Bayesian Learning

A.Christidis, G.Giannoulis, Prof. Ioannis Pitas
Aristotle University of Thessaloniki
pitas@csd.auth.gr
www.aiia.csd.auth.gr
Version 3.3.1

Artificial Intelligence &
Information Analysis Lab

Bayesian Learning

- **Bayesian classification**
- Bayesian clustering

Bayes probability

General Bayesian classification problem: Classify data sample $\mathbf{x} \in \mathbb{R}^n$ to one of the m classes $\mathcal{C}_i, \quad i = 1, \dots, m$.

Definitions:

- $P(\mathcal{C}_i)$: ***A-priori probability*** of class $\mathcal{C}_i$.
- $P(\mathcal{C}_i|\mathbf{x})$: ***A-posteriori probability*** that the class $\mathcal{C}_i$ is adopted, given data sample $\mathbf{x}$.

Bayes probability

- $p(\mathbf{x}|\mathcal{C}_i)$: ***Multidimensional conditional probability distribution*** of data sample $\mathbf{x}_i$, given class $\mathcal{C}_i$.
- $p(\mathbf{x})$: ***Multidimensional probability distribution*** of data sample $\mathbf{x}$.
- $P(\mathbf{x}, \mathcal{C}_i)$: ***Joint probability*** of $\mathbf{x}$ and $\mathcal{C}_i$.

Bayes theorem:

$$p(\mathbf{x}|\mathcal{C}_i)P(\mathcal{C}_i) = P(\mathcal{C}_i|\mathbf{x})p(\mathbf{x}) = P(\mathbf{x}, \mathcal{C}_i).$$

Bayes Decision

General approach: Given m ***hypotheses*** (one per class $\mathcal{C}_i, i = 1, \ldots, M$), ***choose the one having the least cost*** (***risk***).

- L_{ij}: cost of adopting $\mathcal{C}_j$ when choosing $\mathcal{C}_i$ is the correction decision.
- The average cost of adopting $\mathcal{C}_j$ given data vector $\mathbf{x}$, is given by:

$$r_j(\mathbf{x}) = \sum_{i=1}^{m} L_{ij} P(\mathbf{x}, \mathcal{C}_i) = \sum_{i=1}^{m} L_{ij} p(\mathbf{x}|\mathcal{C}_i) P(\mathcal{C}_i).$$

Bayes Decision

Bayes Decision Rule:

- For a given data sample $\mathbf{x}$, if $r_k(\mathbf{x}) < r_j(\mathbf{x})$ for every $j \neq k$, $j, k = 1, \dots m$, then classify $\mathbf{x}$ to class $\mathcal{C}_k$.

- That is, for everydata sample $\mathbf{x}$, the hypothesis (class) $\mathcal{C}_k$ resulting in the ***minimal Bayes cost*** $r_k(\mathbf{x})$ is adopted.

Maximum A-Posteriori Criterion (MAP)

Special case:

- $L_{ii} = 0$ (zero cost for correct decisions).
- $L_{ij} = L$: cost is independent of class pair $\mathcal{C}_i, \mathcal{C}_j$, when $i \neq j$.
- Then Bayes rule is greatly simplified. $\mathcal{C}_k$ is selected if:

$$r_k(\mathbf{x}) = \sum_{i \neq k} p(\mathbf{x}|\mathcal{C}_i)P(\mathcal{C}_i) < \sum_{i \neq j} p(\mathbf{x}|\mathcal{C}_i)P(\mathcal{C}_i) = r_j(\mathbf{x}).$$

MAP Criterion

- By eliminating mutual terms in this inequality, $\mathcal{C}_k$ is selected if:

$$p(\mathbf{x}|\mathcal{C}_k)P(\mathcal{C}_k) > p(\mathbf{x}|\mathcal{C}_j)P(\mathcal{C}_j),$$
$$P(\mathbf{x}, \mathcal{C}_k) > P(\mathbf{x}, \mathcal{C}_j),$$
$$P(\mathcal{C}_k|\mathbf{x})p(\mathbf{x}) > P(\mathcal{C}_j|\mathbf{x})p(\mathbf{x}).$$

Maximum A-Posteriori Criterion (***MAP***):

- $\mathcal{C}_k$ is selected if:

$$P(\mathcal{C}_k|\mathbf{x}) > P(\mathcal{C}_j|\mathbf{x}).$$

ML Criterion

- Special case. ***Equiprobable classes***:

$$P(\mathcal{C}_i) = \frac{1}{m}, \qquad i = 1, \ldots, m.$$

Maximum Likelihood Criterion (***ML***):

- $\mathcal{C}_k$ is selected if:

$$p(\mathbf{x}|\mathcal{C}_k) > p\left(\mathbf{x}\middle|\mathcal{C}_j\right), \qquad \forall j \neq k,$$

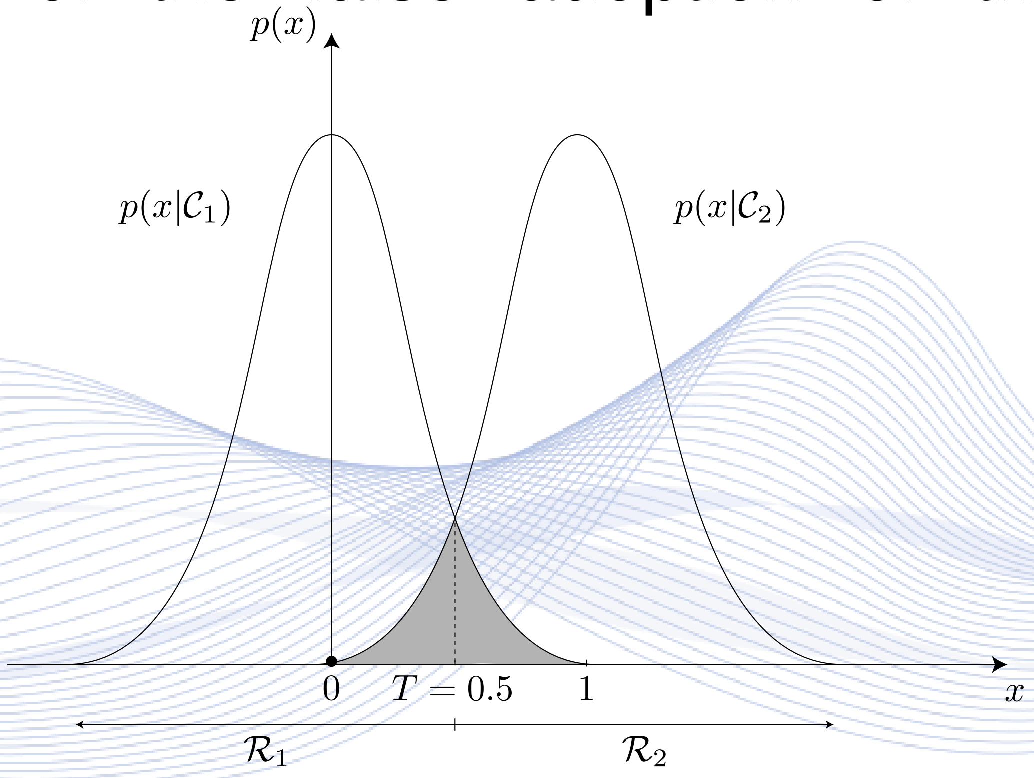
ML Criterion

- In the 1D case, decision regions $\mathcal{R}_1$ and $\mathcal{R}_2$ are defined as:

$$\mathcal{R}_1 = \{x \in \mathbb{R},\ \ p(x|\mathcal{C}_1) > p(x|\mathcal{C}_2)\},$$
$$\mathcal{R}_2 = \{x \in \mathbb{R},\ \ p(x|\mathcal{C}_1) < p(x|\mathcal{C}_2)\}.$$

ML Criterion

In this special case, the costs $r_1(\mathbf{x}), r_2(\mathbf{x})$ are proportional to the possibility of the false adoption of the class $\mathcal{C}_1$, $\mathcal{C}_2$ respectively.



Binary Classifier for two classes having 1D pdfs $N(0,1), N(1,1)$.

