Multivariate Methods

Introduction to Machine Learning – GIF-7015

Professor: Christian Gagné

Week 3

UNIVERSITÉ LAVAL

3.1 Multivariate data

Multivariate data

- Parametric methods as seen last week $\Rightarrow$ estimating a variable X
  - In general, we measure several variables $\{X_1, X_2, \ldots, X_D\}$ for a data

$$\mathcal{X} = \{\mathbf{x}^t, r^t\}_{t=1}^N, \quad \mathbf{x}^t = [x_1^t \; x_2^t \; \cdots \; x_D^t]^\top$$

- Naming for variables (X_i)
  - Inputs
  - Features
  - Attributes
- Naming for data ($\mathbf{x}^t$)
  - Observations
  - Examples
  - Instances

Matrix representation:

$$\mathbf{X} = \begin{bmatrix} x_1^1 & x_2^1 & \cdots & x_D^1 \\ x_1^2 & x_2^2 & \cdots & x_D^2 \\ \vdots & \vdots & \ddots & \vdots \\ x_1^N & x_2^N & \cdots & x_D^N \end{bmatrix}$$

Means and variances, multivariate case

- Mean vector $\boldsymbol{\mu}$ defined as the mean of each column (each variable) of a set $\mathbf{X}$

$$\mathbb{E}[\mathbf{X}] = \boldsymbol{\mu} = [\mu_1 \ \mu_2 \ \cdots \ \mu_D]^\top$$

- Variance of a variable X_i is σ_i^2.
- Covariance of two variables X_i and X_j is noted $\sigma_{i,j}$

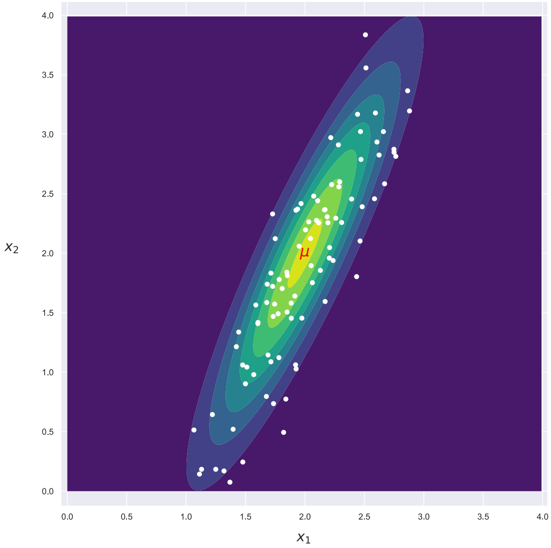
$$\sigma_{i,j} \equiv \mathrm{Cov}(X_i, X_j) = \mathbb{E}[(X_i - \mu_i)(X_j - \mu_j)] = \mathbb{E}[X_i X_j] - \mu_i \mu_j$$

- Covariance matrix $\boldsymbol{\Sigma}$
  - Symmetrical $D \times D$ matrix ($\sigma_{i,j} = \sigma_{j,i}$)
  - Positive values on the diagonal ($\sigma_{i,i} = \sigma_i^2$)

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & \sigma_{1,2} & \cdots & \sigma_{1,D} \\ \sigma_{1,2} & \sigma_2^2 & \cdots & \sigma_{2,D} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{1,D} & \sigma_{2,D} & \cdots & \sigma_D^2 \end{bmatrix}$$

$$\begin{aligned} \boldsymbol{\Sigma} \equiv \mathrm{Cov}(\mathbf{X}) &= \mathbb{E}\left[(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})^\top\right] \\ &= \mathbb{E}\left[\mathbf{X}\mathbf{X}^\top\right] - \boldsymbol{\mu}\boldsymbol{\mu}^\top \end{aligned}$$

Mean and covariance of samples



Estimator of means and variances, multivariate case

- Estimator of the mean based on maximum likelihood

$$\mathbf{m} = \frac{\sum_{t=1}^{N} \mathbf{x}^t}{N}, \text{ where } m_i = \frac{\sum_{t=1}^{N} x_i^t}{N}, \ i = 1, \ldots, D$$

- Let $\mathbf{S}$, the estimator of the covariance matrix $\boldsymbol{\Sigma}$

$$\mathbf{S} = \begin{bmatrix} s_1^2 & s_{1,2} & \cdots & s_{1,D} \\ s_{1,2} & s_2^2 & \cdots & s_{2,D} \\ \vdots & \vdots & \ddots & \vdots \\ s_{1,D} & s_{2,D} & \cdots & s_D^2 \end{bmatrix}$$

$$s_i^2 = \frac{\sum_{t=1}^{N} (x_i^t - m_i)^2}{N}$$

$$s_{i,j} = \frac{\sum_{t=1}^{N} (x_i^t - m_i)(x_j^t - m_j)}{N}$$

  - Developing equations for $\mathbf{S}$ is complex, it requires the application of the spectral theorem

Correlation

- Correlation between variables X_i and X_j

$$\mathrm{Corr}(X_i, X_j) \equiv \rho_{i,j} = \frac{\sigma_{i,j}}{\sigma_i \sigma_j}$$

