Lecture 2: Linear Classification

Machine Learning, Summer Term 2019

Michael Tangermann Frank Hutter Marius Lindauer

University of Freiburg

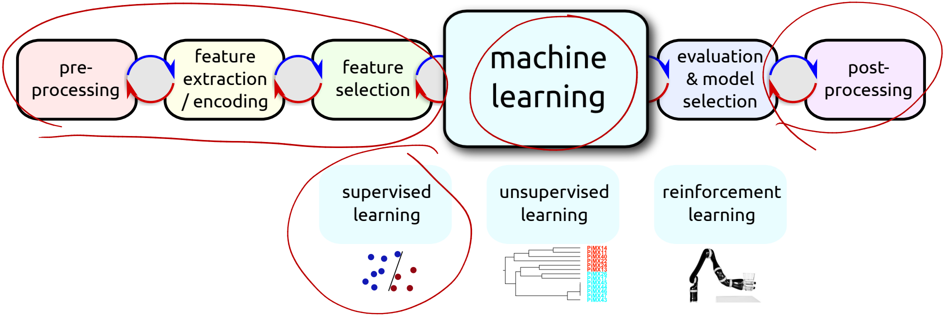
Lecture Overview

1. Brief Recapitulation: Supervised Classification
2. Motivation: Linear Discriminant Functions
3. LDA Assumptions and Data Scenarios
4. Geometric Interpretation
5. How to Derive the Parameters...
6. Wrapup: Summary, Related Topics, Preview

Lecture Overview

1. Brief Recapitulation: Supervised Classification
2. Motivation: Linear Discriminant Functions
3. LDA Assumptions and Data Scenarios
4. Geometric Interpretation
5. How to Derive the Parameters...
6. Wrapup: Summary, Related Topics, Preview

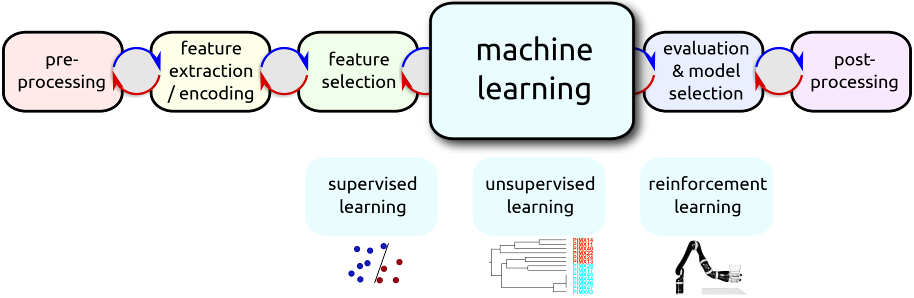
Reminder: ML Design Cycle



Linear discriminant functions are for supervised classification:

- Use past experience to predict the future $x_i \in \mathbb{R}^D$

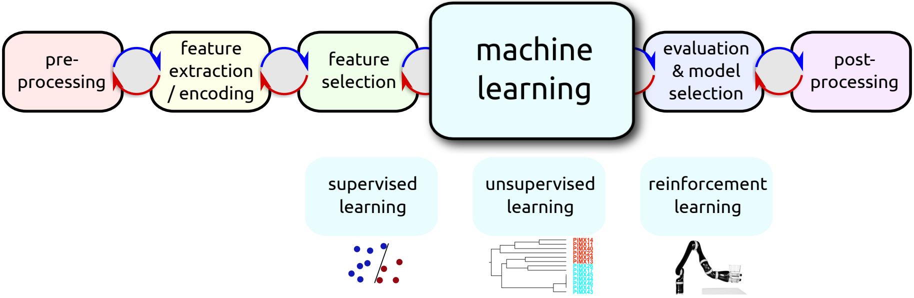
Reminder: ML Design Cycle



Linear discriminant functions are for supervised classification:

- Use past experience to predict the future
- Use labelled data points $\langle(\underline{\mathbf{x}_i}, \underline{y_i})\rangle_{i=1}^{N}$

Reminder: ML Design Cycle



Linear discriminant functions are for supervised classification:

- Use past experience to predict the future
- Use labelled data points $\langle(\mathbf{x}_i, y_i)\rangle_{i=1}^{N}$
- Train a **model** which can predict the label y_{N+1} of a new data point $\mathbf{x}_{N+1}$

Reminder: A Simple Classification Example

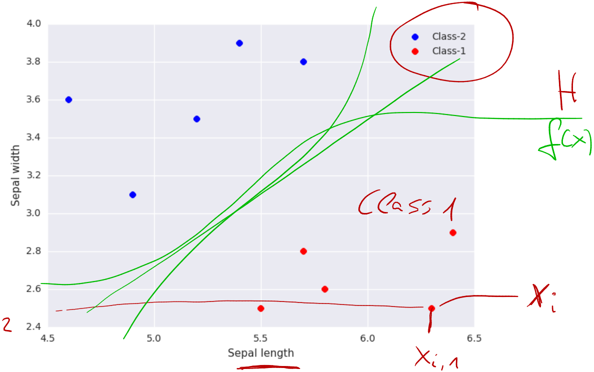
- A classical data set from Botany: classifying Iris flowers
  - feature 1: sepal length
  - feature 2: sepal width

$x_i = \begin{pmatrix} \cdot \\ \cdot \\ \cdot \end{pmatrix}$

features   class label

$\mathbf{x}_{i,1}$	$\mathbf{x}_{i,2}$	y_i
6.40	2.90	2
5.50	2.50	2
5.20	3.50	1
4.60	3.60	1
5.70	3.80	1
6.30	2.50	2
5.80	2.60	2
4.90	3.10	1
5.70	2.80	2
5.40	3.90	1

X



Reminder: Terminology

- A data point $\mathbf{x}_i$ is a column vector in $\mathbb{R}^D$
- A label y is discrete, e.g. $y \in \{0,1\}$

We are given a labeled data set of N examples:

- $\mathbf{X}$ is a $N \times D$ matrix containing the continuous feature values. ($\mathbf{X}$ contains one transposed data point $\mathbf{x}_i^T$ per row.)
- $\mathbf{y}$ is a $N \times 1$ vector containing the discrete labels.