Artificial Intelligence & Information Analysis Lab

ML Criterion

In the case of a two-class problem ($m = 2$), Bayes rule becomes:

- Adopt $\mathcal{C}_1$, if $r_1(\mathbf{x}) < r_2(\mathbf{x})$ or:

$$\Lambda(\mathbf{x}) = \frac{p(\mathbf{x}|\mathcal{C}_1)}{p(\mathbf{x}|\mathcal{C}_2)} > T_{12}.$$

- $\Lambda(\mathbf{x})$: ***likelihood ratio.***
- Decision threshold T_{12}:

$$T_{12} = \frac{P(\mathcal{C}_2)(L_{21} - L_{22})}{P(\mathcal{C}_1)(L_{12} - L_{11})}.$$

ML Criterion

- In the case of MAP criterion:

$$T_{12} = \frac{P(\mathcal{C}_2)}{P(\mathcal{C}_1)}.$$

- In the case of ML criterion:

$$T_{ML} = 1.$$

Bayes Decision

- In the multiclass case ($m > 2$), class $\mathcal{C}_k$ is adopted if:

$$\Lambda_{kj}(\mathbf{x}) = \frac{p(\mathbf{x}|\mathcal{C}_k)}{p(\mathbf{x}|\mathcal{C}_j)} > T_{kj}, \qquad \forall j \neq k, \qquad j,k = 1, \ldots, m.$$

- or equivalently:

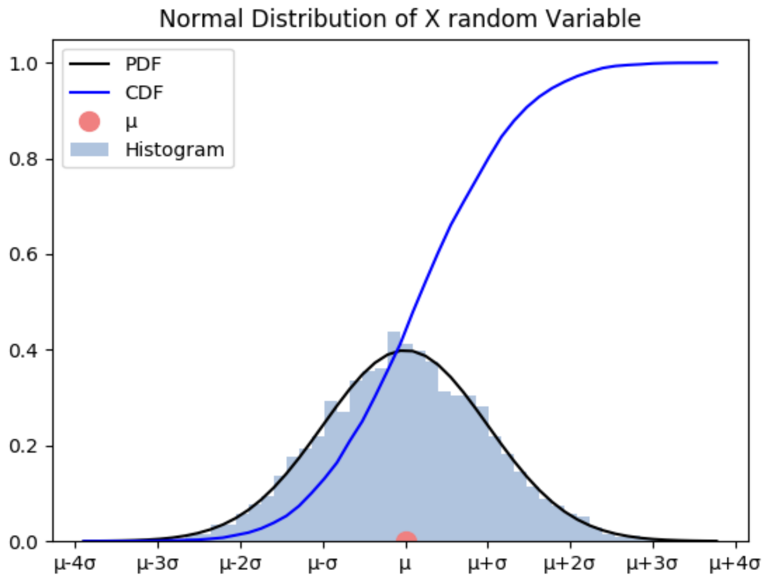
$$\ln \Lambda_{kj}(\mathbf{x}) = \ln p(\mathbf{x}|\mathcal{C}_k) - \ln p(\mathbf{x}|\mathcal{C}_k) > \ln T_{kj}\,.$$

- T_{kj} thresholds depends on the employed MAP/ML criterion, L_{kj} and $P(\mathcal{C}_j)$.

Gaussian Probability Distribution

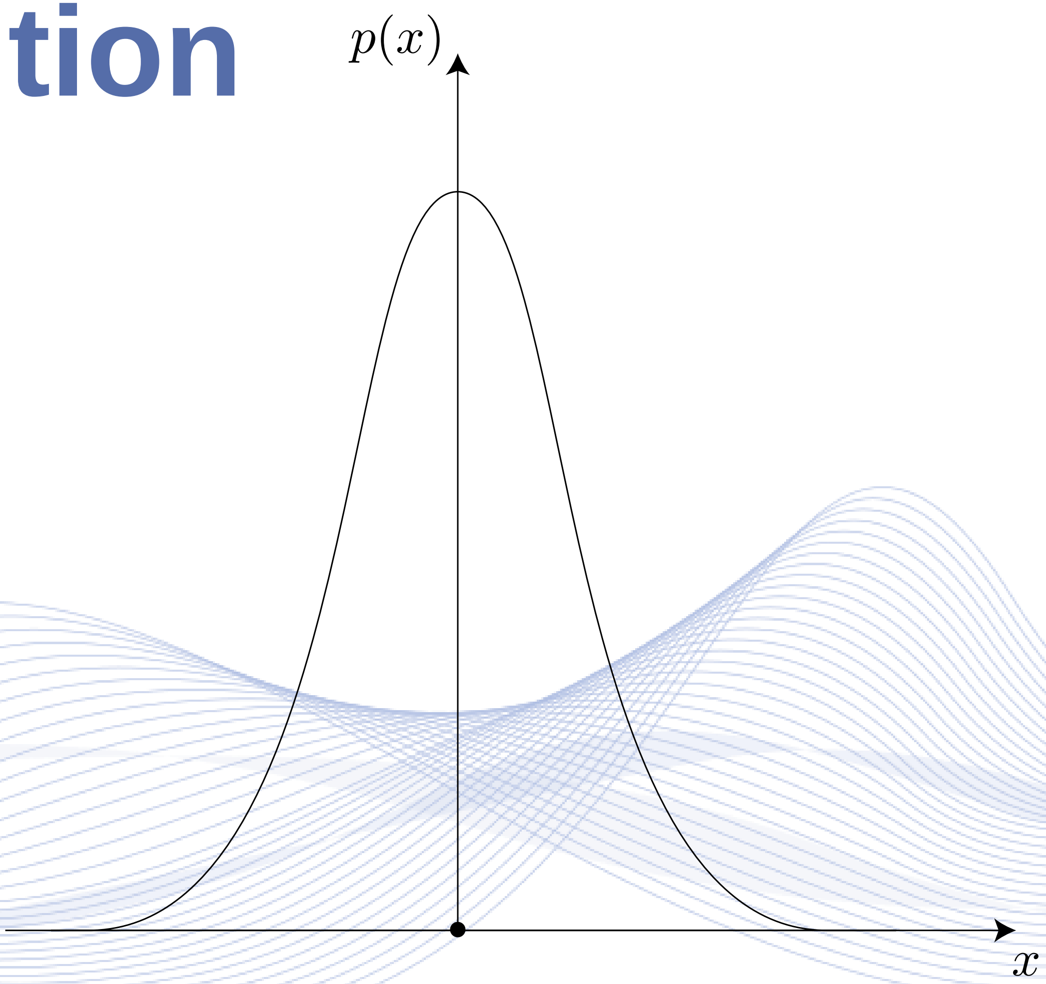
Normal (Gaussian) distribution $N(m, \sigma)$:

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{1}{2}\left(\frac{x-m}{\sigma}\right)^2}.$$



Artificial Intelligence & Information Analysis Lab

Gaussian Probability Distribution



1D Gaussian distribution $N(0,1)$.

Gaussian Probability Distribution

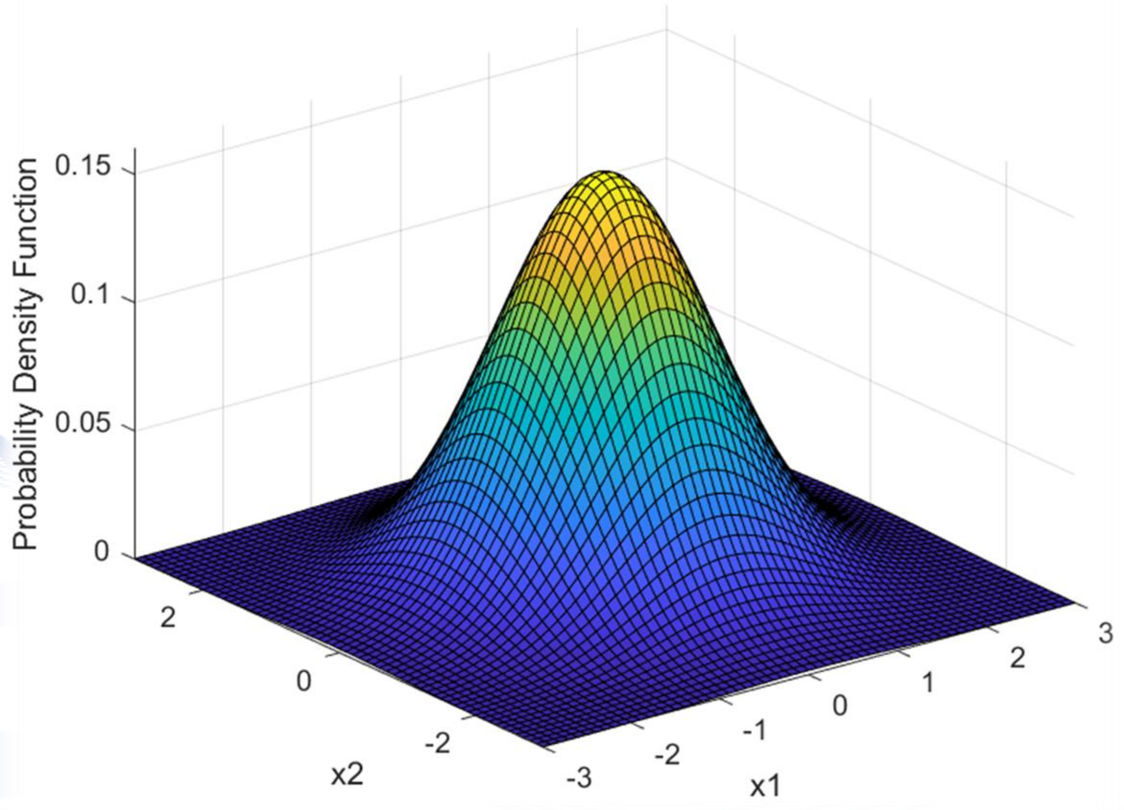
- Gaussian (normal) joint pdf $N(m_1, m_2, \sigma_1, \sigma_2, r)$:

