

MACHINE LEARNING

*ESWC SUMMER
SCHOOL 2016*

*Blaž Fortuna
Jožef Stefan Institute*

Motivation

Problem: Develop a program that identifies all persons in an article

German Chancellor **Angela Merkel** said she “can allow the time” for the U.K. to decide what it wants from the European Union in Brexit talks, signaling that the other 27 member countries will press ahead with “the European project” in the meantime.

Starting off three days of talks with fellow European leaders on the EU’s way forward, **Merkel** suggested there’s no advantage in pressing Prime Minister **Theresa May**’s government for now because the EU can’t negotiate new relations until the U.K. government’s position is clear.

...

Motivation

How to approach the problem?

- Use a list of common personal names
- Look for words beginning with capital characters
- Identify typical expressions:
\$PERSON said, Prime Minister \$PERSON, ...

German Chancellor **Angela Merkel** said she “can allow the time” for the U.K. to decide what it wants from the European Union in Brexit talks, signaling that the other 27 member countries will press ahead with “the European project” in the meantime.

Starting off three days of talks with fellow European leaders on the EU’s way forward, **Merkel** suggested there’s no advantage in pressing Prime Minister **Theresa May**’s government for now because the EU can’t negotiate new relations until the U.K. government’s position is clear.

...

Motivation

Suppose we have a collection of articles with marked mentions of people.

Can we “train” a program to recognize mentions from new articles?

Yes, this is exactly what Machine learning tries to accomplish!

How would that work in our example?

- Break articles into words and describe each word by some characteristics

=> extract features

- Mark each word as \$PERSON or \$OTHER

=> training data

- Identify “rules” that can tell for a word in a new article: \$PERSON or \$OTHER

=> fit a model

Outline

1. What is machine learning?
 2. Standard ML problems: supervised, unsupervised, ...
 3. How to represent data?
 4. What are models?
 5. How to fit a model?
 6. How to evaluate?
-



What is Machine Learning

Machine learning is a subfield of computer science that evolved from the study of pattern recognition and computational learning theory in artificial intelligence.

"Field of study that gives computers the ability to learn without being explicitly programmed". 1959, Arthur Samuel

Machine learning explores the study and construction of algorithms that can learn from and make predictions on data.

Such algorithms operate by building a model from example inputs in order to make data-driven predictions or decisions, rather than following strictly static program instructions.

Apparently, Computer
Scientists are better
than Statisticians
at marketing.



Terminology

Training data:

- Input data used by machine learning algorithm to learn from
- Examples: tweets with labels indicating sentiment, economical forecasting

Model:

- Function created by machine learning model based on the training data
- Examples: decision tree, linear function, neural net

Prediction:

- Application of a model to a new data point
 - Example: sentiment label of a new tweet, GDP figure for next quarter
-

http://analytics.ijs.si/~blazf/programs/svm_gui.zip

DEMO SVM

Standard Problems

Supervised learning:

- We show the program examples of input and expected output

Unsupervised learning:

- No expected output, goal is to find patterns in the input data

Reinforcement learning:

- Program interacts with dynamic environment with a certain goal
- Example: AlphaGo

Semi-supervised:

- Combination of supervised and unsupervised approaches
-

Supervised learning

Training data: inputs and their expected outputs

$$\{(x_i, y_i); x_i \in X, y_i \in Y\}$$

Goal: a function $f: X \rightarrow Y$ such that $f(x_i) = y_i$.

Sub tasks:

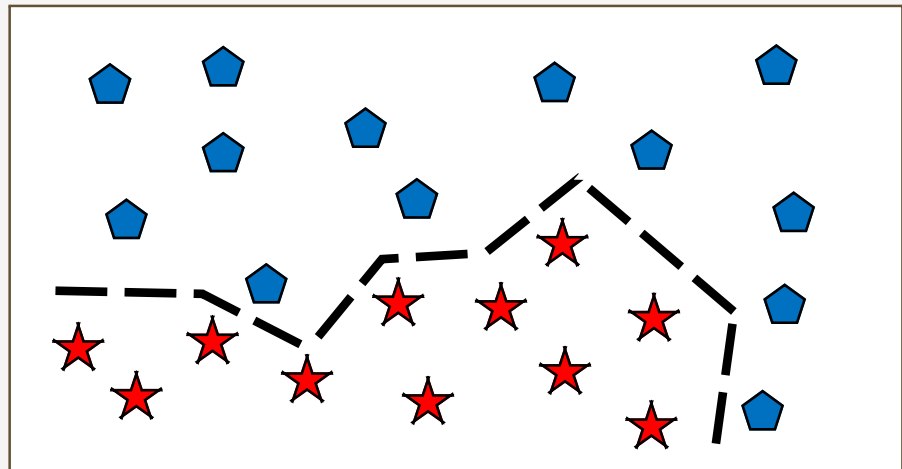
- Classification
 - Regression
 - Ranking
-

Classification

Y is a finite set: $Y = \{y_1, \dots, y_k\}$

Binary classification is special case when $k = 2$:

- Given $\{(x_i, y_i); x_i \in X, y_i \in \{1, -1\}\}$
- Identify $f: X \rightarrow Y$ such that $f(x_i) = y_i$
- f is called *binary classifier*



Classification

Y is a finite set: $Y = \{y_1, \dots, y_k\}$

Multiclass problems ($k > 2$) can be converted to multiple binary problems:

- One-vs-all: train k binary classifier, keep most confident class
- One-vs-one: train $\frac{k(k-1)}{2}$ binary classifiers, keep class with most votes

Some models directly target multi-class problems

- E.g. multinomial logistic regression using log-linear model
-

Classification

Many problems can be translated to (binary) classification task

- Multi label problem:

$$f: X \rightarrow 2^Y$$

- Can be seen as $|Y|$ binary tasks with one binary classifier for each label
- Structured output:
 - Can be seen as binary task where correct (input, output) pairs are positive and rest are negative:

$$f(\phi(x_i, y_i)) = 1$$

$$f(\phi(x_i, y)) = -1, \quad y \neq y_i$$

$$F(x) = \operatorname{argmax}_y \{f(\phi(x, y))\}$$

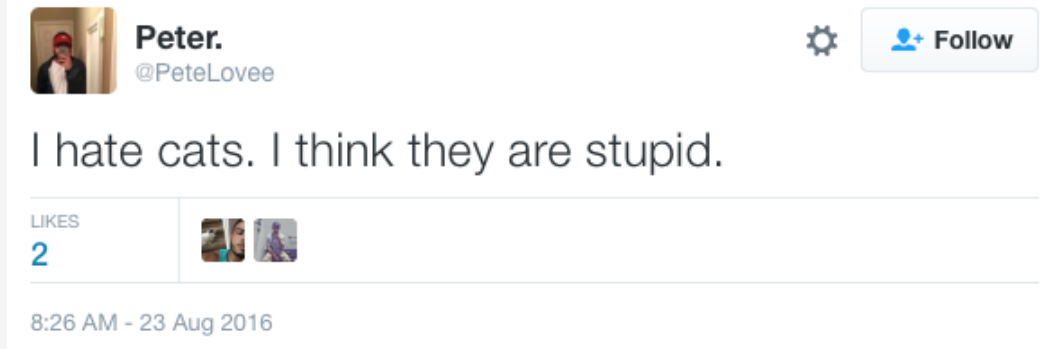
Document categorization

$X = \{\text{documents}\}, Y = \{\text{Technology, Sports, Entertainment, ...}\}$