Classification: The Workflow

(omitting the subscripts "$_i$" for data points for convenience...)

- Learn a function $f(\mathbf{x})$, which shall separate the two classes as *good* as possible.

Classification: The Workflow

(omitting the subscripts "$_i$" for data points for convenience...)

- Learn a function $f(\mathbf{x})$, which shall separate the two classes as *good* as possible.
- Once f(x) has been learned you can make decisions on a novel input vector $\mathbf{x}$:
  - assign $\mathbf{x}$ to class $\mathcal{C}_1$ if $f(\mathbf{x}) \geq 0$
  - assign $\mathbf{x}$ to class $\mathcal{C}_2$ otherwise

Classification: The Workflow

(omitting the subscripts "$_i$" for data points for convenience...)

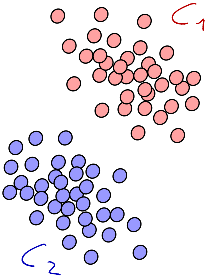
- Learn a function $f(\mathbf{x})$, which shall separate the two classes as *good* as possible.
- Once f(x) has been learned you can make decisions on a novel input vector $\mathbf{x}$:
  - assign $\mathbf{x}$ to class $\mathcal{C}_1$ if $f(\mathbf{x}) \geq 0$
  - assign $\mathbf{x}$ to class $\mathcal{C}_2$ otherwise

Thus the corresponding decision boundary H is defined by the relation $f(\mathbf{x}) = 0$.

Lecture Overview

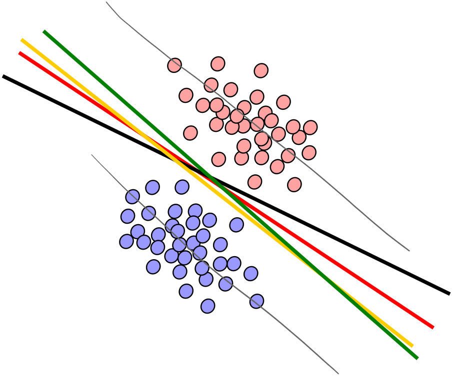
Linear Discriminant Function: The Basic Idea

Example of two linearly separable classes with data points $\mathbf{x} \in \mathbb{R}^2$.
Which is the best decision boundary? Please vote.



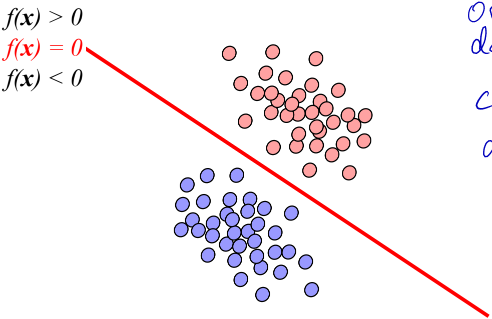
Example of two linearly separable classes with data points $\mathbf{x} \in \mathbb{R}^2$.

Which is the best decision boundary? Please vote.



Linear Discriminant Function: The Basic Idea

What is your intuition based on? 



distances (x_i, H)

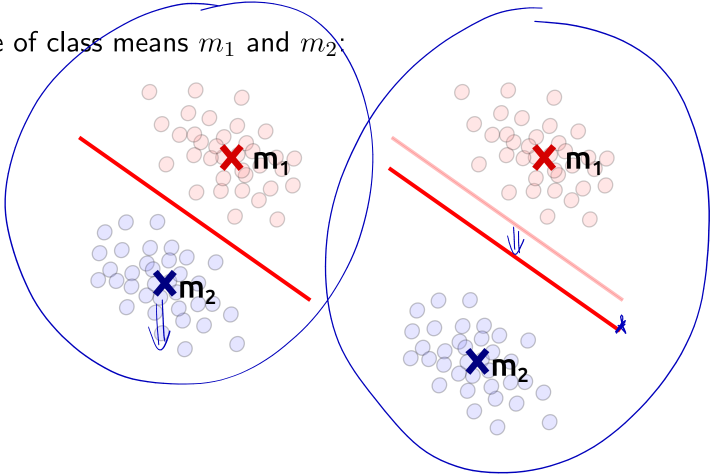
orientation of data distributions

class means

density

What influences the Decision Boundary?

Influence of class means m_1 and m_2:

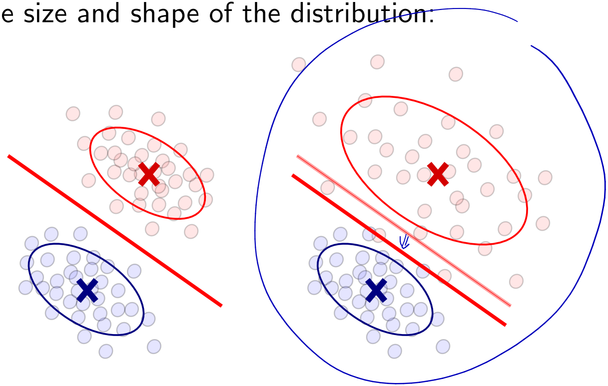


With N_k points in class $\mathcal{C}_k$:

$$\mathbf{m}_1 = \frac{1}{N_1} \sum_{n \in \mathcal{C}_1} \mathbf{x}_n \quad \text{and} \quad \mathbf{m}_2 = \frac{1}{N_2} \sum_{n \in \mathcal{C}_2} \mathbf{x}_n$$

What influences the Decision Boundary?