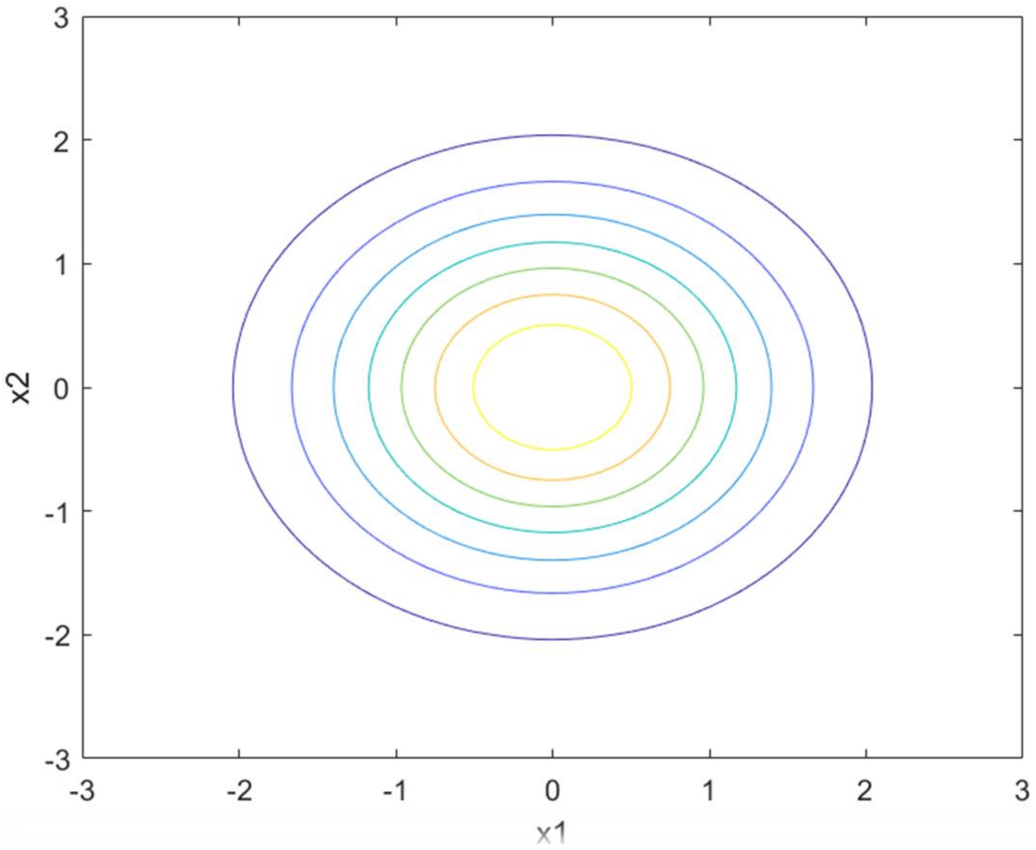
$$f_{XY}(x,y) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-r^2}} e^{A},$$

$$A = -\frac{1}{2(1-r^2)}\left(\left(\frac{x-m_1}{\sigma_1}\right)^2 + \left(\frac{y-m_2}{\sigma_2}\right)^2 - \frac{2r(x-m_1)(y-m_2)}{\sigma_1\sigma_2}\right)$$

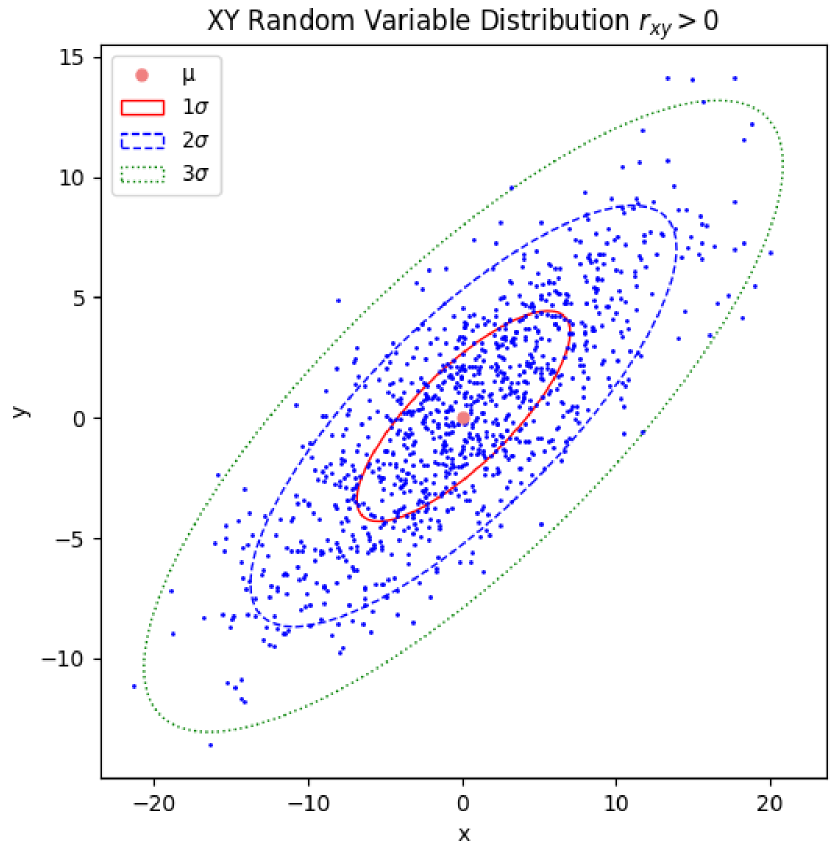
- r: correlation coefficient of X, Y.

