Pattern Recognition - The Formal Setting

- Input Data: *Measurements, Words, Fish lengths,..*
  Each pattern is represented as a set of features.

$$\underline{x} = [x_1, x_2, \ldots, x_l]^T \in X$$

- Label: *Fish type, Fish weight, Spam/Non-Spam……*
  The output label, the property we wish to predict.

$$y \in Y$$

- Pattern Recognition Machine f : Predicts a label for each input data .

$$\hat{y} = f(\underline{x})$$

- Training set: N samples with known labels.

$$D = \{\underline{x}_i, y_i\}_{i=1}^{N}$$

The Problem →

The Formal Setting (2)

The Problem:

Predict a label y of a new input data x not in the training data.

There are two flavors of the labeling problem:
Regression and Classification

- Classification: The label is discrete

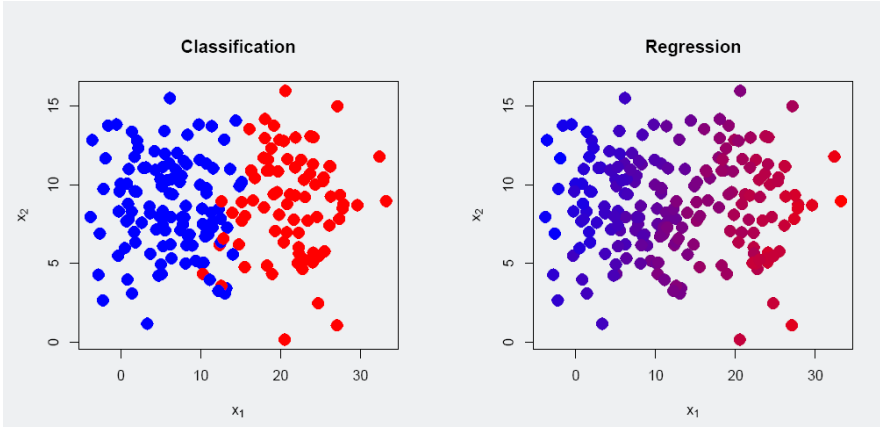
$$y \in \{0, 1, 2, \ldots., K\}$$

Remark: For discrete labels many different notations are used. e.g. for different classes c_i or ω_i are common.

- Regression: The label is real-valued

$$y \in \mathbb{R}$$

Regression and Classification



The Formal Setting + Probability Theory

An inherent difficulty is the ever-changing appearance of input data samples, e.g. a sea bass is not always exactly 1.5 m long – this is what makes the problem challenging!

- A way to formalize this is to model the feature vector as random variable X with a probability distribution $P(X = x)$.
- Consider all data samples drawn independently from the same true distribution $(x, y) \sim P(x, y)$
  (i.i.d. independent and identical distributed)
- **$P(x, y)$ is usually unknown,**
  only a training set with N samples might be available !!!!

Decision Theory:

What is the best prediction we can make if we know $P(x, y)$.

Generative and Discriminative Modelling

4

We need the posterior distribution $P(\mathrm{y} \mid x)$ to make a decision. There are two fundamental ways of getting there:

- Generative Modelling

  Use a conditional and prior probability in Bayes's rule

$$\Rightarrow \quad P(y|x) = \frac{P(x|y)P(y)}{P(x)}$$

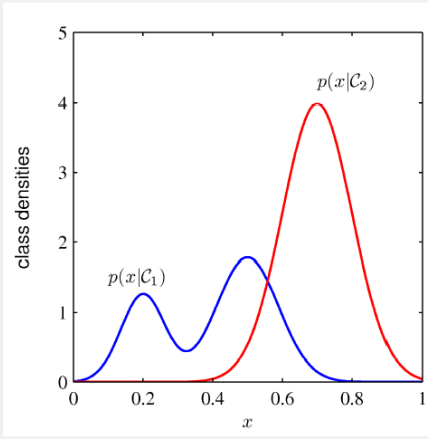
- Discriminative Modelling

  Directly find an expression for $P(\mathrm{y} \mid x)$ as a function of x and y. (In practice this functional form is often relatively easy)

Generative and Discriminative Modelling

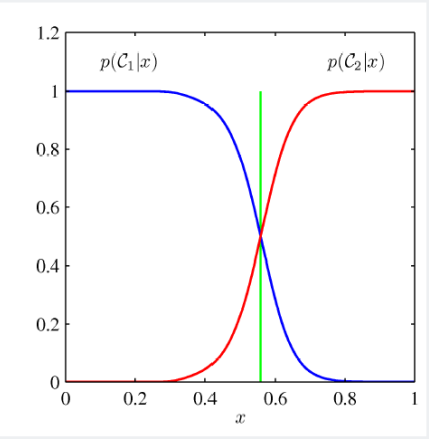
5

An Example: Bishop 1.27



Generative Model

Note the bimodality of $\mathcal{C}_1$



Discriminative Model

The mixture detail is not relevant

Bayes Decision Theory

Bayes Rule

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

Bayes, Thomas (1763) An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London,* **53:370-418**

Remark: In some cases A and B are variables representing Cause and Symptom! Then the conditional probabilities represent:

a) P(C|S) the **diagnostic direction** and
b) P(S|C) the **causal direction** .