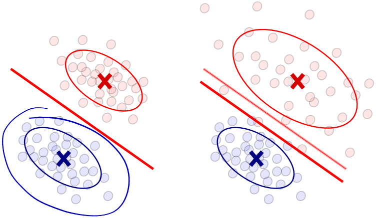
Influence of the size and shape of the distribution:



How can a (normal) distribution be described? 

Gaussian

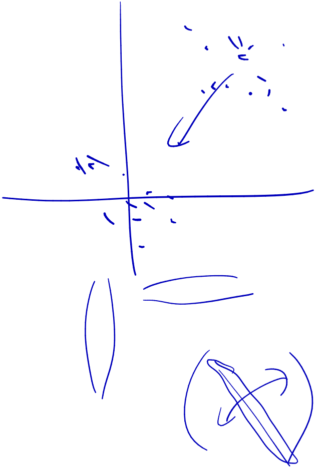
Reminder: Covariance Matrix



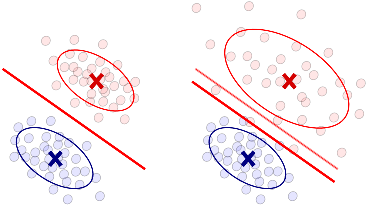
$\mathbf{S}_k$ is the covariance matrix of class $\mathcal{C}_k$:

$\mathbf{S}_k = \frac{1}{N_k - 1} \sum_{n \in \mathcal{C}_k} (\mathbf{x}_n - \mathbf{m}_k)(\mathbf{x}_n - \mathbf{m}_k)^T$

(symmetric and positive semi-definite)



Reminder: Covariance Matrix



$\mathbf{S}_k$ is the covariance matrix of class $\mathcal{C}_k$:

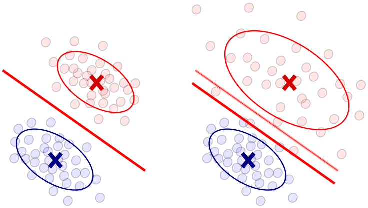
$\mathbf{S}_k = \frac{1}{N_k - 1} \sum_{n \in \mathcal{C}_k} (\mathbf{x}_n - \mathbf{m}_k)(\mathbf{x}_n - \mathbf{m}_k)^T$

(symmetric and positive semi-definite)

Assuming equally large classes, the total within-class covariance matrix is:

$\mathbf{S}_W = \frac{1}{2}(\mathbf{S}_1 + \mathbf{S}_2)$ (for two classes; sometimes factor $\frac{1}{2}$ is ommitted)

Reminder: Covariance Matrix



$\mathbf{S}_k$ is the covariance matrix of class $\mathcal{C}_k$:

$\mathbf{S}_k = \frac{1}{N_k - 1} \sum_{n \in \mathcal{C}_k} (\mathbf{x}_n - \mathbf{m}_k)(\mathbf{x}_n - \mathbf{m}_k)^T$

(symmetric and positive semi-definite)

Assuming equally large classes, the total within-class covariance matrix is:

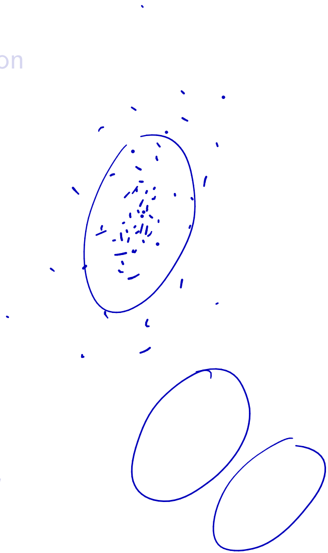
$\mathbf{S}_W = \frac{1}{2}(\mathbf{S}_1 + \mathbf{S}_2)$ (for two classes; sometimes factor $\frac{1}{2}$ is ommitted)

$\mathbf{S}_B$ is the between-class covariance matrix:

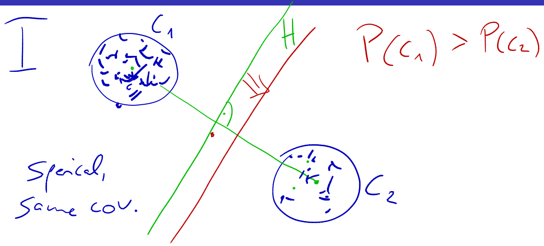
$\mathbf{S}_B = (\mathbf{m}_2 - \mathbf{m}_1)(\mathbf{m}_2 - \mathbf{m}_1)^T$ (for two classes)

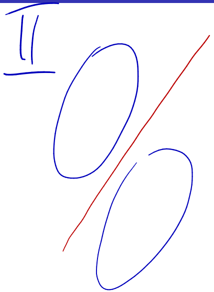
Lecture Overview

1. Brief Recapitulation: Supervised Classification
2. Motivation: Linear Discriminant Functions
3. **LDA Assumptions and Data Scenarios**
4. Geometric Interpretation
5. How to Derive the Parameters...
6. Wrapup: Summary, Related Topics, Preview