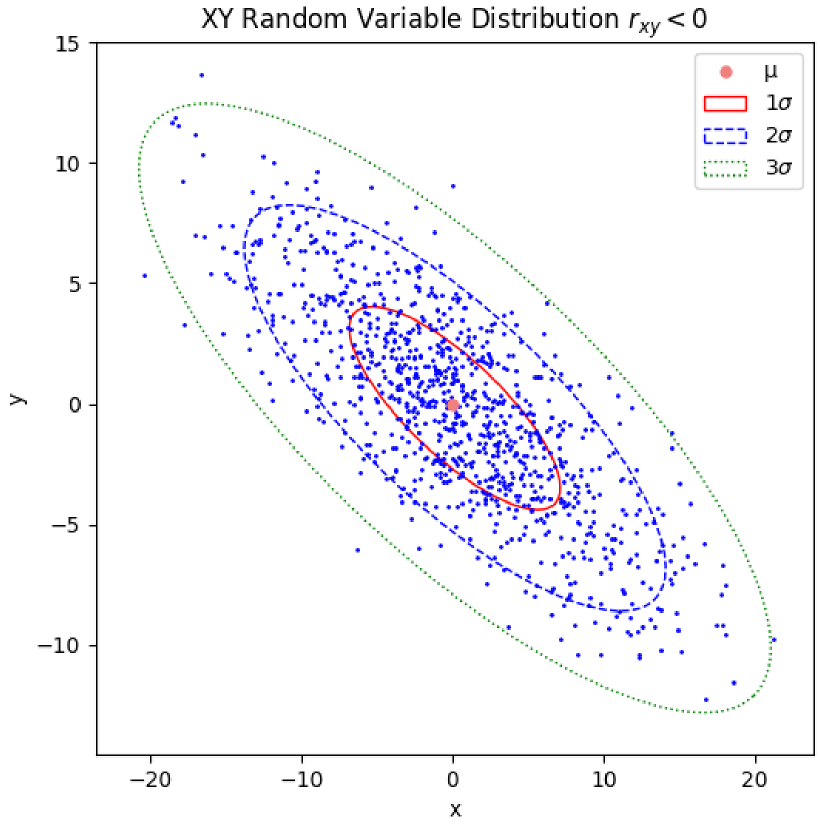
Gaussian Probability Density Function





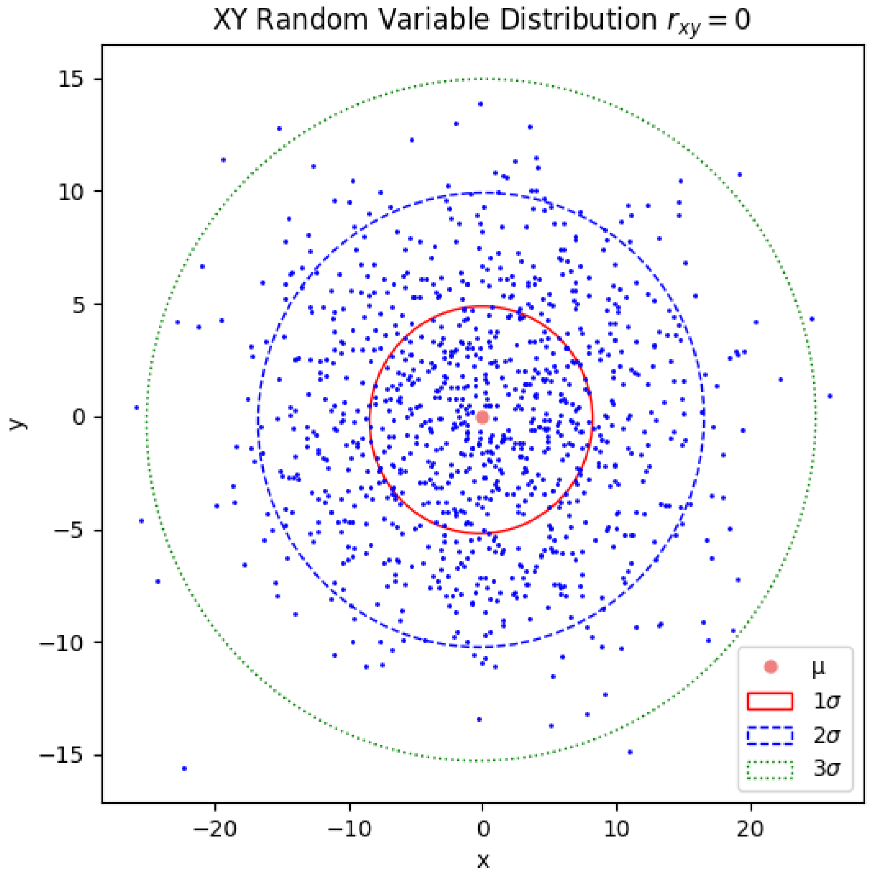
Gaussian Probability Distribution





2D Gaussian pdfs with positive and negative r_{XY}.

Gaussian Probability Distribution



2D Gaussian pdf with $r_{XY} = 0$.

Artificial Intelligence & Information Analysis Lab

Gaussian Probability Density Function

Gaussian (normal) random vectors:

- Variables $X_1, \ldots, X_n$ are jointly normal if:

$$f_{\mathbf{X}}(\mathbf{x}) = \frac{1}{(\sqrt{2\pi})^n \det(\mathbf{C})^{\frac{1}{2}}} e^A,$$

$$A = -\frac{1}{2}(\mathbf{x}-\mathbf{m})^T \mathbf{C}^{-1}(\mathbf{x}-\mathbf{m})$$

- Expected vector: $\mathbf{m} = E\{\mathbf{x}\}$.
- Covariance matrix: $\mathbf{C} = E\{(\mathbf{x}-\mathbf{m})^T(\mathbf{x}-\mathbf{m})\}$.
- $\det(\mathbf{C})$: determinant of $\mathbf{C}$.

Normally Distributed Sample Classification

- In the case where data sample $\mathbf{x} \in \mathbb{R}^n$ belonging to class $\mathcal{C}_k$ follow multivariate normal distribution $N(\mathbf{m}, \mathbf{C})$, we have:

$$\ln p(\mathbf{x}|\mathcal{C}_k) =$$

$$-\frac{1}{2}(\mathbf{x} - \mathbf{m}_k)^T \mathbf{C}_k^{-1}(\mathbf{x} - \mathbf{m}_k) - \frac{1}{2} n \ln(2\pi) - \frac{1}{2}\ln(\det(\mathbf{C}_k)) =$$

$$-\frac{1}{2}\mathbf{x}^T \mathbf{C}_k^{-1}\mathbf{x} + \frac{1}{2}\mathbf{x}^T \mathbf{C}_k^{-1}\mathbf{m}_k + \frac{1}{2}\mathbf{m}_k^T \mathbf{C}_k^{-1}\mathbf{x} - \frac{1}{2}\mathbf{m}_k^T \mathbf{C}_k^{-1}\mathbf{m}_k -$$

$$n \ln(2\pi) - \ln(\det(\mathbf{C}_k)).$$

Normally Distributed Sample Classification

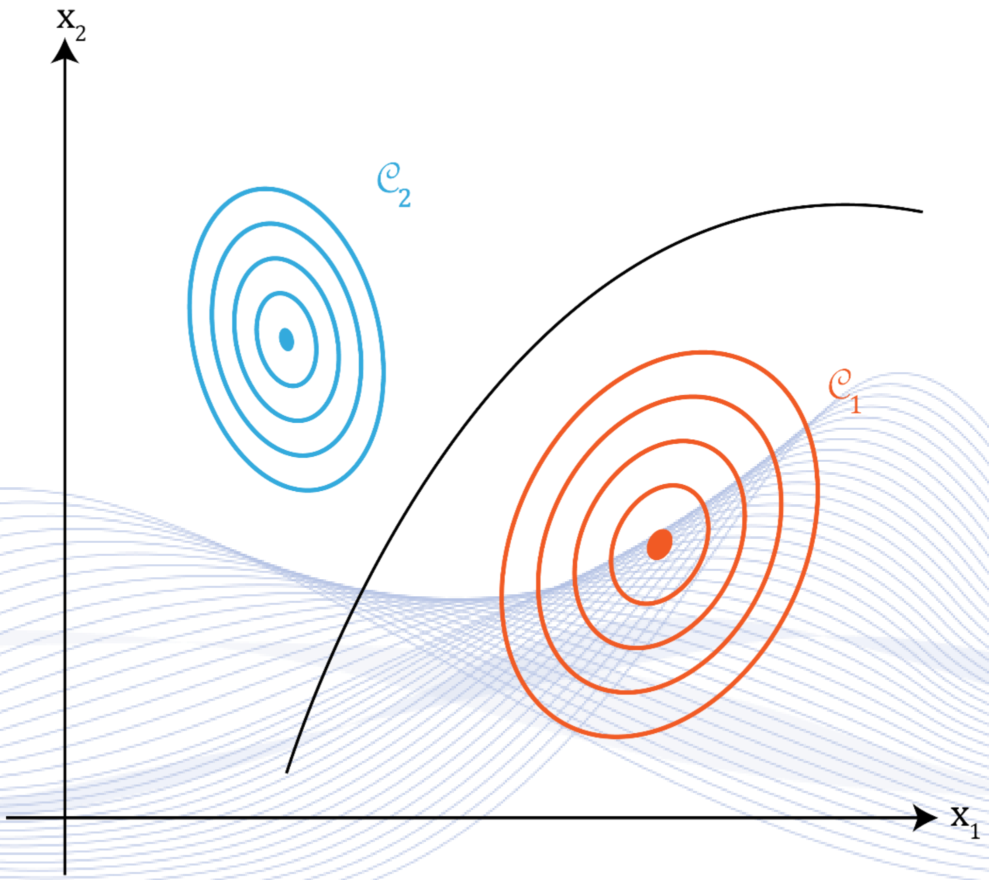
- Thus, $\mathcal{C}_k$ is adopted if:

