Elements of Machine Learning & Data Science

Winter semester 2023/24

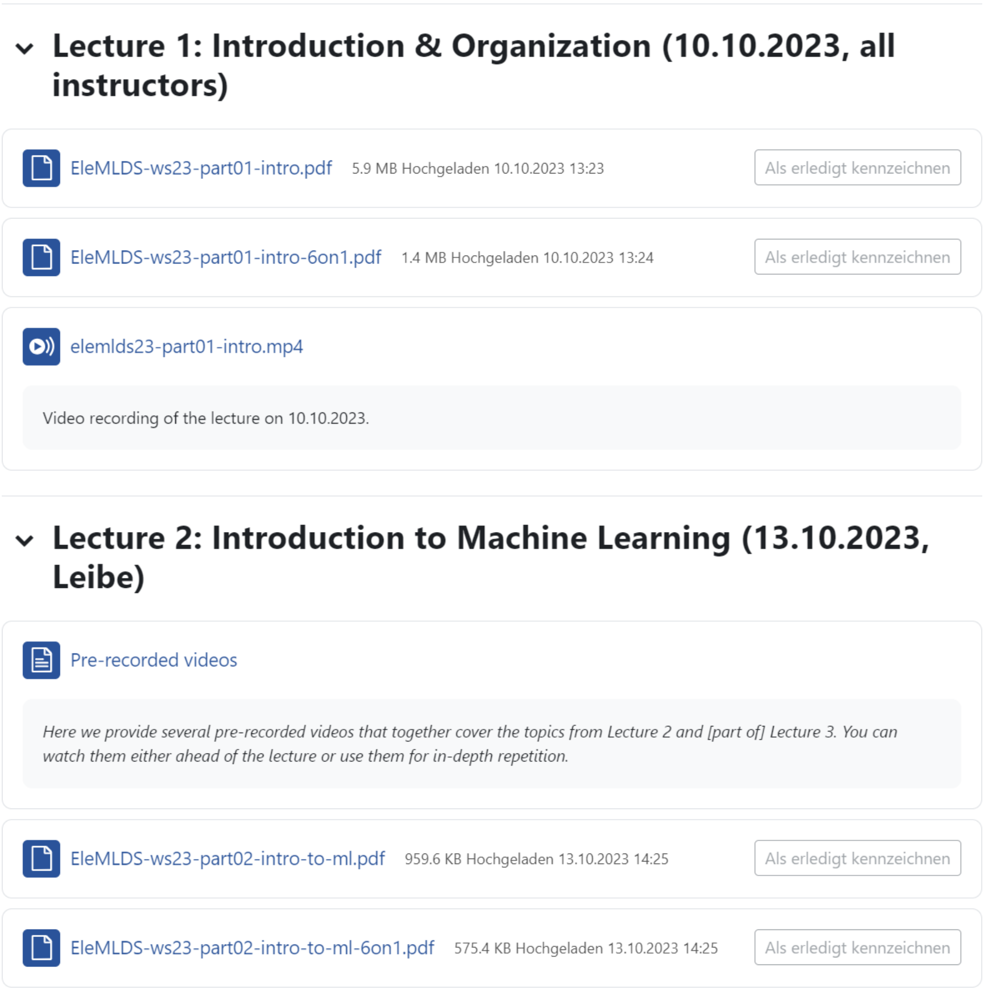
Lecture 2 – Introduction to ML

13.10.2023

Prof. Bastian Leibe

Announcements: Moodle

- **Materials provided on moodle**
  - Pdfs of the slides
  - Video recording of the previous lecture
  - Pre-recorded videos for *this* lecture



- *If you haven't done so yet, please register for the class to get access to the moodle.*

Announcements: Pre-recorded Videos

- **Companion MOOCs**
  - We have created two MOOCs to complement this lecture
    - Basics of ML
    - Basics of Data Science
  - Target: International Master students and students from other degree programs joining Computer Science

- **The pre-recorded videos are part of those MOOCs**
  - Extended explanations of key lecture topics
  - High production value
  - They may not cover the full topic range of the lecture
  - *Please use them as supplementary material*

Announcement: Small-Group Exercises

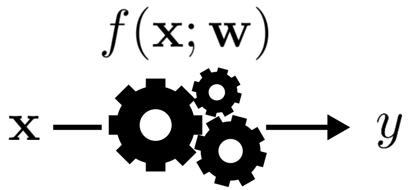
Monday	Tuesday	Wednesday	Thursday	Friday
14:30-18:00h	3x 14:30-16:00h	2x 14:30-16:00h		
3x 16:30-18:00h	3x 16:30-18:00h	2x 16:30-18:00h		
18:30-20:00h	2x 18:30-20:00h			

- **Bi-weekly small-group exercises**
  - We’re currently setting up a poll to collect your preferences for the exercise slots
  - Please enter your choices until Wed, 18.10. evening!
  - Based on the poll results, we will assign you to exercise slots

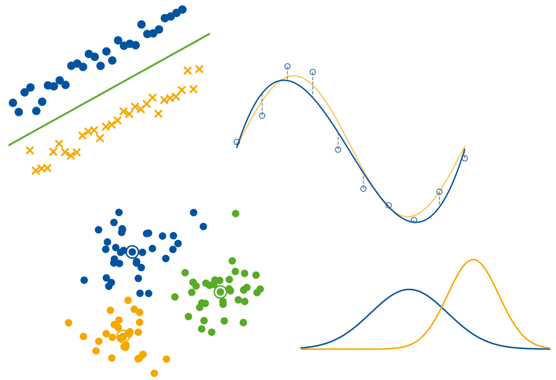
  - *We will send out an email announcement with detailed instructions tonight or on Saturday…*

Machine Learning Topics

1. **Introduction to ML**
2. Probability Density Estimation
3. Linear Discriminants
4. Linear Regression
5. Logistic Regression
6. Support Vector Machines
7. AdaBoost
8. Neural Network Basics



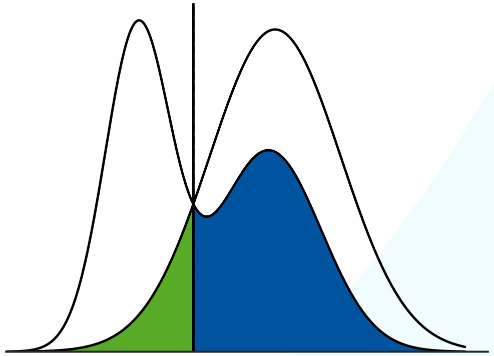
Machine Learning Concepts



Forms of Machine Learning

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

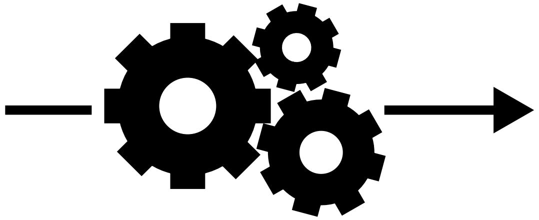
Bayes Decision Theory



Bayes Optimal Classification

Topics for Today

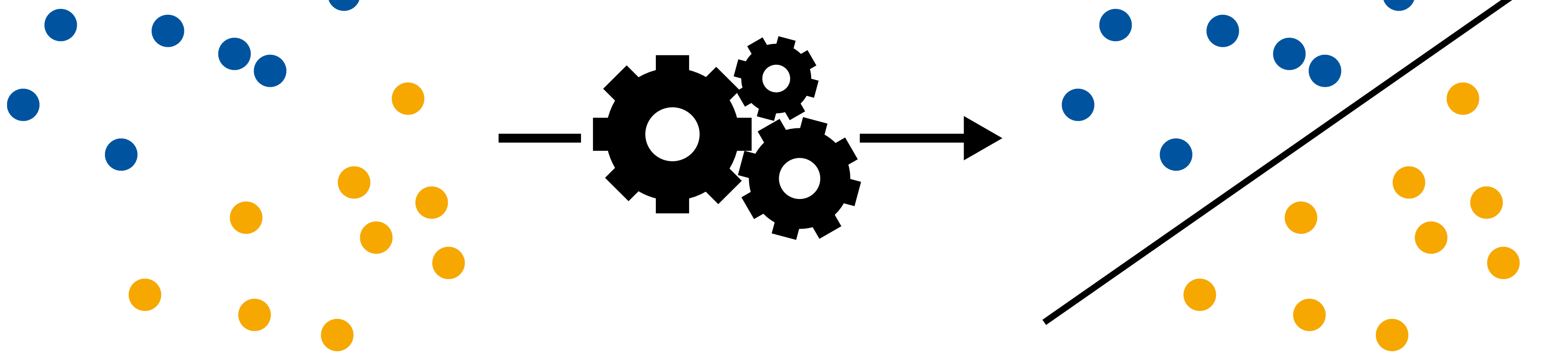
1. **Motivation**
2. Forms of Learning
3. Terms, Concepts, and Notation
4. Bayes Decision Theory



What is Machine Learning?

Machines that learn to perform a task from experience

Performance:
How well is the task executed?