Bayes Decision Theory

- Each pattern is represented by a feature vector $\underline{x} = [x_1, x_2, ..., x_l]^T \in X$ of a random variable X with a probability distribution $P(X = x)$.
- Consider all data samples $\underline{x}$ drawn independently from the same true distribution (i.i.d)

  $(x, y) \sim P(x, y)$
- Assign pattern with feature vector $\underline{x}$ to the most probable of the available classes $\omega_1, \omega_2, ..., \omega_M$

That is, $\underline{x} \rightarrow \omega_i$ if $P(\omega_i | \underline{x}) > P(\omega_j | \underline{x}) \quad \forall\ i \neq j$

So, how do we calculate $P(\omega_i | \underline{x})$?

8

Bayes Decision Theory

- Computation of a-posteriori probabilities $P(\omega_i | \underline{x})$
  - We are looking for $P(\omega_i | \underline{x})$, known as the a-posteriori probability, or posterior, of ω_i given $\underline{x}$.
  - We assume as known
    1.) $P(\omega_1), P(\omega_2)..., P(\omega_M)$ known as the a-priori probability, or prior, of ω_i.
    2.) $p(\underline{x} \mid \omega_i), \quad i = 1, 2..., M$ known as the likelihood of $\underline{x}$ *with respect to* ω_i.

Uppercase P: a probability of a discrete variable
Lowercase p: a probability density function (pdf)

9

Derivation of the Bayes Rule

- Two fundamental rules of probability theory:

sum rule $\quad p(\underline{x}) = \sum_i p(\underline{x}, \omega_i)$

product rule $\quad p(\underline{x}, \omega_i) = p(\underline{x}) P(\omega_i | \underline{x}) = p(\underline{x} | \omega_i) P(\omega_i)$

- From these, derive the BAYES RULE:

$$\Rightarrow \quad P(\omega_i | \underline{x}) = \frac{p(\underline{x} | \omega_i) P(\omega_i)}{p(\underline{x})}$$

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

POSTERIOR = (LIKELIHOOD · PRIOR) / EVIDENCE

Bayes Classification Rule (M=2classes)

❖ Given $\underline{x}$, classify it according to the rule

$$\text{If } P(\omega_1|\underline{x}) > P(\omega_2|\underline{x}) \quad \underline{x} \rightarrow \omega_1$$
$$\text{If } P(\omega_2|\underline{x}) > P(\omega_1|\underline{x}) \quad \underline{x} \rightarrow \omega_2$$

➢ Equivalently: classify $\underline{x}$ according to the rule

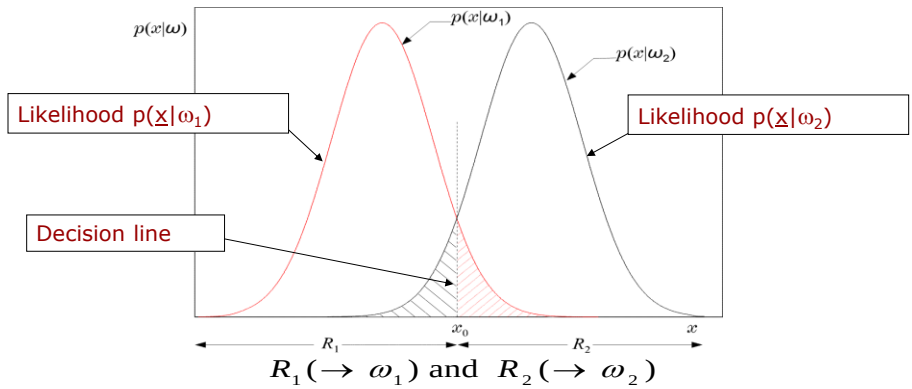
$$p(\underline{x}|\omega_1)P(\omega_1) \quad (><) \quad p(\underline{x}|\omega_2)P(\omega_2)$$

➢ For equiprobable ($P(\omega_1) = P(\omega_2)$) classes, the test is

$$p(\underline{x}|\omega_1) \quad (><) \quad p(\underline{x}|\omega_2)$$

Bayes Classification Rule (M=2)

➢ Graph for two equiprobable classes ω_1, ω_2, with decision line.