$$-\mathbf{x}^T\mathbf{C}_k^{-1}\mathbf{x} + 2\mathbf{m}_k^T\mathbf{C}_k^{-1}\mathbf{x} > -\mathbf{x}^T\mathbf{C}_j^{-1}\mathbf{x} + 2\mathbf{m}_j^T\mathbf{C}_j^{-1}\mathbf{x} + b_{kj},$$

where:

$$b_{kj} = 2\ln \mathrm{T}_{kj} + \ln(\det(\mathbf{C}_k)) - \ln(\det(\mathbf{C}_j)) + \mathbf{m}_k^T\mathbf{C}_k^{-1}\mathbf{m}_k - \mathbf{m}_j^T\mathbf{C}_j^{-1}\mathbf{m}_j.$$

Normally Distributed Sample Classification



Second degree decision boundary for two 2D Gaussian pdfs.

Normally Distributed Sample Classification

- Thus, the optimal classification can be achieved by employing second degree hypersurfaces (e.g., hyper-ellipsoid, hyper-paraboloid, hyper-hyperboloid).
- Hyperplanes are optimal classification surfaces if all classes have ***same covariance matrix***: $\mathbf{C}_k = \mathbf{C}, k = 1, \dots, m$.
- Then first degree (linear) decision surface emerges (***perceptron***). Adopt $\mathcal{C}_k$ if:

$$\mathbf{a}_{kj}^T \mathbf{x} > f_{kj}.$$

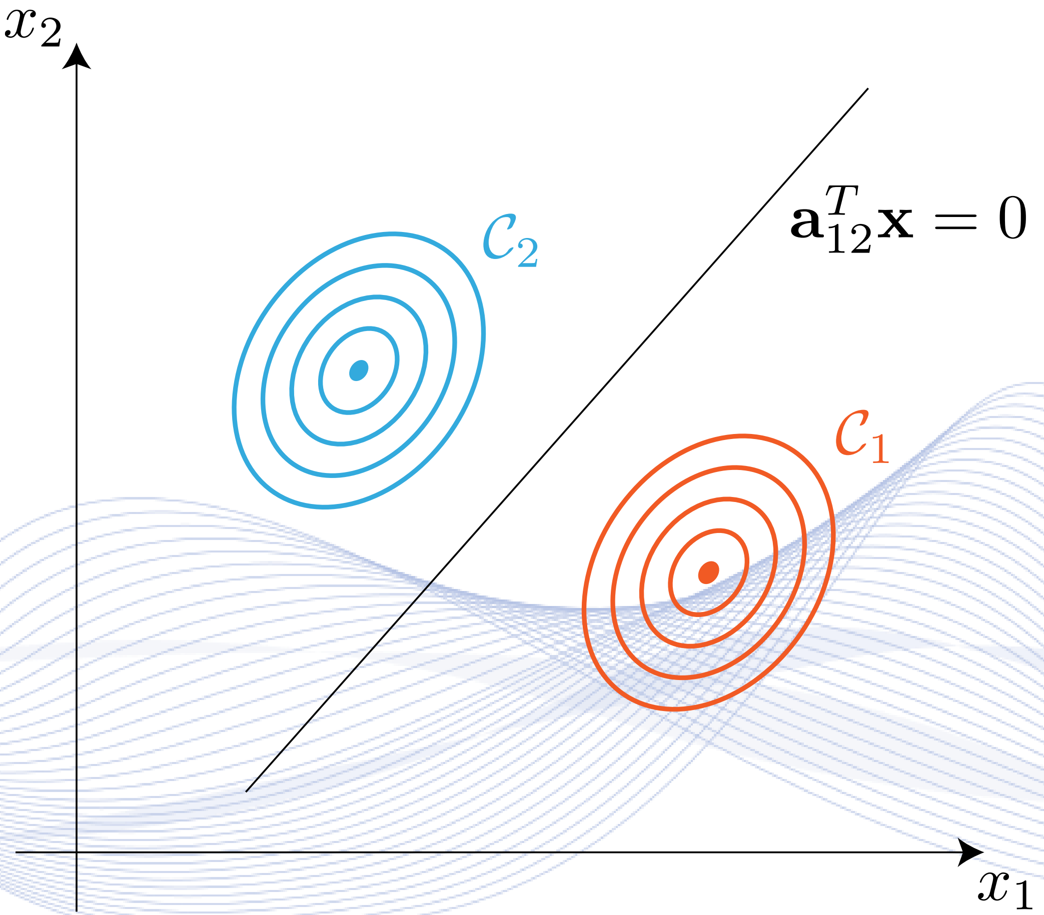
Normally Distributed Sample Classification

- Hyperplane coefficients:

$$\mathbf{a}_{kj}^T = \left(\mathbf{m}_k - \mathbf{m}_j\right)^T \mathbf{C}^{-1},$$

$$f_{kj} = \ln T_{kj} + \frac{1}{2}\mathbf{m}_k^T \mathbf{C}_k^{-1}\mathbf{m}_k - \frac{1}{2}\mathbf{m}_j^T \mathbf{C}_j^{-1}\mathbf{m}_j.$$

Normally Distributed Sample Classification



Linear decision boundary for two 2D Gaussian pdfs having equal $\mathbf{C}$.

Artificial Intelligence & Information Analysis Lab

Normally Distributed Sample Classification

Total classification error probability:

$$\begin{aligned} P_e &= P_{e_1} P(\mathcal{C}_2) + P_{e_2} P(\mathcal{C}_1) \\ &= P\{\mathbf{a}_{12}^T \mathbf{x} > f_{12} | \mathcal{C}_2\} P(\mathcal{C}_2) + P\{\mathbf{a}_{12}^T \mathbf{x} < f_{12} | \mathcal{C}_1\} P(\mathcal{C}_1). \end{aligned}$$

Normally Distributed Sample Classification

- If class $\mathcal{C}_2$ is the correct one, the quantity $\mathbf{a}_{12}^T\mathbf{x}$ has a multivariate normal distribution with expected vector:

$$m = \mathbf{a}_{12}^T E\{\mathbf{x}|\mathcal{C}_2\} = \mathbf{a}_{12}^T\mathbf{m}_2 = (\mathbf{m}_1 - \mathbf{m}_2)^T\mathbf{C}^{-1}\mathbf{m}_2,$$

and variance:

$$\sigma^2 = \mathbf{a}_{12}^T\mathbf{C}\,\mathbf{a}_{12} = (\mathbf{m}_1 - \mathbf{m}_2)^T\mathbf{C}^{-1}(\mathbf{m}_1 - \mathbf{m}_2).$$

Artificial Intelligence & Information Analysis Lab

Normally Distributed Sample Classification

- Error probability is calculated using the erf function:

$$P_{e_1} = P\{\mathbf{a}_{12}^T\mathbf{x} > f_{12} | \mathcal{C}_2\} = \int_{\tau}^{\infty} e^{-\frac{t^2}{2}} dt, \qquad \tau = \frac{f_{12} - \mathbf{a}_{12}^T\mathbf{m}_2}{\sigma}.$$

- If $T_{12} = 1$:

$$\tau = \frac{1}{2}\sqrt{(\mathbf{m}_1 - \mathbf{m}_2)^T\mathbf{C}^{-1}(\mathbf{m}_1 - \mathbf{m}_2)}.$$

Normally Distributed Sample Classification

- Error P_{e_1} is inversely proportional to the ***Mahalanobis distance*** between the two class centers $\mathbf{m}_1, \mathbf{m}_2$:

$$d(\mathbf{m}_1, \mathbf{m}_2) = \sqrt{(\mathbf{m}_1 - \mathbf{m}_2)^T \mathbf{C}^{-1} (\mathbf{m}_1 - \mathbf{m}_2)}.$$

- Error P_{e_2} can be found in the same way.

