Learning and Adaptivity

Introduction - Lecture I

Organization

- Lecture
  - Anastassia Küstenmacher
  - Matias Valdenegro
- Course webpage

Organization

- Structure: weekly 1Lecture + 1Exercise
- Place and Time
- Exam

Organization

Exercises

- Typically 1 exercise sheet every week.
- Mostly Python based exercises.
- Hands-on exercises with the algorithms from the lecture

Teams are encouraged

- You can form teams of up to 3 people for the exercises
- Each team submits one solution with the list of all team members in the submission

Sources

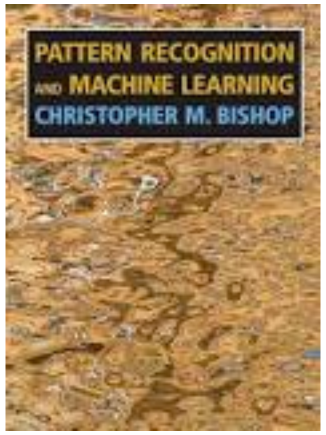
- Textbooks

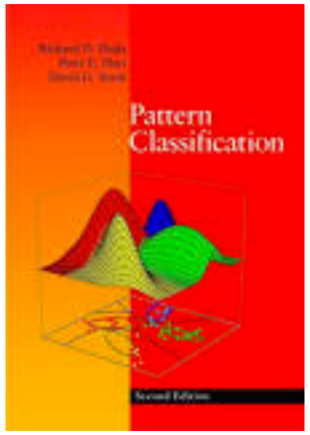
C. M. Bishop Pattern Recognition and Machine Learning
Springer, 2006

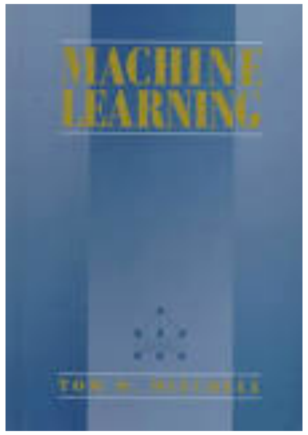
R.O. Duda, P.E. Hart, D.G. Stork Pattern Classification 2nd Ed.,
Wiley-Interscience, 2000

T. Mitchell, Machine Learning, McGraw Hill, 1997

- Research papers will be given out for some topics.







Material

- Slides are based on
    - Prof. Gerhard Kraetzschmar
    - Prof. Bastian Leibe RWTH Aachen

Course Outline

Basic Concepts

- Parametric Method,
- Bayesian Learning and Nonparametrics Methods
- Clustering and Mixture of Gaussians

Supervised Learning, Classification Approaches

- Ensemble Methods and Boosting
- Randomized Trees, Forest

Unsupervised Learning

- Dimensionality Reduction and Manifold Learning (PCA, SNE/t-SNE, MDS, umap)
- Uncertainty Estimation

Reinforcement Learning

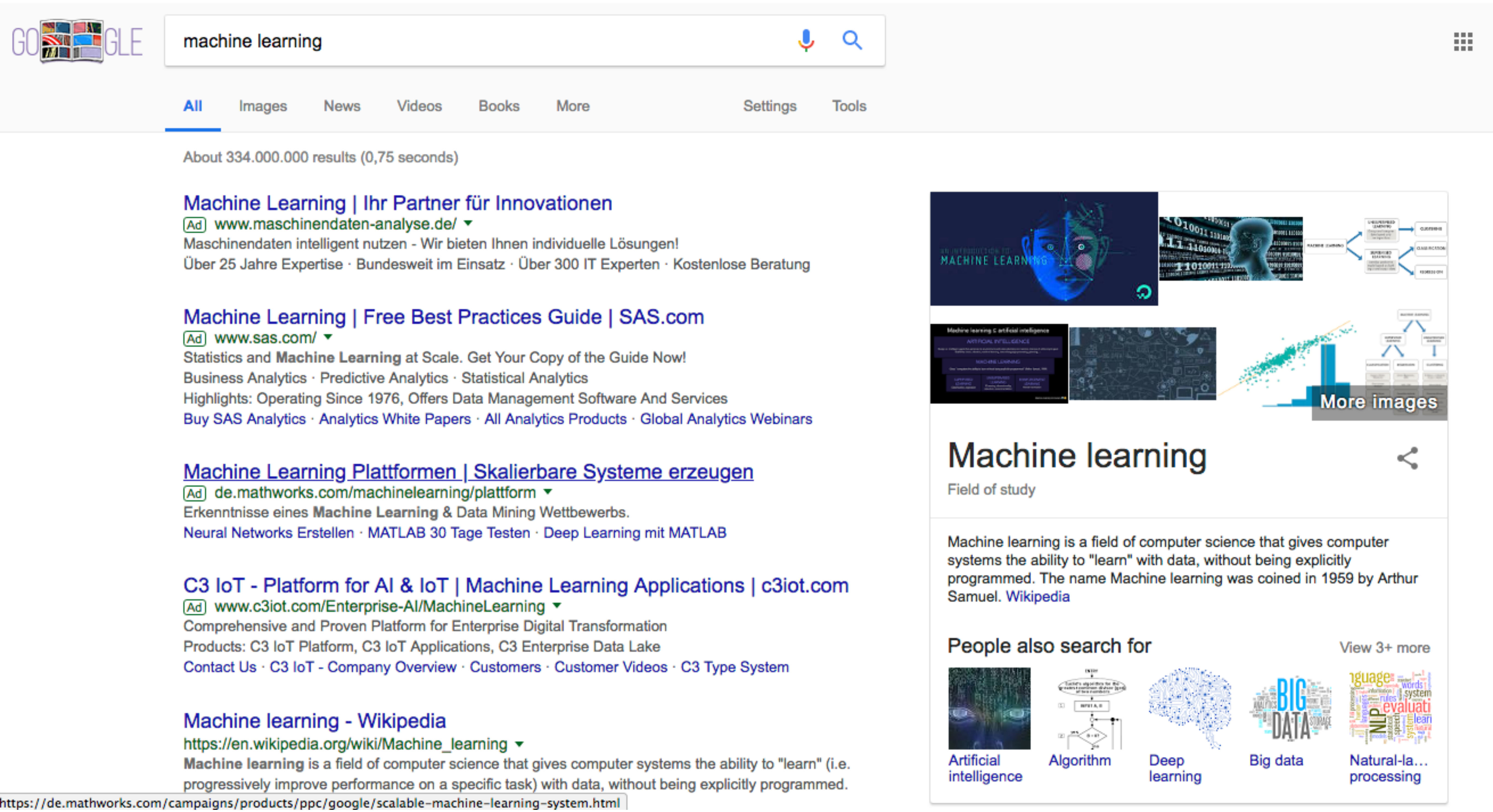
- Classical Reinforcement Learning
- Deep Reinforcement Learning