Experience:
training data

Learning: discover
relevant criteria & adapt
to given situation

Task: problem to solve
(e.g., classification)

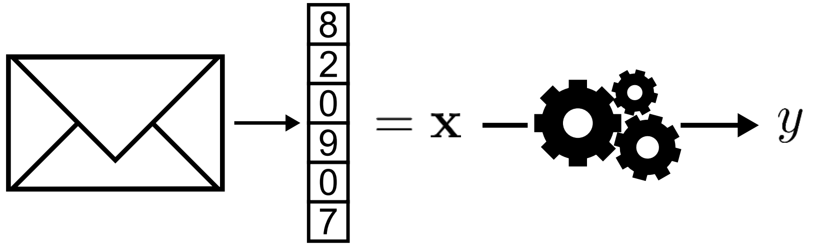
Mathematical Formulation

Machines that learn to perform a task from experience

Often described through a mathematical function:

$$y = f(\mathbf{x}; \mathbf{w})$$

Output y Input $\mathbf{x}$ Parameters $\mathbf{w}$ (learned!)



Discrete targets: Classification

$$y \in \{\text{important}, \text{spam}\}$$

Continuous targets: Regression

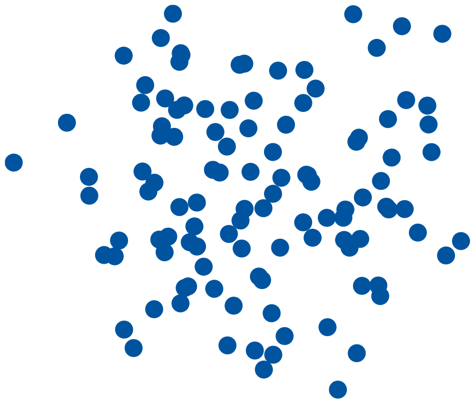
$$y = p(\text{spam}) \in [0, 1]$$

Learning from Data

Machines that learn to perform a task from experience

Learning from collected samples:

$$\mathcal{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$$



Labels available?

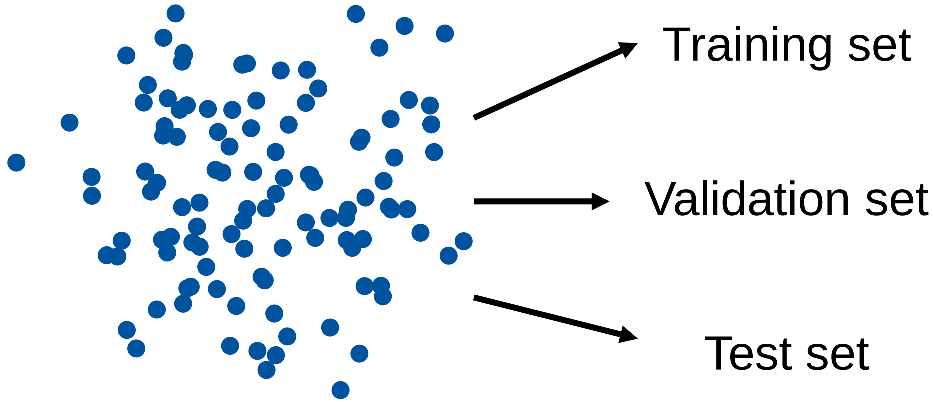
- Yes → Supervised learning
- Some → Semi-supervised learning
- No → Unsupervised learning

Learning via sparse feedback: Reinforcement learning

Measuring Success

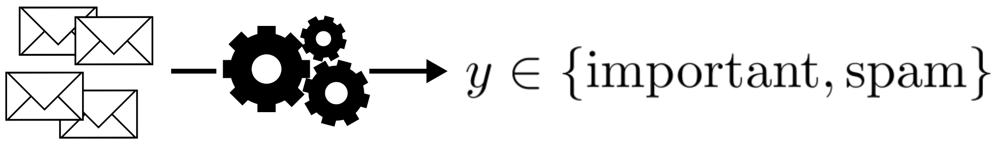
Machines that learn to perform a task from experience

- Performance measure: typically a single number.
  - Calculate with a suitable metric.
- Divide data into disjoint subsets:

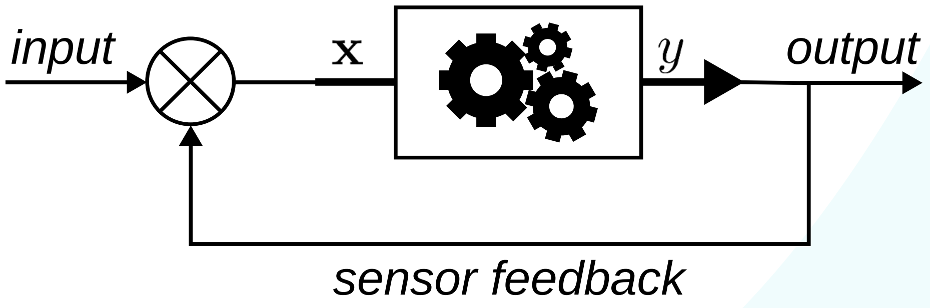


- Measure generalization performance on test set.

E.g., % correctly recognized spam mails



E.g., average distance to desired endpoint

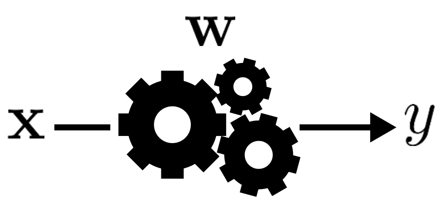


Learning as Optimization

Machines that learn to perform a task from experience

Learning = optimizing $f(\mathbf{x}; \mathbf{w})$

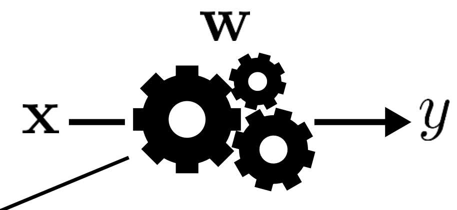
$\mathbf{w}$ *describes the type of model that we use.*