Normally Distributed Sample Classification

Special case:

- Gaussian classes having same diagonal covariance matrix with equal diagonal elements $\mathbf{C} = \sigma^2\mathbf{I}$.
- Then:

$$\ln P(\mathbf{x}|\mathcal{C}_k) = -\frac{1}{2}\mathbf{x}^T\mathbf{C}_k^{-1}\mathbf{x} + \frac{1}{2}\mathbf{x}^T\mathbf{C}_k^{-1}\mathbf{m}_k + \frac{1}{2}\mathbf{m}_k^T\mathbf{C}_k^{-1}\mathbf{x} - $$

$$-\frac{1}{2}\mathbf{m}_k^T\mathbf{C}_k^{-1}\mathbf{m}_k - n\ln(2\pi) - \ln(\det(\mathbf{C}_k)) = -\frac{1}{2\sigma^2}\mathbf{x}^T\mathbf{x} + \frac{1}{\sigma^2}\mathbf{m}_k^T\mathbf{x}$$

$$-\frac{1}{2\sigma^2}\mathbf{m}_k^T\mathbf{m}_k - n\ln(2\pi) - n\ln(\sigma^2).$$

Artificial Intelligence & Information Analysis Lab

Normally Distributed Samples Classification

- In this case, he decision hyperplane takes the form:

$$\ln p(\mathbf{x}|\mathcal{C}_k) + \ln P(\mathcal{C}_k) - \ln p(\mathbf{x}|\mathcal{C}_j) + \ln P(\mathcal{C}_j) =$$
$$= \mathbf{w}^T(\mathbf{x} - \mathbf{x}_0) = 0.$$

- Hyperplane parameters:

$$\mathbf{w} = \mathbf{m}_k - \mathbf{m}_j,$$

$$\mathbf{x}_0 = \frac{1}{2}(\mathbf{m}_k + \mathbf{m}_j) - \sigma^2 \ln\left(\frac{P(\mathcal{C}_k)}{P(\mathcal{C}_j)}\right)\frac{\mathbf{m}_k - \mathbf{m}_j}{\| \mathbf{m}_k - \mathbf{m}_j \|^2}.$$

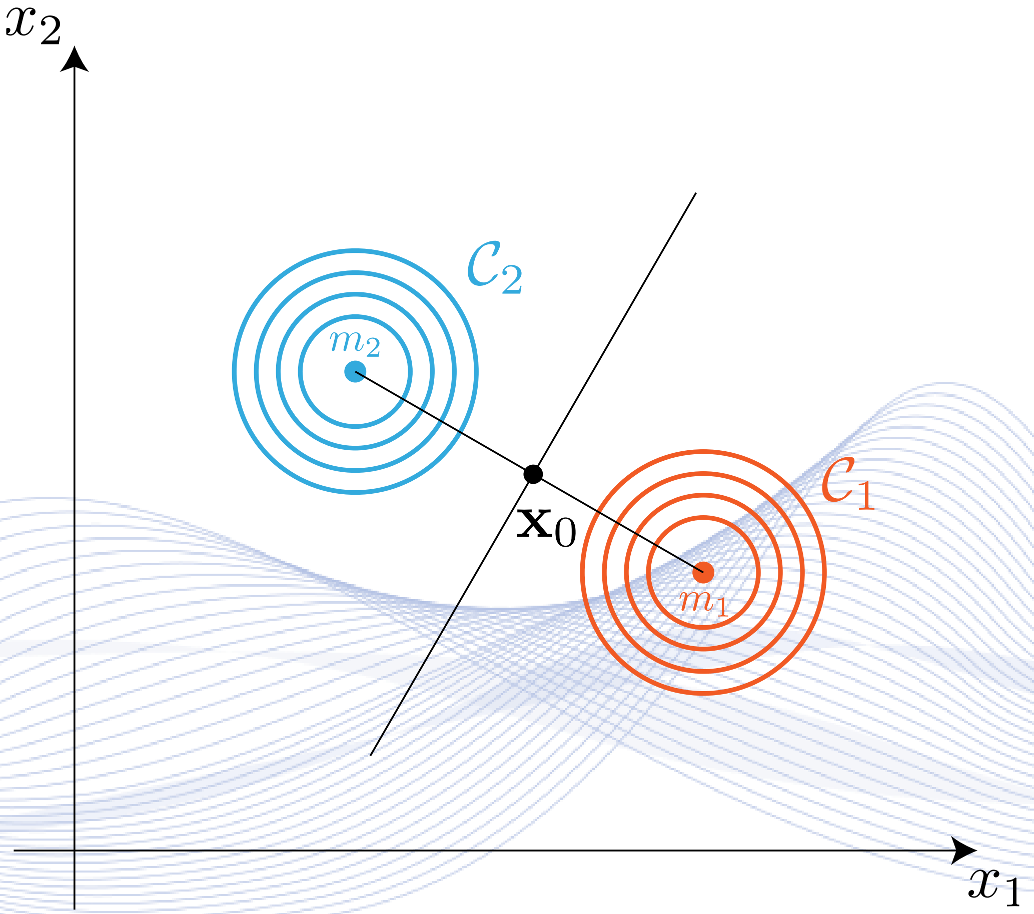
Normally Distributed Samples Classification

- Special cases:
  - If $P(\mathcal{C}_k) = P(\mathcal{C}_j)$, the decision hyperplane is the perpendicular bisector of the line segment connecting class centers $\mathbf{m}_k$ and $\mathbf{m}_j$:

$$\mathbf{x}_0 = \frac{1}{2}(\mathbf{m}_k + \mathbf{m}_j).$$

  - If $P(\mathcal{C}_k) \gg P(\mathcal{C}_j)$, the decision hyperplane approaches $\mathbf{m}_j$.
  - If $P(\mathcal{C}_i) \ll P(\mathcal{C}_j)$, the decision super-surface approaches $\mathbf{m}_k$.

Normally Distributed Samples Classification



Perpendicular bisector of two 2D Gaussian pdfs having equal $\mathbf{C} = \sigma^2\mathbf{I}$.

Artificial Intelligence & Information Analysis Lab

Normally Distributed Samples Classification

- Special cases:

  - In the two class case, if $P(\mathcal{C}_1) = P(\mathcal{C}_2) = 1/2$ and the two classes $\mathcal{C}_1, \mathcal{C}_2$ have 1D data $x \in \mathbb{R}$ follow Gaussian distributions $N(0, \sigma), N(1, \sigma)$ the decision threshold is given by:

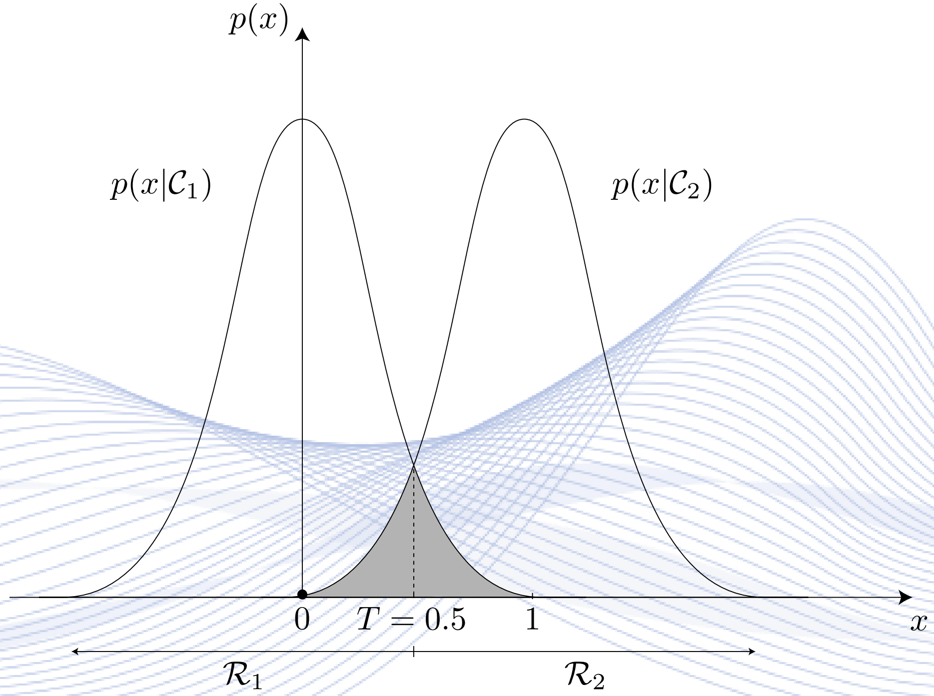
$$T = x_0 = 1/2.$$

  - This describes a routine ***modem*** operation in data communications.