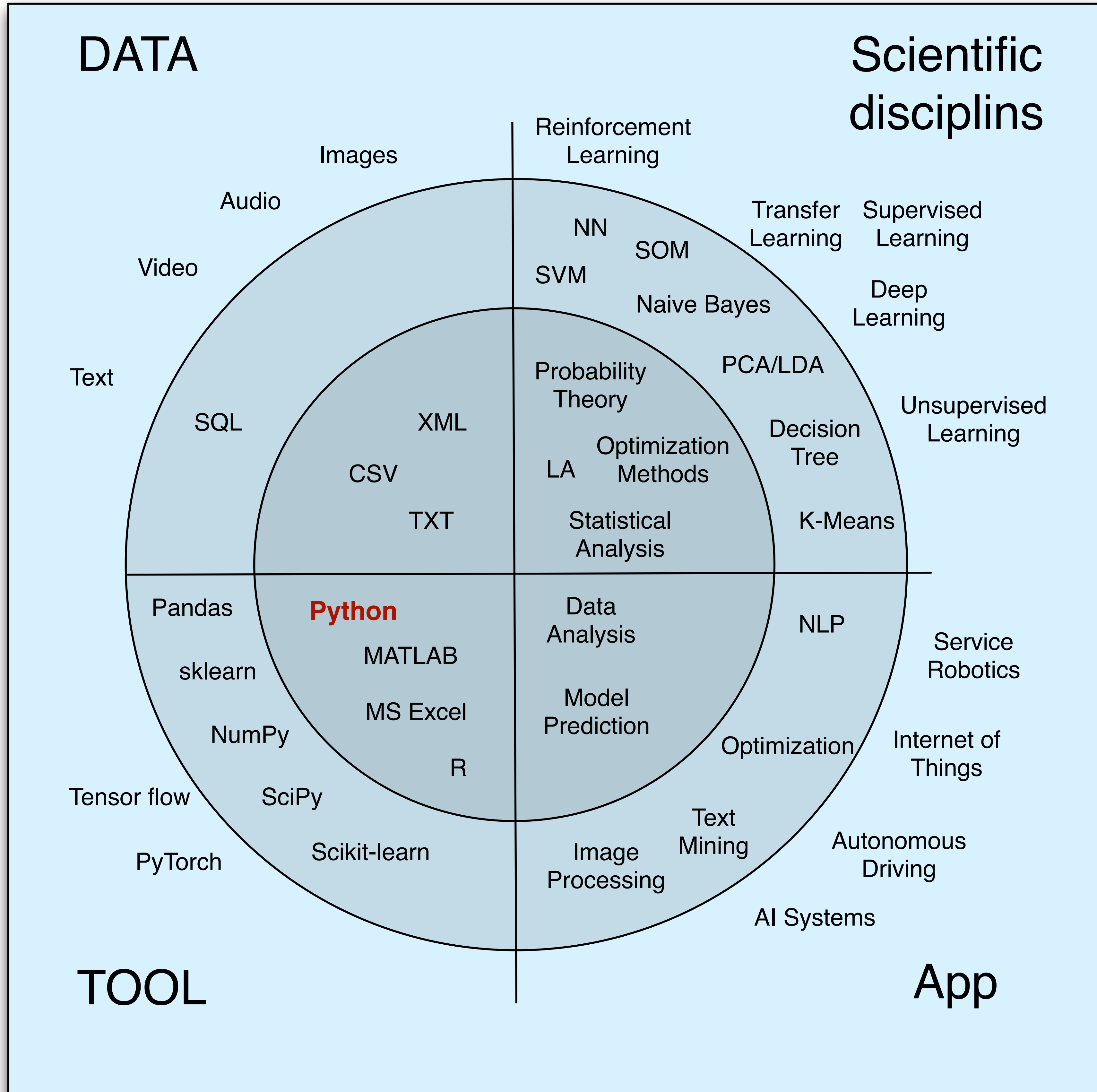
Motivation

Already everywhere

- Data Security
- Speech recognition (e.g. speed-dialing)
- Computer vision (e.g. face detection)
- Hand-written character recognition (e.g. letter delivery)
- Information retrieval (e.g. image & video indexing)
- Operation systems (e.g. caching)
- Fraud detection (e.g. credit cards)
- Text filtering (e.g. email spam filters)
- Game playing (e.g. strategy prediction)
- Robotics (e.g. prediction of battery lifetime)

Information retrieval





Scheme based on Shahzad Cheema scheme

Why Machine Learning?

Goal

- Machines that **learn** to **perform** a **task** from **experience**

Reason?

- Crucial component of every intelligent/autonomous systems
- Important for a system's adaptability
- Important for a system's generalisation capabilities
- Attempt to understand human learning

What is Machine Learning?

Learning to perform a task from experience

Learning

- Most important part here!
- We do not want to encode the knowledge ourselves.
- The machine should **learn** the relevant criteria automatically from past observations and **adapt** to the given situation.

Tools:

- Statistics, Probability theory, Decision theory, Information theory, Optimization theory

What is Machine Learning?

Learning to perform a **task** from experience

Task

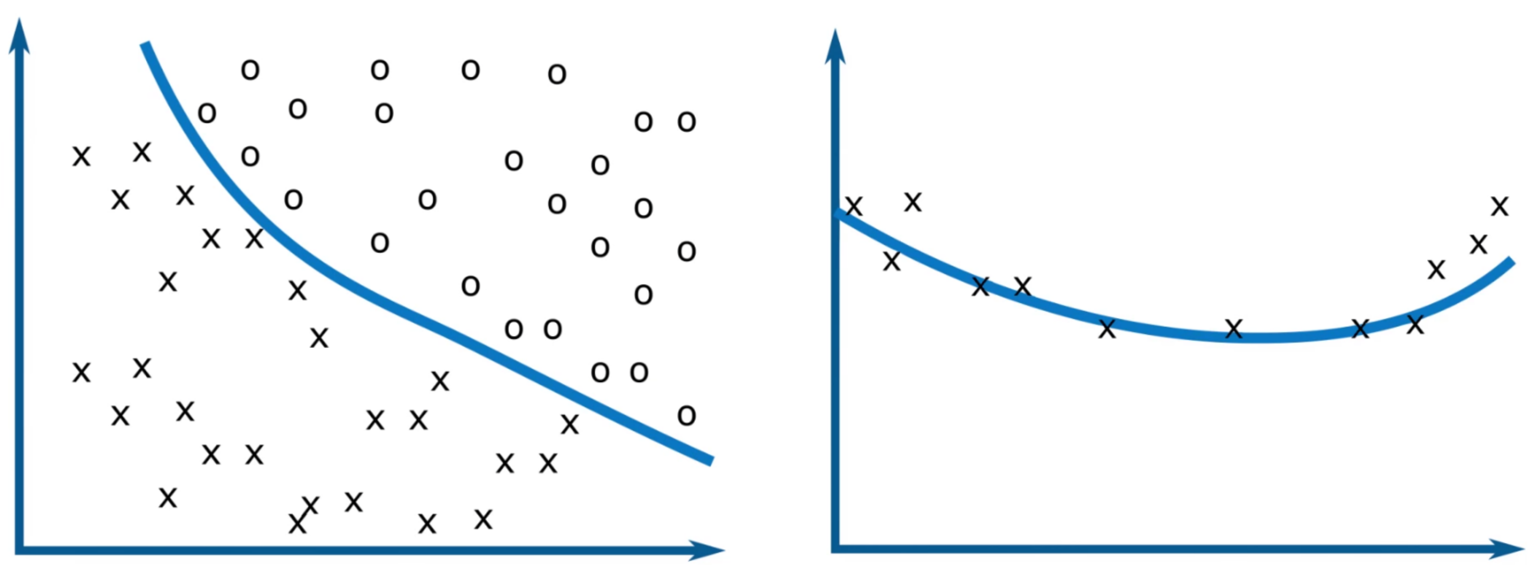
- Can often be expressed through a mathematical function

Output $\rightarrow y = f(x, w) \leftarrow$ *Parameters (what is "learned")*

Input

Classification vs. Regression

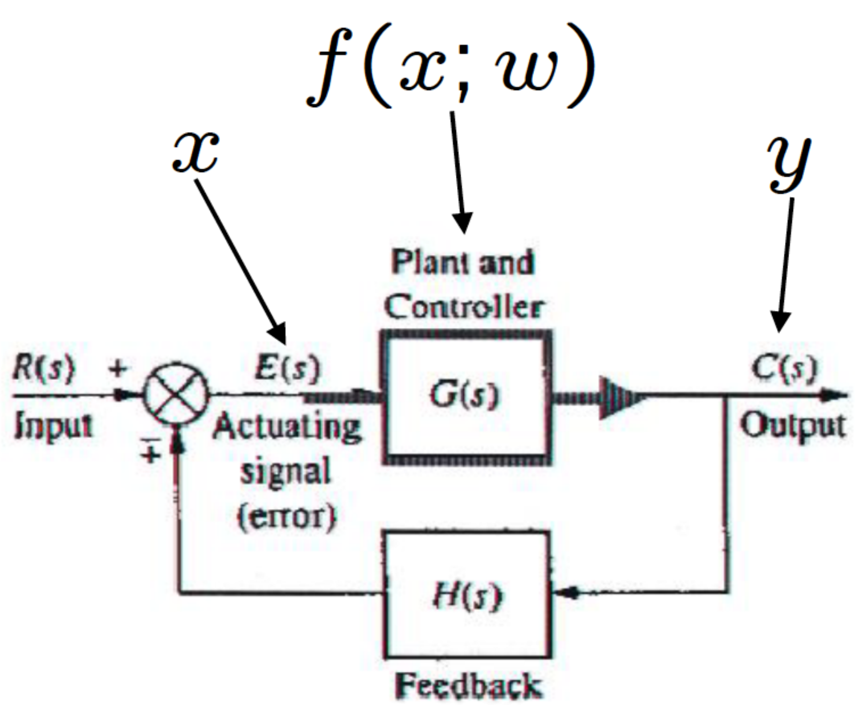
Property	Classification	Regression
Output type	Discrete (labels)	Continuous (number)
What do we search?	Decision boundary	Best fit line'
Evaluation	Accuracy	Sum of squared error



Pictures are taken from 'Machine Learning Introduction: Regression and Classification | Intel Software'

