Elements of Data Science and Artificial Intelligence

Bernt Schiele

Max Planck Institute for Informatics
Saarland Informatics Campus

November 4, 2019

Today's Topics

- Overview Machine Learning
  - What is machine learning ?
  - Different problem settings and examples
- Decision theory, Inference and Decision
- Introduction (Deep) Artificial Neural Networks

Machine Learning

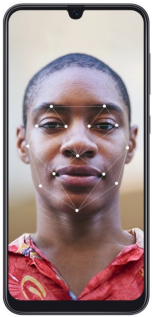
Overview

Machine Learning – what's that?

- Can you think of an application ?
- Do you use machine learning systems already ?

Face Detection

- on your smart phone
- on your digital camera



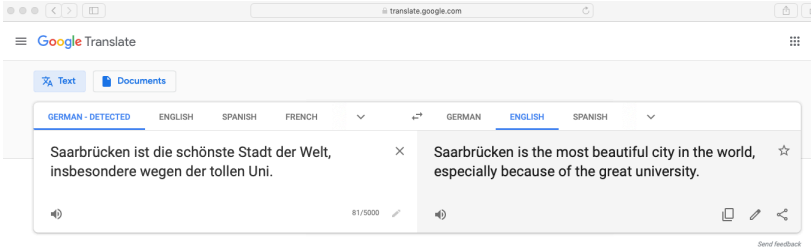
Spam Filtering

image source: quora.com

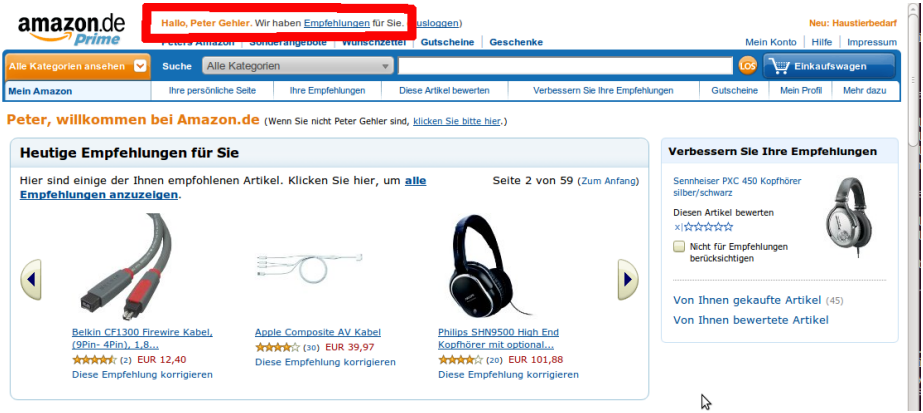
Language Translation

image sources: viesupport.com, translate.google.com

Example:



Product Recommendations e.g. by Amazon



Recommender Systems

Autonomous / Self-Driving Cars

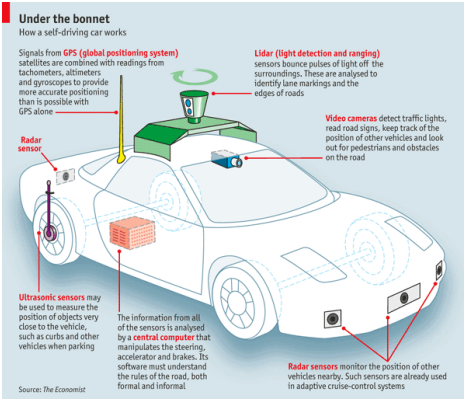


image source: google images, economist.com

Machine Learning – what's that?

- Can you define the term "Machine Learning"?