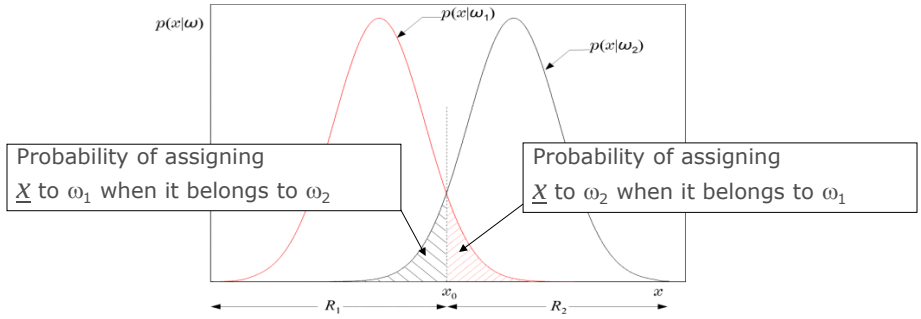
$R_1(\rightarrow \omega_1)$ and $R_2(\rightarrow \omega_2)$

➢ Equivalently in words: Divide space in two regions R1, R2

If

$$If \ x \in R_1 \Rightarrow \underline{x} \ in \ \omega_1$$
$$If \ x \in R_2 \Rightarrow \underline{x} \ in \ \omega_2$$

Bayes: Probability of Error (M=2)



- Total shaded area

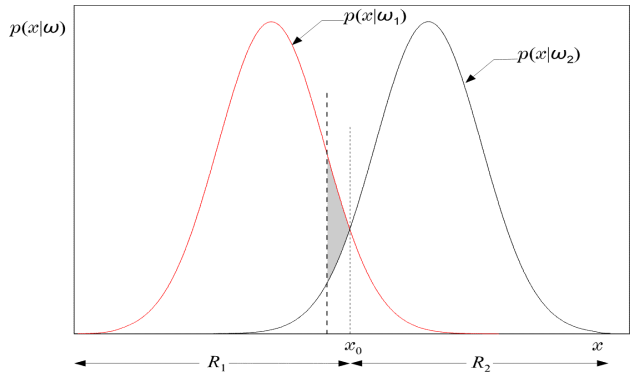
$$P_e = P(\omega_2)\int_{-\infty}^{x_0} p(x|\omega_2)dx + P(\omega_1)\int_{x_0}^{+\infty} p(x|\omega_1)dx$$

= total probability of assigning x to the wrong class

- Bayesian classifier is OPTIMAL with respect to minimizing the classification error probability!!!!

- Bayesian classifier is OPTIMAL with respect to minimizing the classification error probability!!!!

"Proof ?"



"Proof" : If the threshold is moved, the total shaded area INCREASES by the extra "dark" area.

14

Bayes Classification Rule (M>2)

- Now assume there are more than two classes (M>2).
- Given $\underline{x}$ classify it to ω_i if:

$$P(\omega_i|\underline{x}) > P(\omega_j|\underline{x}) \quad \forall j \neq i$$

- This choice, too, minimizes the classification error probability.

15

Minimizing the Risk (Classification)

Some types of classification errors may be more serious than others. If so, we can modify Bayesian classification:

- Assign penalty terms to weight each type of error

For M=2:

- Define the loss matrix $L = \begin{pmatrix} \lambda_{11} & \lambda_{12} \\ \lambda_{21} & \lambda_{22} \end{pmatrix}$
- λ_{12} : penalty term for deciding class ω_2, although the pattern belongs to ω_1, etc.

(usually $\lambda_{ij} > \lambda_{ii}$, and often $\lambda_{ii} = 0$; correct decisions are much less penalized than incorrect ones)

Minimizing the Risk

❖ Define risk r as the expected loss

➢ Risk with respect to ω_1

$$r_1 = \lambda_{11} \int_{R_1} p(\underline{x}|\omega_1) d\underline{x} + \lambda_{12} \int_{R_2} p(\underline{x}|\omega_1) d\underline{x}$$

➢ Risk with respect to ω_2

$$r_2 = \lambda_{21} \int_{R_1} p(\underline{x}|\omega_2) d\underline{x} + \lambda_{22} \int_{R_2} p(\underline{x}|\omega_2) d\underline{x}$$

$\Longrightarrow$ Probabilities of wrong decisions, weighted by the penalty terms

➢ risk $r = r_1 P(\omega_1) + r_2 P(\omega_2)$

Classification under minimal risk

❖ Choose regions R_1 and R_2 so that r is minimized

➢ Assign $\underline{\boldsymbol{x}}$ to ω_1 if

$$\lambda_{11} p(\underline{x}|\omega_1) P(\omega_1) + \lambda_{21} p(\underline{x}|\omega_2) P(\omega_2) < \lambda_{12} p(\underline{x}|\omega_1) P(\omega_1) + \lambda_{22} p(\underline{x}|\omega_2) P(\omega_2)$$

$$\Longrightarrow (\lambda_{11} - \lambda_{12}) p(\underline{x}|\omega_1) P(\omega_1) < (\lambda_{22} - \lambda_{21}) p(\underline{x}|\omega_2) P(\omega_2)$$

➢ Equivalently:

assign $\underline{\boldsymbol{x}}$ to ω_1 if $\ell_{12} \equiv \frac{p(\underline{x}|\omega_1)}{p(\underline{x}|\omega_2)} > \frac{P(\omega_2)}{P(\omega_1)} \frac{\lambda_{21} - \lambda_{22}}{\lambda_{12} - \lambda_{11}}$

ℓ_{12} : likelihood ratio

Now we have a classifier that is OPTIMAL with respect to minimizing the risk or expected loss.