Examples

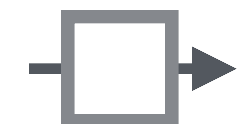
Regression



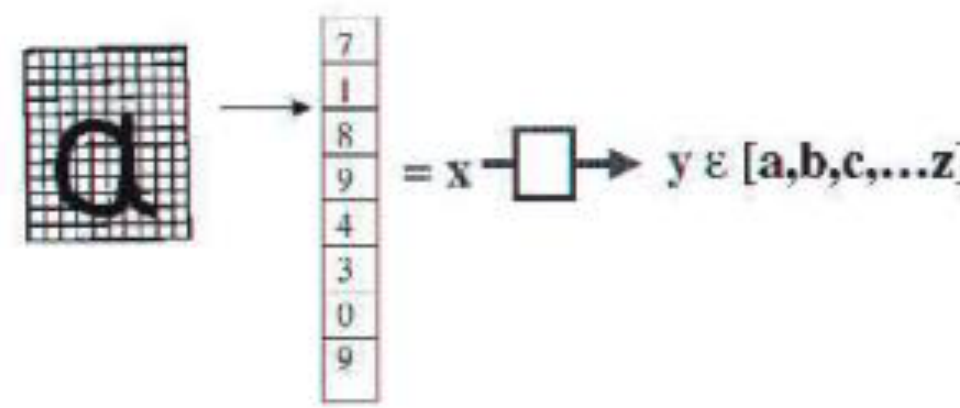
Classification

Email filtering

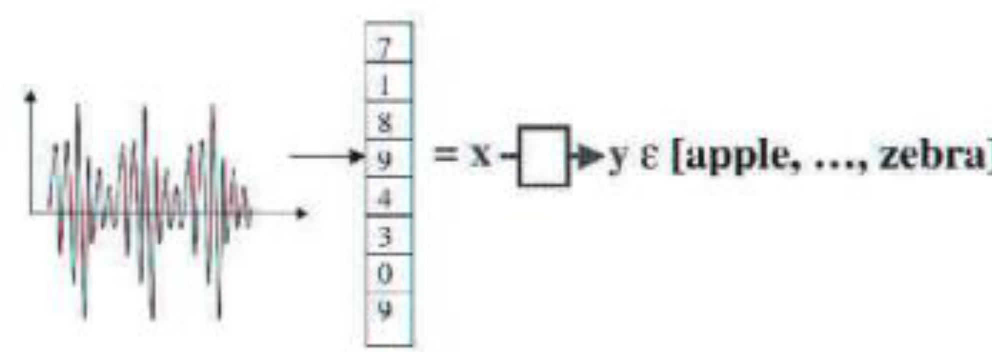
$x \in$ [a-z]
$y \in$ [important, spam]



Character recognition



Speech recognition



What is Machine Learning?

Learning to **perform** a task from experience

Performance: "99% correct classification"

- Of what???
- Characters? Words? Sentences?
- Speaker/writer independent?
- Over what data set?
- …

"The car drives without human intervention 99% of the time on country roads"

What is Machine Learning?

Learning to **perform** a task from experience

Performance measure: Typically one number

- % correctly classified letters
- Average driving distance (until crash…)
- % games won
- % correctly recognized words, sentences, answers

Generalization performance

- Training vs. test
- “All” data

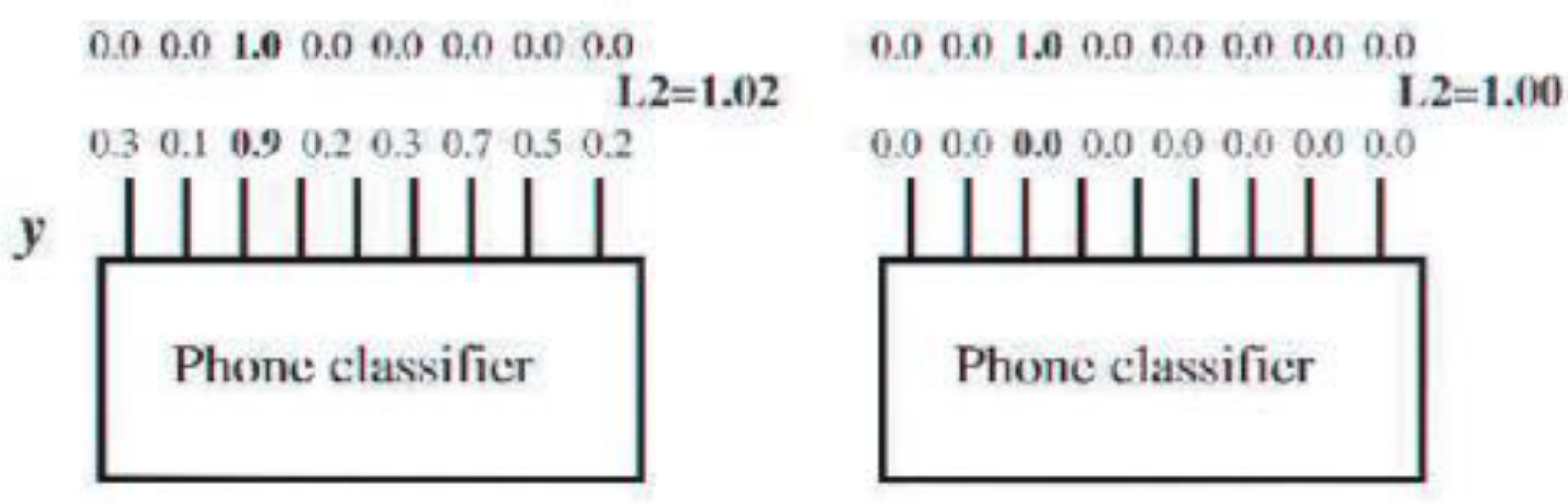
What is Machine Learning?

Learning to **perform** a task from experience

Performance measure: more subtle problem

- Also necessary to compare partially correct outputs.
- How do we weight different kinds of errors?

- Example: L2 norm



What is Machine Learning?

Learning to perform a task from **experience**

What data is given?

Data with labels: **supervised learning**

- Images / speech with target labels
- Car sensor data with target steering signal

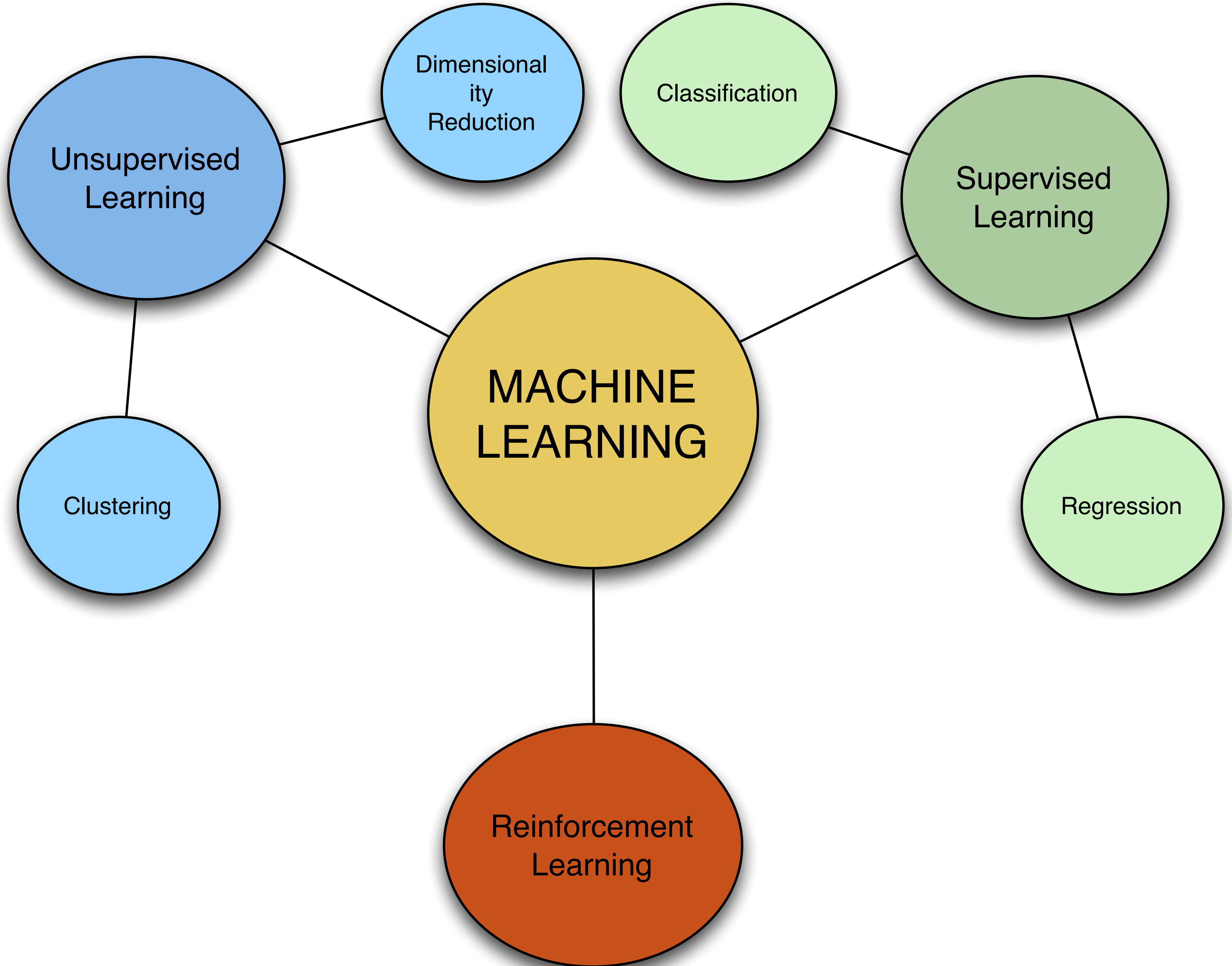
Data without labels: **unsupervised learning**