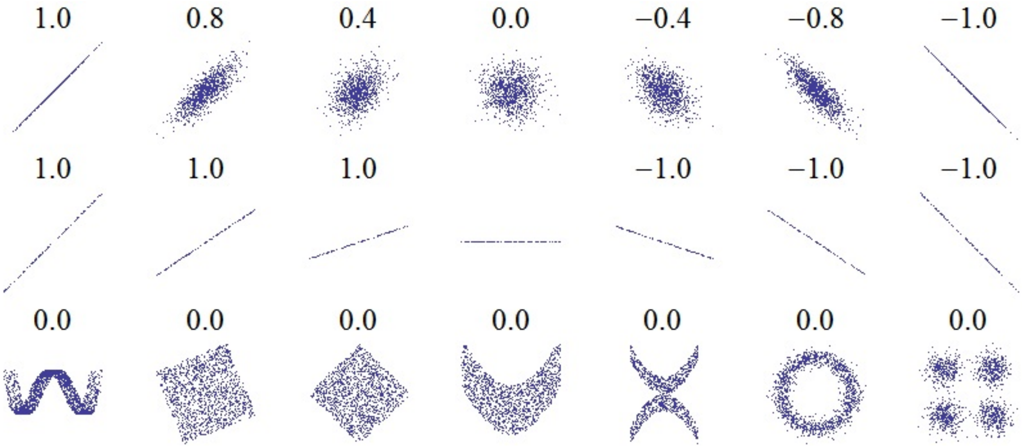
  - Standardized statistical measure, $-1 \leq \rho_{i,j} \leq 1$
  - Two independent variables X_i and X_j $\Rightarrow$ zero correlation
  - The inverse is, however, not true, even if $\rho_{i,j} = 0$, variables X_i and X_j are not necessarily independent (non-linear relation between variables)

- Estimation of correlation

$$r_{i,j} = \frac{s_{i,j}}{s_i s_j}$$

- Matrix **R** is the matrix of the correlation estimator containing the $r_{i,j}$

Correlation and non-linearity



Source: `http://en.wikipedia.org/wiki/File:Correlation_examples.png`

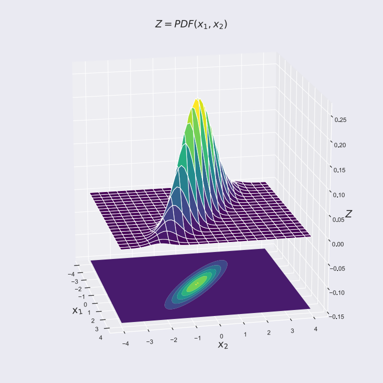
3.2 Multivariate normal distribution

Multivariate normal distribution

- Multidimensional normal distribution $\mathcal{N}_D(\boldsymbol{\mu},\boldsymbol{\Sigma})$

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{0.5D}\,|\boldsymbol{\Sigma}|^{0.5}} \exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right]$$

- Mean vector $\boldsymbol{\mu}$: distribution centre
- Normalization by the inverse of the covariance matrix $\boldsymbol{\Sigma}$



Two-dimensional case

- Two-dimensional normal distribution ($\sigma_{i,j} = \rho\sigma_i\sigma_j$):

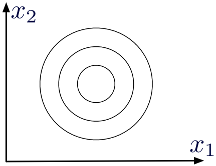
$$\boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}$$

- Four possible cases for $\boldsymbol{\Sigma}$
  1. Diagonal $\boldsymbol{\Sigma}$ ($\rho = 0$) and equal variance for both dimensions (isotropic), $\sigma_1^2 = \sigma_2^2 = \sigma^2$
  2. Diagonal $\boldsymbol{\Sigma}$ ($\rho = 0$) and different variances for the two dimensions, $\sigma_1^2 \neq \sigma_2^2$
  3. Positive correlation between variables, $\rho > 0$
  4. Negative correlation between variables, $\rho < 0$

Two-dimensional case

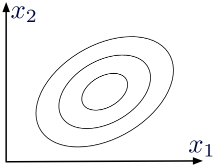
$$\sigma_1^2 = \sigma_2^2 = \sigma^2,\ \rho = 0$$

