

What is "Pattern Recognition"?

The term "Pattern Recognition" ("Mustererkennung") is used for

**Methods for classifying unknown objects based on feature vectors
(narrow sense meaning of Pattern Recognition)**

**Methods of analyzing signals and recognizing interesting patterns
(wide sense meaning of Pattern Recognition)**

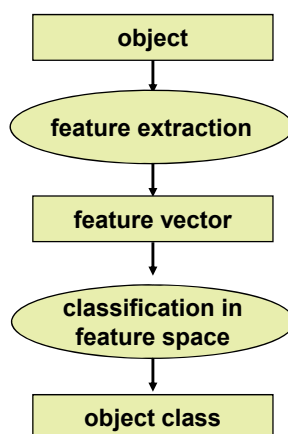
Pattern recognition can be applied to all kinds of signals, e.g.

- images
 - acoustic signals
 - seismographic signals
 - tomographic data
- etc.

The following section deals with Pattern Recognition in the narrow sense.
(see Duda and Hart, Pattern Classification and Scene Analysis, Wiley 73)

1

Basic Terminology for Pattern Recognition



K classes $\omega_1 \dots \omega_K$
 N dimension of feature space
 $\underline{x}^T = [x_1 \ x_2 \ \dots \ x_N]$ feature vector
 $\underline{y}^T = [y_1 \ y_2 \ \dots \ y_N]$ prototype
 (feature vector with known class membership)
 $\underline{y}_i^{(k)}$ i -th prototype of class k
 M_k number of prototypes for class k
 $g_k(\underline{x})$ discriminant function for class k

Problem:

Determine $g_k(\underline{x})$ such that

$$g_k(\underline{x}) > g_j(\underline{x}) \quad \forall \underline{x} \in \omega_k \quad \forall k \neq j$$

2

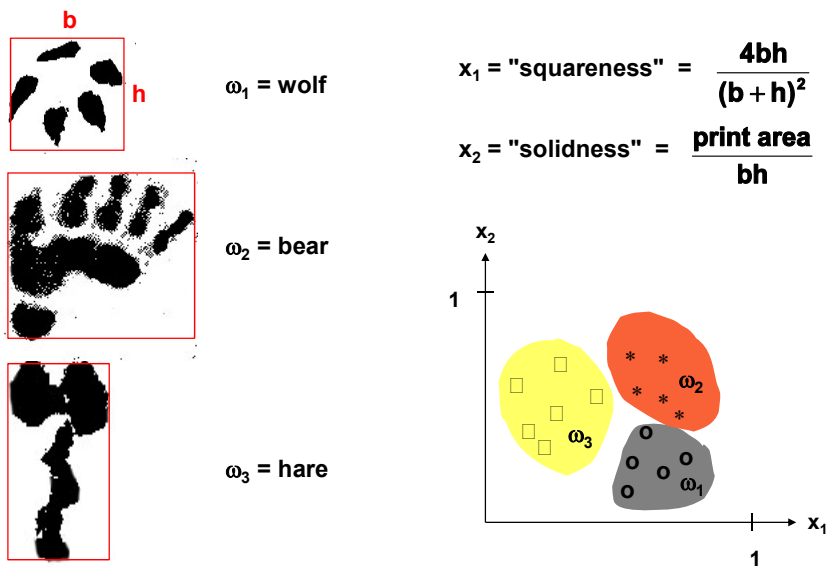
Example: Animal Footprints



What features can be used to distinguish the 3 footprint classes?