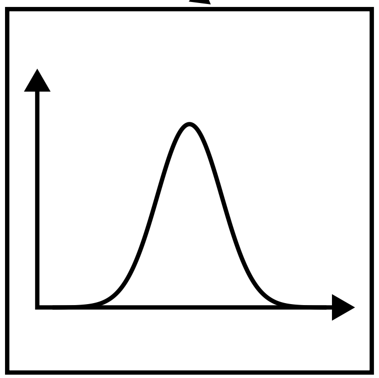
Learning as Optimization

Machines that learn to perform a task from experience

Learning = optimizing $f(\mathbf{x}; \mathbf{w})$



$\mathbf{w}$ *describes the type of model that we use.*

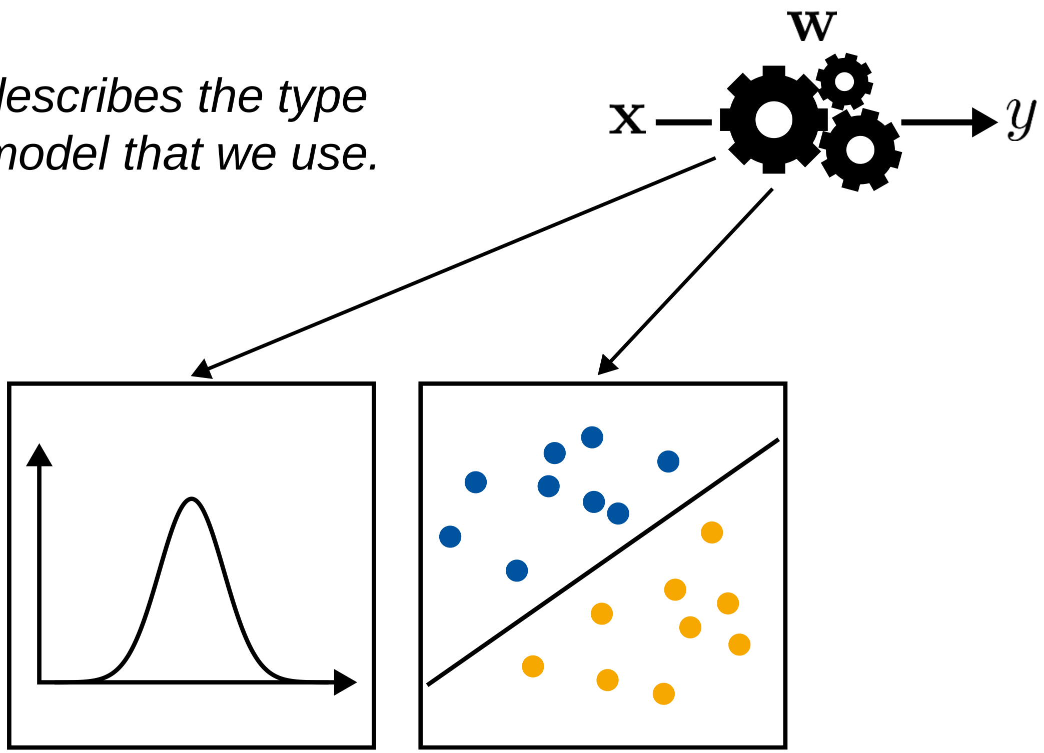


Learning as Optimization

Machines that learn to perform a task from experience

Learning = optimizing $f(\mathbf{x}; \mathbf{w})$

$\mathbf{w}$ *describes the type of model that we use.*

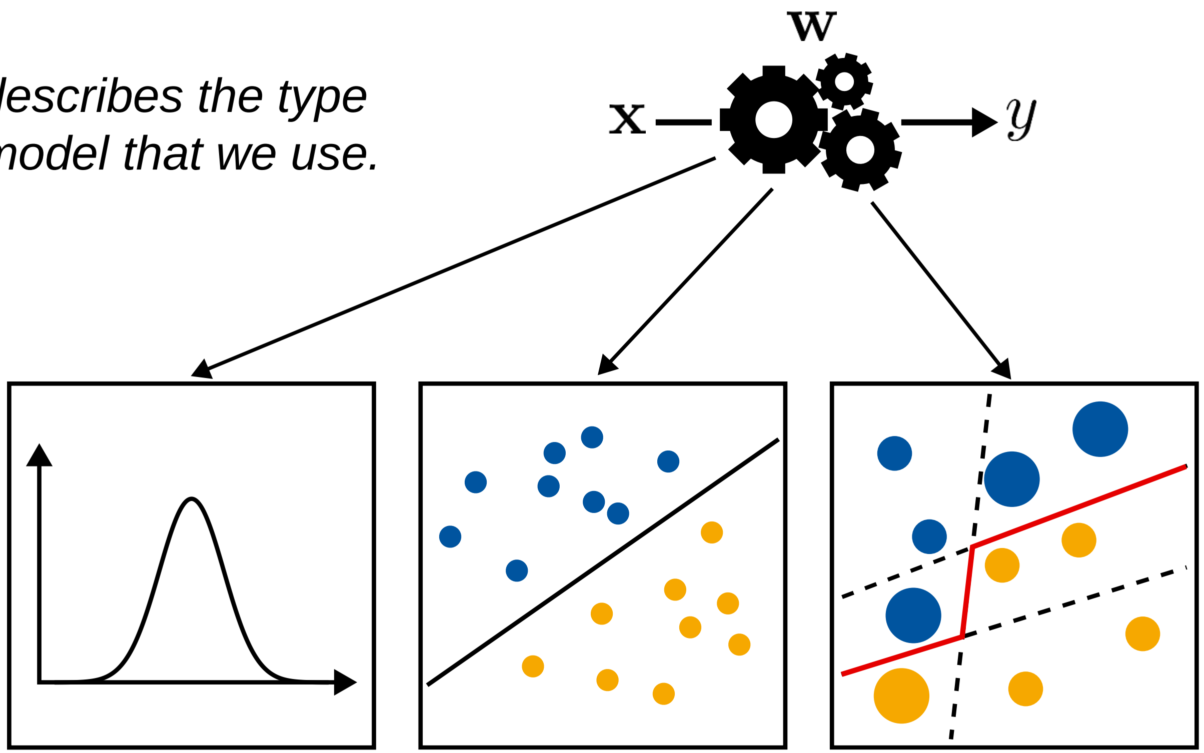


Learning as Optimization

Machines that learn to perform a task from experience

Learning = optimizing $f(\mathbf{x}; \mathbf{w})$

$\mathbf{w}$ *describes the type of model that we use.*

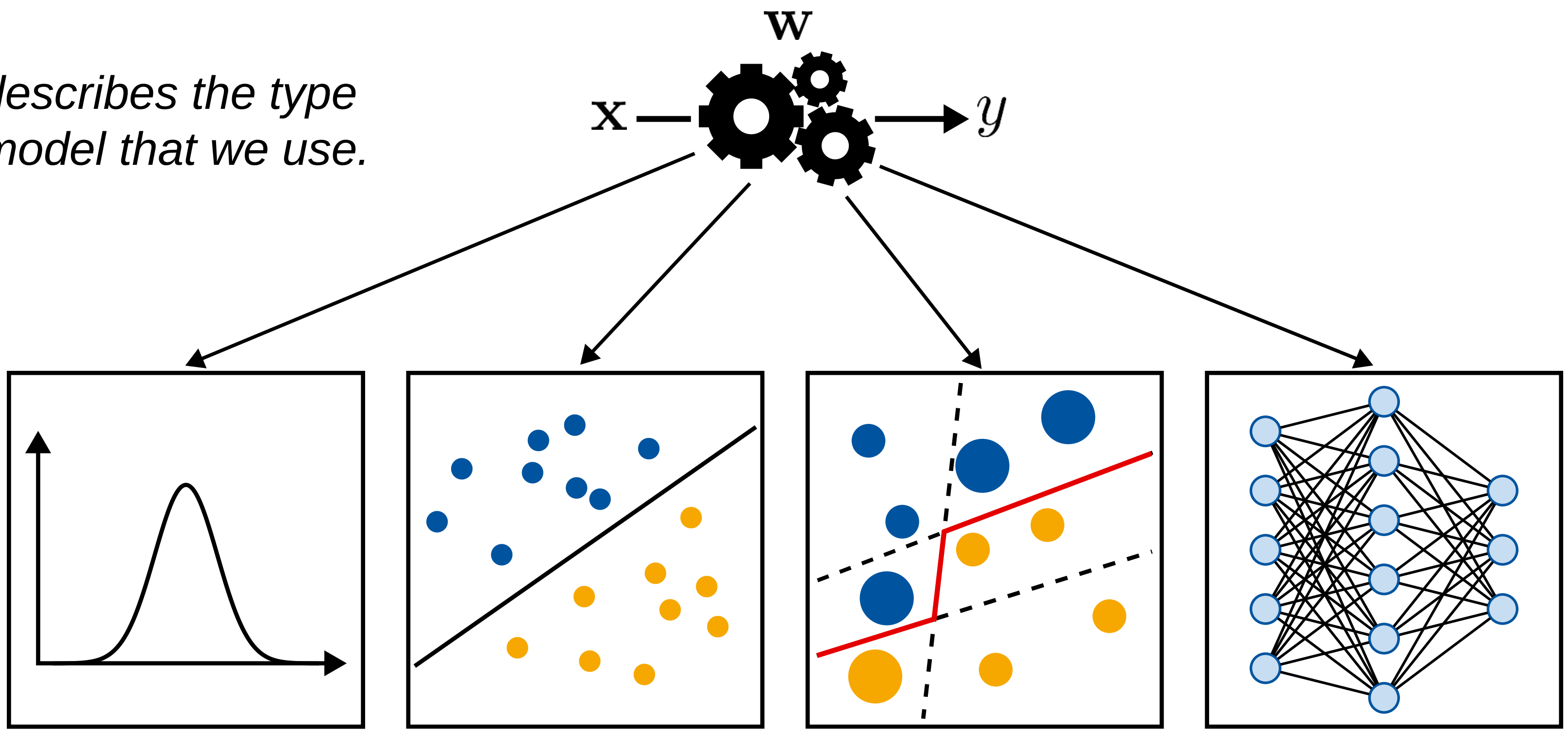


Learning as Optimization

Machines that learn to perform a task from experience

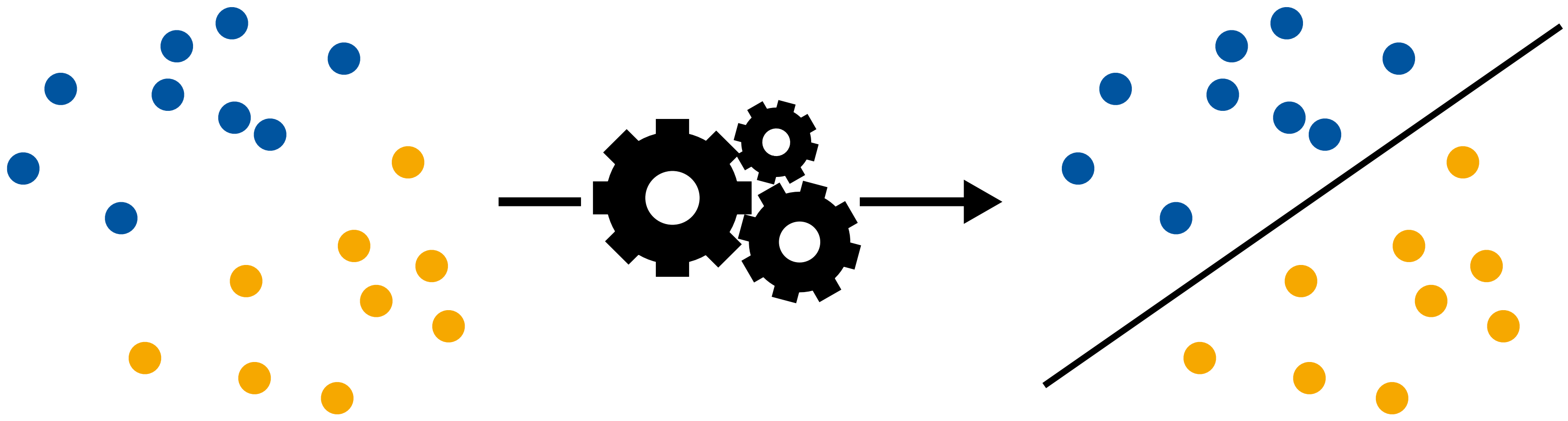
Learning = optimizing $f(\mathbf{x};\mathbf{w})$

$\mathbf{w}$ *describes the type of model that we use.*



What is Machine Learning?