$$\begin{bmatrix} \sigma^2 & 0 \\ 0 & \sigma^2 \end{bmatrix}$$



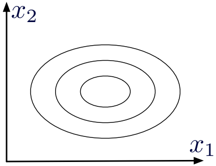
$$\rho > 0$$

$$\begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}$$



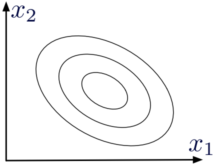
$$\sigma_1^2 > \sigma_2^2,\ \rho = 0$$

$$\begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix}$$



$$\rho < 0$$

$$\begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}$$

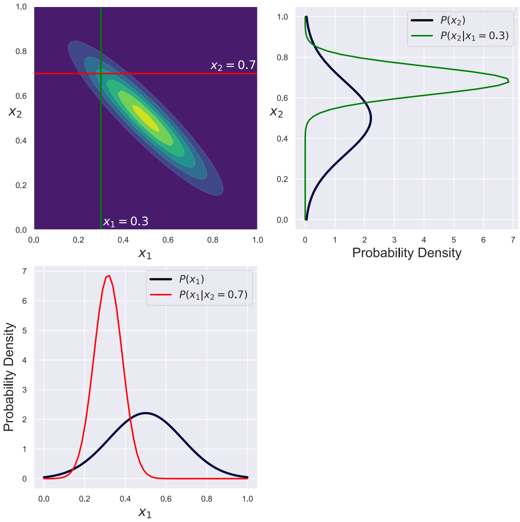


Properties of the multivariate normal distribution

- The value of the determinant $|\boldsymbol{\Sigma}|$ indicates the proximity of samples around $\boldsymbol{\mu}$
  - A low value may indicate high correlation between variables
- Generally, $\boldsymbol{\Sigma}$ is a symmetrical positive-definite matrix
  - Otherwise, $\boldsymbol{\Sigma}$ is singular and $|\boldsymbol{\Sigma}| = 0$
    - $\Rightarrow$ Linear dependence between variables
    - $\Rightarrow$ One of the variables has a variance of 0
- If $\mathbf{x} \sim \mathcal{N}_D(\boldsymbol{\mu},\boldsymbol{\Sigma})$ then $x_i \sim \mathcal{N}(\mu_i,\tilde{\sigma}_i^2)$
  - If x_i are independent ($\sigma_{i,j} = 0,\ \forall i \neq j$), then $x_i \sim \mathcal{N}(\mu_i,\sigma_i^2)$
- A linear projection defined by $\mathbf{W}$ in a space with K dimensions ($K < D$) also follows a multivariate normal distribution

$$\mathbf{W}^\top \mathbf{x} \sim \mathcal{N}_K\left(\mathbf{W}^\top \boldsymbol{\mu},\mathbf{W}^\top \boldsymbol{\Sigma}\mathbf{W}\right)$$

Conditional law of multivariate normal distribution

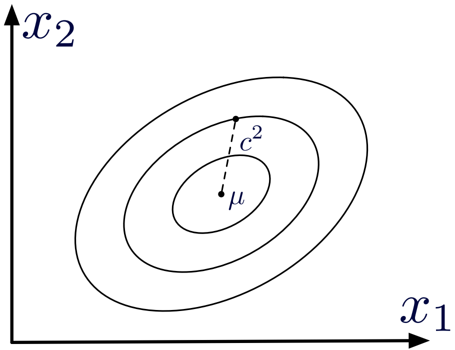


Mahalanobis distance

- Mahalanobis distance

$$D_M(\mathbf{x}) = (\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})$$

  - Distance between the mean vector $\boldsymbol{\mu}$ and a point $\mathbf{x}$, weighted by the covariance matrix $\boldsymbol{\Sigma}$.
  - Contour line corresponds to a constant distance c^2



- 1D case

$$\frac{(x-\mu)^2}{\sigma^2} = (x-\mu)(\sigma^2)^{-1}(x-\mu)$$

3.3 Multivariate classification

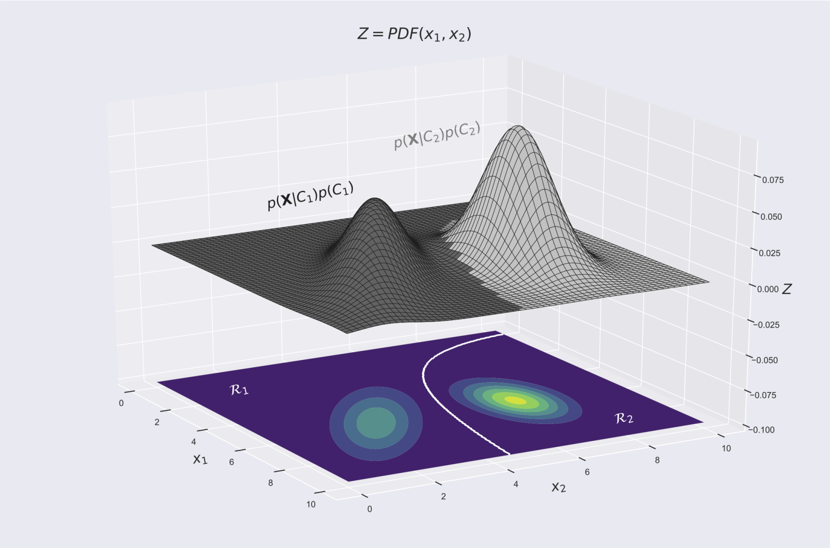
Multivariate classification

- Conditional probability density for classes $p(\mathbf{x}|C_i) \sim \mathcal{N}_D(\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i)$

$$p(\mathbf{x}|C_i) = \frac{1}{(2\pi)^{0.5D}\,|\boldsymbol{\Sigma}_i|^{0.5}} \exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^\top \boldsymbol{\Sigma}_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i)\right]$$