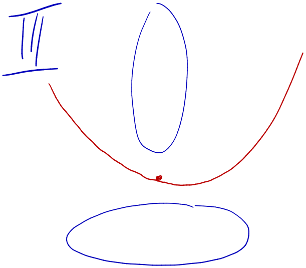


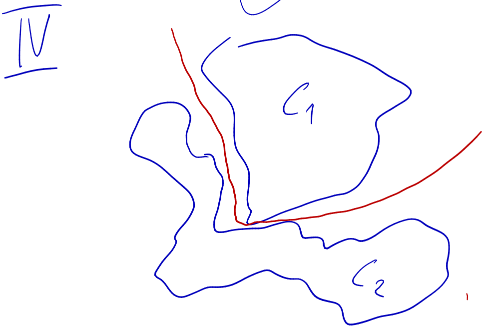
Four Scenarios of Class Distributions





(draw here)





Assumption About the Data

Linear discriminant functions make an important assumption about the data (otherwise, they may fail or underperform!):

Assumption About the Data

Linear discriminant functions make an important assumption about the data (otherwise, they may fail or underperform!):

A1: The data distributions of both classes are Gaussian (normally distributed)

- Data of each class k can be described by a covariance matrix $\mathbf{S}_k$ (also called: scatter matrix)

Assumption About the Data

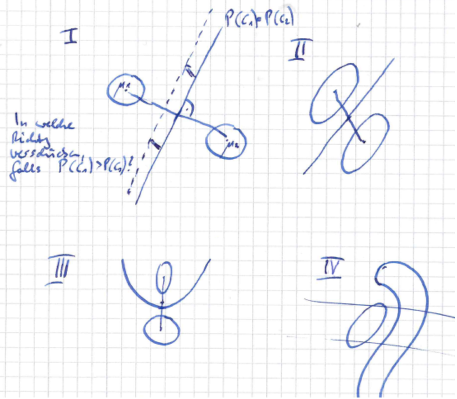
Linear discriminant functions make an important assumption about the data (otherwise, they may fail or underperform!):

A1: The data distributions of both classes are Gaussian (normally distributed)

- Data of each class k can be described by a covariance matrix $\mathbf{S}_k$ (also called: scatter matrix)

A2: The covariance matrices of the classes are equal, i.e. $\mathbf{S}_k = \mathbf{S}$

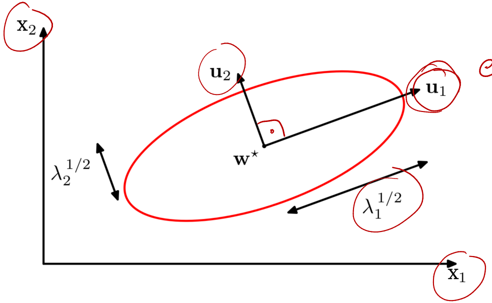
- If you know the covariance matrix of one class, you know also the covariance matrix of every other class.


I
II
III
IV
P(C1)=P(C2)
In welche Richtg verschieben, falls P(C1)>P(C2)?

(draw here)

Reminder: Eigenvalue Decomposition

S



eigenvalue

Eigenvalue decomposition of covariance matrix: $\mathbf{S}\mathbf{u}_i = \lambda_i \mathbf{u}_i$

- requires full rank
- creates novel basis (orthorgonal eigenvectors)
- eigenvectors can be sorted according to eigenvalues
- observe the relation between eigenvectors and variance of a normal distribution
- allows for visual interpretation of covariance matrices

Lecture Overview

Linear Discriminant Function

Using the above assumptions and $k = 2$ classes, we obtain this form of a linear discriminant function:

Skalar

$$f(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + b$$

- $\mathbf{w}^T$ denotes the **transpose** of $\mathbf{w}$.
- $\mathbf{w}^T\mathbf{x}$ is the **inner product** between vectors $\mathbf{w}$ and $\mathbf{x}$.

Terminology:

- $\mathbf{w}$ is called *weight vector*, with $\mathbf{w} \in \mathbb{R}^{D\times 1}$
- $\mathbf{w}$ *projects* the data point $\mathbf{x}$ to a scalar

Linear Discriminant Function

Using the above assumptions and $k = 2$ classes, we obtain this form of a linear discriminant function:

$$f(\mathbf{x}) = \mathbf{w}^T \mathbf{x} + b$$

- $\mathbf{w}^T$ denotes the **transpose** of $\mathbf{w}$.
- $\mathbf{w}^T\mathbf{x}$ is the **inner product** between vectors $\mathbf{w}$ and $\mathbf{x}$.

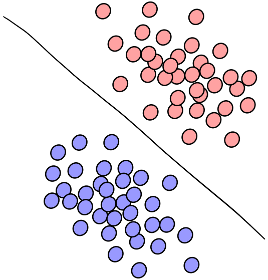
Terminology:

- $\mathbf{w}$ is called *weight vector*, with $\mathbf{w} \in \mathbb{R}^{D\times 1}$
- $\mathbf{w}$ *projects* the data point $\mathbf{x}$ to a scalar
- b is called the *bias* or *threshold weight*.
- (b is also called w_0)

Geometric Interpretation I

Using the above assumptions and $k = 2$ classes, we obtain this form of a linear discriminant function:

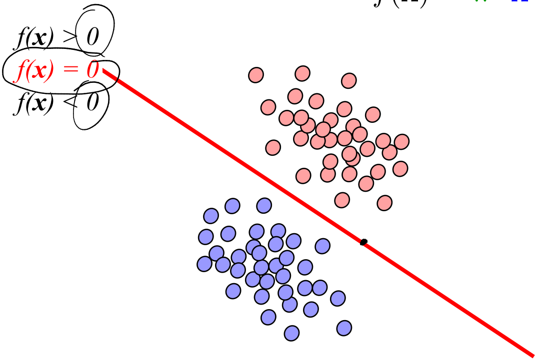
$$f(\mathbf{x}) = \color{green}{\mathbf{w}}^T \color{blue}{\mathbf{x}} + b$$



Geometric Interpretation I

Using the above assumptions and $k = 2$ classes, we obtain this form of a linear discriminant function:

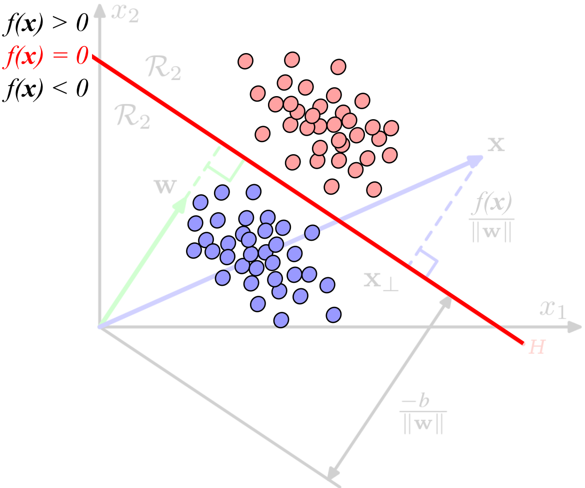
$$f(\mathbf{x}) = \mathbf{w}^T \mathbf{x} + b$$



Geometric Interpretation I

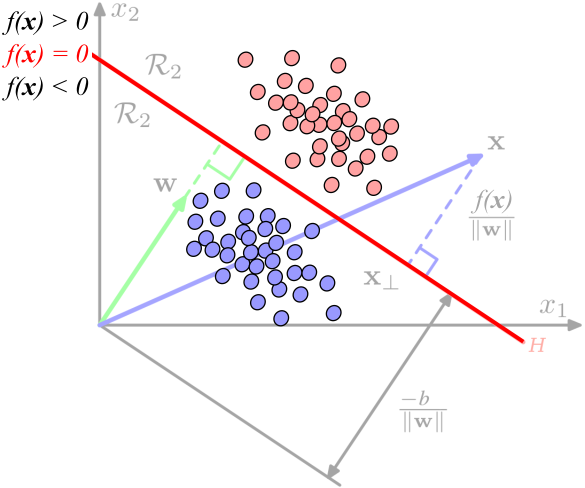
Using the above assumptions and $k = 2$ classes, we obtain this form of a linear discriminant function:

$$f(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + b$$



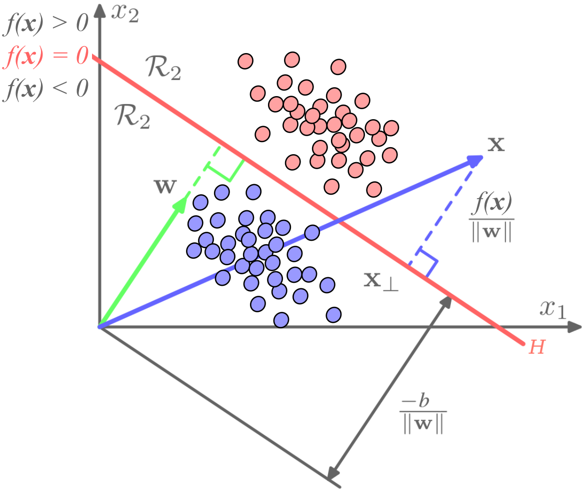
Using the above assumptions and $k = 2$ classes, we obtain this form of a linear discriminant function:

$$f(\mathbf{x}) = \mathbf{w}^T \mathbf{x} + b$$



Using the above assumptions and $k = 2$ classes, we obtain this form of a linear discriminant function:

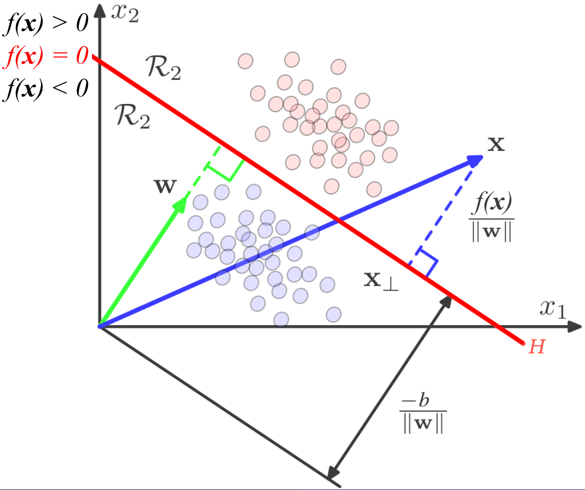
$$f(\mathbf{x}) = \mathbf{w}^T \mathbf{x} + b$$



Geometric Interpretation I

Using the above assumptions and $k = 2$ classes, we obtain this form of a linear discriminant function:

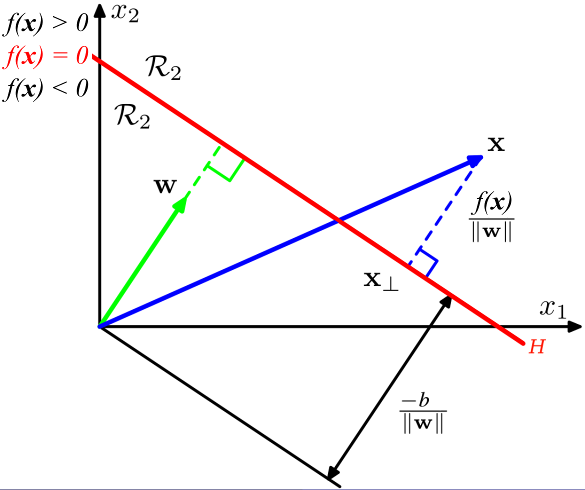
$$f(\mathbf{x}) = \mathbf{w}^T \mathbf{x} + b$$