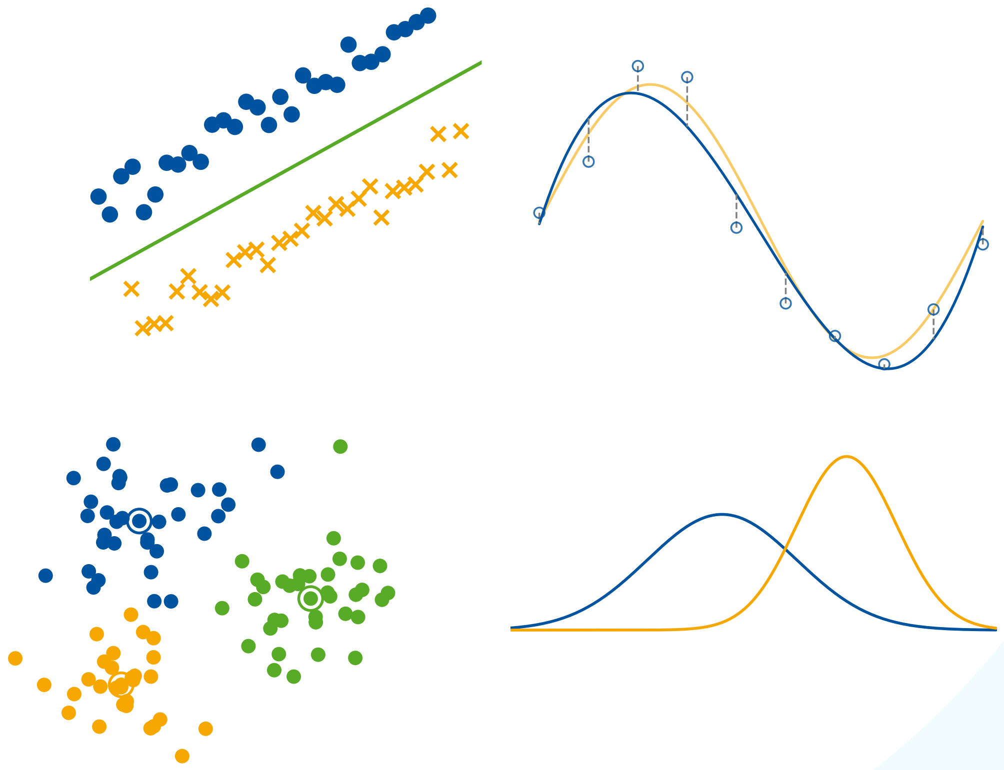
Machines that learn to perform a task from experience



We will focus on statistical Machine Learning.

Topics for Today

1. Motivation
2. **Forms of Learning**
3. Terms, Concepts, and Notation
4. Bayes Decision Theory

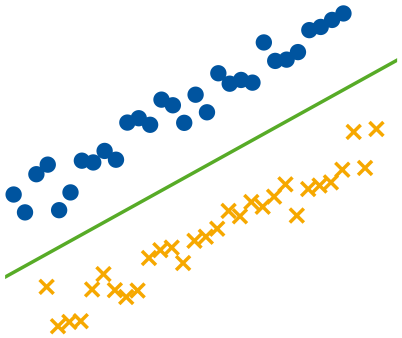


Supervised vs. Unsupervised Learning

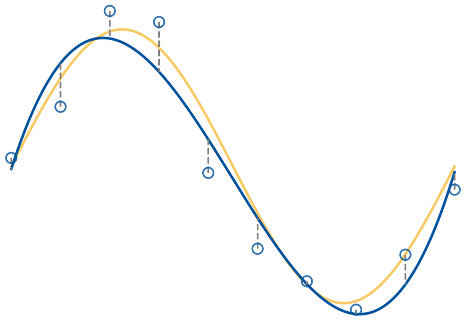
Discrete targets

Continuous targets

Known targets



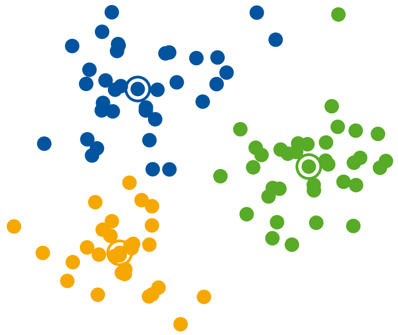
Classification



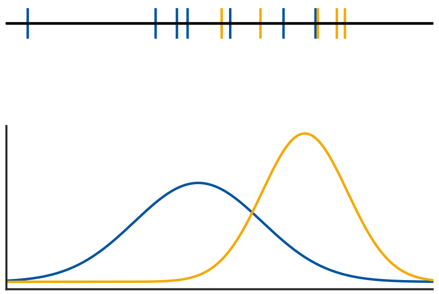
Regression

Supervised learning

Unknown targets



Clustering



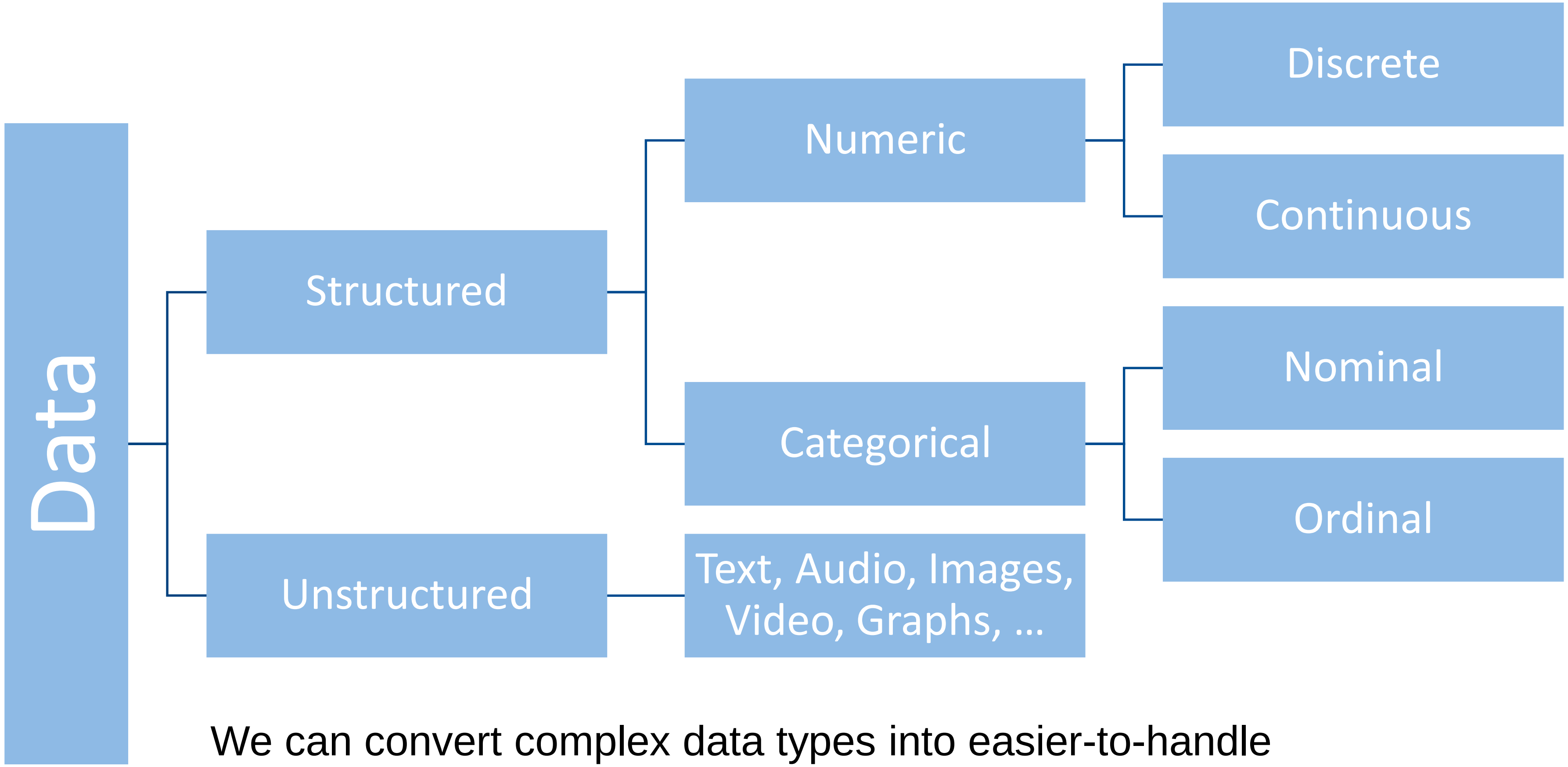
Density estimation

Unsupervised learning

Supervised Learning

- We will mostly focus on supervised learning.
- Given training data with labels: $\mathcal{D} = \{(\mathbf{x}_1, t_1), \dots, (\mathbf{x}_N, t_N)\}$
- The goal is to learn a predictive function $y(\mathbf{x}; \mathbf{w})$ that yields good performance on unseen test data.
- In real-world scenarios, we also need to preprocess our data to handle, e.g.,
  - Missing or wrong values
  - Outliers
  - Inconsistencies

Data Types - Overview

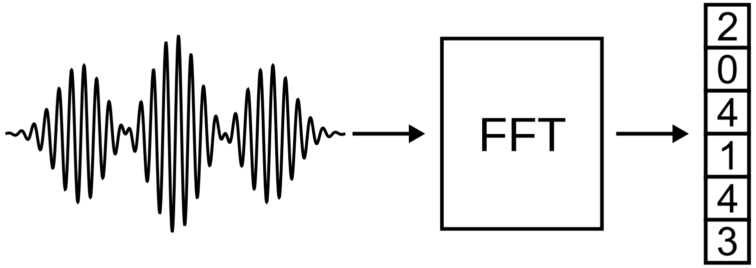


We can convert complex data types into easier-to-handle continuous vector-space data via feature extraction.