- Reasons for using normal distribution in multivariate classification
  - Simplicity of the equation for analytical developments
  - Model that describes many natural phenomena accurately
    - Observations are generally slight variations ($\boldsymbol{\Sigma}$) of a mean observation ($\boldsymbol{\mu}$)
    - Robust model, allows good approximations
  - However, requires data to be grouped together
    - With several groups, we must use a *mixture distribution*, which is a linear combination of several densities (presented later)

Example of multivariate classification



Discriminant function

- Discriminant function with multivariate model

$$\mathrm{h}_i(\mathbf{x}) = \log p(\mathbf{x}|C_i) + \log P(C_i)$$

- For a normal distribution, $p(\mathbf{x}|C_i) \sim \mathcal{N}_D(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$

$$\begin{aligned}
\mathrm{h}_i(\mathbf{x}) &= -\frac{D}{2}\log 2\pi - \frac{1}{2}\log|\boldsymbol{\Sigma}_i| - \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^\top \boldsymbol{\Sigma}_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i) + \log P(C_i) \\
&= -\frac{1}{2}\log|\boldsymbol{\Sigma}_i| - \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^\top \boldsymbol{\Sigma}_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i) + \log P(C_i)
\end{aligned}$$

Parameters estimate

- Parameters estimate based on maximum likelihood
  - Set $\mathcal{X} = \{\mathbf{x}^t, \mathbf{r}^t\}_{t=1}^N$, with $r_i^t = \begin{cases} 1 & \text{if } \mathbf{x}^t \in C_i \\ 0 & \text{otherwise} \end{cases}$

$$\begin{aligned} \hat{P}(C_i) &= \frac{\sum_t r_i^t}{N} \\ \mathbf{m}_i &= \frac{\sum_t r_i^t \mathbf{x}^t}{\sum_t r_i^t} \\ \mathbf{S}_i &= \frac{\sum_t r_i^t (\mathbf{x}^t - \mathbf{m}_i)(\mathbf{x}^t - \mathbf{m}_i)^\top}{\sum_t r_i^t} \end{aligned}$$

Quadratic discriminant function

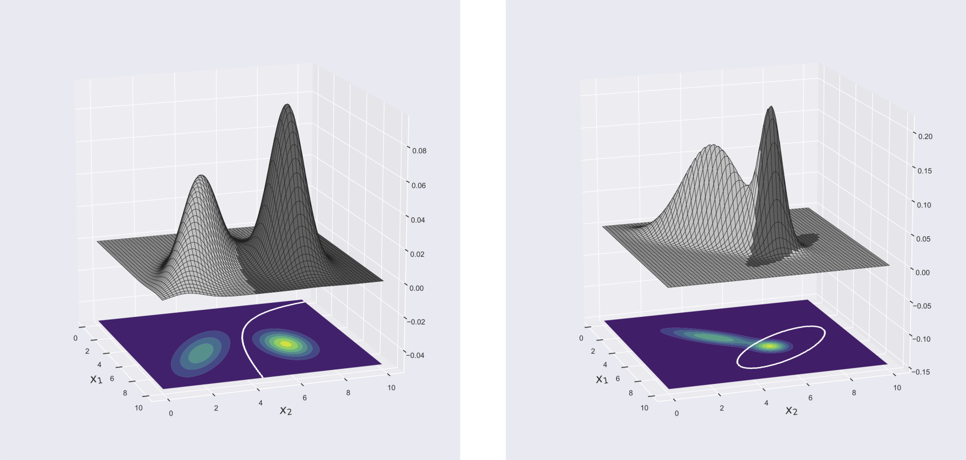
- Include $\hat{P}(C_i)$, $\mathbf{m}_i$ and $\mathbf{S}_i$ into the formula of $\mathrm{h}_i(\mathbf{x})$

$$\mathrm{h}_i(\mathbf{x}) = -\frac{1}{2}\log|\mathbf{S}_i| - \frac{1}{2}(\mathbf{x}-\mathbf{m}_i)^\top \mathbf{S}_i^{-1}(\mathbf{x}-\mathbf{m}_i) + \log \hat{P}(C_i)$$