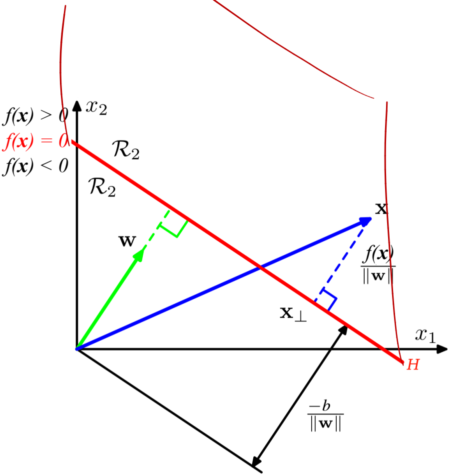
Geometric Interpretation I

Using the above assumptions and $k = 2$ classes, we obtain this form of a linear discriminant function:

$$f(\mathbf{x}) = \mathbf{w}^T \mathbf{x} + b$$

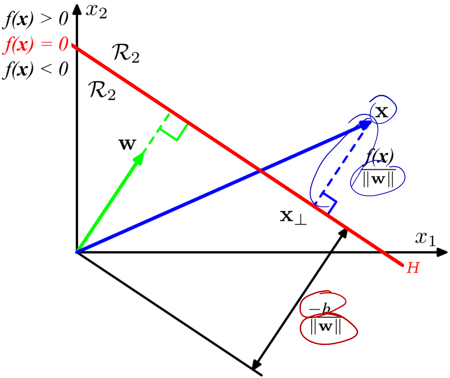


Geometric Interpretation II



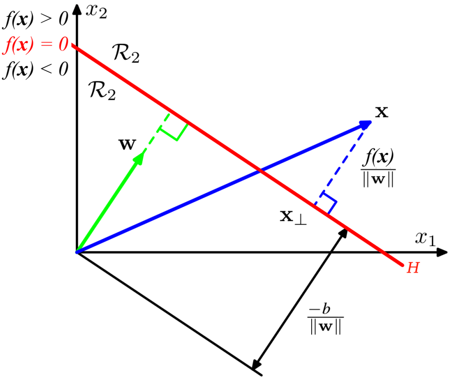
- $\mathbf{w}$ is perpendicular to the decision boundary H
- In general: decision boundary is a (D-1)-dimensional hyperplane.

Geometric Interpretation II



- $\mathbf{w}$ is perpendicular to the decision boundary H
- In general: decision boundary is a (D-1)-dimensional hyperplane.
- Displacement of H from origin is controlled by bias b.

Geometric Interpretation II

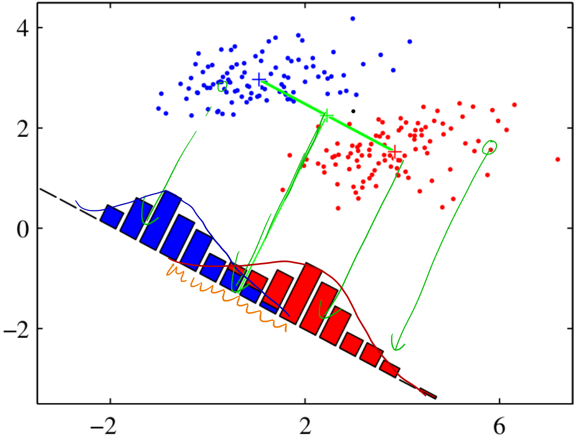


- $\mathbf{w}$ is perpendicular to the decision boundary H
- In general: decision boundary is a (D-1)-dimensional hyperplane.
- Displacement of H from origin is controlled by bias b.
- Signed orthogonal distance of an arbitrary point $\mathbf{x}$ to H is determined by $\frac{f(\mathbf{x})}{\|\mathbf{w}\|}$

Lecture Overview

Two Example Projections

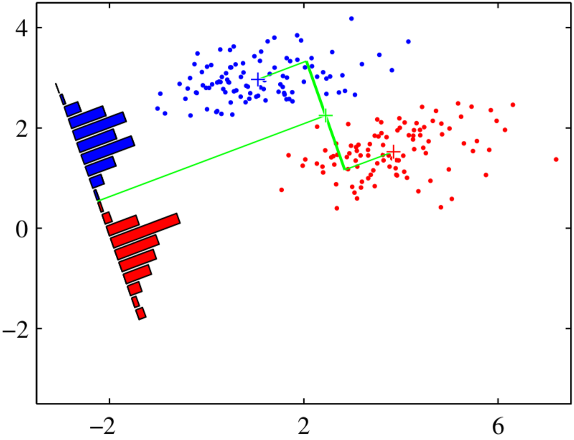
$$f(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + b$$



- Sub-optimal: large overlap of projected distributions

Two Example Projections

$$f(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + b$$

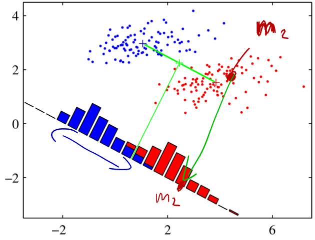


- Pretty good: small overlap of projected distributions

What is a Good Projection $\mathbf{w}$?

Find a $\mathbf{w}$ such, that the projected data maximizes the Fisher criterion:

$$J(\mathbf{w}) = \frac{(m_2 - m_1)^2}{s_1^2 + s_2^2} \qquad \underset{\mathbf{w}}{\operatorname{argmax}} \frac{(m_2 - m_1)^2}{s_1^2 + s_2^2}$$

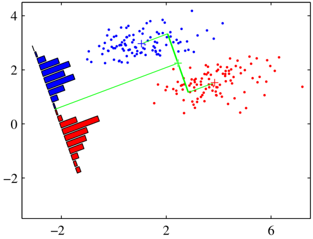


- Projected means $m_k = \mathbf{w}^T \mathbf{m}_k$ should have large distance:
  Thus maximize $(m_2 - m_1)^2$
- For the projected data, the within-class variance s_k^2 of every class $\mathcal{C}_k$ should be small.
  Thus minimize $(s_1^2 + s_2^2)$

What is a Good Projection $\mathbf{w}$?