Features

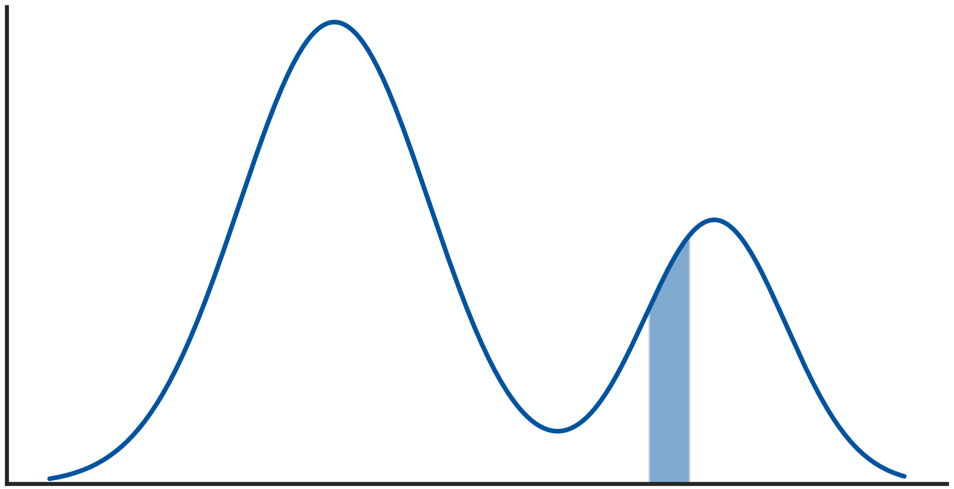
- Feature extraction is the process that creates descriptive vectors from samples.
  - Features should be invariant to irrelevant input variations.
  - Selecting the “right” features is crucial.
  - Usually encode some domain knowledge.
  - Higher-dimensional features are more discriminative.
- Curse of dimensionality: complexity increases exponentially with number of dimensions.

Example: convert audio snippet to feature vector with Fast Fourier Transform (FFT).



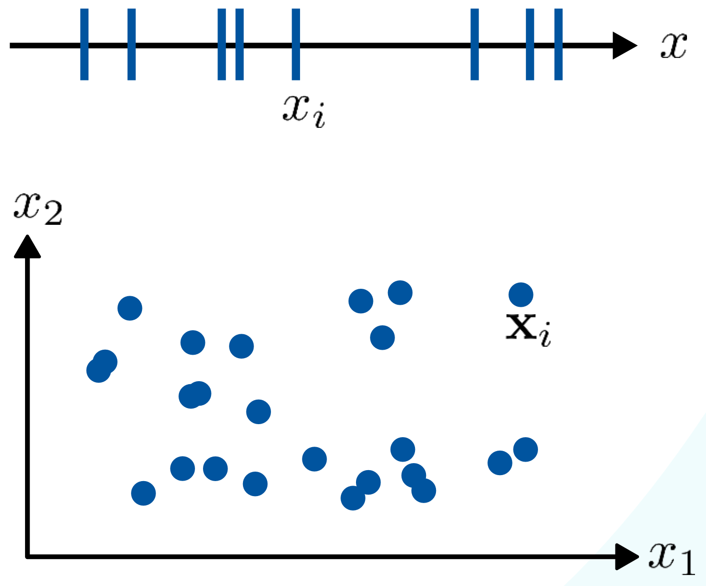
Introduction

1. Motivation
2. Forms of learning
3. **Terms, Concepts, and Notation**
4. Bayes Decision Theory



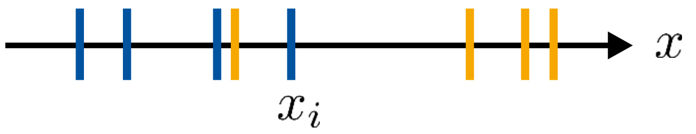
Terms, Concepts, and Notation

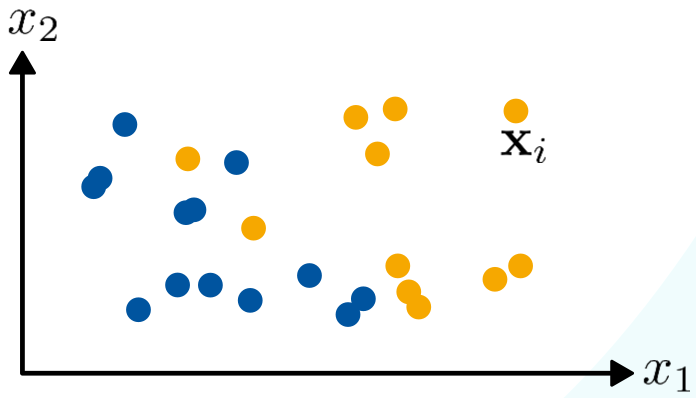
- Most of our tools will be based on statistics and probability theory.
- We will review the most important concepts here.
- Some Notation:
  - Scalar data $x \in \mathbb{R}$
  - Vector-valued data $\mathbf{x} \in \mathbb{R}^D$
  - Datasets $\mathcal{X} = \{x_1, \ldots, x_N\}$



Terms, Concepts, and Notation

- Most of our tools will be based on statistics and probability theory.
- We will only review the most important concepts here.
- Some Notation:
  - Scalar data $x \in \mathbb{R}$
  - Vector-valued data $\mathbf{x} \in \mathbb{R}^D$
  - Datasets $\mathcal{X} = \{x_1, \ldots, x_N\}$
  - Labelled datasets $\mathcal{D} = \{(x_1, t_1), \ldots, (x_N, t_N)\}$
  - Matrices $\mathbf{M} \in \mathbb{R}^{m \times n}$
  - Dot product $\mathbf{w}^\mathsf{T}\mathbf{x} = \sum_{j=1}^{D} w_j x_j$





E.g., class labels: $\mathcal{C}_1$ $\mathcal{C}_2$

Probability Basics

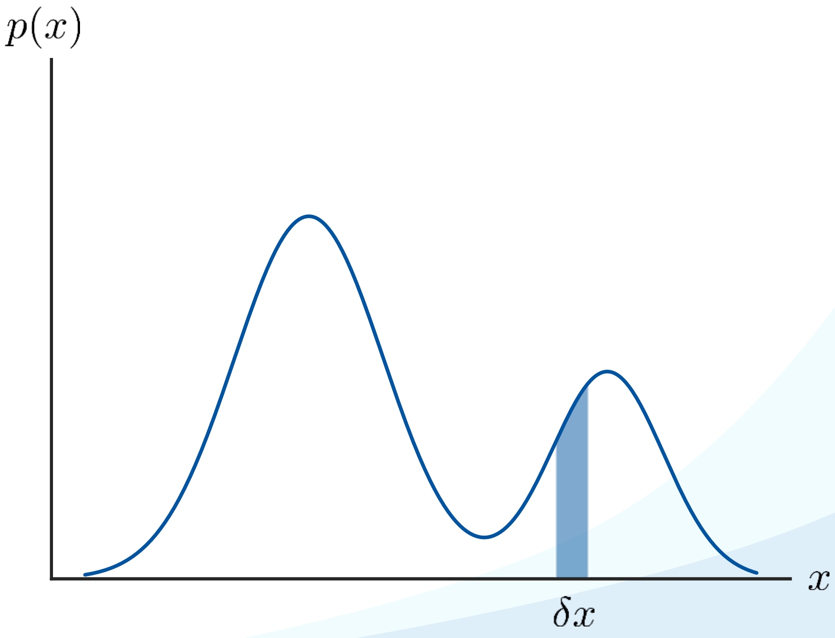
- Probabilities are defined over random variables:
  - Discrete case:

$$p(X = x_j) = \frac{n_j}{N}$$

  - Continuous case:

$$p(X \in (x_1, x_2)) = \int_{x_1}^{x_2} p(x)\,\mathrm{d}x$$

Where $p(x)$ is the probability density function (pdf) of x .