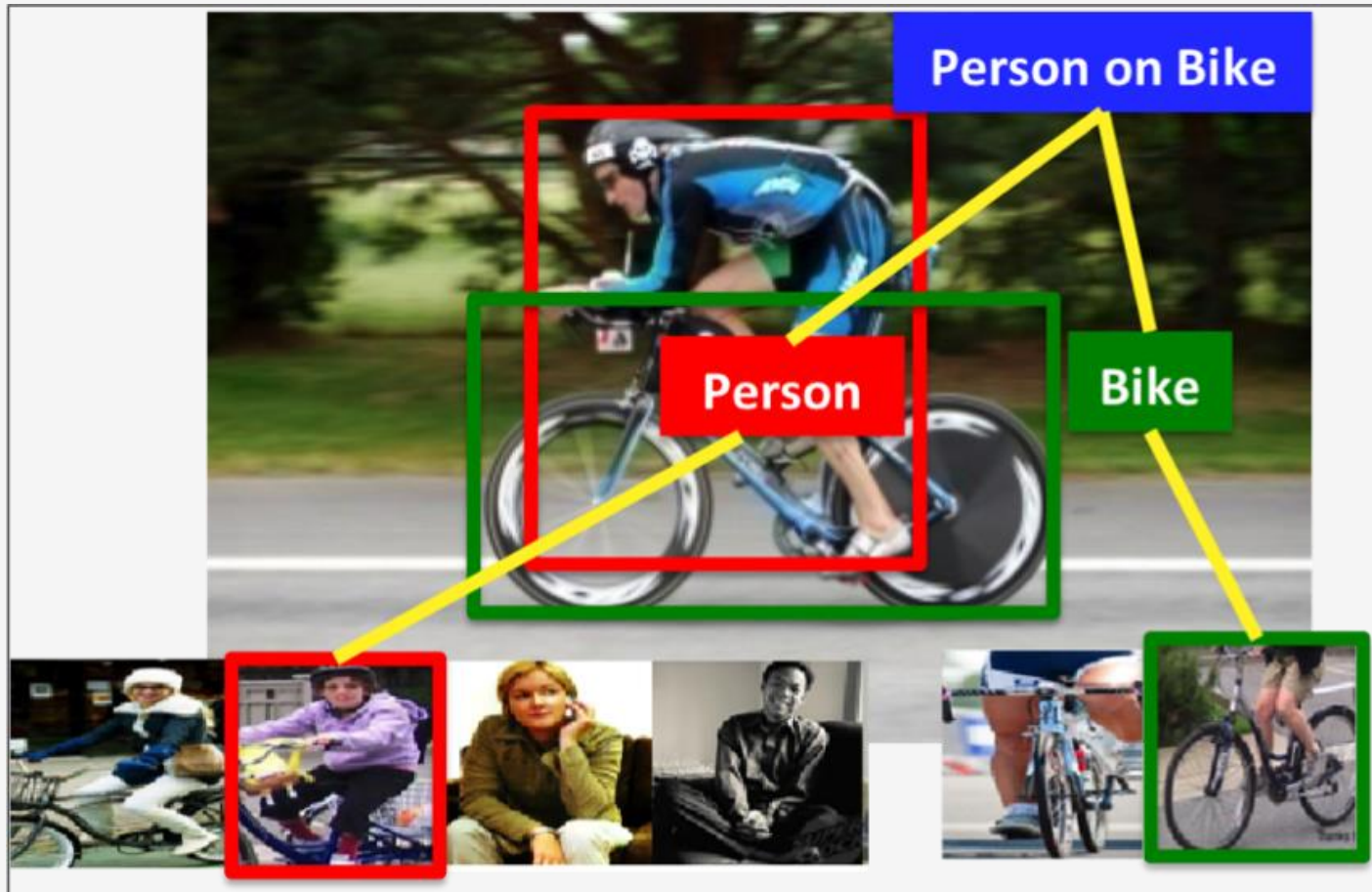


Sentiment analysis

$X = \{\text{tweets}\}, Y = \{\text{positive}, \text{negative}, \text{neutral}\}$



Object detection



Tian Lan, Michalis Raptis, Leonid Sigal, and Greg Mori.
From Subcategories to Visual Composites: A Multi-Level Framework for Object
Detection.

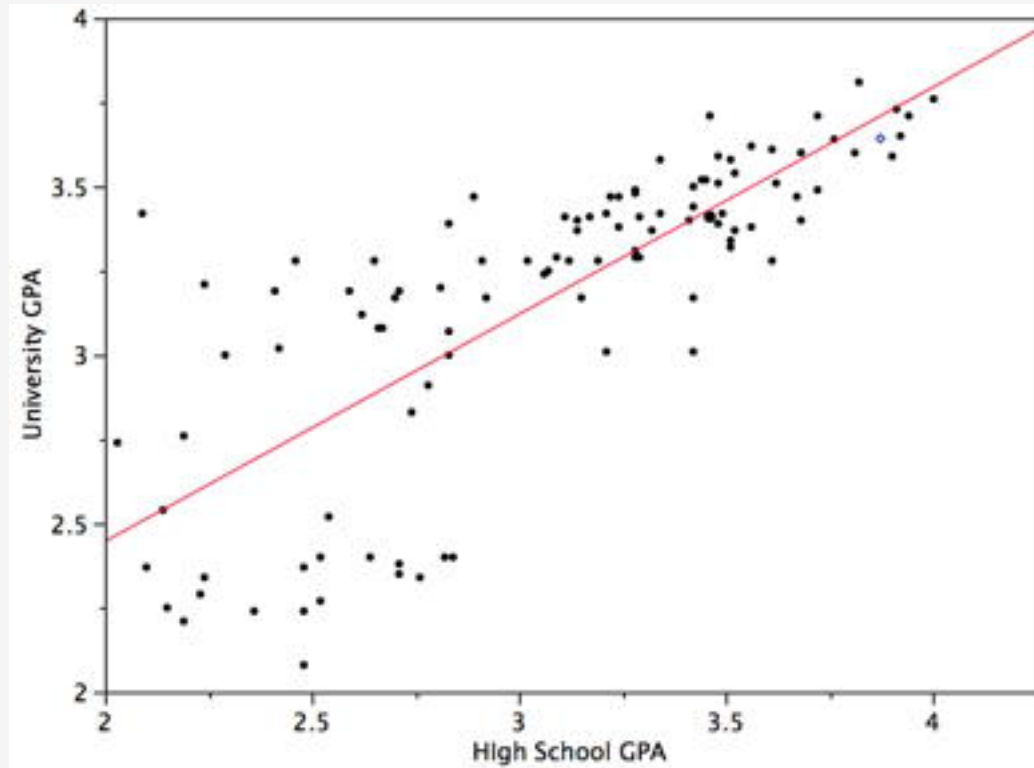
IEEE International Conference on Computer Vision (ICCV), 2013

