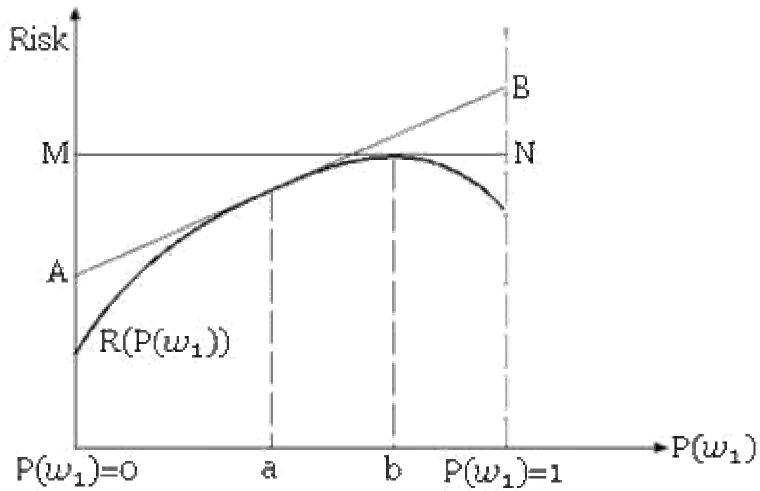
$P(w_2) = 1 - P(w_1)$ and $\int_{R_1} P(x|w_1) dx = 1 - \int_{R_2} P(x|w_2) dx$, then

$$R\left(P(w_1), R_1\right) = \lambda_{22} + (\lambda_{12} - \lambda_{22}) \int_{R_1} P(x|w_2) dx$$
$$+ P(w_1) \left[(\lambda_{11} - \lambda_{22}) - (\lambda_{21} - \lambda_{11}) \int_{R_2} P(x|w_1) dx - (\lambda_{12} - \lambda_{22}) \int_{R1} P(x|w_2) dx \right]$$

Minimax Classification

- For a fix R_1, the overall risk is a linear function of $P(w_1)$, and the maximum error occurs in $P(w_1)=0$, or $P(w_1)=1$.
  - **Why should the line be a tangent to $R(P(w_1),R_1)$?**
- For all possible R_1s, we are looking for the one which minimizes this maximum error, i.e.

$$R_1 = \arg\min_{R_1} \left\{ \max R\left(P(w_1), R_1\right) \right\}$$



Minimax Derivation

- Another way to solve R_1 in minimax is from:

$$R\left(P(w_1), R_1\right) = \lambda_{22} + (\lambda_{12} - \lambda_{22})\int_{R_1} p(x|w_2)dx = R_{mm}, \text{ minimax risk}$$

$$+ P_1 \times \left((\lambda_{11} - \lambda_{22}) + (\lambda_{21} - \lambda_{11})\int_{R_2} p(x|w_1)dx - (\lambda_{12} - \lambda_{22})\int_{R_1} p(x|w_2)dx \right) = 0$$

  - **If you get multiple solutions, choose one that gives you the minimum Risk**

Template designed by Jafar Muhammadi

Neyman-Pearson Criterion

- **If we do not know the prior probabilities, Bayesian optimum classification is not possible.**
  - **Suppose that the goal is maximizing the probability of detection, while constraining the probability of false-alarm to be less than or equal to a certain value.**
    - E.g. in a radar system false alarm (assuming an enemy aircraft is approaching while this is not the case) may be OK but it is very important to maximize the probability of detecting a real attack
  - **Based on this constraint (Neyman-Pearson criterion) we can design a classifier**
    - Typically must adjust boundaries numerically (for some distributions, such as Gaussian, analytical solutions do exist.

Any Question?

End of Lecture 6

Thank you!

Spring 2012

http://ce.sharif.edu/courses/90-91/2/ce725-1/