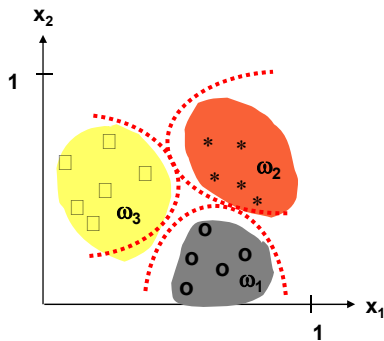
3

A Feature Space for Footprints



4

Discriminant Functions for Footprints

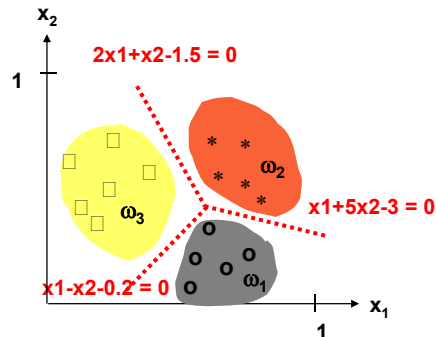


Quadratic discriminant functions:

$$g_1 = -9x_1^2 + 10.8x_1x_2 - 2.84$$

$$g_2 = x_1 + 20x_2^2 - 28x_2 + 9.4$$

$$g_3 = -x_1 + 5.6x_2^2 - 5.6x_2 - 1$$



Piecewise linear discriminant functions:

$$g_1 = 1 \text{ if } (x_1 - x_2 - 0.2 > 0) \wedge (x_1 + 5x_2 - 3 < 0) \text{ else } 0$$

$$g_2 = 1 \text{ if } (x_1 + 5x_2 - 3 > 0) \wedge (2x_1 + x_2 - 1.5 > 0) \text{ else } 0$$

$$g_3 = 1 \text{ if } (2x_1 + x_2 - 1.5 < 0) \wedge (x_1 - x_2 - 0.2 < 0) \text{ else } 0$$

5

Existence of Discriminant Functions

- For given prototypes, discriminant functions always exist as long as no two prototypes belonging to different classes are equal.
- If $g_i(\underline{x})$, $i = 1 \dots K$, are discriminant functions for given prototypes, then

$$g_i'(\underline{x}) = a(\underline{x}) g_i(\underline{x}) + b(\underline{x}), \quad a(\underline{x}) > 0, \quad i = 1 \dots K$$

are also discriminant functions.

- If the classes of a 2-class problem are separable, then there always exists a function $g(\underline{x})$ such that

$$g(\underline{x}) > 0 \quad \forall \underline{x} \in \omega_1$$

$$g(\underline{x}) < 0 \quad \forall \underline{x} \in \omega_2$$

6

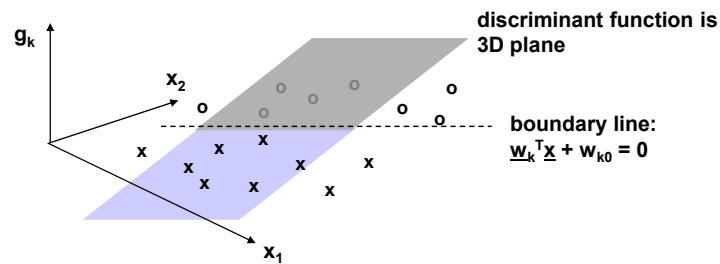
Linear Discriminant Functions

Linear discriminant functions are attractive because they can be

- easily determined from prototypes
- easily analyzed
- easily evaluated

Basic form of linear discriminant function:

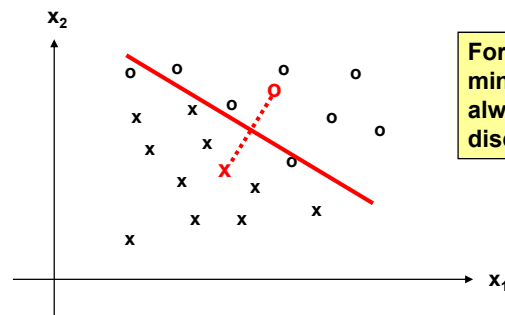
$$g_k(\underline{x}) = \underline{w}_k^T \underline{x} + w_{k0}$$