Regression

Training data: $\{(x_i, y_i); x_i \in X, y_i \in \mathbb{R}\}$

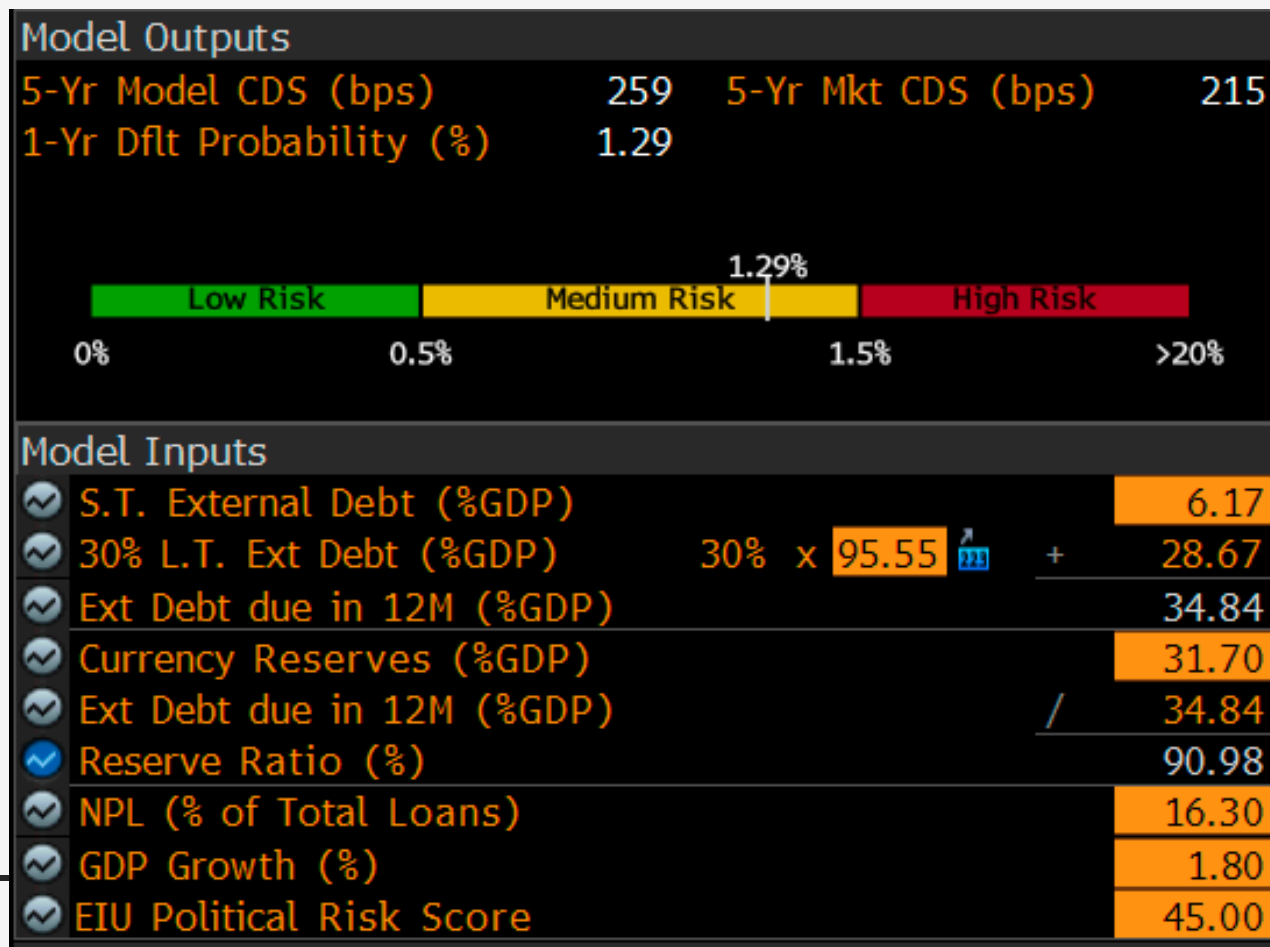
Goal: function $f: X \rightarrow \mathbb{R}$ such that $f(x_i) = y_i$.

Estimate university GPA



Probability of Default

Probability of defaulting in, for example, one year time.



Ranking

Goal: function which can rank a set of objects:

$$f: (q, \{x_1, \dots, x_n\}) \mapsto (x_{i_1}, \dots, x_{i_n})$$

Training data:

- Pointwise: $\{(q, x_i, j); q \in Q, x_i \in X, j \in \mathbb{N}\}$
 - Pairwise: $\{(q, x_i, x_j); q \in Q, x_i \in X, x_j \in X, rank(x_i) > rank(x_j)\}$
-

Search engine ranking


dubrovnik

All Images Maps Videos News More ▾ Search tools

About 28.500.000 results (0,67 seconds)

Dubrovnik - Wikipedia, the free encyclopedia
<https://en.wikipedia.org/wiki/Dubrovnik> ▾
Dubrovnik is a Croatian city on the Adriatic Sea, in the region of Dalmatia. It is one of the most prominent tourist destinations in the Mediterranean Sea, a seaport ...
[Walls of Dubrovnik](#) · [Flag of Dubrovnik](#) · [Dubrovnik-Neretva County](#) · [Slivno](#)

Images for dubrovnik Report images



[More images for dubrovnik](#)

Dubrovnik - Lonely Planet
<https://www.lonelyplanet.com/croatia/dubrovnik> ▾
Regardless of whether you are visiting Dubrovnik for the first time or the hundredth, the sense of awe never fails to descend when you set eyes on...
[Top things to do in Dubrovnik](#) · [Best places to stay in Dubrovnik](#) · [Image gallery](#)

Old City of Dubrovnik - UNESCO World Heritage Centre
whc.unesco.org ▾ [Culture](#) ▾ [World Heritage Centre](#) ▾ [The List](#) ▾
Old City of Dubrovnik. The 'Pearl of the Adriatic', situated on the Dalmatian coast, became an important Mediterranean sea power from the 13th century onwards.

Dubrovnik Online
www.dubrovnik-online.com/ ▾
Dubrovnik Online, the Nr.1 internet guide for Dubrovnik and Dubrovnik Region. All travel and tourist information about Dubrovnik at one place.

dolphin


All Images Videos Maps News More ▾ Search tools

About 131.000.000 results (0,73 seconds)

Dolphin Emulator - GameCube/Wii games on PC
<https://dolphin-emu.org/> ▾
Official website of Dolphin, the GameCube and Wii emulator. Download the latest version (5.0-466) now or ask questions on our forums for help.
[Download](#) · [Compatibility](#) · [Forums](#) · [FAQ](#)

Dolphin - Wikipedia, the free encyclopedia
<https://en.wikipedia.org/wiki/Dolphin> ▾
Dolphins are a widely distributed and diverse group of fully aquatic marine mammals. They are an informal grouping within the order Cetacea, excluding whales ...
[Bottlenose dolphin](#) · [Porpoise](#) · [Oceanic dolphin](#) · [Dolphin \(disambiguation\)](#)

In the news


**Feds Want to Ban Swimming With Hawaii Dolphins**
[NBCNews.com](#) - 23 hours ago
Federal regulators are proposing to ban swimming with dolphins in Hawaii, a move that ...

Meet the Dolphin species that was hidden in Smithsonian's fossil room
[PBS NewsHour](#) - 20 hours ago

Why the Feds want to ban swimming with dolphins in Hawaii
[Inhabitat](#) - 1 day ago

More news for dolphin

Images for dolphin Report images



Recommendation

Merkel Says EU Faces Hard Work to Overcome Brexit Setback

by Arne Delfs Ott Ummelas
ArneDelfs ottummelas

August 25, 2016 – 12:03 PM CEST Updated on August 25, 2016 – 1:44 PM CEST



■ "We have begun a so-called process of reflection to explore in which areas we should develop further as a priority," Merkel told reporters in the Estonian capital of Tallinn. Photographer: Raigo Pajula/AFP via Getty Images

Recommended



Net Migration Into U.K. Remains at Near-Record Levels



U.K. Consumer Boom Set to Flag as Brexit Ignites Inflation



Britain Is About to Take a Great (Battery) Leap Forward



Biden Snubbed During Turkey Visit

Unsupervised learning

Training data: set of unlabeled examples $\{x_i; x_i \in X\}$.

Subtasks:

- Clustering
 - Anomaly Detection or Outlier Detection
 - Dimensionality Reduction
 - Density Estimation
 - Autoencoders
-