Machine Learning – A First Definition

Arthur Samuel (1959)

- Machine Leaning: Field of study that gives computers the ability to learn without being explicitly programmed

In the examples given before:

- Spam Filtering: user labels spam emails – ML should learn from those labels
- Face Detection: position of faces is annotated by Apple/Samsung/... in many pictures – ML should learn from these annotations
- Language Translation: language experts translates many sentences from language X to language Y – ML should learn to automatically translate from these sentences
- Autonomous Driving: Humans drive cars around – ML should learn by observing the drivers

Machine Learning – A slightly more Formal Definition

- Goal of machine learning:
  - Machines that learn to perform a task from experience
- We can formalize this as

$$y = f(x; w) \tag{1}$$

  y is called *output variable*,
  x the *input variable* and
  w the model parameters (typically learned)
- Classification vs regression:
  - regression: y continuous
  - classification: y discrete (e.g. class membership)

Machine Learning – Examples

- Formalization:

$$y = f(x; w) \tag{2}$$

y is called *output variable*,
x the *input variable*

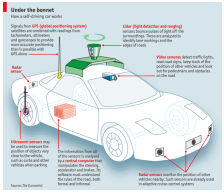
- **Spam Filtering:**
  $y =$ either *Spam* or *No Spam* (binary classification problem)
  $x =$ incoming email

Machine Learning – Examples

- Formalization:

$$y = f(x; w) \tag{3}$$

y is called *output variable*,
x the *input variable*



- **Autonomous Driving:**
  $y =$ is a a combination of several variables
  - steering angle (regression problem), acceleration (regression problem), ...

  $x =$ is also a combination of several variables
  - GPS, images, laser range scanner, map-data, ...

Machine Learning – A Definition

- Goal of machine learning:
  - Machines that learn to perform a task from experience
- Formalization:

$$y = f(x; w) \tag{4}$$

y is called *output variable*,
x the *input variable* and
w the model parameters (typically learned)

- learn... adjust the parameter w
- ... a task ... the function f
- ... from experience using a training dataset $\mathcal{D}$, where either $\mathcal{D} = \{x_1, \ldots, x_n\}$ or $\mathcal{D} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$

Different Scenarios

- Supervised Learning
- Unsupervised Learning
- Reinforcement Learning

- Let's discuss

Supervised Learning

- Given are pairs of training examples from $\mathcal{X} \times \mathcal{Y}$

$$\mathcal{D} = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\} \tag{5}$$

- Goal is to learn the relationship between x and y, that is:

$$y = f(x; w) \tag{6}$$

- Given a new example point x predict y

$$y = f(x; w) \tag{7}$$

- We want to generalize to unseen data

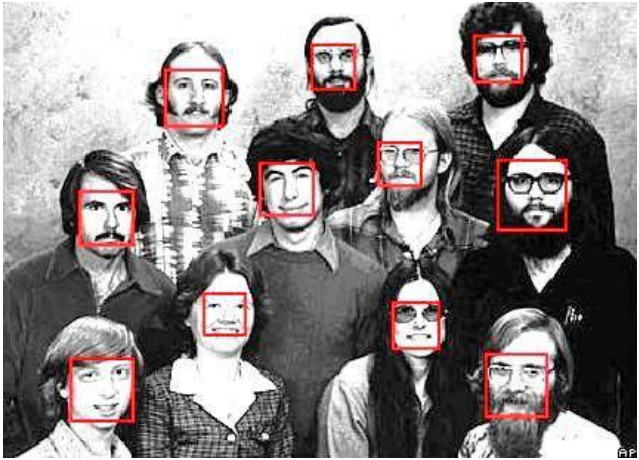
Supervised Learning

Example Supervised Learning

- **Linear classifier**:

$$y = f(x; w) = \begin{cases} +1 & w^T x + b > 0 \\ -1 & otherwise \end{cases} \tag{8}$$

Supervised Learning – Examples



Face Detection

Supervised Learning – Examples

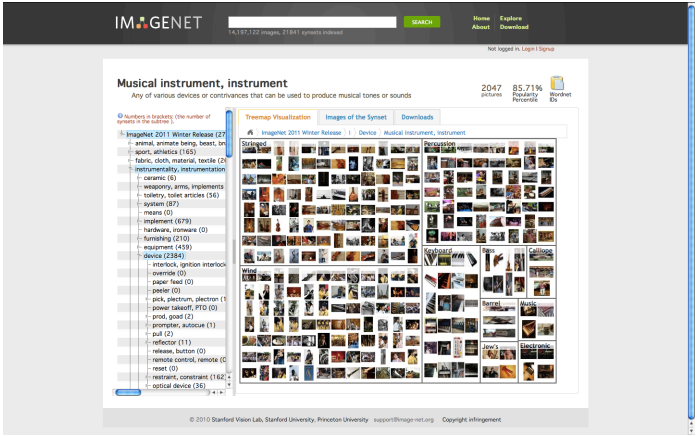


Image Classification

Supervised Learning – Examples

Semantic Image Segmentation

Supervised Learning – Examples

- Person recognition / identification
- Credit card fraud detection
- Speech recognition
- Visual object detection
- Prediction survival rate of a patient
- ...