7

Class Average Minimal Distance Classification

- Represent prototypes by class averages
- Assign object to class with minimum distance between object and class average



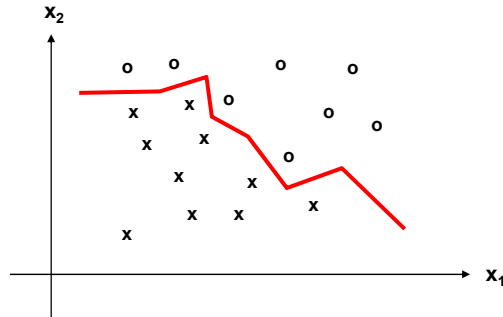
For a 2-class problem, the minimal distance criterion always results in a linear discriminant function

Class average minimal distance classification may not separate prototypes even if they are linearly separable!

8

Nearest Neighbour Classification

Assign object to class with nearest prototype

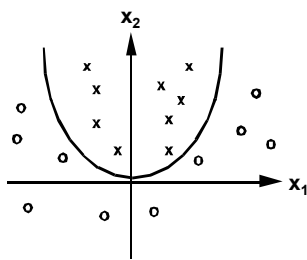


Piece-wise linear discriminant function

The nearest neighbour criterion classifies all prototypes correctly (except equal prototypes of different classes). The decision regions are not necessarily coherent.

9

Generalized Linear Discriminant Functions



Example:

Prototypes are not linearly separable

A quadratic discriminant function may work:

$$g(\mathbf{x}) = a_1x_1 + a_2x_2 + b_{11}x_1^2 + b_{22}x_2^2 + b_{12}x_1x_2 + c$$

Transformation of prototypes into higher-dimensional feature space may allow linear discriminant functions.

Transformation for the example: $z_1 = x_1$ $z_2 = x_2$ $z_3 = x_1^2$ $z_4 = x_2^2$ $z_5 = x_1x_2$

Linear discriminant function in z-space: $g(\mathbf{z}) = a_1z_1 + a_2z_2 + b_{11}z_3 + b_{22}z_4 + b_{12}z_5 + c$

Advantage: Linear separation algorithms may be applied

Disadvantage: Dimensionality of feature space is drastically increased

10

Linear Discriminant Functions for 2-Class Problems

Normalize prototypes such that

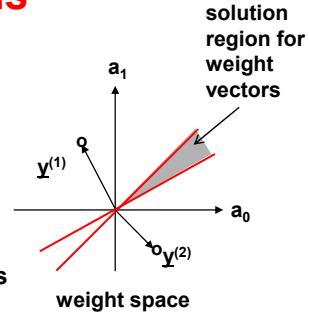
$$\underline{y}^T = [1 \ y_1 \ y_2 \ \dots \ y_N]$$

Discriminant function g can be expressed as

$$g(\underline{x}) = \underline{a}^T \underline{x} \quad \text{with} \quad \underline{a}^T = [a_0 \ a_1 \ \dots \ a_N]$$

Prototypes of class w_2 are negated such that

$$\underline{a}^T \underline{y} > 0 \Rightarrow \text{correct classification of both classes}$$



Solution region in weight space (if it exists) is the space at the positive side of all hyperplanes $\underline{a}^T \underline{y} = 0$. Any weight vector \underline{a} in this solution region gives a correct discriminant function.

Possible further constraints on solution vector \underline{a} :

$$\|\underline{a}\| = 1 \quad \text{and} \quad \underline{a}^T \underline{y} > b \quad \text{for all } \underline{y}$$

b is "margin", i.e. minimal distance of a correctly classified point from the hyperplanes defined by the prototypes.

11

Perceptron Learning Rule

A solution vector \underline{a} can be determined iteratively by minimizing a criterion function $J(\underline{a})$ by gradient descent.