Clustering

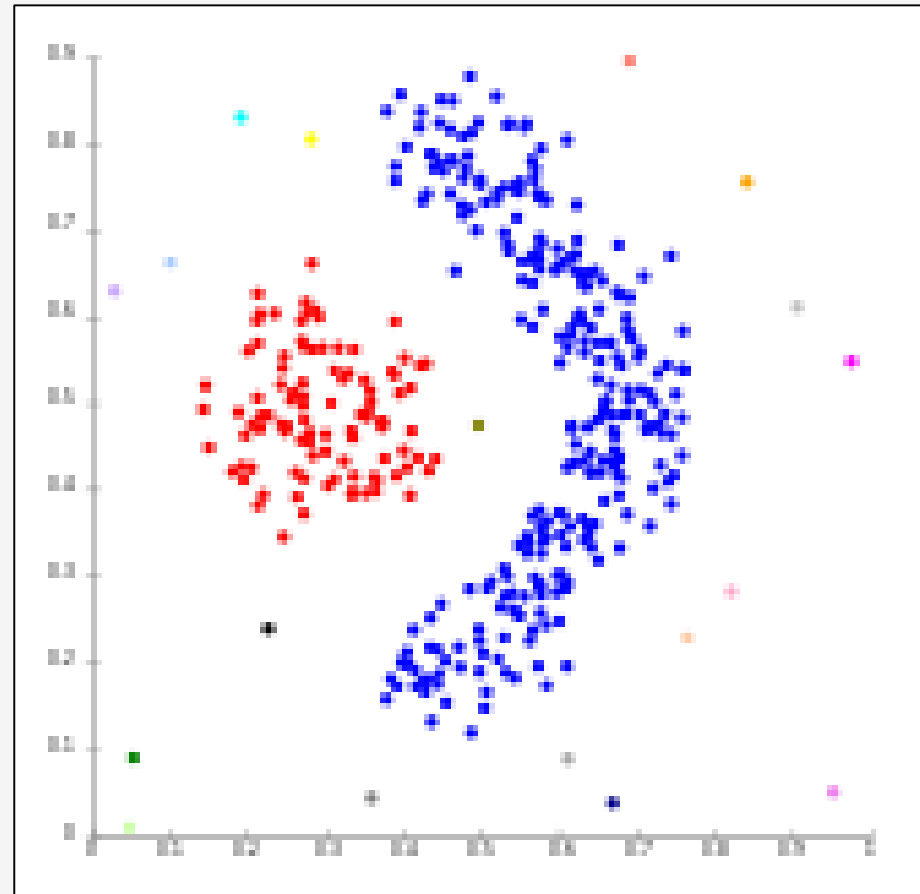
Group objects $\{x_i; x_i \in X\}$ into groups (i.e. clusters) such that objects from the same groups are “more similar” to each other than to other objects.

One object can belong to:

- exactly one cluster (e.g. k-means)
- several clusters (e.g. LDA)













Typically hard to evaluate

- what is correct clustering?



Clustering news articles

$X = \{\text{news articles}\}$, cluster = event

 <p>TRUMP</p> <p>The Latest: Trump says no one immune from enforcement</p> <p>2732 articles in 7 languages</p>	 <p>SAMSUNG</p> <p>Samsung 'to recall Galaxy Note 7 phones after explosion claims'</p> <p>1032 articles in 7 languages</p>	 <p>FLORIDA, UNITED STATES</p> <p>The Latest: Forecasters issue hurricane warning in Florida</p> <p>1018 articles in 1 language</p>	 <p>MEXICO</p> <p>LO ULTIMO: Trump insiste en que México pague muro</p> <p>958 articles in 1 language</p>
 <p>CUBA</p> <p>Historic flight touches down in Cuba</p> <p>720 articles in 6 languages</p>	 <p>CARACAS, VENEZUELA</p> <p>Amid threats of violence Venezuelan opposition tests resolve</p> <p>572 articles in 7 languages</p>	 <p>FRANZ BECKENBAUER</p> <p>Swiss open 2006 World Cup criminal case against Franz</p> <p>510 articles in 7 languages</p>	 <p>TURKEY</p> <p>Turkey vows to keep attacking US-backed Syrian Kurd forces</p> <p>469 articles in 4 languages</p>
 <p>DILMA ROUSSEFF</p> <p>El Senado destituye a Rousseff y confirma a Temer como presidente...</p> <p>443 articles in 1 language</p>	 <p>SAMSUNG</p> <p>First look: Samsung Gear S3 Smartwatch [VIDEO]</p> <p>398 articles in 6 languages</p>	 <p>COLIN Kaepernick</p> <p>Colin Kaepernick protests national anthem again, is joined by...</p> <p>390 articles in 1 language</p>	 <p>CHELSEA F.C.</p> <p>Chelsea have re-signed David Luiz from Paris St-Germain for £30m</p> <p>380 articles in 5 languages</p>

BELFAST TELEGRAPH

Fri, 02 Sep, 04:30

Samsung 'to recall Galaxy Note 7 phones after explosion claims'

Samsung will issue a global recall of the Galaxy Note 7 smartphone as soon as this weekend after its investigation into explosion claims found batteries were at fault, according to South Korea's Yonhap News.

Samsung Electronics declined to comment on the report on Friday, but said it was conducting ...

NEWS.COM.AU

Fri, 02 Sep, 02:18

Samsung to recall phones after explosions

Samsung will reportedly issue a global recall of the Galaxy Note 7 smartphone as soon as this weekend after its investigation on explosion claims found batteries were at fault.

Samsung Electronics declined to comment on the report in South Korea's Yonhap News on Friday, but said it was conducting ...

MANILA BULLETIN

Fri, 02 Sep, 03:22

Samsung to recall phones after explosion claims -- report

SEOUL, South Korea -- Samsung will issue a global recall of the Galaxy Note 7 smartphone as soon as this weekend after its investigation on explosion claims found batteries were at fault, according to South Korea's Yonhap News.

Samsung Electronics declined to comment on the report on Friday, but ...

ZEE NEWS

Fri, 02 Sep, 05:20

Samsung to recall Galaxy Note 7 phones after explosion claims: Report

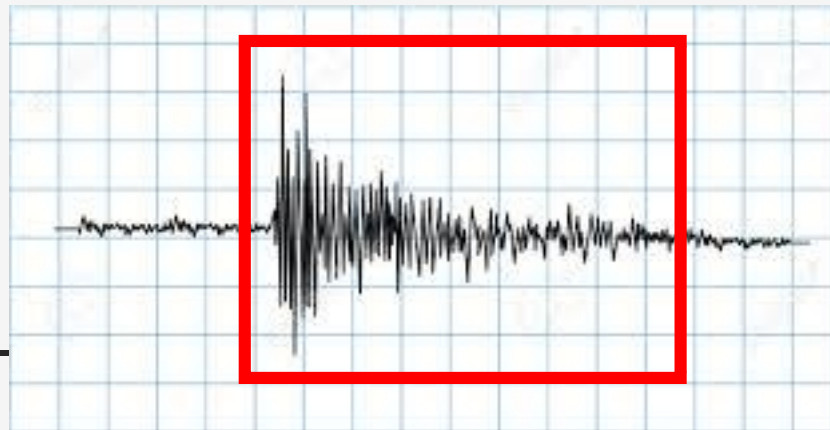
Seoul: Samsung will issue a global recall of the Galaxy Note 7 smartphone as soon as this weekend after its investigation on explosion claims found batteries were at fault, according to South Korea's Yonhap News.

Samsung Electronics declined to comment on the report today, but said it was conducting ...

Anomaly Detection

Identify events or observations that do not conform to expected.

"Anomaly" is a very domain specific term:



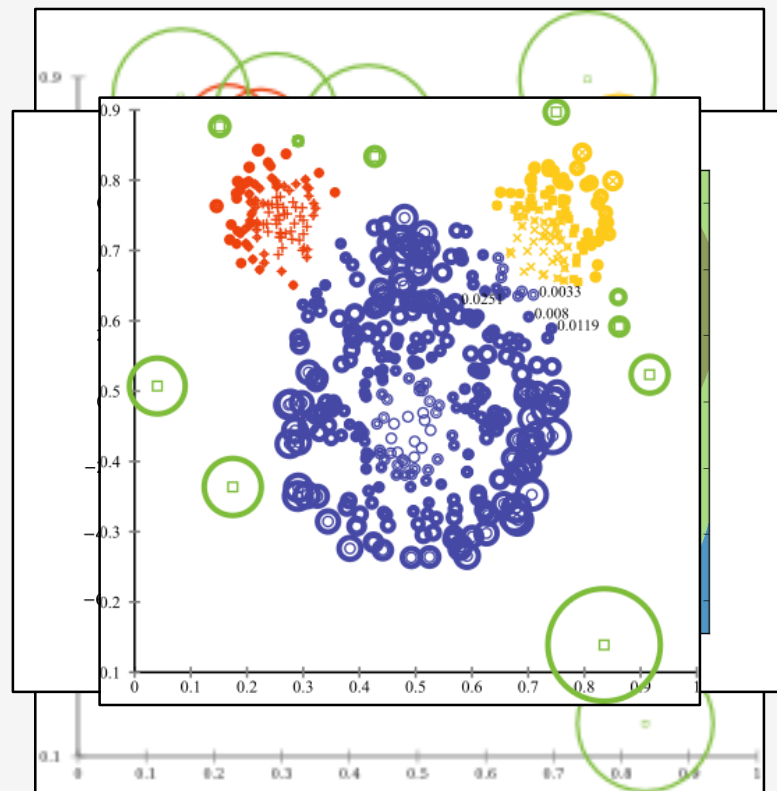
Density-based Approaches

Model probability distribution over observations

Anomalies = observations with low likelihood (e.g. $< 0.1\%$)