Probability Basics

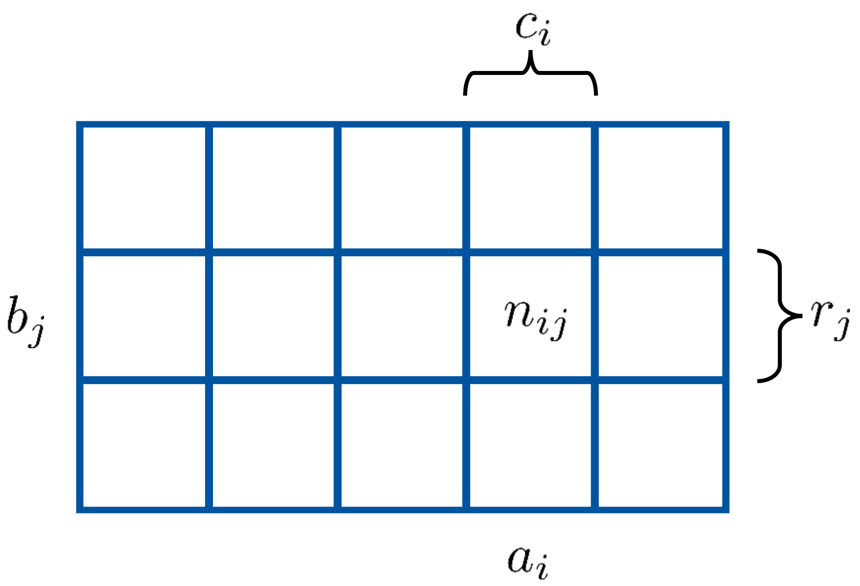
- Random variables $A \in \{a_i\}, B \in \{b_j\}$
- Consider N trials:

$$n_{ij} = \#\{A = a_i \wedge B = b_j\}$$

$$c_i = \#\{A = a_i\}$$

$$r_j = \#\{B = b_j\}$$

- Derive from this:
  - Joint probability $p(A = a_i, B = b_j) = \frac{n_{ij}}{N}$
  - Marginal probability $p(A = a_i) = \frac{c_i}{N}$
  - Conditional probability $p(B = b_j | A = a_i) = \frac{n_{ij}}{c_i}$

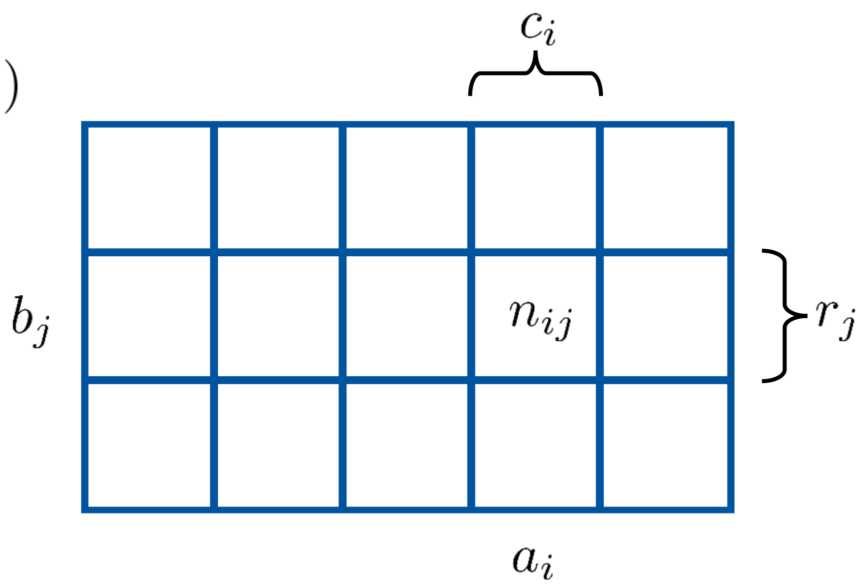


- Sum rule:

$$p(A = a_i) = \frac{c_i}{N} = \frac{1}{N}\sum_j n_{ij} = \sum_{b_j} p(A = a_i, B = b_j)$$

- Product rule:

$$p(A = a_i, B = b_j) = \frac{n_{ij}}{N} = \frac{n_{ij}}{c_i} \cdot \frac{c_i}{N}$$
$$= p(B = b_j | A = a_i) p(A = a_i)$$



Rules of Probability - Summary

- Sum rule:

$$p(A) = \sum_B p(A, B)$$

- Product rule:

$$p(A, B) = p(B|A)p(A)$$

- Combine into Bayes' Theorem:

$$p(A|B) = \frac{p(B|A)p(A)}{p(B)}$$
$$= \frac{p(B|A)p(A)}{\sum_A p(B|A)p(A)}$$

This is the most important equation in this course!

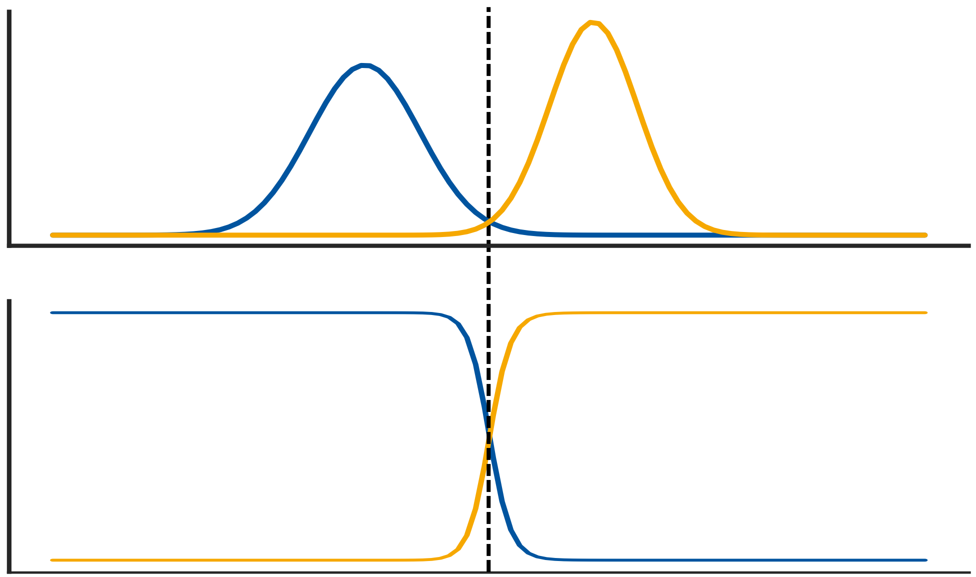
Introduction

1. Motivation

2. Forms of learning

3. Terms, Concepts, and Notation

4. **Bayes Decision Theory**



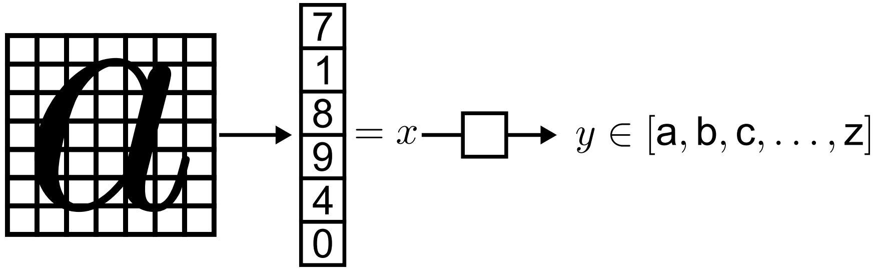
Bayes Decision Theory

- Goal: predict an output class $\mathcal{C}$ from measurements $\mathbf{x}$, by minimizing the probability of misclassification.
- *How can we make such decisions optimally?*
- Bayes Decision Theory gives us the tools for this
  - Based on Bayes' Theorem:

$$p(Y|X) = \frac{p(X|Y)p(Y)}{p(X)}$$

  - In the following, we will introduce its basic concepts…

Example: handwritten character recognition



$\mathbf{x}$: e.g., pixel values

Core Concept: Priors

- What can we tell about the outcome of an experiment *before* making any measurements?
- The a-priori probability $p(\mathcal{C})$ captures the probability distribution over the different class outcomes
  - Based on previously observed data
  - i.e., independent of the actual measurement
- The prior probabilities over all possible class outcomes sum to one.

Example: in English text, the letter "e" makes up ~13% of all letters:

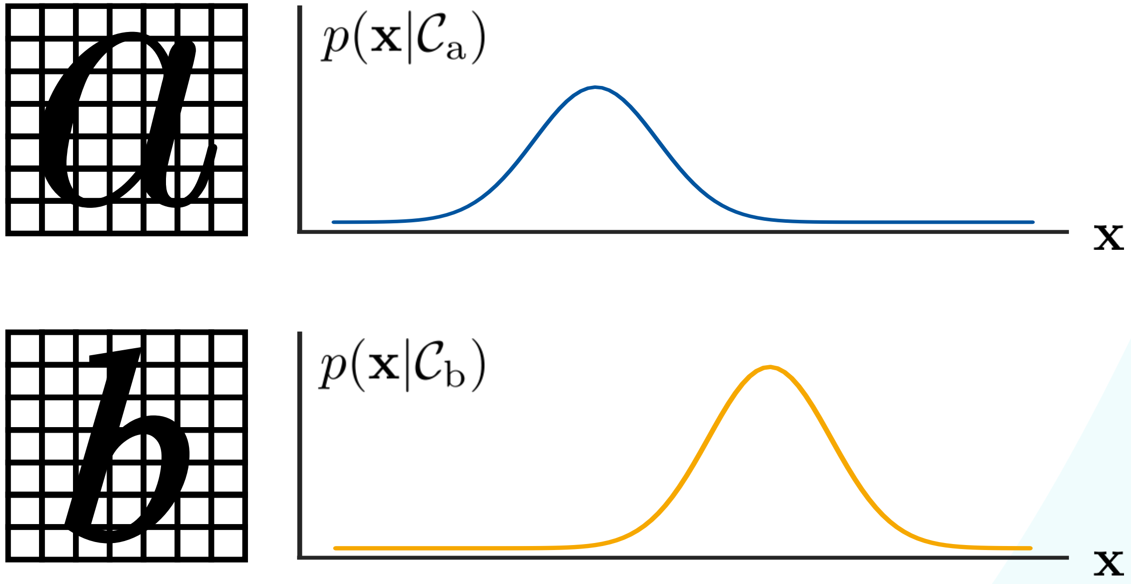
$$p(\mathcal{C}_{\mathrm{e}}) = 0.13$$

And there are 26 letters in the English alphabet:

$$\sum_{\alpha \in \{\mathrm{a},\ldots,\mathrm{z}\}} p(\mathcal{C}_\alpha) = 1$$

Core Concept: Likelihood

- How *likely* is it that we *observe* a certain measurement $\mathbf{x}$ *given* an example of class $\mathcal{C}$?
- This is expressed by the likelihood $p(\mathbf{x}|\mathcal{C})$
  - It is called a *class-conditional distribution*, since it specifies the distribution of $\mathbf{x}$ conditioned on the class $\mathcal{C}$.
  - We can estimate the likelihood from the distribution of measurements $\mathbf{x}$ observed on the given training data.