Perceptron criterion function:

$$J_p(\underline{a}) = \sum_{\underline{y} \in B} (-\underline{a}^T \underline{y})$$

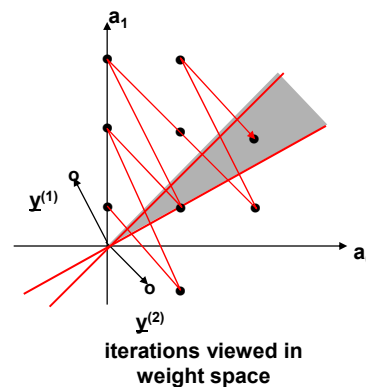
with $B = \{\text{all misclassified prototypes}\}$

Gradient: $\nabla J_p(\underline{a}) = \sum_{\underline{y} \in B} (-\underline{y})$

Basic gradient descent algorithm:

$$\underline{a}_{k+1} = \underline{a}_k + \rho_k \sum_{\underline{y} \in B} (\underline{y})$$

Weight vector \underline{a} is modified in negative gradient direction.



Example (see illustration) with $y_1^T = [-1 \ 2]$, $y_2^T = [1 \ -1]$, $\rho = 2$:

k	0	1	2	3	4	5	6	7	8
\underline{a}_k	0	2	0	2	0	2	4	2	4
	1	-1	3	1	5	3	1	5	3

4
3 solution

12

Statistical Decision Theory

Generating decision functions from a statistical characterization of classes (as opposed to a characterization by prototypes)

Advantages:

1. The classification scheme may be designed to satisfy an objective optimality criterion:
Optimal decisions minimize the probability of error.
2. Statistical descriptions may be much more compact than a collection of prototypes.
3. Some phenomena may only be adequately described using statistics, e.g. noise.

13

Example: Medical Screening (1)

Health test based on some measurement x (e.g. ECG evaluation)

It is known that every 10th person is sick (prior probability):

ω_1 class of healthy people $P(\omega_1) = 9/10$
 ω_2 class of sick people $P(\omega_2) = 1/10$

Task 1: Classify without taking any measurements (to save money)

Decision rule 1a: Classify every 10th person as sick

$$P(\text{error}) = P(\text{decide sick if healthy}) + P(\text{decide healthy if sick}) \\ = 1/10 \cdot 9/10 + 9/10 \cdot 1/10 = 0.18$$

Decision rule 1b: Classify all persons as healthy

$$P(\text{error}) = P(\text{decide healthy if sick}) = 1/10 = 0.1$$

- Decision rule 1b is better because it gives lower probability of error
- Decision rule 1b is optimal because no other decision rule can give a lower probability of error (try "every n -th" in 1a and minimize over n)

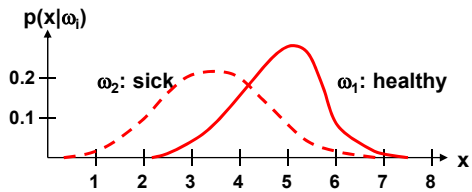
14

Example: Medical Screening (2)

Task 2: Classify after taking a measurement x

Assume that the statistics of prototypes are given as $p(x|\omega_i)$, $i = 1, 2$

Person No.	x	indication
⋮	⋮	⋮
⋮	⋮	⋮
134	7.4	neg
135	6.8	neg
136	4.2	pos
137	5.6	neg
138	5.8	pos
139	7.2	neg
⋮	⋮	⋮
⋮	⋮	⋮



$P(e|x) = P(\text{error given } x) = P(\omega \neq \omega' | x) = 1 - P(\omega|x)$
 where ω' is the class assigned to x by the decision rule.

**$P(e|x)$ is minimized by choosing the class which maximizes $P(\omega|x)$.
 Hence $g_i(x) = P(\omega_i|x)$ are discriminant functions.**

How do we get the "posterior" probabilities $P(\omega_i|x)$?