Examples:

- k-nearest neighbors (kNN)
- local density \sim distance to NN
- One-class SVM
- Clustering-based

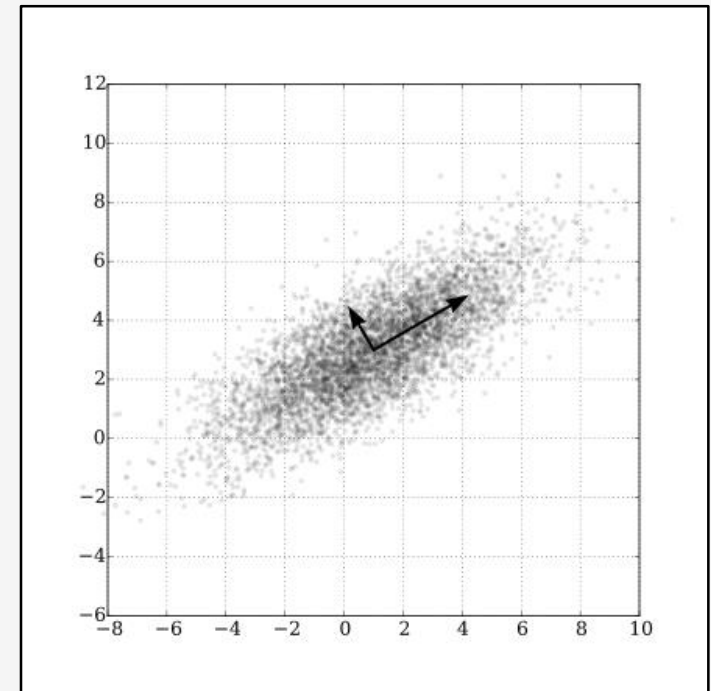


Dimensionality Reduction

Identify core components of higher-dimensional data

Examples:

- Principal Component Analysis:
 - Identify linearly uncorrelated principal components
 - First component should have highest variance, etc.
- Singular Value Decomposition:
 - Approximate input data with low rank matrices
- Canonical Correlation Analysis
 - Works with aligned multi-view data
 - Identifies main correlated dimensions



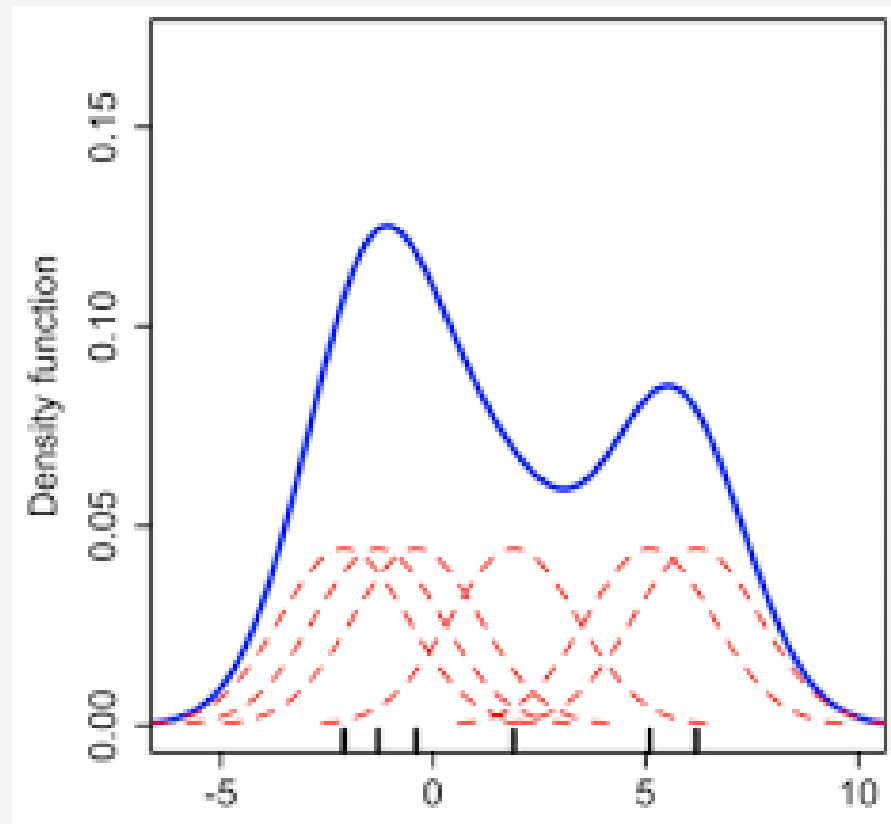
Density Estimation

Estimate underlying probability distribution from the training data

- Training data => observations

Example:

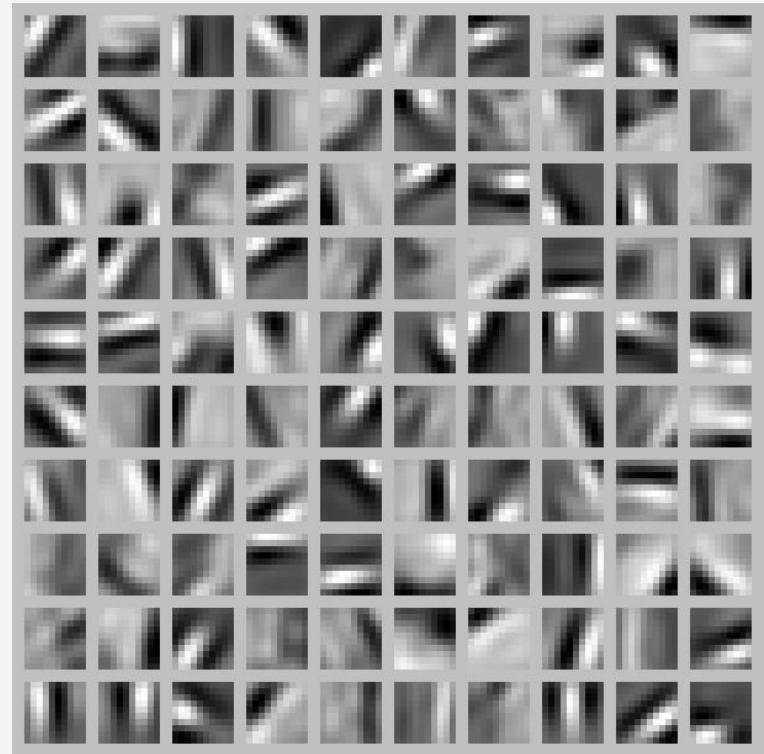
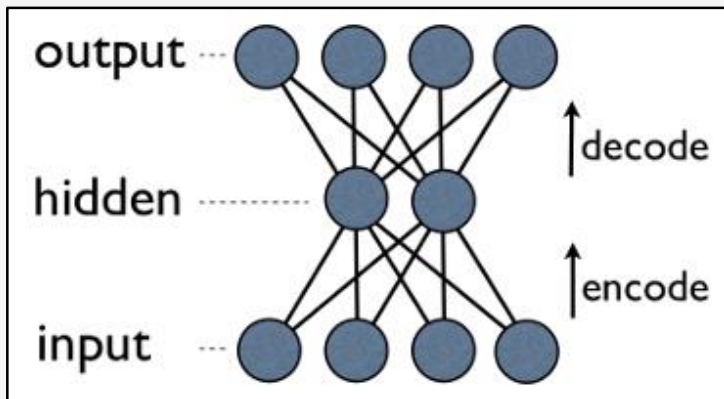
- Kernel density estimation