Supervised Learning - Models

Flashing more keywords

- Linear Classifier
- Multilayer Perceptron (Backpropagation)
- (Deep) Convolutional Neural Networks (Backpropagation)

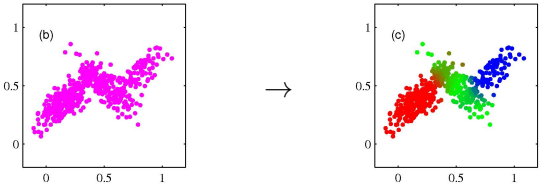
- Support Vector Machine (SVM)
- Linear Regression, Logistic Regression
- Boosting
- Graphical models
- ...

Unsupervised Learning

- We are given some input data points

$$\mathcal{D} = \{x_1, x_2, \ldots, x_n\} \tag{9}$$

- Goals:
  - Determine the data distribution $p(x) \rightarrow$ density estimation
  - Visualize the data by projections $\rightarrow$ dimensionality reduction
  - Find groupings of the data $\rightarrow$ clustering



Unsupervised Learning – Examples



Image Priors for Denoising

Unsupervised Learning – Examples

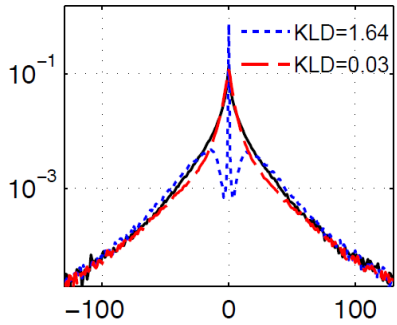


Image Priors for Inpainting

black line: statistics form original images, blue and red: statistics after applying two different algorithms
Image from *"A generative perspective on MRFs in low-level vision"*, Schmidt et al., CVPR2010

Unsupervised Learning – Examples

- Clustering scientific publications according to topics
- Clustering flickr images
- Clustering Youtube videos
- Novelty detection, predicting outliers
  - Anomaly detection in visual inspection
  - Video surveillance

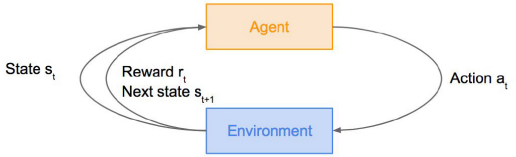
Unsupervised Learning – Models

Just *flashing* some keywords ($\rightarrow$ Machine Learning)

- Mixture Models
- K-Means clustering
- Kernel Density Estimation
- Neural Networks, e.g. Auto-Encoder Networks
- Principal Component Analysis (PCA)
- ...

Reinforcement Learning

Problems involving an **agent** interacting with an **environment**, which provides numeric **reward** signals



Goal: Learn how to take actions in order to maximize reward

Reinforcement Learning

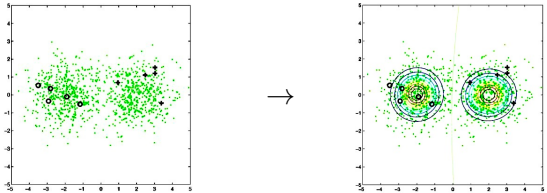
- Setting: given a situation, find an action to maximize a reward function
- Feedback:
  - we only get feedback of how well we are doing
  - we do *not* get feedback what the best action would be ("indirect teaching")
- Feedback given as **reward**:
  - each action yields reward, or
  - a reward is given at the end (e.g. robot has found his goal, computer has won game in Backgammon)
- **Exploration:** try out new actions
- **Exploitation:** use known actions that yield high rewards
- Find a good trade-off between exploration and exploitation

Variations of the general theme

- All problems fall in these broad categories
- But your problem will surely have some extra twists
- Many different variations of the aforementioned problems are studied separately
- Let's look at some ...