Find a $\mathbf{w}$ such, that the projected data maximizes the Fisher criterion:

$$J(\mathbf{w}) = \frac{(m_2 - m_1)^2}{s_1^2 + s_2^2} \qquad \underset{\mathbf{w}}{\operatorname{argmax}} \frac{(m_2 - m_1)^2}{s_1^2 + s_2^2}$$

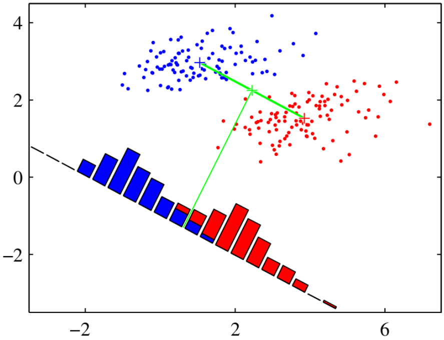


- Projected means $m_k = \mathbf{w}^T \mathbf{m}_k$ should have large distance:
  Thus maximize $(m_2 - m_1)^2$
- For the projected data, the within-class variance s_k^2 of every class $\mathcal{C}_k$ should be small.
  Thus minimize $(s_1^2 + s_2^2)$

Fisher Criterion with Explicit $\mathbf{w}$

More useful: Fisher criterion formulated in the original space with explicit projections by $\mathbf{w}$.
(See [Bishop, Section 4.1.4] for details on the conversion steps)

$$J(\mathbf{w}) = \frac{\mathbf{w}^T \mathbf{S}_B \mathbf{w}}{\mathbf{w}^T \mathbf{S}_W \mathbf{w}}$$



- How should the covariance matrices behave? 🤚🏾🤚🏽

Fisher Criterion with Explicit $\mathbf{w}$

Search the maximum of $J(\mathbf{w})$ by differentiation with respect to $\mathbf{w}$. Make use of the two assumptions [Bishop, Section 4.1.4] to yield:

$$\mathbf{w} = \mathbf{S}_W^{-1}(\mathbf{m}_2 - \mathbf{m}_1)$$

and

$$b = -\tfrac{1}{2}\mathbf{w}(\mathbf{m}_1 + \mathbf{m}_2)$$

$x_i \in \mathbb{R}^D$

$i = 1 \ldots N$

- Nice: weight vector $\mathbf{w}$ and bias b can be computed analytically! (Key: within-class covariance matrices are identical and Gaussian.)
- Please observe: the total within-class covariance matrix $\mathbf{S}_W$ needs to be estimated. (This can be tricky, see assignment 2)

How many free parameters need to be determined for $\mathbf{S}_W$?
Please vote: D, N, $D(D+1)/2$, D^2

Fisher Criterion with Explicit $\mathbf{w}$

Search the maximum of $J(\mathbf{w})$ by differentiation with respect to $\mathbf{w}$. Make use of the two assumptions [Bishop, Section 4.1.4] to yield:

$$\mathbf{w} = \mathbf{S}_W^{-1}(\mathbf{m}_2 - \mathbf{m}_1)$$

and

$$b = -\frac{1}{2}\mathbf{w}(\mathbf{m}_1 + \mathbf{m}_2)$$

- Nice: weight vector $\mathbf{w}$ and bias b can be computed analytically! (Key: within-class covariance matrices are identical and Gaussian.)
- Please observe: the total within-class covariance matrix $\mathbf{S}_W$ needs to be estimated. (This can be tricky, see assignment 2)

  How many free parameters need to be determined for $\mathbf{S}_W$? Please vote: D, N, $D(D+1)/2$, D^2
- Please discuss with your neighbour for 1 min and vote again.

Fisher Criterion with Explicit $\mathbf{w}$

Looking for the maximum of $J(\mathbf{w})$, differentiation with respect to $\mathbf{w}$ and use of our assumptions yields [Bishop, Section 4.1.4]

$$\mathbf{w} = \mathbf{S}_W^{-1}(\mathbf{m}_2 - \mathbf{m}_1)$$

and

$$b = -\tfrac{1}{2}\mathbf{w}(\mathbf{m}_1 + \mathbf{m}_2)$$

analytic

Results:

- Con: covariance matrix needs to be inverted!
  (Runtime? Stability? Approximations like pseudo-inverse?)

Typical Use of Linear Discriminant Analysis

(image: courtesy of Robert Bosch)

- As a first shot method.
- When the noise model of your sensors is known.
- When the class means are informative.

Lecture Overview

Summary by learning goals

Having heard this lecture, you can now . . .

- describe a classification problem and explain linear separability
- give different criteria for good class separability

Summary by learning goals

Having heard this lecture, you can now ...

- describe a classification problem and explain linear separability
- give different criteria for good class separability
- formulate the decision function for a linear discriminant
- explain the meaning and characteristics of the weight vector, decision boundary, covariance matrices, the bias etc.

Summary by learning goals

Having heard this lecture, you can now ...

- describe a classification problem and explain linear separability
- give different criteria for good class separability
- formulate the decision function for a linear discriminant
- explain the meaning and characteristics of the weight vector, decision boundary, covariance matrices, the bias etc.
- explain (different) assumptions for linear discriminant analysis

Summary by learning goals

Having heard this lecture, you can now ...