- Automatic clustering of sounds and phonemes
- Automatic clustering of web sites

Some data with, some without labels: **semi-supervised learning**

No examples: **learning by doing**

Feedback/rewards: **reinforcement learning**



Learning to perform a task from **experience**

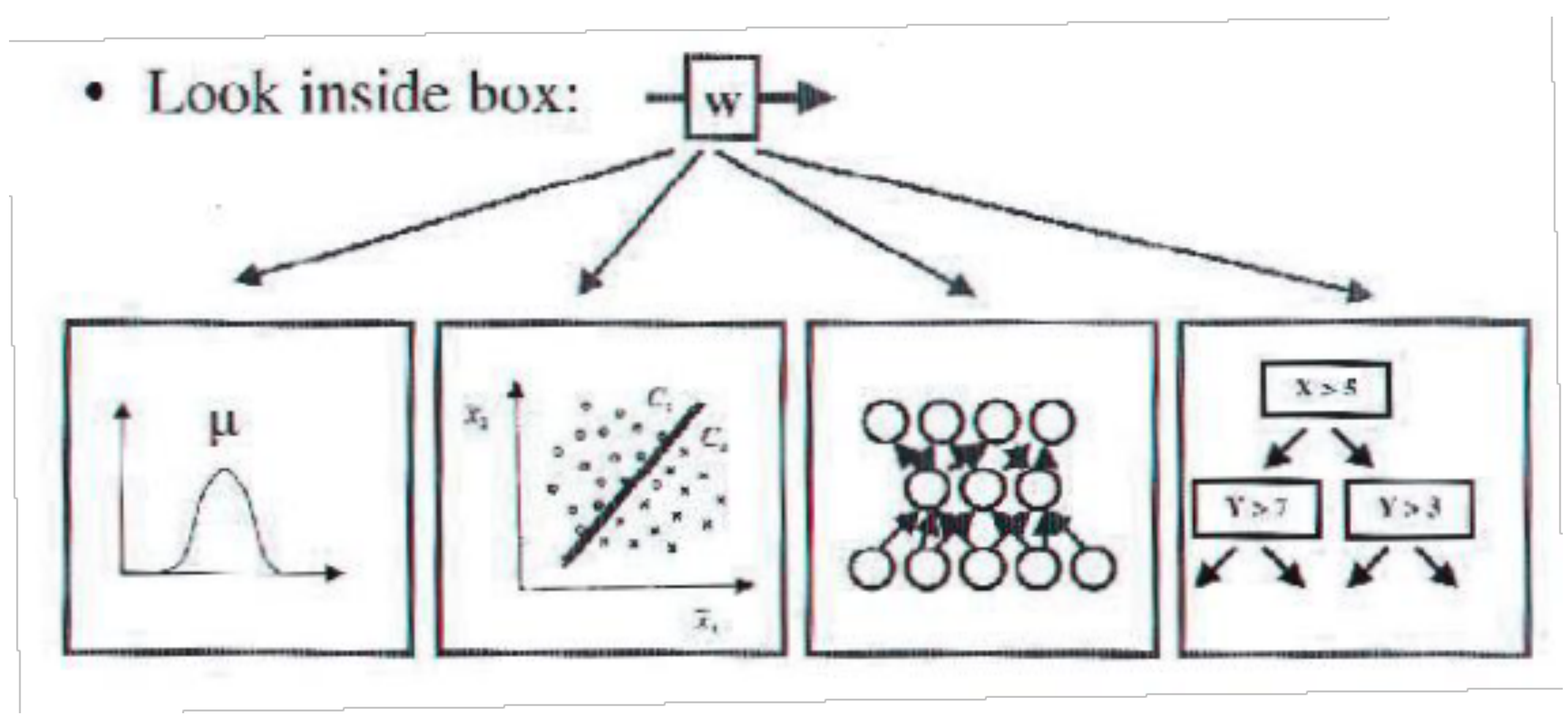
What is Machine Learning?

$$y = f(x; w)$$

w: characterizes the family of functions

w: indexes the space of hypotheses

w: vector, connection matrix, graph, …

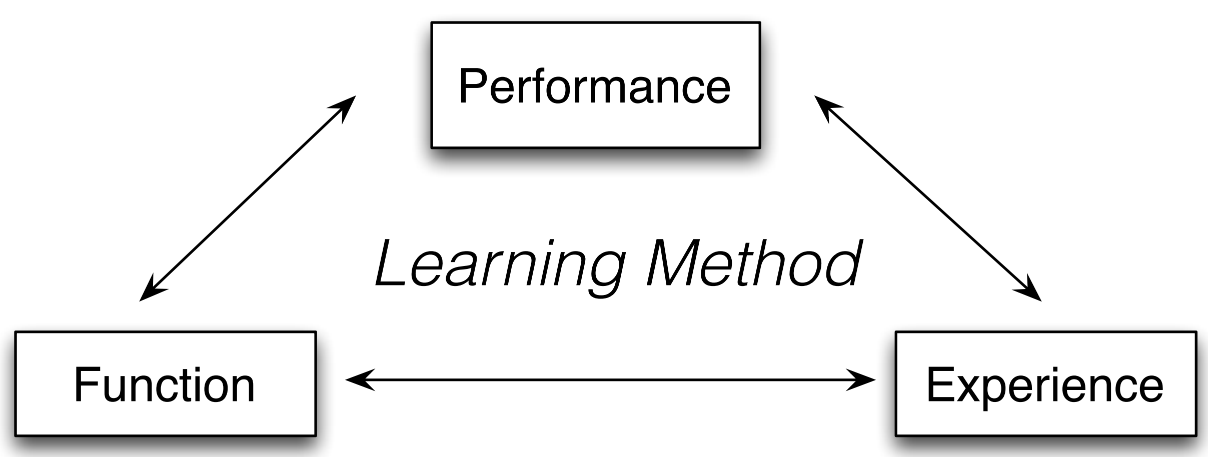


What is Machine Learning?

Learning to perform a task from experience

Learning

- Most often learning = optimization
- Search in hypothesis space
- Search for the "best" function / model parameter w
  - I.e. maximize $y = f(x; w)$ w.r.t. the performance measure



Course Outline

Basic Concepts

- Parametric Method,
- Bayesian Learning and Nonparametrics Methods
- Clustering and Mixture of Gaussians

Supervised Learning, Classification Approaches

- Ensemble Methods and Boosting
- Randomized Trees, Forest

Unsupervised Learning

- Dimensionality Reduction and Manifold Learning (PCA, SNE/t-SNE, MDS, umap)
- Uncertainty Estimation

Reinforcement Learning

- Classical Reinforcement Learning
- Deep Reinforcement Learning

Today's Lecture

Fundamentals

- Recall of Probability Theory
- Probabilities
- Probability densities
- Expectations and covariances

Bayes Decision Theory

- Basic concepts
- Minimizing the misclassification rate
- Minimizing the expected loss

Probability Theory

"Probability theory is nothing but common sense reduced to calculation."

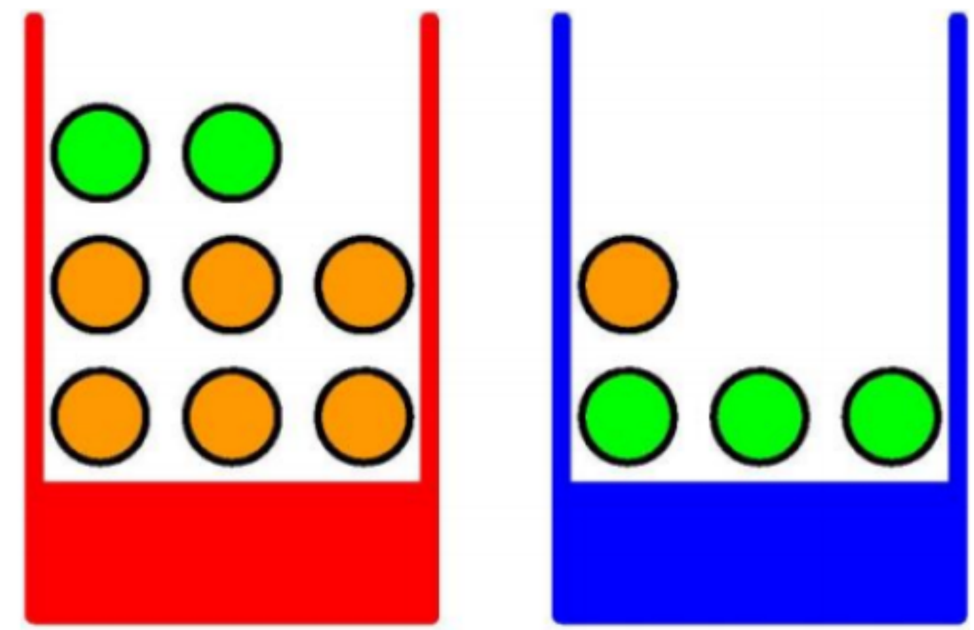
Pierre-Simon de Laplace, 1749-1827

Slide source: B. Leibe Slides ML, 2015

Probability Theory

Example: apples and oranges

- We have two boxes to pick from.
- Each box contains both types of fruit.
- What is the probability of picking an apple?