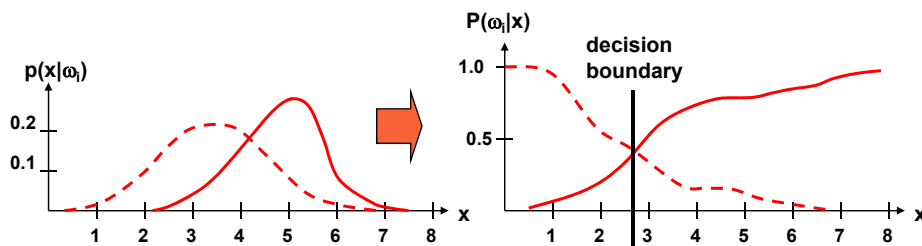
15

Example: Medical Screening (3)

The posterior probabilities $P(\omega_i|x)$ can be computed from the "likelihood" $p(x|\omega_i)$ using Bayes' formula:

$$P(\omega_i|x) = \frac{p(x|\omega_i) \cdot P(\omega_i)}{p(x)} = \frac{p(x|\omega_i) \cdot P(\omega_i)}{\sum_j p(x|\omega_j) P(\omega_j)}$$

For the example, using Bayes' Formula, one could get:



16

General Framework for Bayes Classification

Statistical decision theory minimizes the probability of error for classifications based on uncertain evidence

$\omega_1 \dots \omega_K$ K classes
 $P(\omega_k)$ prior probability that an object of class k will be observed
 $\underline{x} = [x_1 \dots x_N]$ N-dimensional feature vector of an object
 $p(\underline{x}|\omega_k)$ conditional probability ("likelihood") of observing \underline{x} given that the object belongs to class ω_k
 $P(\omega_k|\underline{x})$ conditional probability ("posterior probability") that an object belongs to class ω_k given \underline{x} is observed

Bayes decision rule:

Classify given evidence \underline{x} as class ω' such that ω' minimizes the probability of error $P(\omega \neq \omega' | \underline{x})$

=> Choose ω' which maximizes the posterior probability $P(\omega | \underline{x})$

$g_i(\underline{x}) = P(\omega_i|\underline{x})$ are discriminant functions.

17

Bayes 2-class Decisions

If the decision is between 2 classes ω_1 and ω_2 , the decision rule can be simplified:

Choose ω_1 if $\frac{p(\underline{x} | \omega_1)}{p(\underline{x} | \omega_2)} > \frac{P(\omega_2)}{P(\omega_1)}$

$\frac{p(\underline{x} | \omega_1)}{p(\underline{x} | \omega_2)}$ is called the "likelihood ratio"

Several alternative forms are possible for a discriminant function:

$$g(\underline{x}) = P(\omega_1|\underline{x}) - P(\omega_2|\underline{x}) \qquad \frac{p(\underline{x} | \omega_1)}{p(\underline{x} | \omega_2)}$$

For exponential and Gaussian distributions it is useful to take the logarithm:

$$g(\underline{x}) = \log \frac{P(\omega_1 | \underline{x})}{P(\omega_2 | \underline{x})} = \log \frac{P(\underline{x} | \omega_1)P(\omega_1)}{P(\underline{x} | \omega_2)P(\omega_2)} = \log \frac{P(\underline{x} | \omega_1)}{P(\underline{x} | \omega_2)} - \log \frac{P(\omega_2)}{P(\omega_1)}$$

18

Normal Distributions

Gaussian ("normal") multivariate distribution:

$$p(\underline{x}) = \frac{1}{(2\pi)^{\frac{N}{2}} |\Sigma|^{\frac{1}{2}}} \exp \left[-\frac{1}{2} (\underline{x} - \underline{\mu})^T \Sigma^{-1} (\underline{x} - \underline{\mu}) \right]$$

$$\Sigma = E[(\underline{x} - \underline{\mu})(\underline{x} - \underline{\mu})^T] \quad \text{N-by-N covariance matrix}$$