It is closely related to the Bayes classifier that optimally minimizes error probability.

If $P(\omega_1) = P(\omega_2) = \frac{1}{2}$ and $\lambda_{11} = \lambda_{22} = 0$

$$\underline{x} \rightarrow \omega_1 \text{ if } P(\underline{x}|\omega_1) > P(\underline{x}|\omega_2)\frac{\lambda_{21}}{\lambda_{12}}$$

$$\underline{x} \rightarrow \omega_2 \text{ if } P(\underline{x}|\omega_2) > P(\underline{x}|\omega_1)\frac{\lambda_{12}}{\lambda_{21}}$$

if $\lambda_{21} = \lambda_{12} \Rightarrow$ risk minimization is equivalent to error probability minimization

An Example:

- $p(x|\omega_1) = \frac{1}{\sqrt{\pi}}\exp(-x^2)$
- $p(x|\omega_2) = \frac{1}{\sqrt{\pi}}\exp(-(x-1)^2)$
- $P(\omega_1) = P(\omega_2) = \frac{1}{2}$
- $L = \begin{pmatrix} 0 & 0.5 \\ 1.0 & 0 \end{pmatrix}$

Let's compare a) error probability minimization and
b) average risk minimization

a) Compute threshold x_0 for minimum P_{error} :

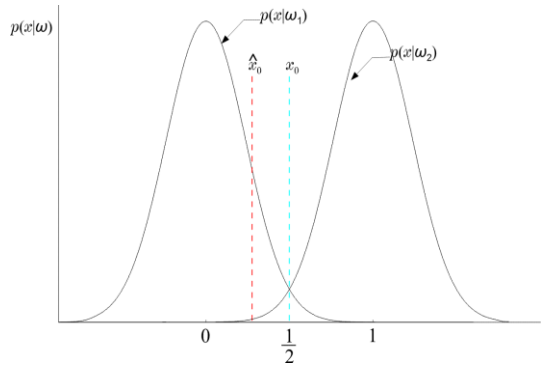
Solve $\exp(-x^2) = \exp(-(x-1)^2) \Rightarrow x_0 = \frac{1}{2}$

An Example (2):

b) Compute threshold $\hat{x}_0$ for minimum risk r

Solve $\exp(-x^2) = 2\exp(-(x-1)^2) \Rightarrow \hat{x}_0 = \frac{(1-ln\,2)}{2} < \frac{1}{2}$

➢ Thus, $\hat{x}_0$ lies to the left of $x_0 = \frac{1}{2}$



Minimizing the Risk for Regression

When solving a regression problem the risk/loss function is important: We can make infinitely many miss predictions:

Again: Minimize expected loss, with respect to $f(x)$.

$$E[L] = \int L(y, f(x))\, p(x, y)\, dx\, dy$$

Here the loss function is not a table anymore, it's a function of the continuous label y and its predicted value $f(x)$.
We need variational calculus to minimize this expression with respect to the function $f(x)$..... *More in the Machine Learning course!*

On the next slide we will show the results for some very common loss functions (without proofs)! →

Loss Functions: Some Examples
(proof Bishop 1.5.5)

$$L(y, f(x)) = (y - f(x))^2 \quad \rightarrow \quad f(x) = E[y|x] = \int y \; p(y|x) dy$$

- The conditional expectation value of the label, given the data, is the best possible prediction if we assume a quadratic loss function
- The function $f(x) = E[y|x]$ is called the Regression function.

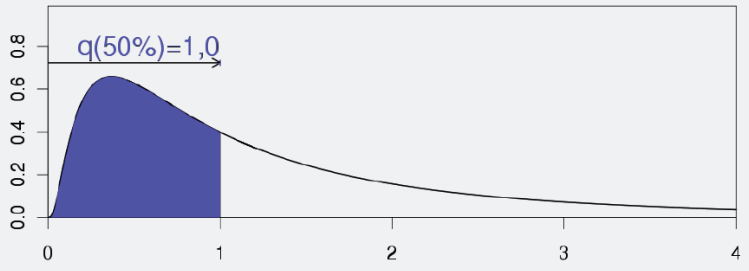
$$L(y, f(x)) = |y - f(x)| \quad \rightarrow \quad f(x) = \text{MEDIAN}[y|x]$$

$$L(y, f(x)) = \begin{cases} 0 & f(x) = y \\ 1 & f(x) \neq y \end{cases} \quad \rightarrow \quad f(x) = \text{MODE}[y|x]$$

- The Median is a robust estimator.
- The 0/1 loss function assigns the single most probable label
- Compare the last case with classification -> is also a 0/1 loss function