Autoencoders

Artificial Neural Network approach for dimensionality reduction

- Low-dimensional approximation
- Unsupervised feature learning

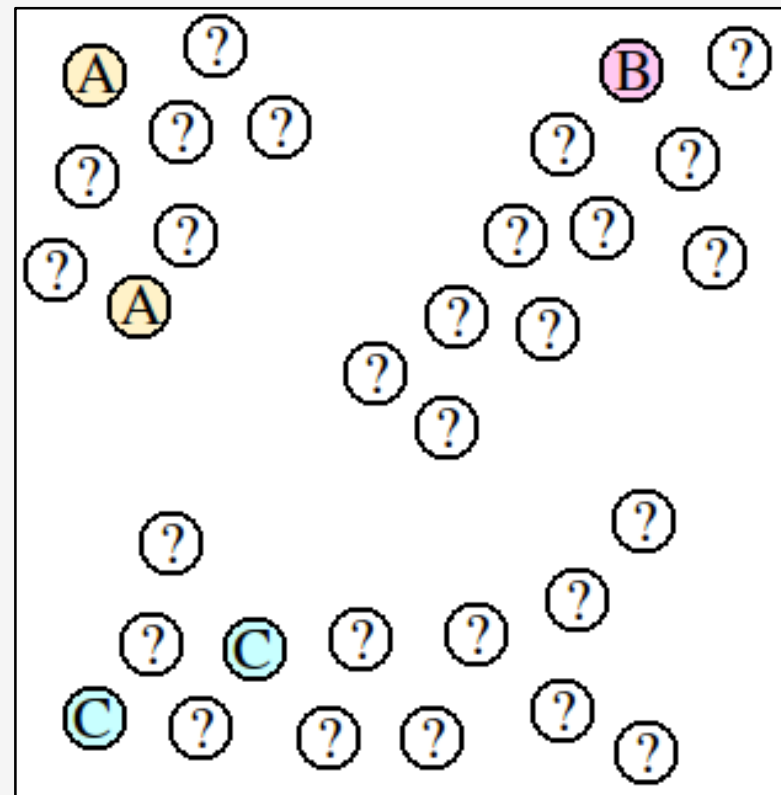


Semi-supervised Learning

Dealing with combination of labeled and unlabeled data

Transduction

- Reasoning from training to specific test cases
- Result is extending labels to test cases
- No resulting model or classifier



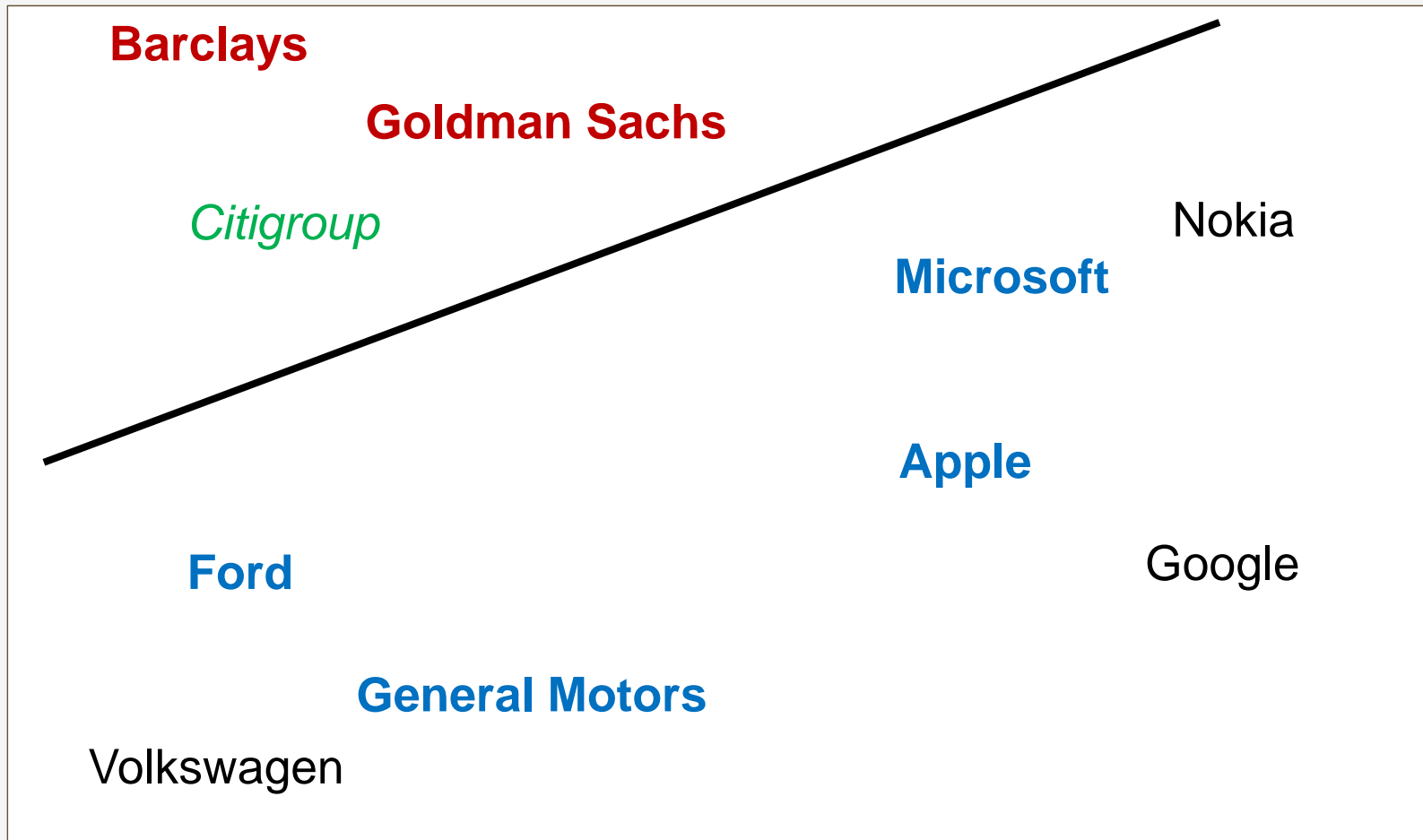
Semi-supervised Learning

Active-learning

- Interactive label acquisition, try to reduce the cost of labeling
- Identify unlabeled example(s) which would “best contribute” to the model
- “best contribute” can mean many things: Uncertainty Sampling, Expected Error Reduction



Active Learning Example



Data Representation

Data points are represented by a set of features or attributes

Actual representation depends on the algorithm:

- Decision trees work with individual attributes
 - differentiate between discrete and continuous attributes
 - Linear models (Logistic Regression, SVD) work with vectors
 - data points encoded as vectors $x \in \mathbb{R}^n$
 - Some approaches only require similarity measure between data points
 - Kernel methods, some clustering approaches
 - Algorithms for learning representations
 - Example: word2vec
-

Representing Text Documents

Vector space model:

- Identify all unique words in the document, each word becomes one dimension
- Document is represented by a vector x with
 - $x_i = 1$ when i -th word occurs in the document, and
 - $x_i = 0$ when i -th word does not occur in the document



Doc1

German Chancellor **Angela Merkel** said she “can allow the time” for the U.K. to decide what it wants from the European Union in **Brexit** talks, signaling that the other 27 member countries will press ahead with “the European project” in the meantime.

Doc2

Not long ago, **Angela Merkel's** dominant **position** in Germany and her status as the most influential leader in **Europe** seemed **secure**.

Now voters in her home state of Mecklenburg-Vorpommern appear poised to inflict a humiliating defeat to the **German ...**

Doc3

Angela Merkel has expressed fears the €13bn **Apple** tax ruling will hurt investment in **Europe**, putting her on the same side as **Ireland** in a looming showdown over the limits of national sovereignty and the rights of the federally minded European Commission.