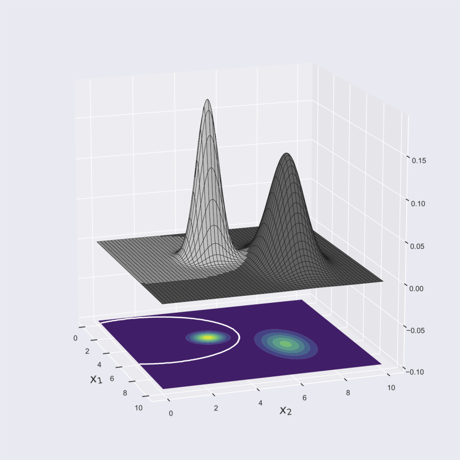
- Equivalent formulation

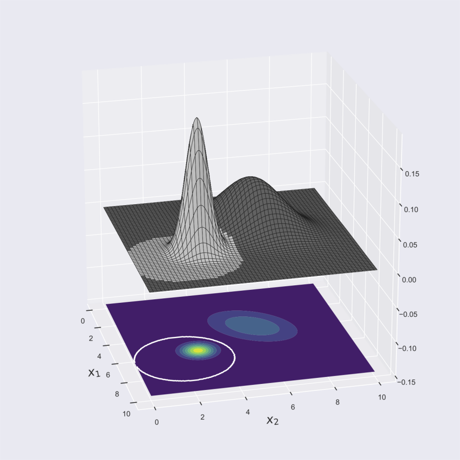
$$\begin{aligned}
\mathrm{h}_i(\mathbf{x}) &= \mathbf{x}^\top \mathbf{W}_i \mathbf{x} + \mathbf{w}_i^\top \mathbf{x} + w_i^0 \\
\mathbf{W}_i &= -\frac{1}{2}\mathbf{S}_i^{-1} \\
\mathbf{w}_i &= \mathbf{S}_i^{-1}\mathbf{m}_i \\
w_i^0 &= -\frac{1}{2}\mathbf{m}_i^\top \mathbf{S}_i^{-1}\mathbf{m}_i - \frac{1}{2}\log|\mathbf{S}_i| + \log \hat{P}(C_i)
\end{aligned}$$

Examples of quadratic discriminant function (1/3)

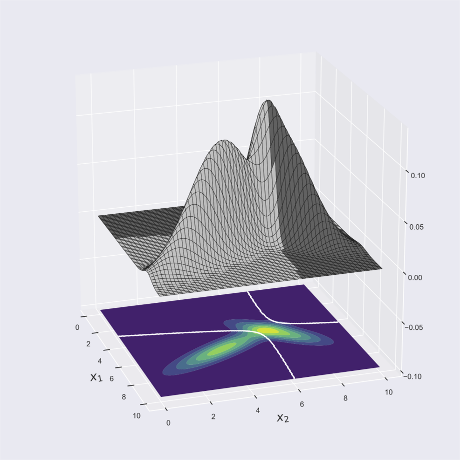


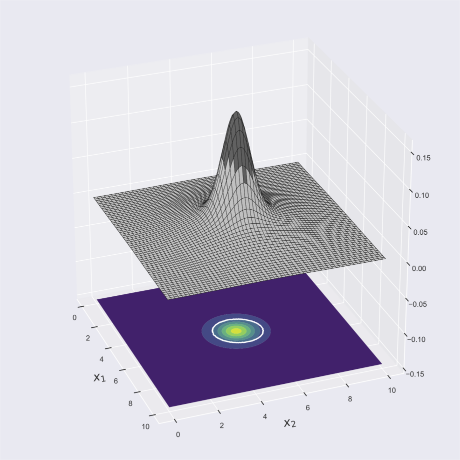
Examples of quadratic discriminant function (2/3)





Examples of quadratic discriminant function (3/3)

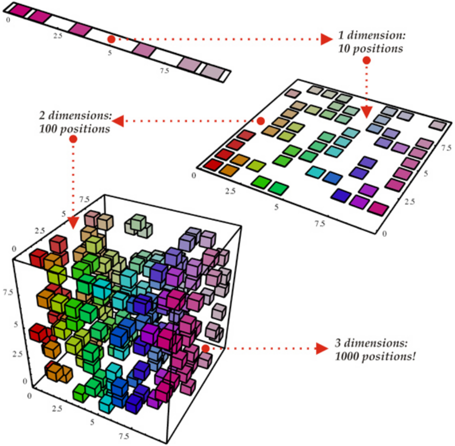




The curse of dimensionality

- The curse of dimensionality
  - The addition of a dimension creates an exponential increase of the mathematical space
  - If 100 points are equidistant from 0.01 in one dimension $\Rightarrow 10^{20}$ points are needed in 10 dimensions to keep the same sampling density
- High number of parameters to be estimated with quadratic discriminant function
  - $K \times D$ for means and $K \times \frac{D(D+1)}{2}$ for covariance matrices
- With a high dimensionality (large D) and few data (small N), high risk of singular matrices $\mathbf{S}_i$
  - Even if $|\mathbf{S}_i| \neq 0$, a small change can cause a large variation of $\mathbf{S}_i^{-1} \Rightarrow$ instabilities
- Solution: dimensionality reduction by feature selection or projection (seen at the end of the semester)

The curse of dimensionality



Source: Y. Bengio, `http://www.iro.umontreal.ca/~bengioy/yoshua_en/research_files/CurseDimensionality.jpg`, accessed October 2, 2016.

3.4 Model simplifications for classification

Sharing the covariance matrix

- Simplification 1: sharing the covariance matrix

$$\mathbf{S} = \sum_t \hat{P}(C_i)\, \mathbf{S}_i$$

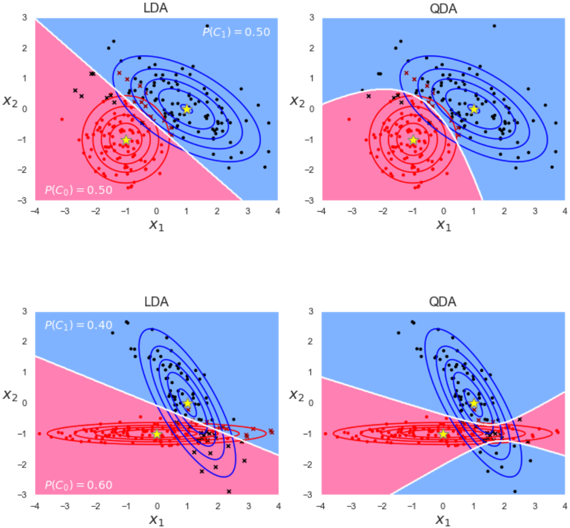
  - $K \times D$ parameters for means
  - $\frac{D(D+1)}{2}$ parameters for shared covariance matrix
  - Corresponding discriminant function