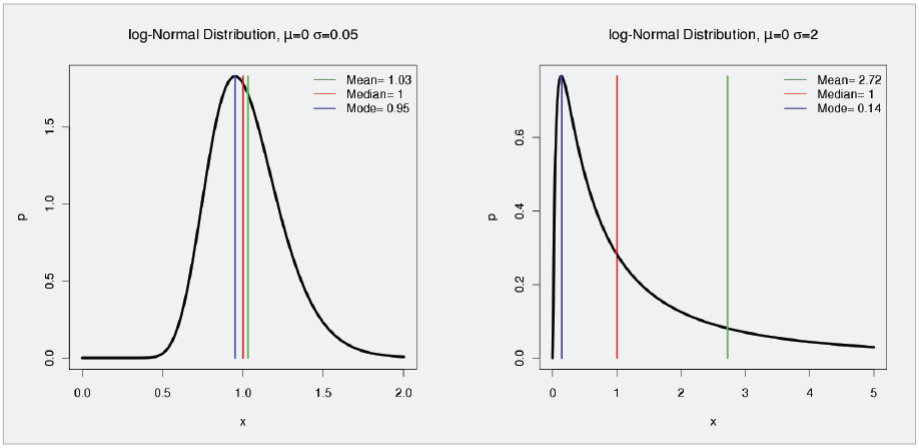
In the Gaussian distribution the mean, the mode and the median coincide!!

The median



$$\int_{-\infty}^{M} f(x) \, \mathrm{d}x = \frac{1}{2}$$

Mean, Median, Mode



Discriminant Functions

General form of a pattern classifier:

Assign $\underline{x}$ to ω_i if $g_i(\underline{x}) > g_j(\underline{x}) \quad \forall i \neq j \qquad j = 0, \ldots, M$

➢For each class ω_i, there is a discriminant function $g_i(x)$

Examples of Classifiers seen so far:

a) Bayesian minimum error classifier, in various equivalent forms

$g_i(\underline{x}) = p(\omega_i \mid \underline{x})$ or

$g_i(\underline{x}) = p(\underline{x} \mid \omega_i)\, P(\omega_i)$

b) Bayesian minimum risk classifier, see slide 17

$g_i(x) = (\lambda_{jj} - \lambda_{ji})\, p(x|\omega_j)\; P(\omega_j)$

Discriminant Functions (2)

Discriminant functions are never unique.

Equivalent functions always exist that produce the same classification result.

$g_i(\underline{x}) > g_j(\underline{x})$ can be replaced by $f(g_i(\underline{x})) > f(g_j(\underline{x}))$
if f is a monotonically increasing function

Example: $ln(.)$ is monotonically increasing.

=> The discriminant function $g_i(\underline{x}) = P(\underline{x} \mid \omega_i)\, P(\omega_i)$

can be replaced by $g_i(\underline{x}) = ln(P(\underline{x} \mid \omega_i)\, P(\omega_i))$

$= lnP(\underline{x} \mid \omega_i) + ln\, P(\omega_i)$

Taking the logarithm often makes computations easier

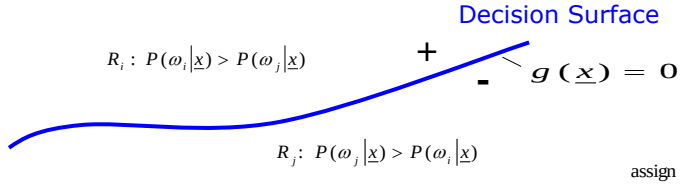
Decision Surfaces (1): Bayes classifier

Space is divided into regions $R_1,\ldots, R_M$ by the discriminant functions $g_i(\underline{x}) = P(\omega_i|\underline{x})$, $i = 1\ldots M$.

If regions R_i, R_j are contiguous, the surface separating them is

$$g(\underline{x}) \equiv P(\omega_i \mid \underline{x}) - P(\omega_j \mid \underline{x}) = 0$$

On one side of the surface, $g(x)$ is positive (+), on the other negative (-).



assign $\underline{x} \rightarrow \omega_i$ if $\underline{x} \in R_i$

Decision Surfaces (2): general case

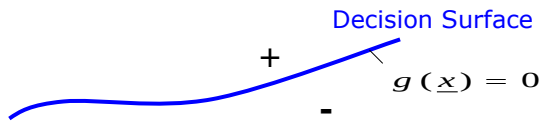
❖ General form of a decision surface (two-class case):

$$g(\underline{x}) \equiv g_1(\underline{x}) - g_2(\underline{x}) = 0$$

where $g_1(\underline{x})$, $g_2(\underline{x})$ are discriminant functions

❖ Now we can use $g(\underline{x})$ to rewrite the classification rule.

$$\underline{x} \rightarrow \omega_1 \quad \text{if} \quad g(\underline{x}) > 0 \ ,$$
$$\underline{x} \rightarrow \omega_2 \quad \text{if} \quad g(\underline{x}) < 0$$



Discriminant Functions (3)

Discriminant functions can also be defined independently of the Bayesian rule.