	Angela	Angela_Merkel	Brexit	Europe	German	Ireland	position	secure	Apple	...
Doc1	1	1	1	0	1	0	0	0	0	
Doc2	1	1	0	1	1	0	1	1	0	
Doc3	1	1	0	1	0	1	0	0	1	

Representing Text Documents

- Bag-of-words: ignore word order, can compensate with n-grams
- Different options on how to weight words, e.g. TFIDF
- Stop-words: can ignore some common words ("the", "a", ...)
- Stemming or Lemmatization: normalize words

E.g. "banks" and "banking" => "bank"

- High dimensionality: can compensate with "hashing trick"



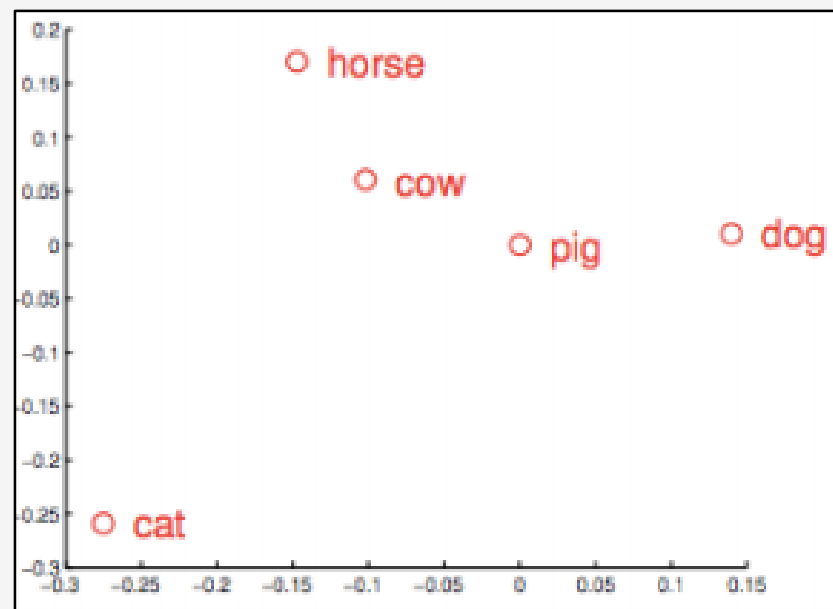
Learning Representation

Can we learn features?

- Yes we can with Deep Learning!
- Example: word embedding using methods such as word2vec
- We want to solve the task: predict word given its context

investment in Europe putting her

- Solve the problem with neural network, where each word is mapped to n -dimensional vector in the first layer
- n -dimensional vectors are good feature representations for the words

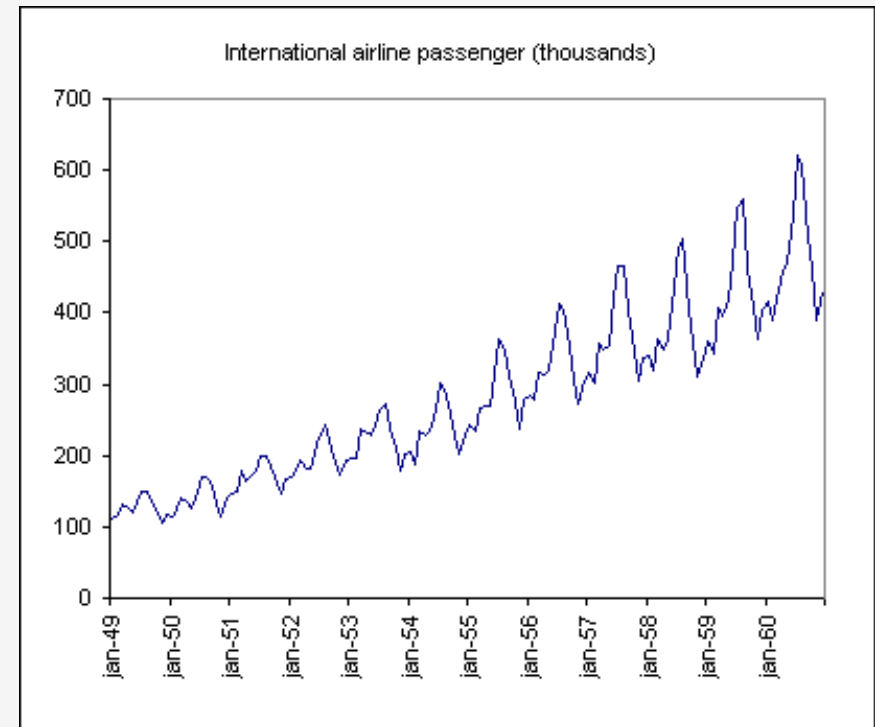


Time Series

A series of data points in time order.

Features:

- Simple statistic features: mean, deviation, ...
- Time-delay embedding: last n -values
- Frequency domain features: DFT



Machine Learning Models

Model is the what we learn from training data

- i.e. function $f: X \rightarrow Y$

Different types of model:

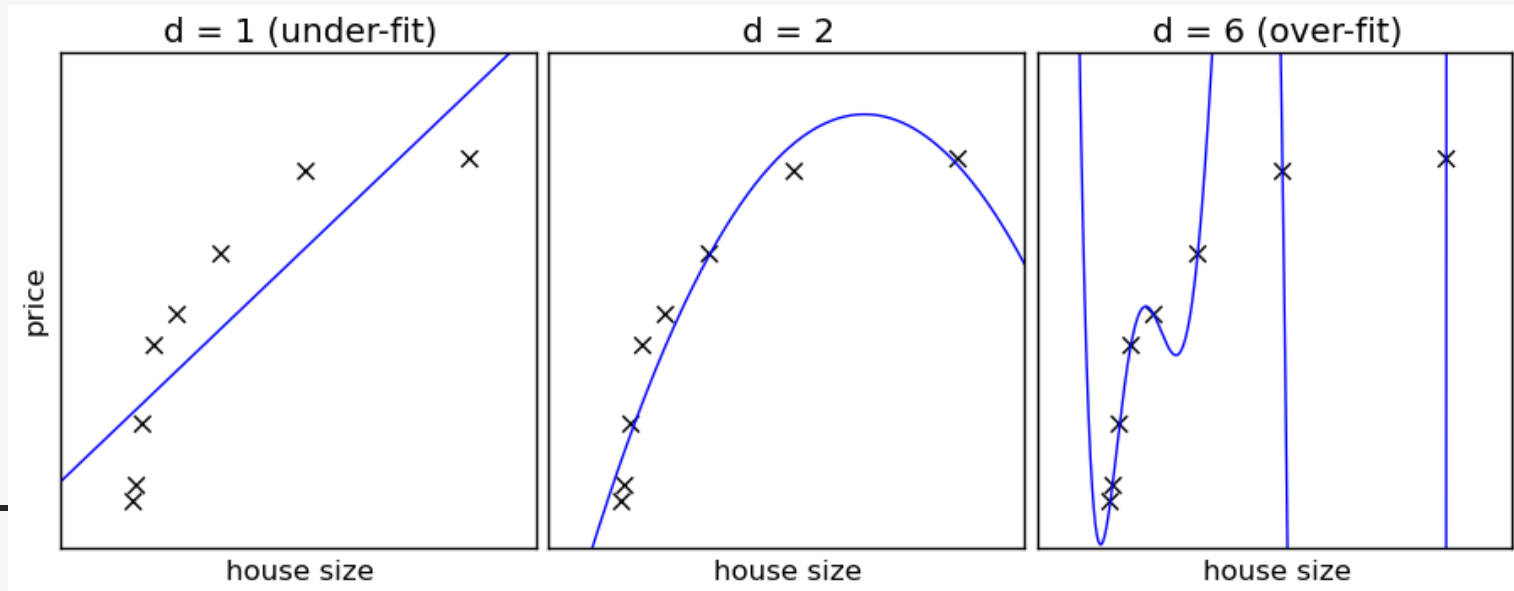
- Discriminative: can only tell the class, no implications on the distribution
 - Example: Support Vector Machine
- Generative: can be used to generate data similar to training data
 - Example: LDA, Restricted Boltzmann Machine, Hidden Markov Model