- Here, $\mathbf{x}$ measures certain properties of the input data.
  - E.g., the fraction of black pixels
  - We simply treat it as a vector $\mathbf{x} \in \mathbb{R}^D$.

Core Concept: Posterior

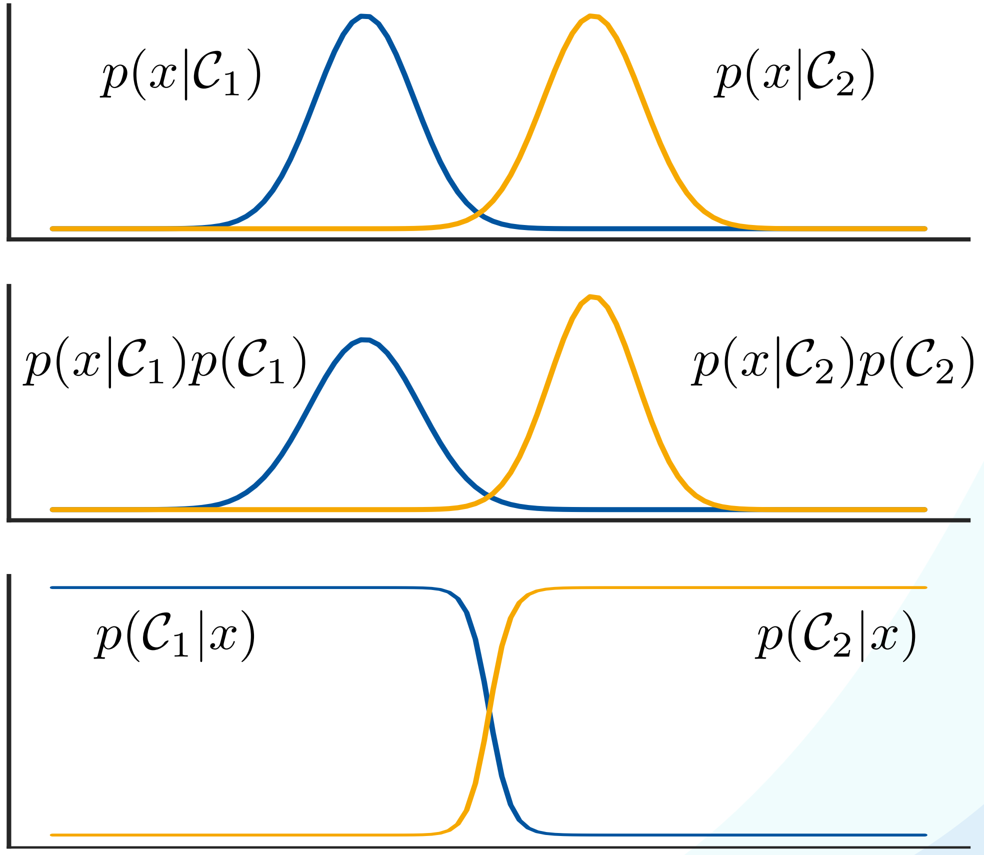
- What is the probability for class $\mathcal{C}_k$ if we made a measurement $\mathbf{x}$?
- This a-posteriori probability $p(\mathcal{C}_k|\mathbf{x})$ can be computed via Bayes' Theorem after we observed $\mathbf{x}$:

$$p(\mathcal{C}_k|\mathbf{x}) = \frac{p(\mathbf{x}|\mathcal{C}_k)p(\mathcal{C}_k)}{p(\mathbf{x})} = \frac{p(\mathbf{x}|\mathcal{C}_k)p(\mathcal{C}_k)}{\sum_j p(\mathbf{x}|\mathcal{C}_j)p(\mathcal{C}_j)}$$

- *This is usually what we're interested in!*

- Interpretation

$$posterior = \frac{likelihood \cdot prior}{normalization\ factor}$$



Making Optimal Decisions

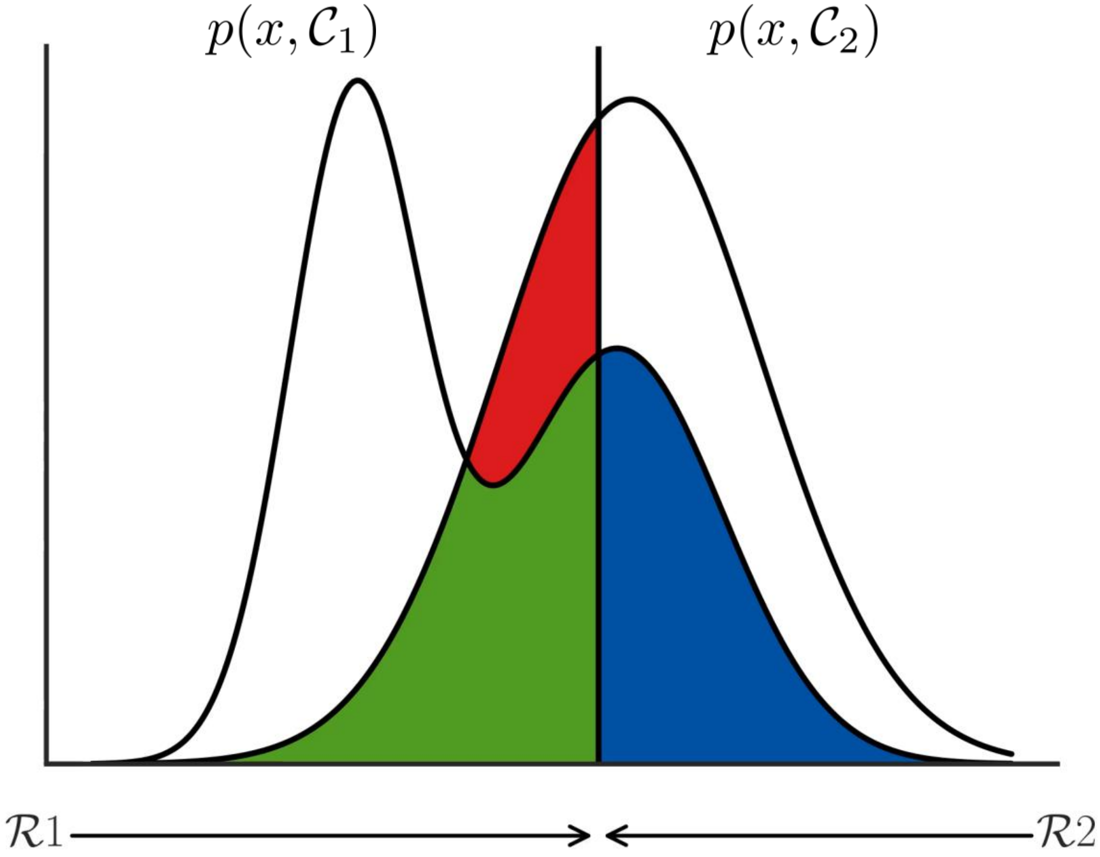
- Goal: minimize the probability of misclassification.

$$p(\text{mistake}) = p(x \in \mathcal{R}_1, \mathcal{C}_2) + p(x \in \mathcal{R}_2, \mathcal{C}_1)$$
$$= \int_{\mathcal{R}_1} p(x, \mathcal{C}_2)\,\mathrm{d}x + \int_{\mathcal{R}_2} p(x, \mathcal{C}_1)\,\mathrm{d}x$$
$$= \int_{\mathcal{R}_1} p(\mathcal{C}_2|x)p(x)\,\mathrm{d}x + \int_{\mathcal{R}_2} p(\mathcal{C}_1|x)p(x)\,\mathrm{d}x$$

- Note:

+ = constant

We can only reduce



$\mathcal{R}_1$ and $\mathcal{R}_2$ are the decision regions after setting a decision threshold.

Making Optimal Decisions

- Goal: minimize the probability of misclassification.