$$\mathrm{h}_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x} - \mathbf{m}_i)^\top \mathbf{S}^{-1}(\mathbf{x} - \mathbf{m}_i) + \log \hat{P}(C_i)$$

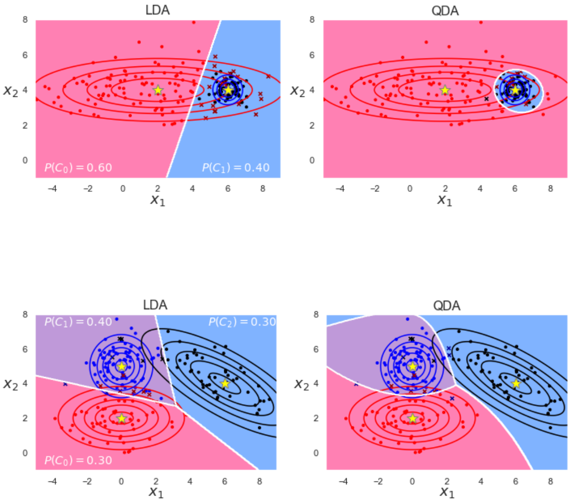
    - $\mathbf{x}^\top \mathbf{S}^{-1}\mathbf{x}$ common for all $\mathrm{h}_i(\mathbf{x})$
  - Reformulation as a linear discriminant function

$$\mathrm{h}_i(\mathbf{x}) = \mathbf{w}_i^\top \mathbf{x} + w_i^0, \quad \mathbf{w}_i = \mathbf{S}^{-1}\mathbf{m}_i, \quad w_i^0 = -\frac{1}{2}\mathbf{m}_i^\top \mathbf{S}^{-1}\mathbf{m}_i + \log \hat{P}(C_i)$$

Linear and quadratic discriminant functions (1/2)



Linear and quadratic discriminant functions (2/2)



Naive Bayes classifier

- Simplification 2: elements out of the diagonal of $\mathbf{S}$ have a value of 0

$$\mathbf{S} = \begin{bmatrix} s_1^2 & 0 & \cdots & 0 \\ 0 & s_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & s_D^2 \end{bmatrix}$$

- Corresponding discriminant function (naive Bayes classifier)

$$\mathrm{h}_i(\mathbf{x}) = -\frac{1}{2}\sum_{j=1}^{D}\left(\frac{x_j - m_{i,j}}{s_j}\right)^2 + \log \hat{P}(C_i)$$

- Number of parameters for the covariance matrix: D
  - Reduction from a quadratic to a linear order

Nearest mean classifier

- Simplification 3: isotropic covariance matrix, with all variances equal ($\sigma_i = \sigma, \forall i$)
- Reduction from a Mahalanobis distance to a Euclidean distance

$$(\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) = \sigma^{-2} (\mathbf{x} - \boldsymbol{\mu})^\top (\mathbf{x} - \boldsymbol{\mu}) = \frac{\|\mathbf{x} - \boldsymbol{\mu}\|^2}{\sigma^2}$$

- Corresponding discriminant function

$$\mathrm{h}_i(\mathbf{x}) = -\frac{\|\mathbf{x} - \boldsymbol{\mu}_i\|^2}{2s^2} + \log \hat{P}(C_i) = -\frac{1}{2s^2} \sum_{j=1}^{D} (x_j - m_{i,j})^2 + \log \hat{P}(C_i)$$

- Simplification 4: a priori equal probabilities ($P(C_i) = P(C_j), \forall i,j$)
  - Nearest mean classifier

$$\mathrm{h}_i(\mathbf{x}) = -\|\mathbf{x} - \mathbf{m}_i\|^2$$

Nearest mean classifier

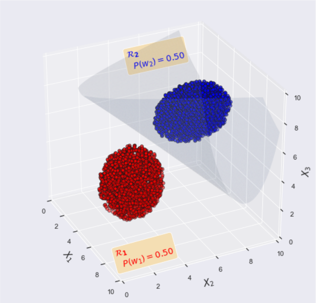
$$
\begin{aligned}
\mathrm{h}_i(\mathbf{x}) &= -\|\mathbf{x}-\mathbf{m}_i\|^2 \\
&= -(\mathbf{x}-\mathbf{m}_i)^\top(\mathbf{x}-\mathbf{m}_i) \\
&= -(\mathbf{x}^\top\mathbf{x} - 2\mathbf{m}_i^\top\mathbf{x} + \mathbf{m}_i^\top\mathbf{m}_i)
\end{aligned}
$$