Formalization

- Let $B \in \{r,b\}$ be a random variable for the box we pick.
- Let $F \in \{a,o\}$ be a random variable for the type of fruit we get.
- Suppose we pick the red box 40% of the time. We write this as

$$p(B=r)=0.4 \qquad p(B=b)=0.6$$

- The probability of picking an apple given a choice for the box is

$$p(F=a|B=r)=0.25 \qquad p(F=a|B=r)=0.25$$

- What is the probability of picking an apple?

$$p(F=a)=?$$

Slide source: B. Leibe Slides ML, 2015

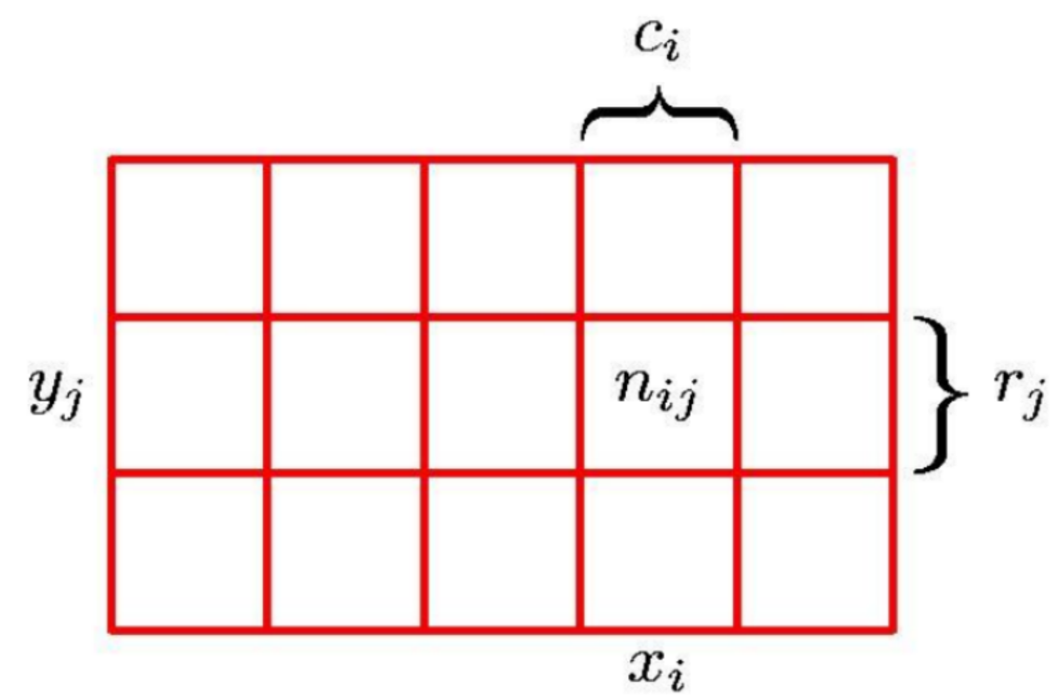
Probability Theory

More general case

Consider two random variables $X \in \{x_i\}$ and $Y \in \{y_j\}$

Consider N trials and let

$$\begin{aligned} n_{ij} &= \#\{X = x_i \wedge Y = y_j\} \\ c_i &= \#\{X = x_i\} \\ r_j &= \#\{Y = y_j\} \end{aligned}$$



Then we can derive

Joint probability

$$p(X = x_i, Y = y_j) = \frac{n_{ij}}{N}$$

Marginal probability

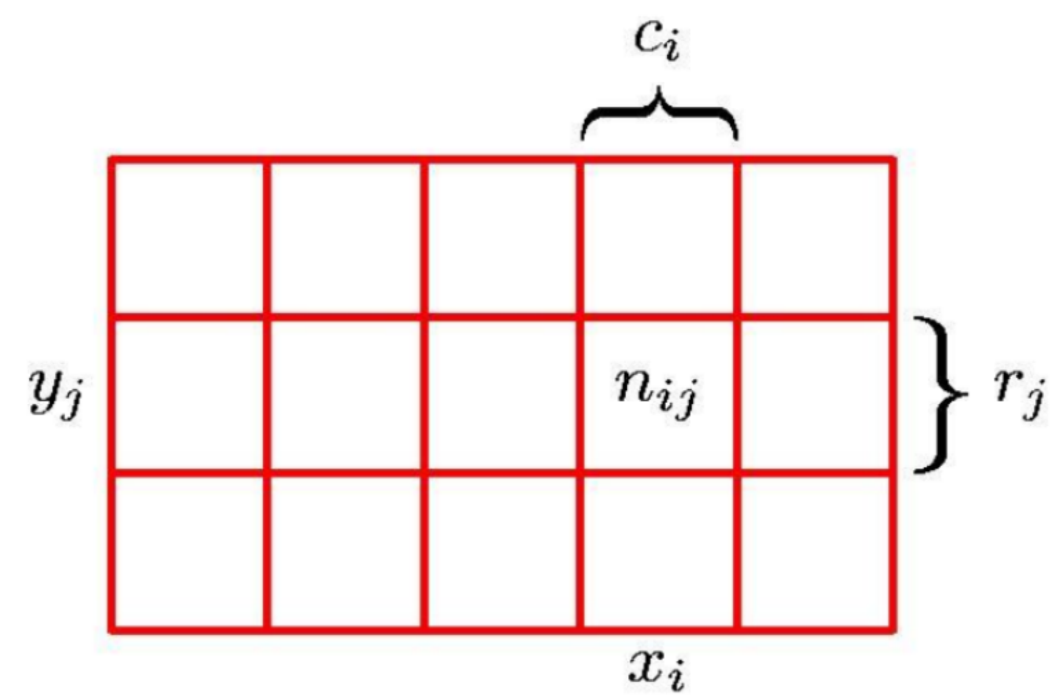
$$p(X = x_i) = \frac{c_i}{N}.$$

Conditional probability

$$p(Y = y_j | X = x_i) = \frac{n_{ij}}{c_i}$$

Slide source: B. Leibe Slides ML, 2015

Probability Theory



Rules of probability

Sum Rule

$$p(X = x_i) = \frac{c_i}{N} = \frac{1}{N} \sum_{j=1}^{L} n_{ij} = \sum_{j=1}^{L} p(X = x_i, Y = y_j)$$

Product Rule

$$\begin{aligned} p(X = x_i, Y = y_j) &= \frac{n_{ij}}{N} = \frac{n_{ij}}{c_i} \cdot \frac{c_i}{N} \\ &= p(Y = y_j | X = x_i) p(X = x_i) \end{aligned}$$