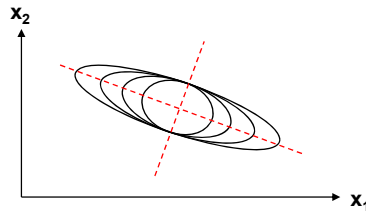
$$\underline{\mu} = E[\underline{x}] \quad \text{mean vector}$$

For decision problems, loci of points of constant density are interesting. For Gaussian multivariate distributions, these are hyperellipsoids:

$$(\underline{x} - \underline{\mu})^T \Sigma^{-1} (\underline{x} - \underline{\mu}) = \text{constant}$$

Eigenvectors of Σ determine directions of principal axes of the ellipsoids, eigenvalues determine lengths of the principal axes.

$d^2 = (\underline{x} - \underline{\mu})^T \Sigma^{-1} (\underline{x} - \underline{\mu})$ is called "squared Mahalanobis distance" of \underline{x} from $\underline{\mu}$.



19

Discriminant Function for Normal Distributions

General form:

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

For $p(\underline{x}|\omega_i) \sim N(\underline{\mu}_i, \Sigma_i)$:

$$g_i(\underline{x}) = -1/2 (\underline{x} - \underline{\mu}_i)^T \Sigma_i^{-1} (\underline{x} - \underline{\mu}_i) - 1/2 \log |\Sigma_i| + \log P(\omega_i)$$

We consider the discriminant functions for three interesting special cases:

- univariate distribution $N=1$
- statistically independent, equal variance variables x_i
- equal covariance matrices $\Sigma_i = \Sigma$

20

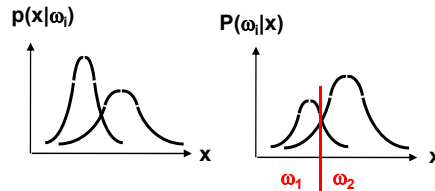
Univariate Distribution

$p(x|\omega_i)$ are univariate Gaussian distributions.

Example: 2 classes ω_1 and ω_2

$$p(x | \omega_1) = \frac{1}{\sqrt{2\pi}\sigma_1} \exp\left[-\frac{(x - \mu_1)^2}{2\sigma_1^2}\right]$$

$$p(x | \omega_2) = \frac{1}{\sqrt{2\pi}\sigma_2} \exp\left[-\frac{(x - \mu_2)^2}{2\sigma_2^2}\right]$$



Decision rule:

$$g_i(x) = \log P(\omega_i|x)$$

$$g_i(x) = -1/(2\Sigma_i^2) (x - \mu_i)^2 - 1/2 \log \Sigma_i + \log P(\omega_i)$$

21

Statistically Independent, Equal Variance Variables

In case of insufficient statistical data, variables are sometimes assumed to be statistically independent and of equal variance.

$$\Sigma_i = \Sigma^2 I$$

$$g_i(\underline{x}) = -1/(2\Sigma^2) \|\underline{x} - \underline{\mu}_i\|^2 + \log P(\omega_i)$$

If $P(\omega_i) = 1/N$, then the decision rule is equivalent to the minimum-distance classification rule.

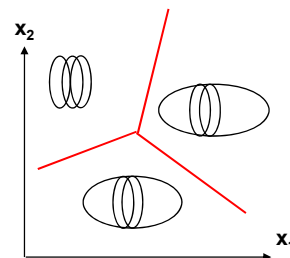
By expanding $g_i(\underline{x})$ and dropping the $\underline{x}^T \underline{x}$ term, one gets the decision rule:

$$g_i(\underline{x}) = -1/(2\Sigma^2)[-2\underline{\mu}_i^T \underline{x} + \underline{\mu}_i^T \underline{\mu}_i] + \log P(\omega_i)$$

which is linear in \underline{x} and can be written

$$g_i(\underline{x}) = \underline{w}_i^T \underline{x} + w_{i0}$$

The decision surface is composed of hyperplanes.