As $\mathbf{x}^\top\mathbf{x}$, common $\forall \mathrm{h}_i(\mathbf{x})$

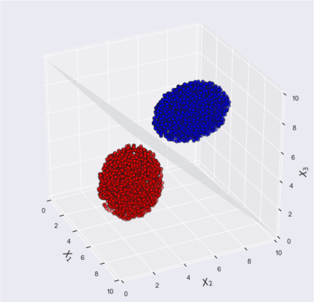
$$
\begin{aligned}
\mathrm{h}_i(\mathbf{x}) &= \mathbf{w}_i^\top\mathbf{x} + w_i^0 \\
\mathbf{w}_i &= \mathbf{m}_i \\
w_i^0 &= -\frac{1}{2}\|\mathbf{m}_i\|^2
\end{aligned}
$$

3D examples with simplifications

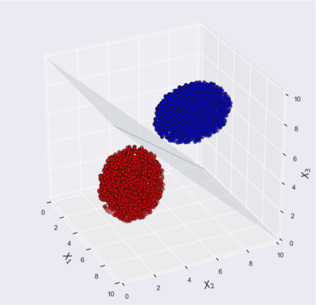
Quadratic



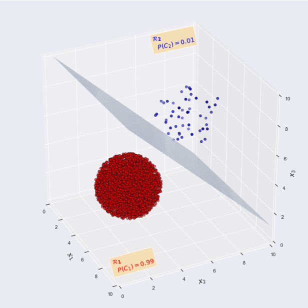
Linear



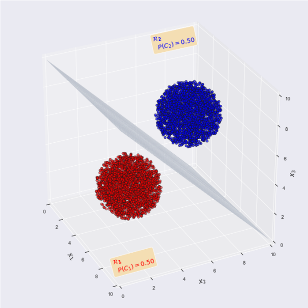
Nearest mean



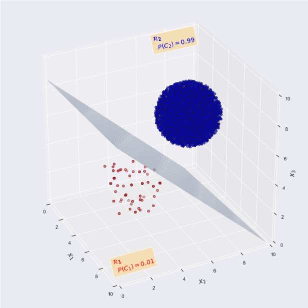
Effect of the a priori probabilities (linear case)



$P(C_1) = 0.99,\ P(C_2) = 0.01$



$P(C_1) = 0.50,\ P(C_2) = 0.50$



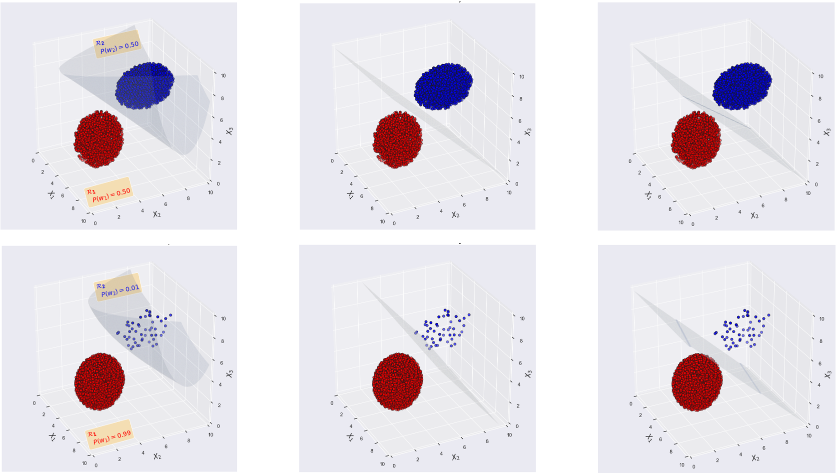
$P(C_1) = 0.01,\ P(C_2) = 0.99$

Effect of the a priori probabilities

Quadratic | Linear | Nearest mean

$P(C_1) = P(C_2) = 0.5$

$P(C_1) = 0.99$
$P(C_2) = 0.01$



Summary of variants

Densities	Covariance matrices	Number of parameters
shared $\mathbf{\Sigma}$, hypersphere densities (isotropic)	$\mathbf{\Sigma}_i = \mathbf{\Sigma} = \sigma^2\mathbf{I}$	1
shared $\mathbf{\Sigma}$, densities aligned on the axes	$\mathbf{\Sigma}_i = \mathbf{\Sigma}$ and $\sigma_{i,j} = 0$	D
shared $\mathbf{\Sigma}$, hyperellipsoidal densities	$\mathbf{\Sigma}_i = \mathbf{\Sigma}$	$\frac{D(D+1)}{2}$
different $\mathbf{\Sigma}$, hyperellipsoidal densities	$\mathbf{\Sigma}_i$	$K\frac{D(D+1)}{2}$

Discriminant analysis with regularization

- Rewriting of the covariance matrix

$$\boldsymbol{\Sigma}_i' = \alpha\sigma^2\mathbf{I} + \beta\boldsymbol{\Sigma} + (1-\alpha-\beta)\boldsymbol{\Sigma}_i$$

  - $\alpha = \beta = 0 \Rightarrow$ quadratic discriminant
  - $\alpha = 0$ and $\beta = 1 \Rightarrow$ linear discriminant with shared covariance matrix
  - $\alpha = 1$ and $\beta = 0 \Rightarrow$ linear discriminant with shared isotropic covariance matrix (nearest mean classifier if a priori probabilities are equal)
  - Variety of classifiers with α and β between these extreme values
- Possible regularization by an optimization criterion taking into account the values of α and β.