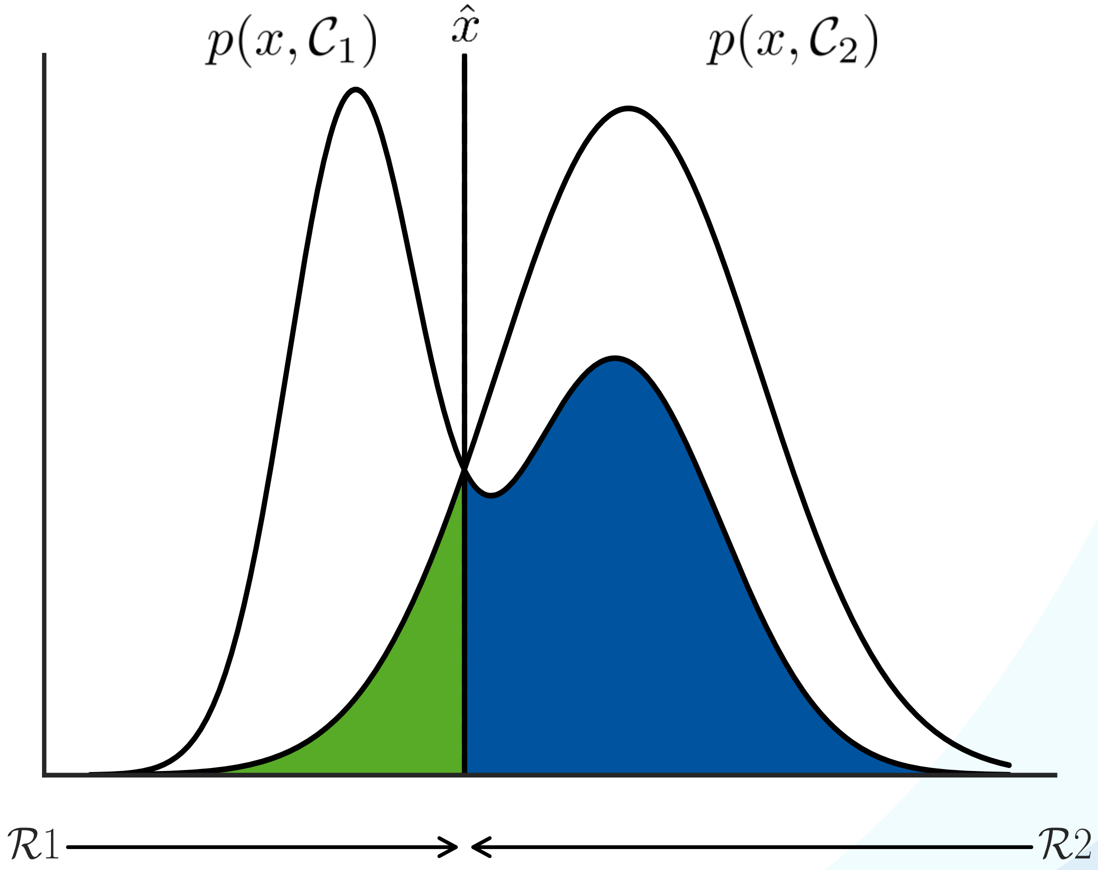
$$p(\text{mistake}) = p(x \in \mathcal{R}_1, \mathcal{C}_2) + p(x \in \mathcal{R}_2, \mathcal{C}_1)$$
$$= \int_{\mathcal{R}_1} p(x, \mathcal{C}_2)\,\mathrm{d}x + \int_{\mathcal{R}_2} p(x, \mathcal{C}_1)\,\mathrm{d}x$$
$$= \int_{\mathcal{R}_1} p(\mathcal{C}_2|x)p(x)\,\mathrm{d}x + \int_{\mathcal{R}_2} p(\mathcal{C}_1|x)p(x)\,\mathrm{d}x$$

- Note:

 + = constant

We can only reduce

- *Minimal error at the intersection* $\hat{x}$



$\mathcal{R}_1$ and $\mathcal{R}_2$ are the decision regions after setting a decision threshold.

Making Optimal Decisions

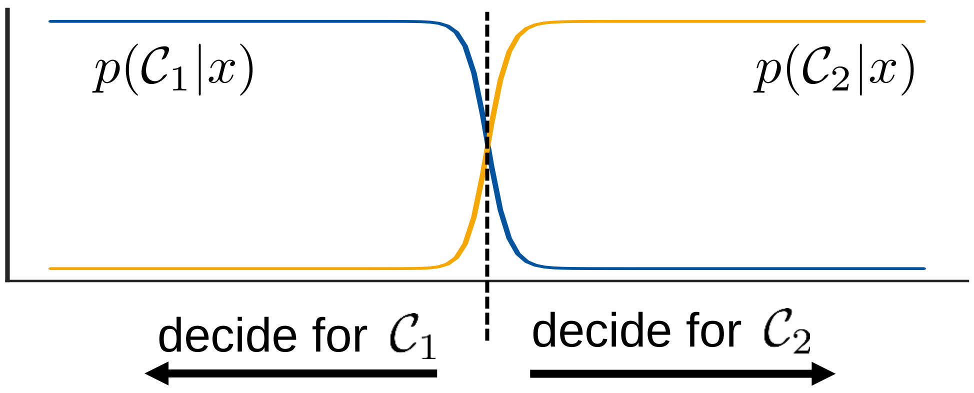
- Our goal is to minimize the probability of a misclassification.
- The optimal decision rule is: decide for $\mathcal{C}_1$ iff

$$p(\mathcal{C}_1|\mathbf{x}) > p(\mathcal{C}_2|\mathbf{x})$$

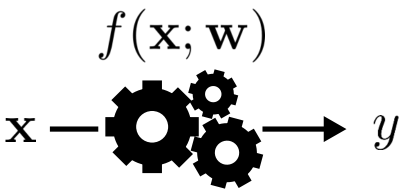
- Or for multiple classes: decide for $\mathcal{C}_k$ iff

$$p(\mathcal{C}_k|\mathbf{x}) > p(\mathcal{C}_j|\mathbf{x}) \ \forall j \neq k$$

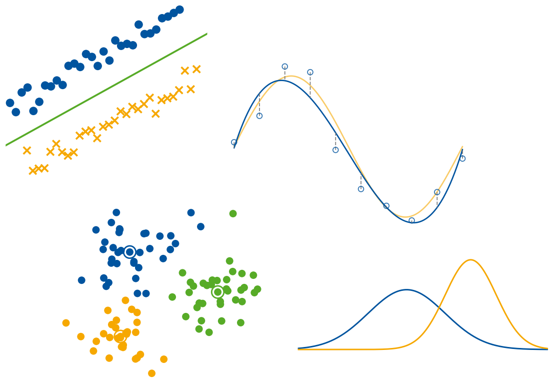
- *Once we can estimate posterior probabilities, we can use this rule to build classifiers.*



Summary: Introduction to ML



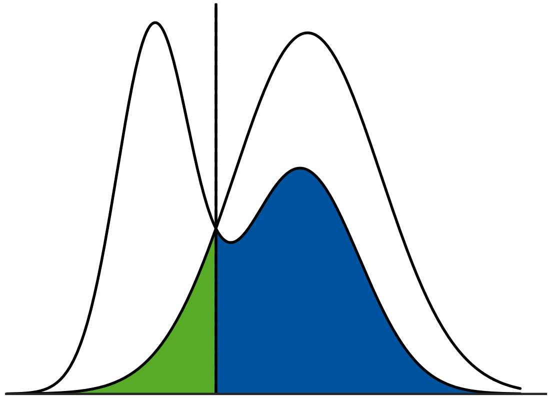
Machine Learning



Forms of Machine Learning

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

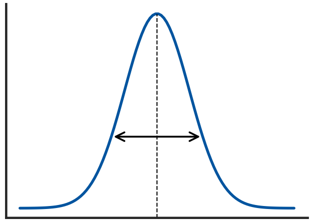
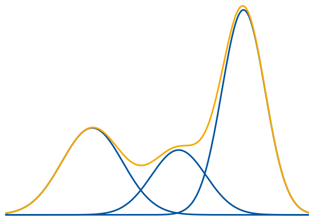
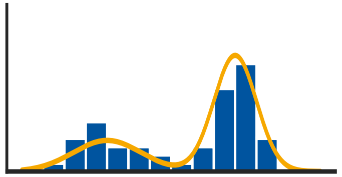
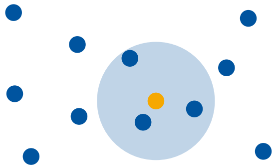
Bayes Theorem



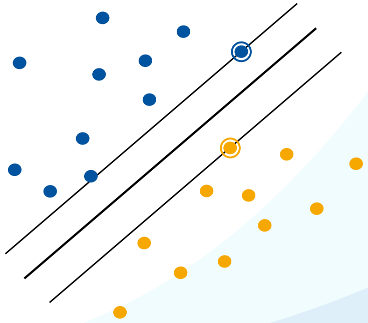
Bayes Optimal Classification

Next Lectures…

- Ways how to estimate the probability densities $p(\mathbf{x}|\mathcal{C})$
  - Parametric methods
    - Gaussian distribution
    - Mixtures of Gaussians
  - Non-parametric methods
    - Histograms
    - k-Nearest Neighbor
    - Kernel Density Estimation

- Ways to directly model the posteriors $p(\mathcal{C}_k|\mathbf{x})$
  - Linear discriminants
  - Logistic regression, SVMs, Neural Networks, …



Machine Learning Topics

1. Introduction to ML
2. **Probability Density Estimation**
3. Linear Discriminants
4. Linear Regression
5. Logistic Regression
6. Support Vector Machines
7. AdaBoost
8. Neural Network Basics

References and Further Reading

- More information, including a short review of Probability theory and a good introduction in Bayes Decision Theory can be found in Chapters 1.1, 1.2 and 1.5 of

Christopher M. Bishop
Pattern Recognition and Machine Learning
Springer, 2006