- This leads to `suboptimal´ solutions, with no guarantee to minimize the classification error probability.
- Yet if chosen appropriately, can be **computationally more tractable**, especially when the pdf´s can not be computed correctly.
- Examples will follow in subsequent lectures.

Bayesian Classifier for Normal Distributions

- Often, the correct pdf $p(\underline{x}|\omega_i)$ for a dataset is not known.

- Let's assume a multivariate Gaussian distribution:

$$\underline{x} \in \mathbb{R}^{\ell}$$

$$p(\underline{x}|\omega_i) = \frac{1}{(2\pi)^{\frac{\ell}{2}} |\Sigma_i|^{\frac{1}{2}}} \exp\left(-\frac{1}{2} (\underline{x} - \underline{\mu}_i)^{\mathrm{T}} \Sigma_i^{-1} (\underline{x} - \underline{\mu}_i) \right)$$

$$\underline{\mu}_i = E[\underline{x}]$$

$$\Sigma_i = E\left[(\underline{x} - \underline{\mu}_i)(\underline{x} - \underline{\mu}_i)^{\mathrm{T}} \right] \quad l \times l \text{ matrix called covariance matrix}$$

Bayesian Classifier for Normal Distributions

- $ln(\ )$ is a monotonic function.

We define:

$$g_i(\underline{x}) = \ln(\ p(\underline{x}|\omega_i) \cdot P(\omega_i)\ )$$

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

$$g_i(\underline{x}) = -\frac{1}{2} (\underline{x} - \underline{\mu}_i)^T \Sigma_i^{-1} (\underline{x} - \underline{\mu}_i) + \ln P(\omega_i) + c_i$$

$$c_i = -(\frac{\ell}{2}) \ln 2\pi - (\frac{1}{2}) \ln |\Sigma_i|$$

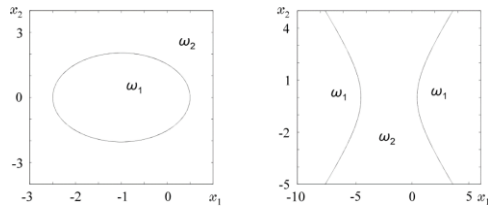
$$g_i(\underline{x}) = -\frac{1}{2} \underline{x}^T \Sigma_i^{-1} \underline{x} + \frac{1}{2} \underline{x}^T \Sigma_i^{-1} \underline{\mu}_i - \frac{1}{2} \underline{\mu}_i^T \Sigma_i^{-1} \underline{\mu}_i + \frac{1}{2} \underline{\mu}_i^T \Sigma_i^{-1} \underline{x} + \ln P(\omega_i) + c_i$$

Bayesian Classifier for Normal Distributions

- Example with : $\ell = 2$ and $\Sigma_i = \begin{pmatrix} \sigma_i^2 & 0 \\ 0 & \sigma_i^2 \end{pmatrix}$

$$g_i(\underline{x}) = -\frac{1}{2\sigma_i^2}(x_1^2 + x_2^2) + \frac{1}{\sigma_i^2}(\mu_{i1}x_1 + \mu_{i2}x_2) - \frac{1}{2\sigma_i^2}(\mu_{i1}^2 + \mu_{i2}^2) + \ln P(\omega_i) + c_i$$

That is, $g_i(\underline{x})$ is quadratic and surfaces $g_i(\underline{x}) - g_j(\underline{x}) = 0$ are quadrics, ellipsoids, parabolas, hyperbolas, pairs of lines.



Bayesian Classifier for Normal Distributions

- Decision Hyperplanes $g_i(\underline{x}) - g_j(\underline{x}) = 0$
  - Case: $\Sigma_j = \Sigma_i = \Sigma$

ALL the quadratic terms $\underline{x}^T \Sigma_i^{-1} \underline{x}$ and c_i are not of interest. They are not involved in the comparisons. Then we can write equivalently:

$$g_i(\underline{x}) = -\frac{1}{2}\underline{x}^T\Sigma_i^{-1}\underline{x} + \frac{1}{2}\underline{x}^T\Sigma_i^{-1}\underline{\mu}_i + \frac{1}{2}\underline{\mu}_i^T\Sigma_i^{-1}\underline{x} - \frac{1}{2}\underline{\mu}_i^T\Sigma_i^{-1}\underline{\mu}_i + \ln P(\omega_i) + c_i$$

$$g_i(\underline{x}) = \underline{w}_i^T\underline{x} + w_{io}$$

$$\text{with} \quad \underline{w}_i = \Sigma^{-1}\underline{\mu}_i \quad \text{and} \quad w_{i0} = \ln P(\omega_i) - \frac{1}{2}\underline{\mu}_i\Sigma^{-1}\underline{\mu}_i$$