Artificial Intelligence & Information Analysis Lab

Bayes Decision

In this special case, the costs $r_1(\mathbf{x}), r_2(\mathbf{x})$ are proportional to the possibility of the false adoption of the class $\mathcal{C}_1$, $\mathcal{C}_2$ respectively.



Binary Classifier for two classes having 1D pdfs $N(0,1), N(1,1)$.

Normally Distributed Samples Classification

Special case:

- Gaussian classes having same non-diagonal covariance matrix $\mathbf{C}$.
- A decision hyperplane results:

$$\mathbf{w}^T(\mathbf{x} - \mathbf{x}_0) = 0,$$

Having parameters

$$\mathbf{w} = \mathbf{C}^{-1}(\mathbf{m}_k - \mathbf{m}_j),$$

$$\mathbf{x}_0 = \frac{1}{2}(\mathbf{m}_k + \mathbf{m}_j) - \ln\left(\frac{P(\mathcal{C}_k)}{P(\mathcal{C}_j)}\right)\frac{\mathbf{m}_k - \mathbf{m}_j}{d(\mathbf{m}_k, \mathbf{m}_j)}.$$

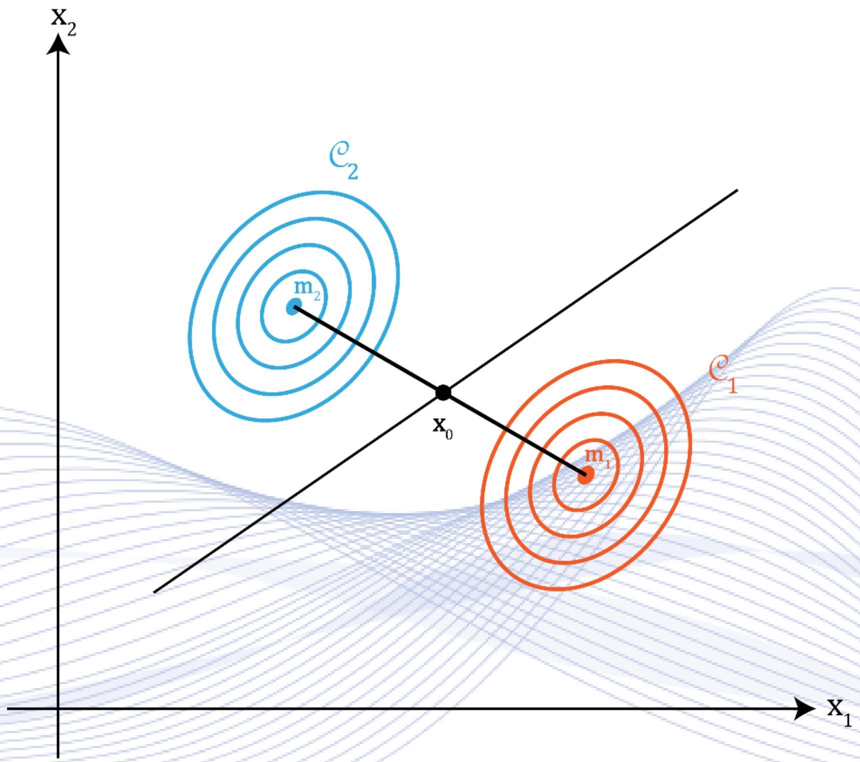
Normally Distributed Sample Classification

- where:

$$d(\mathbf{m}_k, \mathbf{m}_j) = \sqrt{\left(\mathbf{m}_k - \mathbf{m}_j\right)^T \mathbf{C}^{-1}\left(\mathbf{m}_k - \mathbf{m}_j\right)}.$$

is the Mahalanobis distance between the two class centers $\mathbf{m}_k, \mathbf{m}_j$.

Normally Distributed Samples Classification



Linear decision boundary or two equiprobable 2D Gaussian pdfs having equal $\mathbf{C}$.

Artificial Intelligence & Information Analysis Lab

Bayesian Learning

- Bayesian classification
- **Bayesian clustering**

Bayesian clustering

- It follows Bayesian philosophy for clustering.
- The data set must be partitioned in m clusters, $\mathcal{C}_j, j = 1, \dots, m$.
- Each vector $\mathbf{x}_i \in \mathbb{R}^n$, $i = 1, \dots, N$, belongs to a cluster $\mathcal{C}_j$ with probability $P(\mathcal{C}_j|\mathbf{x}_i)$.
- A vector $\mathbf{x}_i$ is assigned to a cluster $\mathcal{C}_k$ if:

$$P(\mathcal{C}_k|\mathbf{x}_i) > P\left(\mathcal{C}_j\middle|\mathbf{x}_i\right), \qquad j = 1, \dots, m, \qquad k \neq j.$$

Bayesian clustering

- Clustering using ***Expectation Maximization (EM)*** algorithm.
- E-step of EM algorithm.
  - Entropy functional to be optimized at iteration step t using an iterative algorithm:

$$E(\boldsymbol{\Theta}; \hat{\boldsymbol{\Theta}}_t) = \sum_{i=1}^{N} \sum_{j=1}^{m} P(\mathcal{C}_j | \mathbf{x}_i; \hat{\boldsymbol{\Theta}}_t) \ln \left(p(\mathbf{x}_i | \mathcal{C}_i ; \boldsymbol{\theta}) P(\mathcal{C}_j) \right).$$

Bayesian clustering

- $\boldsymbol{\theta} = [\boldsymbol{\theta}_1^T, \dots, \boldsymbol{\theta}_m^T]^T$, $\boldsymbol{\theta}_k$: the parameter vector corresponding cluster k.
- Typical case: $\boldsymbol{\theta}_k = \left[\mathbf{m}_k^T \mathbf{c}_k^T\right]^T$, containing the cluster k location and dispersion parameters.
- $\mathbf{P} = [P(\mathcal{C}_1), \dots, P(\mathcal{C}_m)]^T$ with $P(\mathcal{C}_k)$ the a priori probability for cluster k.
- $\boldsymbol{\Theta} = [\boldsymbol{\theta}^T, \mathbf{P}^T]^T$.

- M-step of the EM algorithm:

$$\hat{\boldsymbol{\Theta}}_{t+1} = \arg\ max_{\boldsymbol{\Theta}} E\left(\boldsymbol{\Theta}; \hat{\boldsymbol{\Theta}}_t\right).$$

Bayesian clustering

- Estimate $\boldsymbol{\theta}_j$ by cost function E differentiation:

$$\sum_{i=1}^{N}\sum_{j=1}^{m} P\left(\mathcal{C}_j \middle| \mathbf{x}_i; \widehat{\boldsymbol{\Theta}}_t\right)\frac{\partial}{\partial \boldsymbol{\theta}_j} \ln\left(p(\mathbf{x}_i|\mathcal{C}_i\,; \boldsymbol{\theta})\right) = 0.$$

- Maximization under constraints:

$$P\left(\mathcal{C}_j\right) \geq 0, \qquad j = 1, 2, \dots, m,$$

$$\sum_{j=1}^{m} P\left(\mathcal{C}_j\right) = 1.$$

Bayesian clustering

- Lagrangian function:

$$E_\lambda(\mathbf{P};\lambda) = E\left(\boldsymbol{\Theta};\widehat{\boldsymbol{\Theta}}_t\right) - \lambda\left(\sum_{j=1}^{m} P\left(\mathcal{C}_j\right) - 1\right).$$

- $\{\boldsymbol{\theta}_k, \boldsymbol{\theta}_j\}, k \neq j$ pairs are assumed to be independent.
- Setting partial derivatives of $E_\lambda(\mathbf{P};\lambda)$ with respect to $P\left(\mathcal{C}_j\right)$ equal to zero results in:

$$P\left(\mathcal{C}_j\right) = \frac{1}{\lambda}\sum_{i=1}^{N} P\left(\mathcal{C}_j \middle| \mathbf{x}_i;\widehat{\boldsymbol{\Theta}}_t\right), \qquad j = 1, 2, \dots, m.$$

Bayesian clustering

- By summing $P(\mathcal{C}_j), j = 1, \ldots, m,$ we obtain:

$$\lambda = \sum_{i=1}^{N} \sum_{j=1}^{m} P(\mathcal{C}_j | \mathbf{x}_i; \hat{\boldsymbol{\Theta}}_t) = N.$$

- And conclude that:

$$P(\mathcal{C}_j) = \frac{1}{N} \sum_{i=1}^{N} P(\mathcal{C}_j | \mathbf{x}_i; \hat{\boldsymbol{\Theta}}_t), j = 1, \ldots, m.$$

Bayesian clustering

- Suitable convergence criterion:

$$\left\|\hat{\boldsymbol{\Theta}}_{t+1} - \hat{\boldsymbol{\Theta}}_t\right\| < \epsilon.$$

where

- $\|.\|$: the appropriate vector norm.
- ϵ: a “small” user-defined constant.