Slide source: B. Leibe Slides ML, 2015

The Rules of Probability

Thus we have

Sum Rule $$p(X) = \sum_Y p(X, Y)$$

Product Rule $$p(X, Y) = p(Y|X)p(X)$$

From these we can derive

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

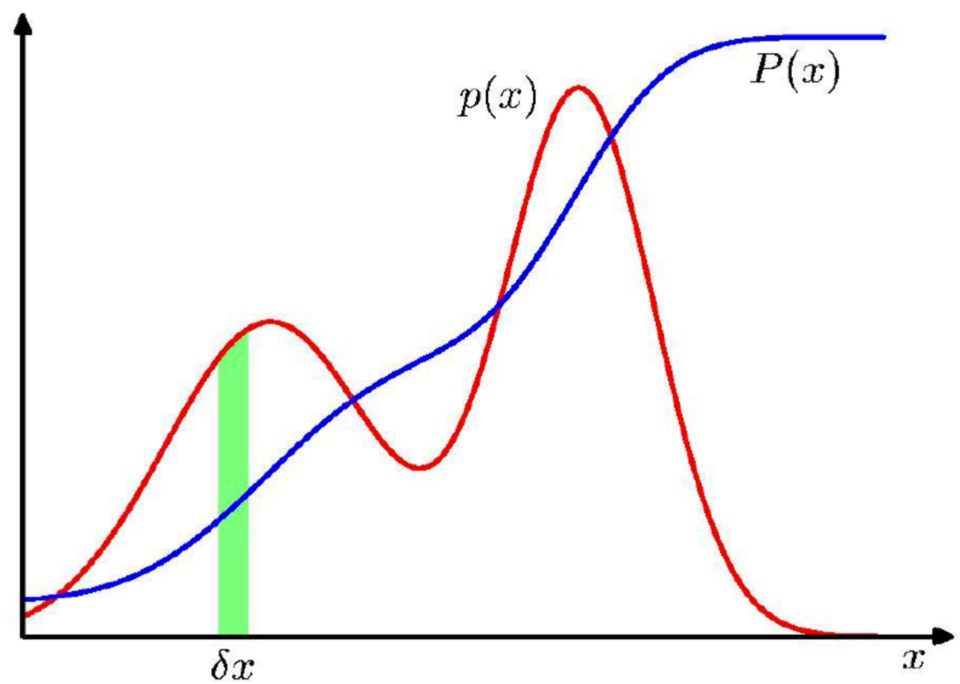
where $$p(X) = \sum_Y p(X|Y)p(Y)$$

Slide source: B. Leibe Slides ML, 2015

Probability Densities

Probabilities over continuous variables are defined over their probability density function (pdf) *p(x)*

$$p(x \in (a, b)) = \int_a^b p(x)\,\mathrm{d}x$$



The probability that *x* lies in the interval *(-inf,z)* is given by the cumulative distribution function

$$P(z) = \int_{-\infty}^{z} p(x)\,\mathrm{d}x$$

Slide source: B. Leibe Slides ML, 2015

Expectations

The average value of some function *f(x)* under a probability distribution *p(x)* is called its **expectation**

$$\mathbb{E}[f] = \sum_x p(x) f(x)$$

discrete case

$$\mathbb{E}[f] = \int p(x) f(x) \, \mathrm{d}x$$

continuous case

If we have a finite number *N* of samples drawn from a pdf, then the expectation can be approximated by

$$\mathbb{E}[f] \simeq \frac{1}{N} \sum_{n=1}^{N} f(x_n)$$

We can also consider a **conditional expectation**

$$\mathbb{E}_x[f|y] = \sum_x p(x|y) f(x)$$

Variances and Covariances

The **variance** provides a measure how much variability there is in *f(x)* around its mean value E[*f(x)*].

$$\mathrm{var}[f] = \mathbb{E}\left[(f(x) - \mathbb{E}[f(x)])^2\right] = \mathbb{E}[f(x)^2] - \mathbb{E}[f(x)]^2$$

For two random variables *x* and *y*, the **covariance** is defines by

$$\begin{aligned}\mathrm{cov}[x, y] &= \mathbb{E}_{x,y}\left[\{x - \mathbb{E}[x]\}\{y - \mathbb{E}[y]\}\right] \\ &= \mathbb{E}_{x,y}[xy] - \mathbb{E}[x]\mathbb{E}[y]\end{aligned}$$

If x and y are vectors then the result is **covariance matrix**

$$\begin{aligned}\mathrm{cov}[\mathbf{x}, \mathbf{y}] &= \mathbb{E}_{\mathbf{x},\mathbf{y}}\left[\{\mathbf{x} - \mathbb{E}[\mathbf{x}]\}\{\mathbf{y}^{\mathrm{T}} - \mathbb{E}[\mathbf{y}^{\mathrm{T}}]\}\right] \\ &= \mathbb{E}_{\mathbf{x},\mathbf{y}}[\mathbf{x}\mathbf{y}^{\mathrm{T}}] - \mathbb{E}[\mathbf{x}]\mathbb{E}[\mathbf{y}^{\mathrm{T}}]\end{aligned}$$

Slide source: B. Leibe Slides ML, 2015

Bayes Decision Theory

Thomas Bayes, 1701-1761

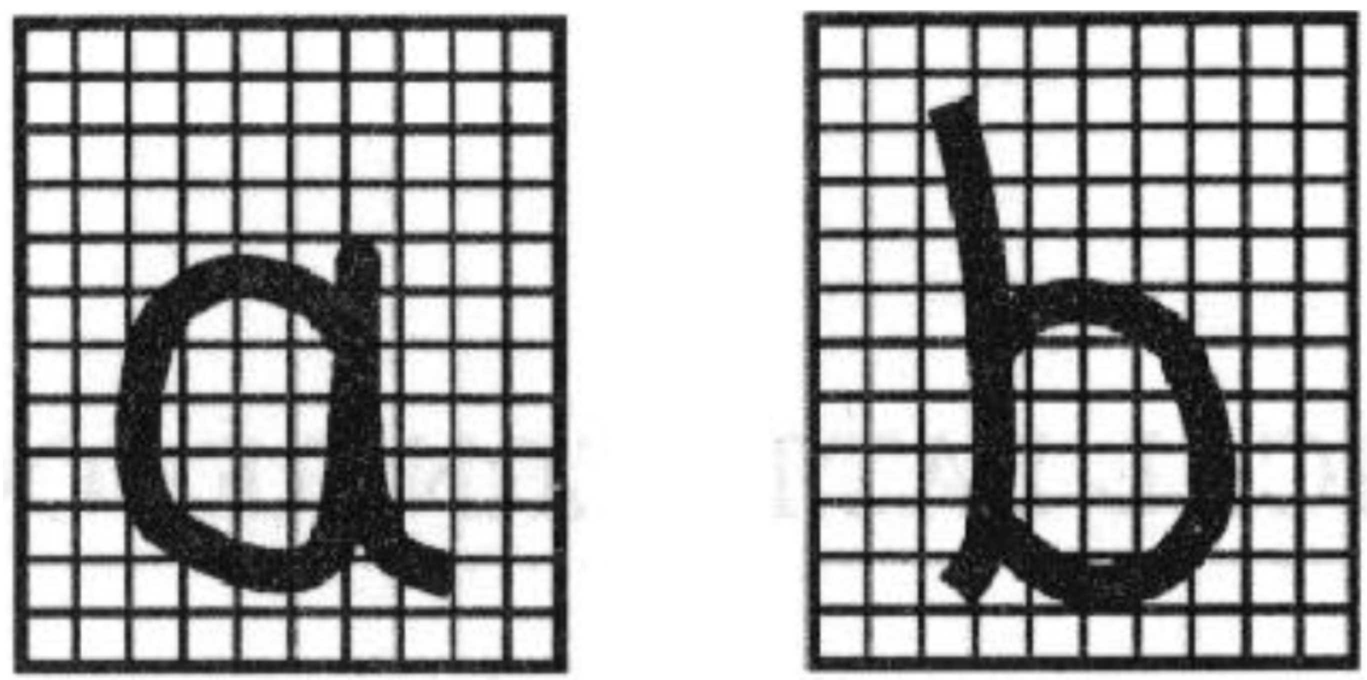
"The theory of inverse probability is founded upon an error, and must be wholly rejected."

R.A. Fisher, 1925

Slide source: B. Leibe Slides ML, 2015

Bayes Decision Theory

Example: handwritten character recognition



Goal:

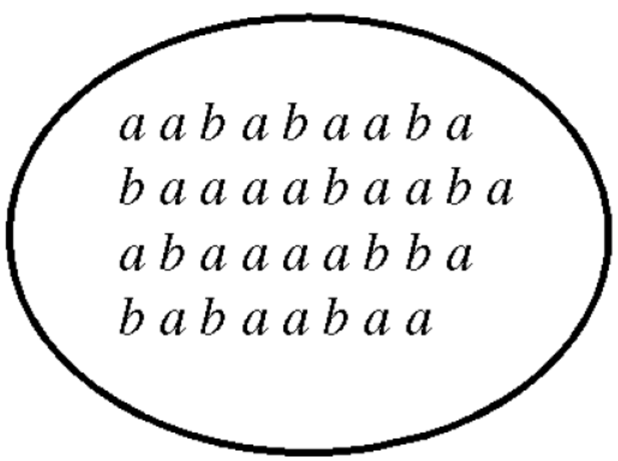
Classify a new letter such that the probability of misclassification is minimised