=> If $\Sigma_i = \Sigma_j$, the discriminant functions are LINEAR

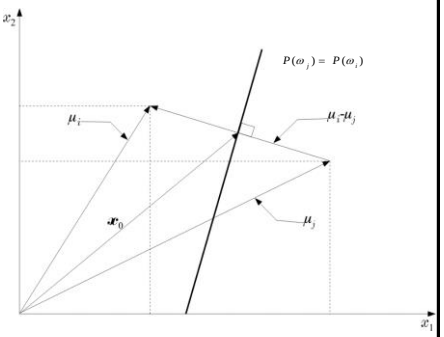
Bayesian Classifier for Normal Distributions

- Subcase A: $\Sigma_i = \Sigma_j = \Sigma = \sigma^2 I$ (multiple of the identity matrix)

Disciminant function:

$$g_i(\underline{x}) = -\frac{1}{2}\underline{x}^T\Sigma_i^{-1}\underline{x} + \frac{1}{2}\underline{x}^T\Sigma_i^{-1}\underline{\mu}_i + \frac{1}{2}\underline{\mu}_i^T\Sigma_i^{-1}\underline{x} - \frac{1}{2}\underline{\mu}_i^T\Sigma_i^{-1}\underline{\mu}_i + \ln P(\omega_i) + c_i$$

$$g_i(\underline{x}) = \underbrace{\frac{1}{\sigma^2}\underline{\mu}_i^T}_{\underline{w}_i^T}\underline{x} + w_{i0}$$



Decision Hyperplane:

$$g_{ij}(\underline{x}) = g_i(\underline{x}) - g_j(\underline{x}) = 0$$

$$g_{ij}(\underline{x}) = \frac{1}{\sigma^2}(\underline{\mu}_i^T - \underline{\mu}_j^T)\underline{x} + w_{i0} - w_{j0}$$

$$g_{ij}(\underline{x}) = \underline{w}^T(\underline{x} - \underline{x}_o)$$

with

$$\underline{w} = \underline{\mu}_i - \underline{\mu}_j,$$

$$\underline{x}_o = \frac{1}{2}(\underline{\mu}_i + \underline{\mu}_j) - \sigma^2 \ln\left(\frac{P(\omega_i)}{P(\omega_j)}\right) \frac{\underline{\mu}_i - \underline{\mu}_j}{\left\|\underline{\mu}_i - \underline{\mu}_j\right\|^2}$$

Bayesian Classifier for Normal Distributions

- again subcase A: $\Sigma_i = \Sigma_j = \Sigma = \sigma^2 I.$

$$g_{ij}(\underline{x}) = g_i(\underline{x}) - g_j(\underline{x}) = 0$$

Then

$$g_{ij}(\underline{x}) = \underline{\mu}_i^T\Sigma^{-1}\underline{x} - \underline{\mu}_j^T\Sigma^{-1}\underline{x} + \frac{1}{2}\underline{\mu}_j^T\Sigma^{-1}\underline{\mu}_j - \frac{1}{2}\underline{\mu}_i^T\Sigma^{-1}\underline{\mu}_i + \ln\frac{P(\omega_i)}{P(\omega_j)}$$

$$g_{ij}(\underline{x}) = (\underline{\mu}_i^T\Sigma^{-1} - \underline{\mu}_j^T\Sigma^{-1})\underline{x} + \frac{1}{2}\underline{\mu}_j^T\Sigma^{-1}\underline{\mu}_j - \frac{1}{2}\underline{\mu}_i^T\Sigma^{-1}\underline{\mu}_i + \ln\frac{P(\omega_i)}{P(\omega_j)}$$

$$g_{ij}(\underline{x}) = \frac{1}{\sigma^2}(\underline{\mu}_i^T - \underline{\mu}_j^T)\underline{x} - \frac{1}{2\sigma^2}(\underline{\mu}_i^2 - \underline{\mu}_j^2) + \ln\frac{P(\omega_i)}{P(\omega_j)}$$

$$g_{ij}(\underline{x}) = \underline{w}^T(\underline{x} - \underline{x}_o)$$

with

$$\underline{w} = \underline{\mu}_i - \underline{\mu}_j,$$

$$\underline{x}_o = \frac{1}{2}(\underline{\mu}_i + \underline{\mu}_j) - \sigma^2 \ln\left(\frac{P(\omega_i)}{P(\omega_j)}\right) \frac{\underline{\mu}_i - \underline{\mu}_j}{\left\|\underline{\mu}_i - \underline{\mu}_j\right\|^2}$$

Bayesian Classifier for Normal Distributions

- Subcase B: $\Sigma_i = \Sigma_j \neq \sigma^2 \mathrm{I}$ (arbitrary nondiagonal matrix)

Then $g_{ij}(\underline{x}) = \underline{w}^T (x - x_0) = 0$

with

$$\underline{w} = \Sigma^{-1}(\underline{\mu}_i - \underline{\mu}_j)$$

$$\underline{x}_0 = \frac{1}{2}(\underline{\mu}_i + \underline{\mu}_j) - \ln\left(\frac{P(\omega_i)}{P(\omega_j)}\right) \frac{\underline{\mu}_i - \underline{\mu}_j}{\left\|\underline{\mu}_i - \underline{\mu}_j\right\|^2_{\Sigma^{-1}}}$$

$$\left\|\underline{x}\right\|_{\Sigma^{-1}} \equiv (\underline{x}^T \Sigma^{-1} \underline{x})^{\frac{1}{2}}$$