Bayesian clustering

- Choose initial estimates at iteration $t=0$: $\boldsymbol{\theta}_0$ and $\mathbf{P}_0$.
- Repeat until convergence, with respect to $\boldsymbol{\Theta}$ is achieved:
  - Compute:

$$P(\mathcal{C}_k|\mathbf{x}_i;\widehat{\boldsymbol{\Theta}}_t)=\frac{p(\mathbf{x}_i|\mathcal{C}_k;\widehat{\boldsymbol{\theta}}_{kt})P(\mathcal{C}_k)_t}{\sum_{j=1}^{m}p(\mathbf{x}_i|\mathcal{C}_j;\widehat{\boldsymbol{\theta}}_{jt})P(\mathcal{C}_j)_t},\qquad i=1,\dots N,\qquad k=1,\dots,m.$$

Bayesian clustering

- Set $\hat{\boldsymbol{\theta}}_{jt+1}$ equal to the solution of:

$$\sum_{i=1}^{N}\sum_{j=1}^{m} P\left(\mathcal{C}_j \middle| \mathbf{x}_i; \hat{\boldsymbol{\Theta}}_t\right) \frac{\partial}{\partial \theta_j} \ln\left(p\left(\mathbf{x}_i \middle| \mathcal{C}_i ; \boldsymbol{\theta}_j\right)\right) = 0, \qquad j = 1, \dots, m.$$

  with respect to $\boldsymbol{\theta}_j$.
- Set:

$$P(\mathcal{C}_k)_{t+1} = \frac{1}{N}\sum_{i=1}^{N} P(\mathcal{C}_k|\mathbf{x}_i; \hat{\boldsymbol{\Theta}}_t), \qquad k = 1, \dots, m.$$

- Repeat for $t + 1$.

Gaussian Clusters

- Multivariate Gaussian cluster probability distribution:

$$p(\mathbf{x}|\mathcal{C}_k;\boldsymbol{\theta}_k) = \frac{1}{\sqrt{(2\pi)^n \det(\mathbf{C}_k)}} e^{-\frac{1}{2}(\mathbf{x}-\mathbf{m}_k)^T \mathbf{C}_k^{-1}(\mathbf{x}-\mathbf{m}_k)}, \quad k = 1, \dots, m.$$

- By taking the logarithm, we obtain:

$$\ln p(\mathbf{x}|\mathcal{C}_k;\boldsymbol{\theta}_k) = \ln \sqrt{\frac{\det(\mathbf{C}_k)}{(2\pi)^n}} - \frac{1}{2}(\mathbf{x}-\mathbf{m}_k)^T \mathbf{C}_k^{-1}(\mathbf{x}-\mathbf{m}_k),$$

$$k = 1, \dots, m.$$

Gaussian Clusters

- Each vector $\boldsymbol{\theta}_k = \left[\mathbf{m}_k^T \mathbf{c}_k^T\right]^T$ consists of n parameters for the location vector $\mathbf{m}_k$ the $\frac{n(n+1)}{2}$ independent parameters of the covariance matrix $\mathbf{C}_k$.
- Therefore, $\boldsymbol{\Theta}$ consists of $mn + m\frac{n(n+1)}{2}$ parameters.

$$P\left(\mathcal{C}_k \middle| \mathbf{x}_i; \widehat{\boldsymbol{\Theta}}_t\right) = \frac{\sqrt{\det(\mathbf{C}_{kt})}\, \mathrm{e}^{\left(-\frac{1}{2}(\mathbf{x}-\mathbf{m}_{kt})^T \mathbf{C}_{kt}^{-1}(\mathbf{x}-\mathbf{m}_{kt})\right)} P(\mathcal{C}_k)_t}{\sum_{j=1}^{m} \sqrt{\det(\mathbf{C}_{jt})}\, \mathrm{e}^{\left(-\frac{1}{2}(\mathbf{x}-\mathbf{m}_{jt})^T \mathbf{C}_{jt}^{-1}(\mathbf{x}-\mathbf{m}_{jt})\right)} P\left(\mathcal{C}_j\right)_t}.$$

Gaussian Clusters

- Updating equation for $\mathbf{m}_k$ and $\mathbf{C}_k$.
  - M-step of the EM algorithm:

$$\mathbf{m}_{k,t+1} = \frac{\sum_{j=1}^{N} P\left(\mathcal{C}_k \middle| \mathbf{x}_j; \widehat{\boldsymbol{\Theta}}_t\right)\mathbf{x}_j}{\sum_{j=1}^{N} P\left(\mathcal{C}_k \middle| \mathbf{x}_j; \widehat{\boldsymbol{\Theta}}_t\right)}.$$

- Cluster centers are weighted averages of cluster data points.

Gaussian Clusters

- Updating equation for $\mathbf{C}_k$.

$$\mathbf{C}_{k,t+1} = \frac{\sum_{j=1}^{N} P(\mathcal{C}_k|\mathbf{x}_j;\widehat{\mathbf{\Theta}}_t)(\mathbf{x}_j-\mathbf{m}_{kt})(\mathbf{x}_j-\mathbf{m}_{kt})^T}{\sum_{j=1}^{N} P(\mathcal{C}_j|\mathbf{x}_j;\widehat{\mathbf{\Theta}}_t)}, \quad k = 1, \dots, m.$$

Bayesian clustering

- The conditional probability $P(\mathcal{C}_j|\mathbf{x}_i)$ indicates how likely it is that $\mathbf{x}_i \in \mathbb{R}^n$ belongs to cluster $\mathcal{C}_j$, $i = 1, \dots, N$.
- The constraint:

$$\sum_{j=1}^{m} P(\mathcal{C}_j|\mathbf{x}_i) = 1$$

describes an $(m-1)$-dimensional hyperplane:

$$\mathbf{a}^T\mathbf{p} = 1,$$

- $\mathbf{p} = [P(\mathcal{C}_1|\mathbf{x}_i), \dots, P(\mathcal{C}_m|\mathbf{x}_i)]^T$.

A Geometrical Interpretation

- Since $0 \le P(\mathcal{C}_j|\mathbf{x}_i) \le 1, \;\; j = 1, \dots, m, \mathbf{p}$ lies inside the unit hypercube $[0,1]^m$.
- Noisy feature vectors or outliers:
  - Let $\mathbf{x}_i$ be such a vector.
  - At least one of $P(\mathcal{C}_j|\mathbf{x}_i), j = 1, \dots, m$ is significant and lies in the interval $\left[\frac{1}{m}, 1\right]$.
  - $\mathbf{x}_i$ will affect at least the estimates for the corresponding cluster $\mathcal{C}_j$,

Resulting in clustering sensitivity to outliers.

A Geometrical Interpretation

- Noisy feature vectors or outliers:
  - Let $\mathbf{x}_i$ be such a vector.
  - At least one of the y_i's $j = 1, \dots, m$ is significant and lies in the interval $\left[\frac{1}{m}, 1\right]$.
  - $\mathbf{x}_i$ will affect at least the estimates for the corresponding cluster C_j.
  - This makes GMDAS sensitive to outliers.

References

[STR1999] M.G. Strintzis, Pattern Recognition, 1999.

[THE2003] S. Theodoridis, K. Koutroumbas, Pattern Recognition, Elsevier, 2003.

Q & A

Thank you very much for your attention!

Contact: Prof. I. Pitas
pitas@csd.auth.gr