3.5 Mixture distribution

Mixture distribution

- Parametric classification with normal distribution: one group per class
  - With several modes in a single class, a normal distribution model is difficult to apply
- Mixture distribution: linear combination of density functions associated with several groups

$$p(\mathbf{x}) = \sum_{i=1}^{K} p(\mathbf{x}|\mathcal{G}_i)P(\mathcal{G}_i)$$

  - Groups must be known and identified in the data
  - Alternative: use an unsupervised approach (*clustering*) to learn the groups
- Mixture distribution of components based on a multivariate normal distribution
  - Component density: $(\mathbf{x}|\mathcal{G}_i) \sim \mathcal{N}_D(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$
  - Parametrization: $\Phi = \{P(\mathcal{G}_i), \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\}_{i=1}^{K}$

Probabilities for mixture distribution

- Mixture distribution

$$p(\mathbf{x}) = \sum_{i=1}^{K} p(\mathbf{x}|\mathcal{G}_i)P(\mathcal{G}_i)$$

- Proportion of the group $\mathcal{G}_i$ in the mixture, $P(\mathcal{G}_i)$

$$\sum_i P(\mathcal{G}_i) = 1$$

- Probability that $\mathbf{x}$ belongs to the group $\mathcal{G}_i$, $P(\mathcal{G}_i|\mathbf{x})$

$$P(\mathcal{G}_i|\mathbf{x}) = \frac{P(\mathcal{G}_i)p(\mathbf{x}|\mathcal{G}_i)}{\sum_j P(\mathcal{G}_j)p(\mathbf{x}|\mathcal{G}_j)}$$

3.6 Multivariate regression

Multivariate regression

- Model for a multivariate linear regression function

$$r^t = \mathrm{h}(\mathbf{x}|w_0,w_1,\ldots,w_D) + \epsilon = w_0 + w_1 x_1^t + w_2 x_2^t + \cdots + w_D x_D^t + \epsilon$$

- White Gaussian noise centred at 0, $\epsilon \sim \mathcal{N}(0,\sigma^2)$
- Minimization of the quadratic error (maximum likelihood)

$$E(w_0,w_1,\ldots,w_D|\mathcal{X}) = \frac{1}{2}\sum_t \left(r^t - w_0 - w_1 x_1^t - w_2 x_2^t - \cdots - w_D x_D^t\right)^2$$

- Solution based on partial derivatives

$$\frac{\partial E}{\partial w_j} = 0,\ \forall j$$

Normal equations for multivariate regression

$$
\begin{aligned}
\sum_t r^t &= Nw_0 + w_1 \sum_t x_1^t + w_2 \sum_t x_2^t + \cdots + w_D \sum_t x_D^t \\
\sum_t x_1^t r^t &= Nw_0 \sum_t x_1^t + w_1 \sum_t (x_1^t)^2 + w_2 \sum_t x_1^t x_2^t + \cdots + w_D \sum_t x_1^t x_D^t \\
\sum_t x_2^t r^t &= Nw_0 \sum_t x_2^t + w_1 \sum_t x_1^t x_2^t + w_2 \sum_t (x_2^t)^2 + \cdots + w_D \sum_t x_2^t x_D^t \\
&\vdots \\
\sum_t x_D^t r^t &= Nw_0 \sum_t x_D^t + w_1 \sum_t x_1^t x_D^t + w_2 \sum_t x_2^t x_D^t + \cdots + w_D \sum_t (x_D^t)^2
\end{aligned}
$$

- Matrix version: $\mathbf{X}^\top \mathbf{r} = \mathbf{X}^\top \mathbf{X} \mathbf{w}$

$$
\mathbf{X} = \begin{bmatrix} 1 & x_1^1 & x_2^1 & \cdots & x_D^1 \\ 1 & x_1^2 & x_2^2 & \cdots & x_D^2 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_1^N & x_2^N & \cdots & x_D^N \end{bmatrix}, \mathbf{w} = \begin{bmatrix} w_0 \\ w_1 \\ \vdots \\ w_D \end{bmatrix}, \mathbf{r} = \begin{bmatrix} r^1 \\ r^2 \\ \vdots \\ r^N \end{bmatrix}
$$

- Solving the system of linear equations

$$
\mathbf{w} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{r}
$$

Notes on multivariate regression

- Normal equations: polynomials of order 1.
  - Resolution with higher order polynomials is rare, except for low D
- Analysis by inspection of w_i values
  - w_i gives the importance of the variable X_i, it allows to classify the variables by order of importance
    - Remove the variables where $w_i \to 0$
    - Interesting for dimensionality reduction (will be seen at the end of the semester)
  - Sign of w_i gives an idea of the effect of the variable X_i.
- Multiple output values $\Rightarrow$ set of independent regression problems