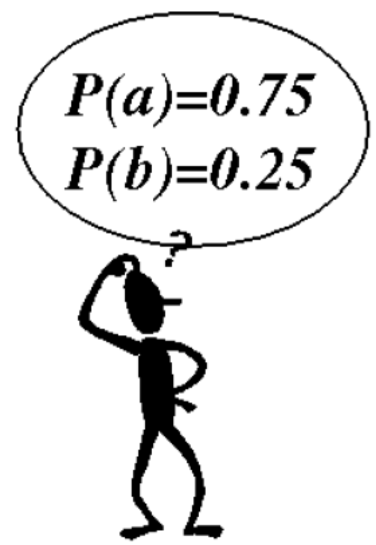
Slide source: B. Leibe Slides ML, 2015

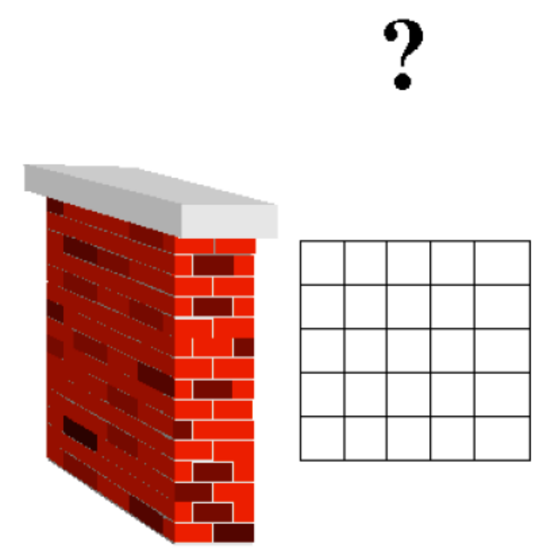
Bayes Decision Theory

Concept 1: **Priors** (a priori probabilities) $\boxed{p\left(C_k\right)}$

- What we can tell about the probability before seeing the data.
- Example:







$$C_1 = a \qquad p\left(C_1\right) = 0.75$$

$$C_2 = b \qquad p\left(C_2\right) = 0.25$$

In general: $\sum_k p(C_k) = 1$

Slide source: B. Leibe Slides ML, 2015

Bayes Decision Theory

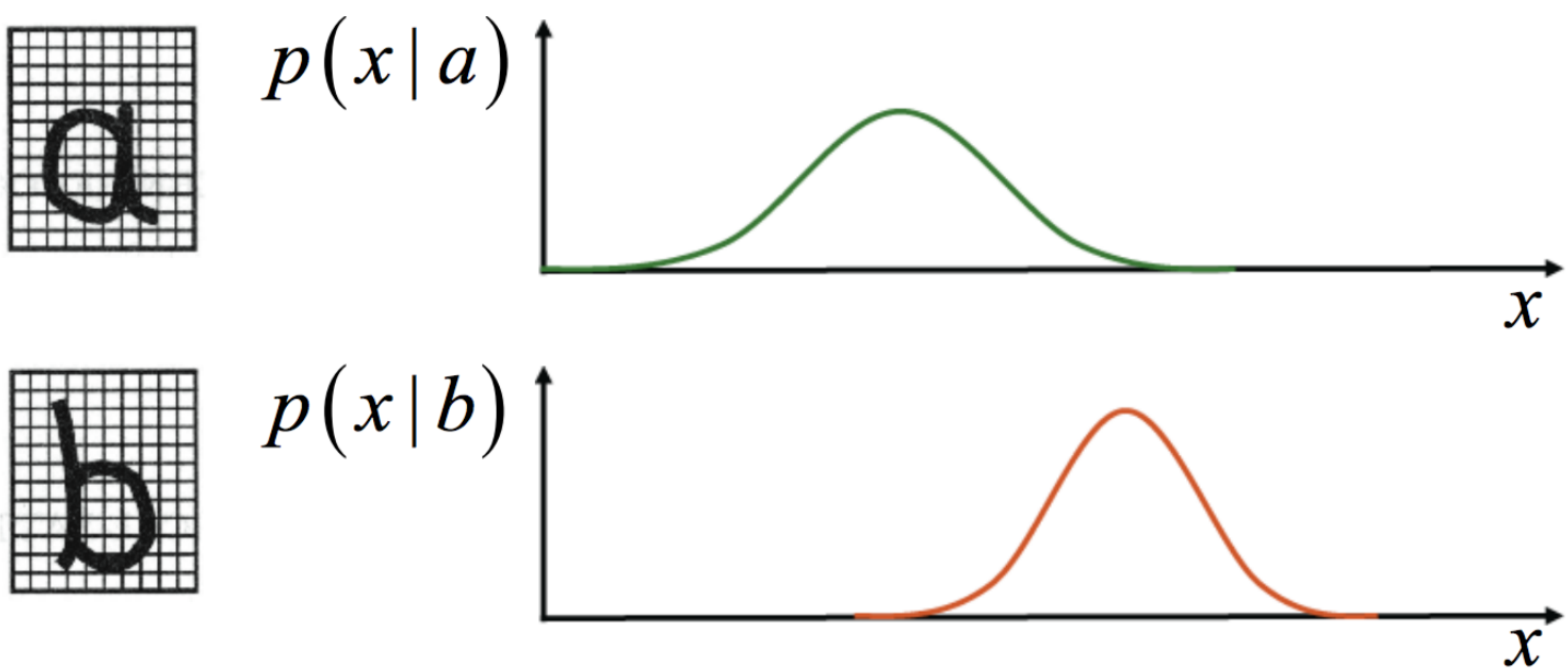
$$\boxed{p(x|C_k)}$$

Concept 2: **Conditional probabilities**

- Let x be a feature vector.
- x measures/describes certain properties of the input.

–E.g. number of black pixels, aspect ratio, …

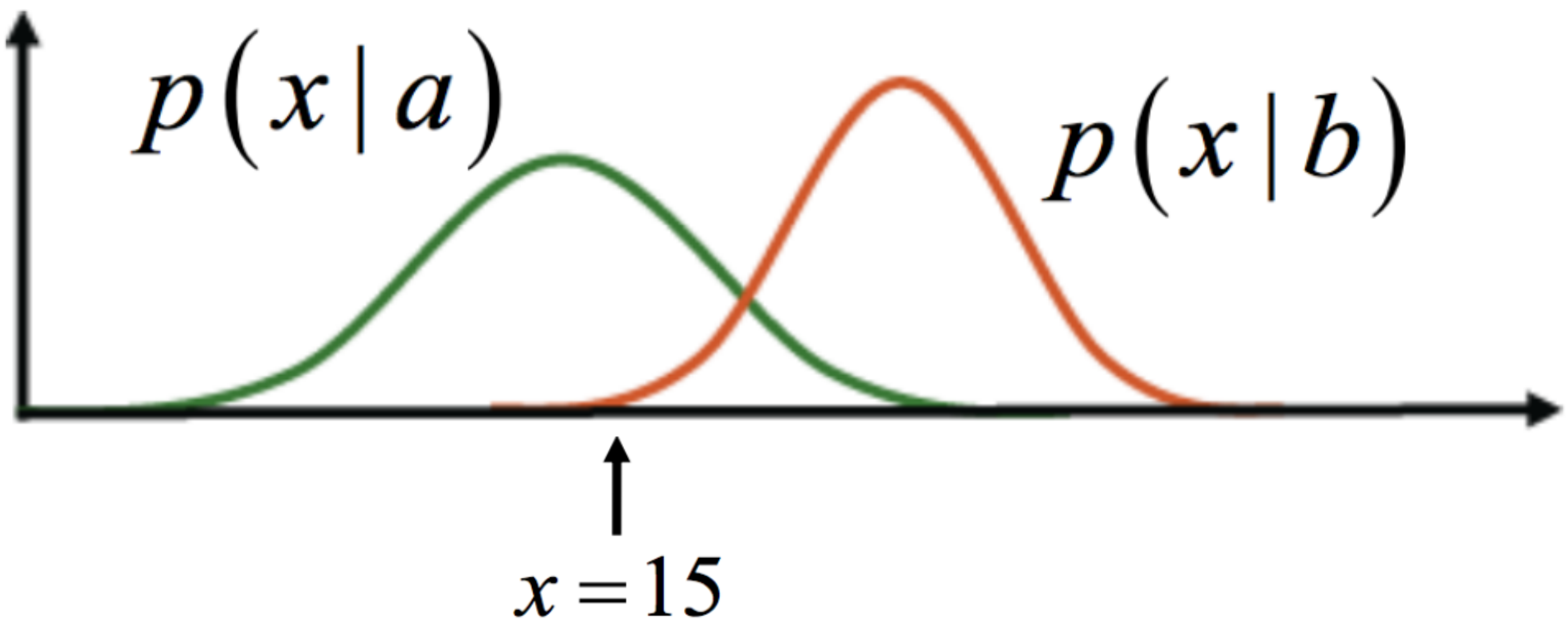
- $p(x|C_k)$ describes its **likelihood** for class C_k.