- describe a classification problem and explain linear separability
- give different criteria for good class separability
- formulate the decision function for a linear discriminant
- explain the meaning and characteristics of the weight vector, decision boundary, covariance matrices, the bias etc.
- explain (different) assumptions for linear discriminant analysis
- formulate the Fisher criterion in the feature space and the projected space

Summary by learning goals

Having heard this lecture, you can now ...

- describe a classification problem and explain linear separability
- give different criteria for good class separability
- formulate the decision function for a linear discriminant
- explain the meaning and characteristics of the weight vector, decision boundary, covariance matrices, the bias etc.
- explain (different) assumptions for linear discriminant analysis
- formulate the Fisher criterion in the feature space and the projected space
- derive a linear discriminant model from given data and apply it to new data

Related Topics and Further Reading

Linear discriminant functions generalize to multiple classes

- Worst solution: one-against-rest
- Slightly better: one-against-one
- Best option for k classes: inherent multiclass formulation
  $y_k(\mathbf{x}) = \mathbf{w}_k^T \mathbf{x} + w_{k0}$
- [Bishop, Section 4.2.1], [Duda, Hart, Stork, Section 5.2.2]

Related Topics and Further Reading

Linear discriminant functions generalize to multiple classes

- Worst solution: one-against-rest
- Slightly better: one-against-one
- Best option for k classes: inherent multiclass formulation $y_k(\mathbf{x}) = \mathbf{w}_k^T\mathbf{x} + w_{k0}$
- [Bishop, Section 4.2.1], [Duda, Hart, Stork, Section 5.2.2]

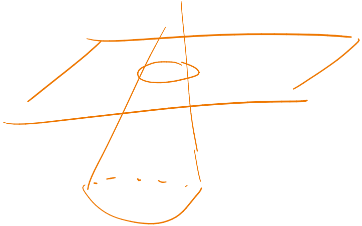
Linear discriminant functions under the easier condition $S_k = \sigma^2\mathbb{I}$

- Search for nearest class mean
- Decision boundary is perpendicular to $m_2 - m_1$
- [Duda,Hart,Stork, Section 2.6]

Related Topics and Further Reading

Linear discriminant functions under the harder condition S_k = arbitrary normal distribution

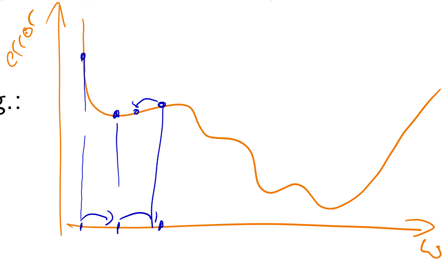
- covariance matrices are different for each class
- decision surfaces are *hyperquadratics*
- Decision regions are not necessarily simply connected any more
- [Duda,Hart,Stork, Chapter 2.6]



Related Topics and Further Reading

Linear discriminant functions under the harder condition $S_k =$ arbitrary normal distribution

- covariance matrices are different for each class
- decision surfaces are *hyperquadratics*
- Decision regions are not necessarily simply connected any more
- [Duda,Hart,Stork, Chapter 2.6]



There are various LDA formulations e.g.:

- as a gradient descent problem
- as an eigenvalue problem
- [Duda,Hart,Stork, Section 2.6]
- as an incremental LDA for online learning [Aliyari Ghassabeh et al. (2015), Pattern Recognition 48(6)]

Related Topics and Further Reading

Bias b can be removed by using augmented vectors:

- slightly enlarging dimensionality: D $\rightarrow$ D+1
- augmented feature vector: $\mathbf{x} \rightarrow \begin{bmatrix} 1 \\ \mathbf{x} \end{bmatrix}$
- augmented weight vector: $\mathbf{w} \rightarrow \begin{bmatrix} b \\ \mathbf{w} \end{bmatrix}$
- or $\mathbf{w} \rightarrow \begin{bmatrix} w_0 \\ \mathbf{w} \end{bmatrix}$
- decision boundary will pass through origin
- [Duda,Hart,Stork, Chapter 5.3]

Related Topics and Further Reading

Generalized linear discriminant functions allow for a non-linear feature pre-processing ϕ by

$$f(\mathbf{x}) = \sum_{i=1}^{D} \mathbf{w}_i \phi(\mathbf{x})$$

log(x)

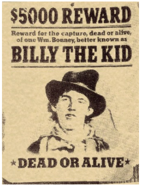
x^2

$x_3 = x_1^2 + \log(x_2)$

but are still linear in $\phi(\mathbf{x})$

- non-linearity in the features, not the classification method
- powerful, if expert knowledge about features is available
- $\rightarrow$ relation to pre-processing (see ML design cycle)
- may enlarge the dimensionality
- [Duda,Hart,Stork, Chapter 5.3]

Hint for the Exam: Overview Sheet for each Method



Typical use case for the method

- supervised / unsupervised / reinforcement learning / dimensionality reduction / ...

Assumptions made by the method about the data, e.g.

- Gaussian data
- equal covariances for each class

Runtime + memory requirements:

- O(N) – data points N
- O(D) – dimensionality D
- at training time
- at recall time (using the model)

Strenghts and weaknesses e.g.

- very sensitive to outliers
- can be used for online learning
- difficult to find good hyperparameters
- easy to interpret a trained model

Related methods, improvements

Regularization approaches, number of free parameters...

Preview of Assignment 2

Objectives of assignment 2:

- Polish up some math concepts
- Get acquainted with covariance matrices (visualize them, analyze their eigenvalue spectrum)

Preview of Assignment 2

Objectives of assignment 2:

- Polish up some math concepts
- Get acquainted with covariance matrices (visualize them, analyze their eigenvalue spectrum)
- Implement LDA for a two-class problem, test it
- Experiment with the assumptions of LDA, break it