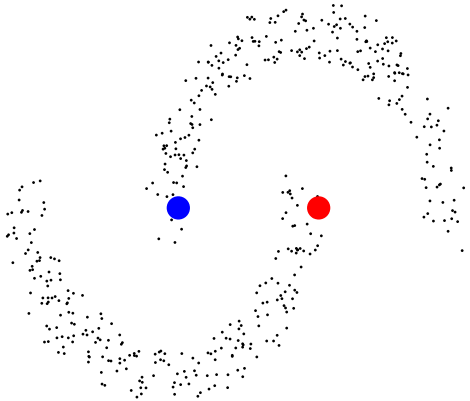
Semi-Supervised Learning

- We are given a dataset of l labeled examples
$$\mathcal{D}_l = \{(x_1, y_1), \ldots, (x_l, y_l)\}$$
as in supervised learning
- Additionally we are given a set of u unlabeled examples
$$\mathcal{D}_u = \{x_{l+1}, \ldots, x_{l+u}\}$$
as in unsupervised learning
- Goal is $y = f(x; w)$
- Question: how can we utilize the extra information in $\mathcal{D}_u$?



Semi-Supervised Learning: Two Moons

- Two labeled examples (red and blue) and additional unlabeled black dots



Two moons

On-line Learning

- The training data is presented step-by-step and is never available entirely
- At each time-step t we are given a new datapoint x_t – (or (x_t, y_t))
- When is online learning a sensible scenario?
  - We want to continuously update the model – we can train a model with little data, but the model should become better over time when more data is available (similar to how humans learn)
  - We have limited storage for data and the model – a viable setting for large-scale datasets (e.g. the size of the internet)

Large-Scale Learning

- Learning with millions of examples
- Study fast learning algorithms (e.g. parallelizable, special hardware)
- Problems of storing the data, computing the features, etc.
- There is no strict definition for "large-scale"
  - Small-scale learning: limiting factor is number of examples
  - Large-scale learning: limited by maximal time for computation (and/or maximal storage capacity)

Some final comments

- All topics are under active development and research
- Supervised classification: basically understood
- Broad range of applications, many exciting developments
- Adopting a "ML view" has far reaching consequences, it touches problems of empirical sciences in general

Decision Theory

Letter Classification

- Classify letters “a” versus “b”

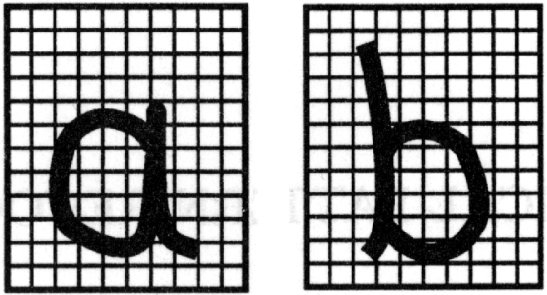


Figure: The letters “a” and “b”

- Goal: classify new letters such that the error probability is minimized

Letter Classification – Priors

Prior Distribution

- How often do the letters “a” and “b” occur ?
- Let us assume

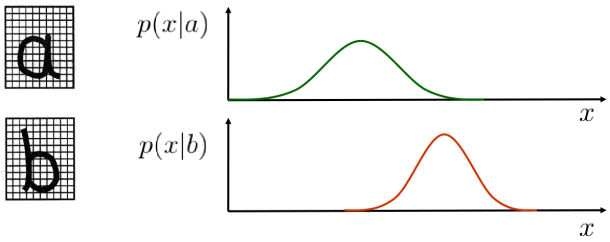
$$C_1 = a \quad p(C_1) = 0.75 \tag{10}$$
$$C_2 = b \quad p(C_2) = 0.25 \tag{11}$$