- Decision hyperplane
  - not normal to $\underline{\mu}_i - \underline{\mu}_j$
  - normal to $\Sigma^{-1}(\underline{\mu}_i - \underline{\mu}_j)$

Bayesian Classifier for Normal Distributions

Different interpretation: Minimum Distance Classifiers

- Case : equiprobable $P(\omega_i) = \frac{1}{M}$ and $\Sigma_i = \Sigma_j = \Sigma$

$$\Rightarrow \quad g_i(\underline{x}) = -\frac{1}{2}(\underline{x} - \underline{\mu}_i)^T \Sigma^{-1}(\underline{x} - \underline{\mu}_i)$$

- subcase A: $\Sigma = \sigma^2 I$: Assign $\underline{x} \rightarrow \omega_i$
  if the Euclidean Distance $d_E \equiv \left\|\underline{x} - \underline{\mu}_i\right\|$ is smaller.

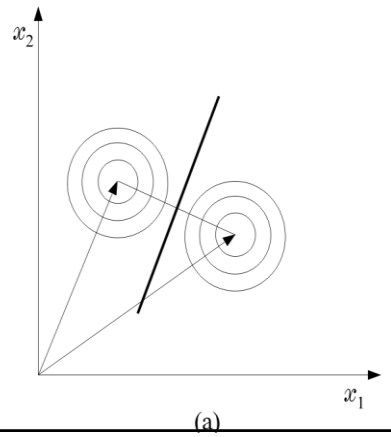
- subcase B: $\Sigma \neq \sigma^2 I$: Assign $\underline{x} \rightarrow \omega_i$
  if the Mahalanobis Distance $d_M = \left\|\underline{x} - \underline{\mu}_i\right\|_{\Sigma^{-1}}$ is smaller.

$$d_M = ((\underline{x} - \underline{\mu}_i)^T \Sigma^{-1}(\underline{x} - \underline{\mu}_i))^{\frac{1}{2}}$$

Bayesian Classifier for Normal Distributions

Subcase A
Use Euclidian Distance:

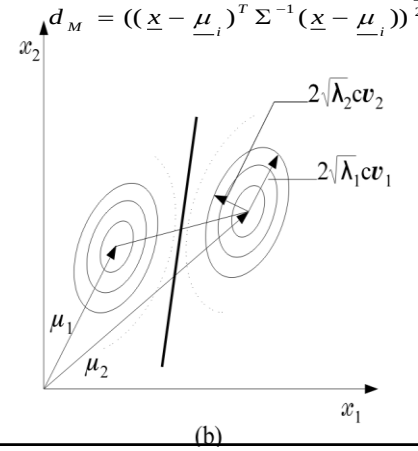
$$d_E \equiv \left\| \underline{x} - \underline{\mu}_i \right\|$$



Subcase B
Use Mahalanobis Distance:

$$d_M = \left\| \underline{x} - \underline{\mu}_i \right\|_{\Sigma^{-1}}$$

$$d_M = ((\underline{x} - \underline{\mu}_i)^T \Sigma^{-1} (\underline{x} - \underline{\mu}_i))^{\frac{1}{2}}$$



Bayesian Classifier for Normal Distributions

Example:

Given two classes ω_1, ω_2:

with $P(\omega_1) = P(\omega_2)$

and $p(\underline{x}|\omega_1) = N(\mu_1, \Sigma)$

$p(\underline{x}|\omega_2) = N(\mu_2, \Sigma)$

$$\underline{\mu}_1 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \quad \underline{\mu}_2 = \begin{bmatrix} 3 \\ 3 \end{bmatrix}, \quad \Sigma = \begin{bmatrix} 1.1 & 0.3 \\ 0.3 & 1.9 \end{bmatrix}$$

Task:

classify the vector $\underline{x} = \begin{bmatrix} 1.0 \\ 2.2 \end{bmatrix}$

using Bayesian classification:

Solution:

- Compute Mahalanobis Distance d_M from μ_1, μ_2

with $\Sigma^{-1} = \begin{bmatrix} 0.95 & -0.15 \\ -0.15 & 0.55 \end{bmatrix}$

$$d_M^2(\mu_i, x) = (x - \mu_i)^T \Sigma^{-1} (x - \mu_i)$$

$$\Rightarrow d^2_{M_1} = [1.0, \quad 2.2] \Sigma^{-1} \begin{bmatrix} 1.0 \\ 2.2 \end{bmatrix} = 2.952$$

$$\Rightarrow d^2_{M_2} = [-2.0, \quad -0.8] \Sigma^{-1} \begin{bmatrix} -2.0 \\ -0.8 \end{bmatrix} = 3.672$$

- classify $\underline{x} \rightarrow \omega_1$. Observe that $d_{E_2} < d_{E_1}$