Slide source: B. Leibe Slides ML, 2015

Bayes Decision Theory

Example:

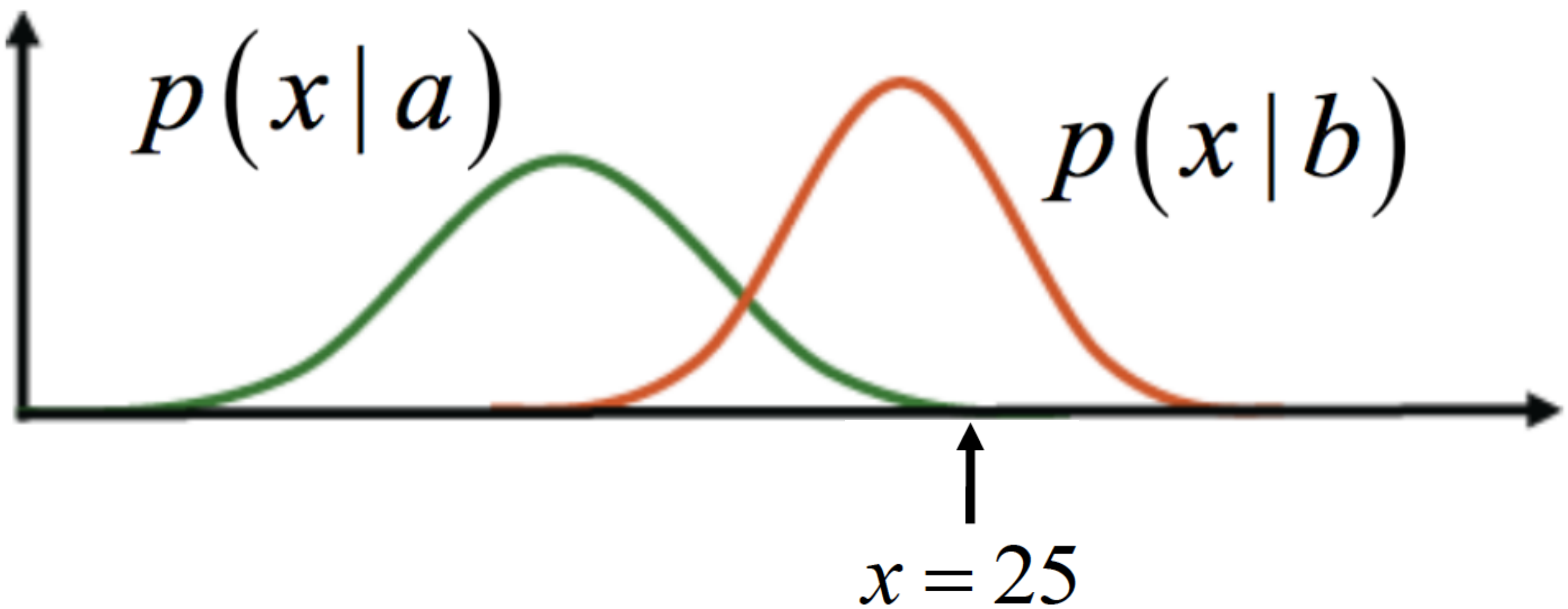


Question:

- Which class?
- Since $p(x|b)$ is much smaller than $p(x|a)$, the decision should be ‘a’ here.

Bayes Decision Theory

Example:



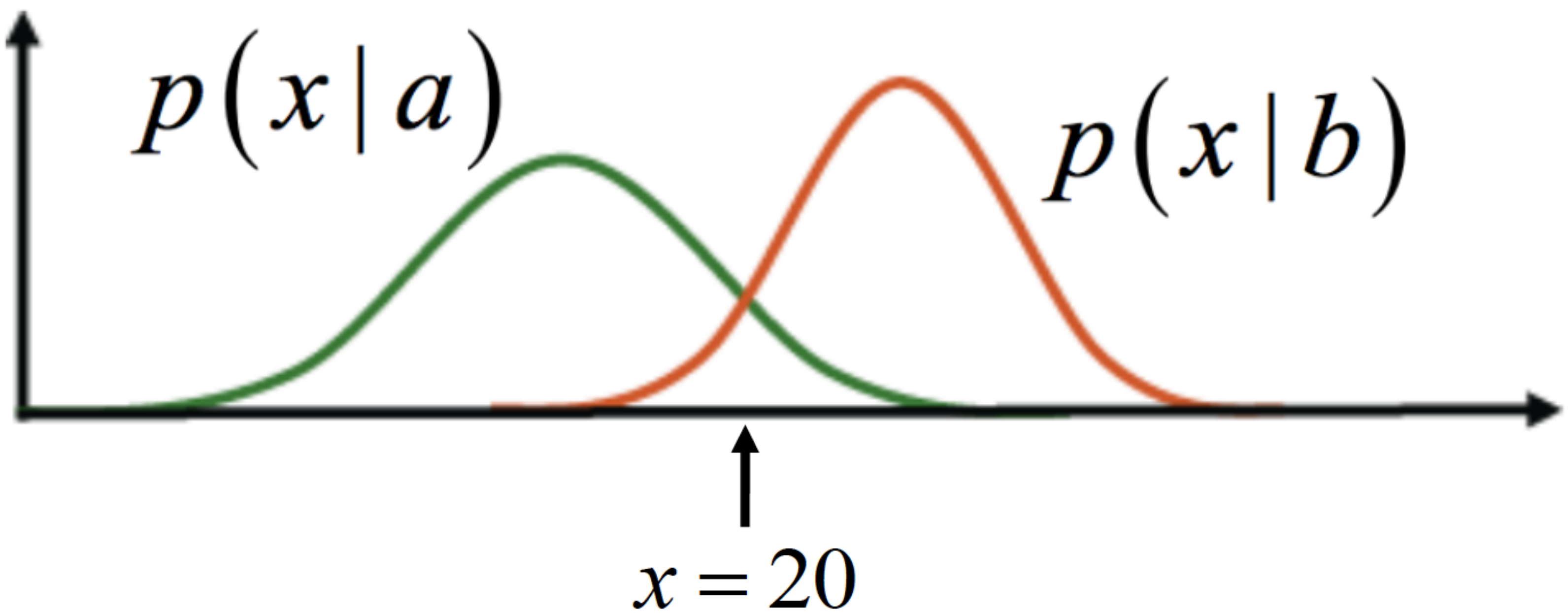
Question:

- Which class?
- Since $p(x|a)$ is much smaller than $p(x|b)$, the decision should be '*b*' here.

Slide source: B. Leibe Slides ML, 2015

Bayes Decision Theory

Example:



Question:

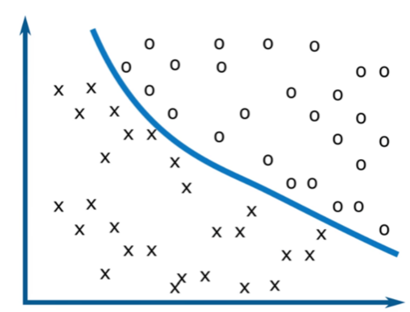
- Which class?
- Remember that $p(a)$ = 0.75 and $p(b)$ = 0.25…
- I.e., the decision should be again 'a'.

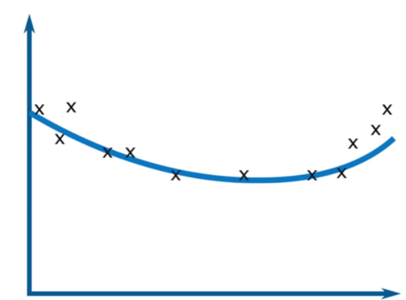
=> How can we formalize this?.

Summary

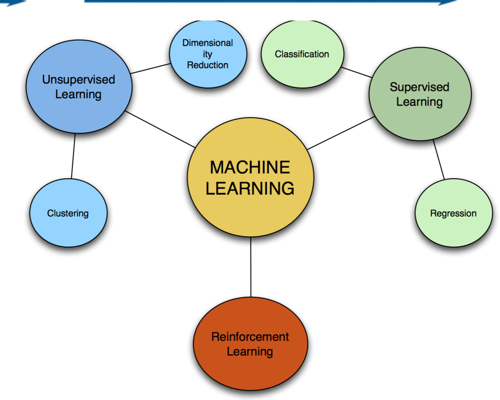
Machines that **learn** to **perform** a **task** from **experience**

Classification vs. Regression

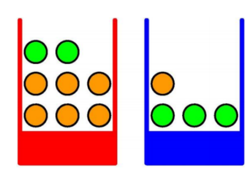




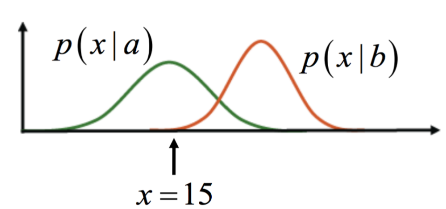
Type of data -> type of ML algorithm



Probability theory



Bayes Decision Theory



Next Lecture

Bayes Decision Theory

- Basic concepts
- Minimizing the misclassification rate
- Minimizing the expected loss

Probability Density Estimation

- General concepts
- Gaussian distribution

Parametric Methods

- Maximum Likelihood approach
- Bayesian vs. Frequentist view on probability
- Bayesian Learning

Readings

- 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