The *prior* has to be a distribution, in particular

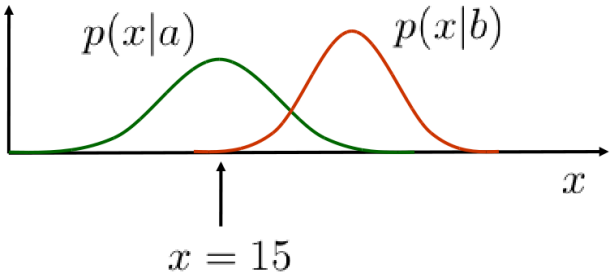
$$\sum_{k=1,2} p(C_k) = 1 \tag{12}$$

Letter Classification - Class Conditionals

- We describe every letter using some feature vector, e.g.
  - the number of black pixels in each box
  - relation between width and height
- Likelihood: How likely has x been generated from $p(\cdot \mid a)$, respectively $p(\cdot \mid b)$?

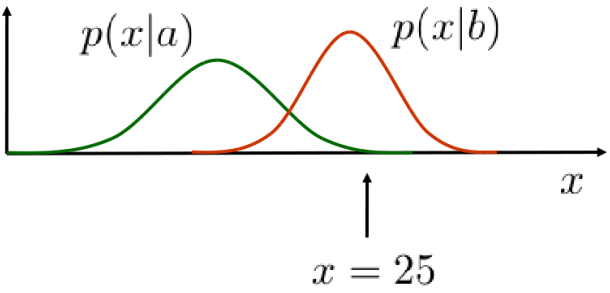


Letter Classification



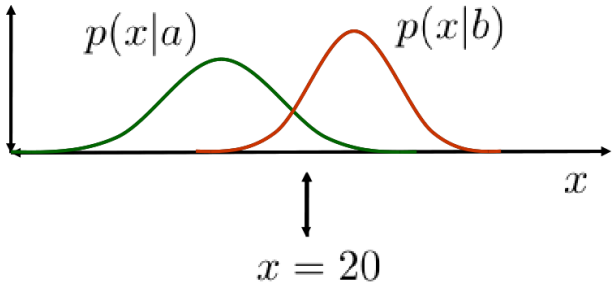
- Which class should we assign x to ?
- The answer
- Class a

Letter Classification



- Which class should we assign x to ?
- The answer
- Class b

Letter Classification



- Which class should we assign x to ?
- The answer
- Class a, since p(a)=0.75

Bayes Theorem

- How do we formalize this?
- We use the (hopefully well known) Bayes Theorem

$$p(Y|X) = \frac{p(X,Y)}{p(X)} = \frac{p(X|Y)p(Y)}{p(X)} \tag{13}$$

- Now we apply it

$$p(C_k|x) = \frac{p(x|C_k)p(C_k)}{p(x)} = \frac{p(x|C_k)p(C_k)}{\sum_j p(x|C_j)p(C_j)} \tag{14}$$

Bayes Theorem

- *Some terminology !*
- Repeated from last slide:

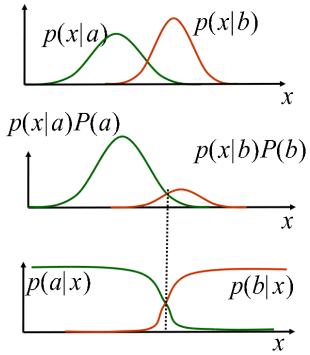
$$p(C_k|x) = \frac{p(x|C_k)p(C_k)}{p(x)} = \frac{p(x|C_k)p(C_k)}{\sum_j p(x|C_j)p(C_j)} \tag{15}$$

- We use the following names

$$\text{Posterior} = \frac{\text{Likelihood} \times \text{Prior}}{\text{Normalization Factor}} \tag{16}$$

- Here the normalization factor is easy to compute.
- It is also called the Partition Function, common symbol Z