22

Equal Covariance Matrices

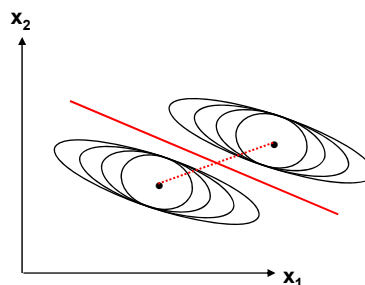
If $\Sigma_i = \Sigma$, the decision rule can be simplified:

$$g_i(\underline{x}) = -1/2 (\underline{x} - \underline{\mu}_i)^T \Sigma^{-1} (\underline{x} - \underline{\mu}_i) + \log P(\omega_i)$$

By expanding the quadratic form and dropping $\underline{x}^T \Sigma^{-1} \underline{x}$ one gets again a linear decision rule which can be written:

$$g_i(\underline{x}) = \underline{w}_i^T \underline{x} + w_{i0}$$

If the a-priori probabilities are equal, the decision rule assigns \underline{x} to the class where the Mahalanobis distance to the mean $\underline{\mu}_i$ is minimal.

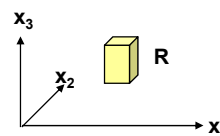


23

Estimating Probability Densities

Let R be a region in feature space with volume V.
Let k out of N samples lie in R.

$$\int_R p(\underline{x}') d\underline{x}' \approx \frac{k}{N} \approx p(\underline{x})V$$



$$p(\underline{x}) \approx \frac{k/N}{V}$$

relative frequency of samples per volume

A sequence of approximations $p_n(\underline{x})$ may be obtained by changing the volume V_n as the number of samples n increases.

Examples:

$V_n < 1/\sqrt{n}$ Parzen Windows
 $k_n \approx \sqrt{n}$ adjust volume for k nearest neighbours

Conditions for a converging sequence of estimates $p_n(\underline{x})$:

1. $\lim_{n \rightarrow \infty} V_n = 0$
2. $\lim_{n \rightarrow \infty} k_n = \infty$
3. $\lim_{n \rightarrow \infty} \frac{k_n}{n} = 0$

24

Estimating the Mean in a Univariate Normal Density

$$p(x|\mu) = N(\mu, \Sigma^2)$$

known normal probability density for x except of unknown mean μ

$$p(\mu) = N(\mu_0, \Sigma_0)$$

prior knowledge about μ in terms of a normal density with known μ_0 and Σ_0

$$X = \{x_1 \dots x_n\}$$

samples drawn from $p(x)$

Estimation using Bayes Rule:

$$p(\mu | X) = \frac{p(X|\mu)p(\mu)}{\int p(X|\mu)p(\mu) d\mu} = \alpha \prod_{k=1}^n p(x_k | \mu) p(\mu) \quad \langle \text{is scale factor independent of } \mu$$

$$= \alpha \prod_{k=1}^n \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\left(\frac{x_k - \mu}{\sigma}\right)^2\right] \frac{1}{\sqrt{2\pi}\sigma_0} \exp\left[-\frac{1}{2}\left(\frac{\mu - \mu_0}{\sigma_0}\right)^2\right] = \frac{1}{\sqrt{2\pi}\sigma_n} \exp\left[-\frac{1}{2}\left(\frac{\mu - \mu_n}{\sigma_n}\right)^2\right]$$

$$\text{with } \mu_n = \frac{n\sigma_0^2}{n\sigma_0^2 + \sigma^2} \left(\frac{1}{n} \sum_{k=1}^n x_k\right) + \frac{\sigma^2}{n\sigma_0^2 + \sigma^2} \mu_0 \quad \text{and} \quad \sigma_n^2 = \frac{\sigma_0^2 \sigma^2}{n\sigma_0^2 + \sigma^2}$$

Best estimate of mean μ after observing n samples