Bayes Theorem



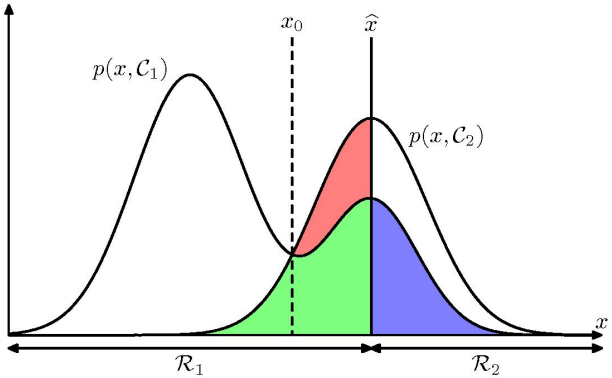
Likelihood

Likelihood $\times$ Prior

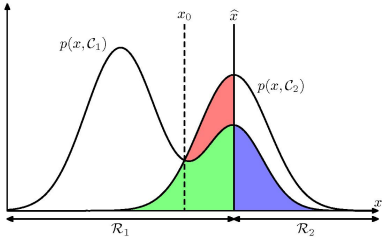
$$\text{Posterior} = \frac{\text{Likelihood} \times \text{Prior}}{\text{Normalization Factor}}$$

How to Decide?

- Two class problem C_1, C_2, plotting Likelihood $\times$ Prior



Minmizing the Error



$$
\begin{aligned}
p(\text{error}) &= p(x \in R_2, C_1) + p(x \in R_1, C_2) && (17)\\
&= p(x \in R_2 | C_1) p(C_1) + p(x \in R_1 | C_2) p(C_2) && (18)\\
&= \int_{R_2} p(x|C_1) p(C_1) \mathrm{d}x + \int_{R_1} p(x|C_2) p(C_2) \mathrm{d}x && (19)
\end{aligned}
$$

General Loss Functions

- So far we considered misclassification error only
- This is also referred to as 0/1 loss
- Now suppose we are given a more general loss function

$$\Delta : \quad \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+ \tag{20}$$

$$(y, \hat{y}) \mapsto \Delta(y, \hat{y}) \tag{21}$$

- How do we read this?
- $\Delta(y, \hat{y})$ is the cost we have to pay if y is the true class but we predict $\hat{y}$ instead

Example: Predicting Cancer

$$\Delta : \quad \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+ \tag{22}$$

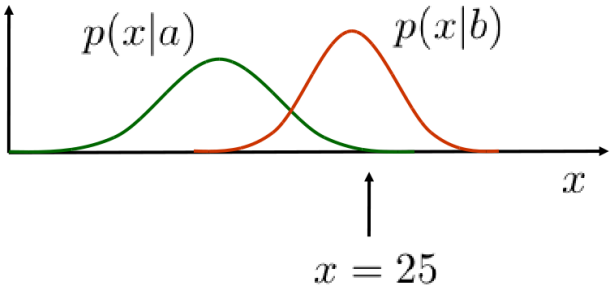
$$(y, \hat{y}) \mapsto \Delta(y, \hat{y}) \tag{23}$$

- Given: X-Ray image, Question: Cancer yes or no?
  Should we have another medical check of the patient?

		$diagnosis:$	
		cancer	normal
$truth:$	cancer	0	1000
	normal	1	0

- For discrete sets $\mathcal{Y}$ this is a loss matrix

Letter Classification

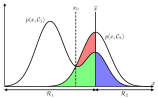


- Which class should we assign x to? $(p(a) = p(b) = 0.5)$
- The answer
- **It depends on the loss**

Minmizing Expected Loss (or Error)

- The expected loss for x (averaged over all decisions)

$$\mathbb{E}[\Delta] = \sum_{k=1,\dots,K} \sum_{j=1,\dots,K} \int_{R_j} \Delta(C_k, C_j) p(x, C_k) \mathrm{d}x \tag{24}$$



- And how do we predict? Decide on one y!

$$\begin{aligned} y^* &= \operatorname*{argmin}_{y \in \mathcal{Y}} \sum_{k=1,\dots,K} \Delta(C_k, y) p(C_k|x) \\ &= \operatorname*{argmin}_{y \in \mathcal{Y}} \mathbb{E}_{p(\cdot|x)}[\Delta(\cdot, y)] \end{aligned}$$

Inference and Decision

- We broke down the process into two steps
  - Inference: obtaining the probabilities $p(C_k|x)$
  - Decision: obtain optimal class assignment
- Two steps !!
- The probabilites $p(\cdot|x)$ represent our belief of the world
- The loss Δ tells us what to do with it!
- $0/1$ loss implies deciding for max probability

Next. . .

we will discuss

- (Deep) Artificial Neural Networks