Linear Model

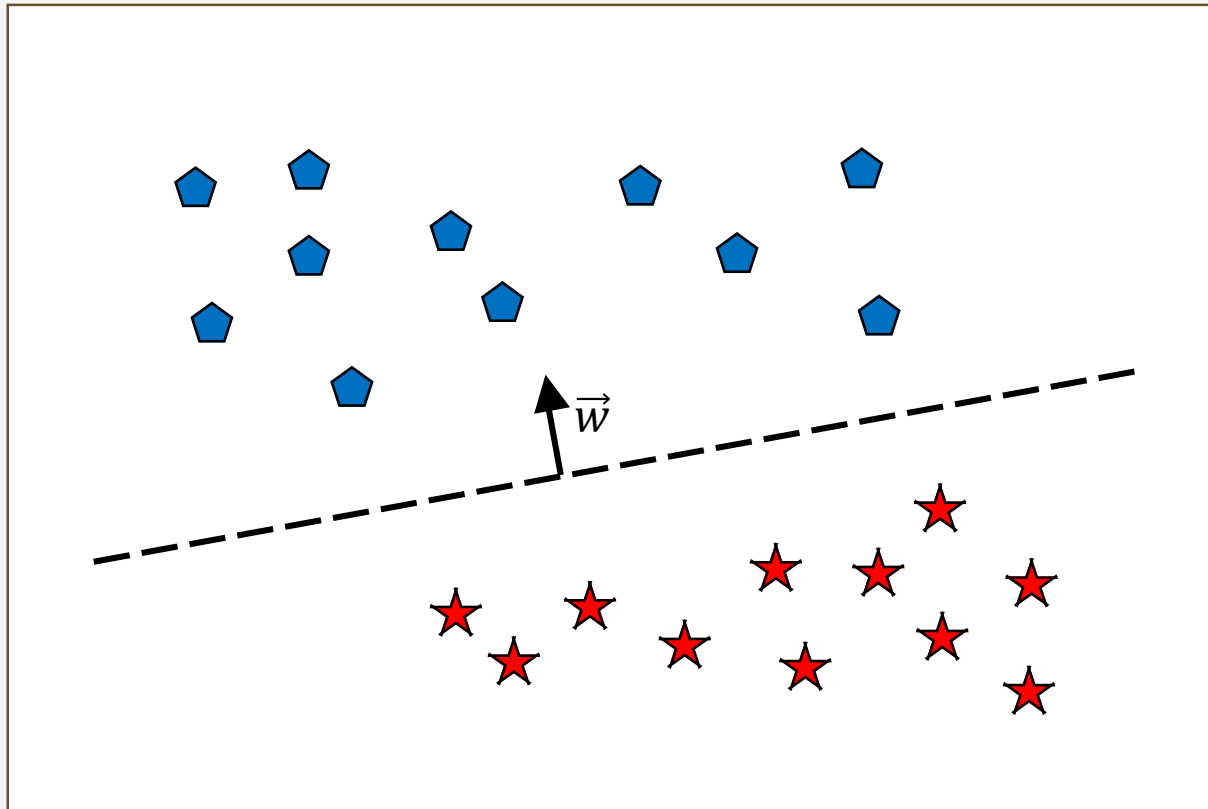
Model is a linear functional $f(x) = w^T x$ where $w \in \mathbb{R}^n$

- Learning linear model boils down to finding “optimal” w
- Typically inexpensive to learn, always inexpensive to apply
- Prediction is just a linear combination of features, $O(n)$
- Cannot capture non-linear patterns



Linear Model for Classification

Geometric interpretation:



Linear Model for Classification

Another interpretation:

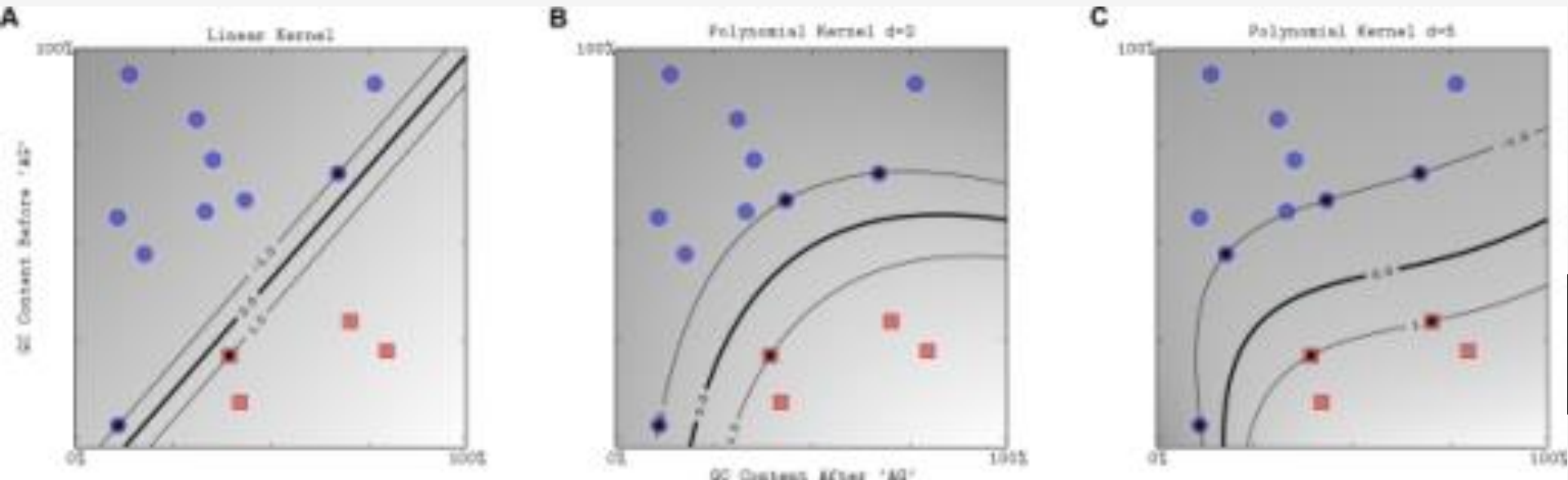
- Each feature gets assigned a weight
- Positive features with positive weights vote for positive class
- Same for negative weights/class
- Larger the $|\text{weight}|$, stronger the vote



Kernel Methods

Core idea: map input data into higher dimensional space: $\phi: X \rightarrow V$

- We can compute dot product in V without explicitly mapping data to V
$$\langle \phi(x_i), \phi(x_j) \rangle = k(x_i, x_j)$$
- Modify learning algorithms to only access input data through kernel data
- Linear model in V can capture complex non-linear patterns in X



Popular Kernels: Polynomial Kernel

Polynomial kernel of degree d :

$$k(x, y) = (x^T y + c)^d$$

- Example for $n = 2, d = 2$ and $c = 1$:

$$\begin{aligned} k([x_1, x_2], [y_1, y_2]) &= (x_1 y_1 + x_2 y_2 + 1)^2 = \\ &= x_1 y_1 (x_1 y_1 + x_2 y_2 + 1) + x_2 y_2 (x_1 y_1 + x_2 y_2 + 1) + (x_1 y_1 + x_2 y_2 + 1) = \\ &= x_1^2 y_1^2 + 2x_1 x_2 y_1 y_2 + x_2^2 y_2^2 + 2x_1 y_1 + 2x_2 y_2 + 1 = \\ &= [x_1^2, \sqrt{2}x_1 x_2, x_2^2, \sqrt{2}x_1, \sqrt{2}x_2, 1]^T [y_1^2, \sqrt{2}y_1 y_2, y_2^2, \sqrt{2}y_1, \sqrt{2}y_2, 1]^T \end{aligned}$$

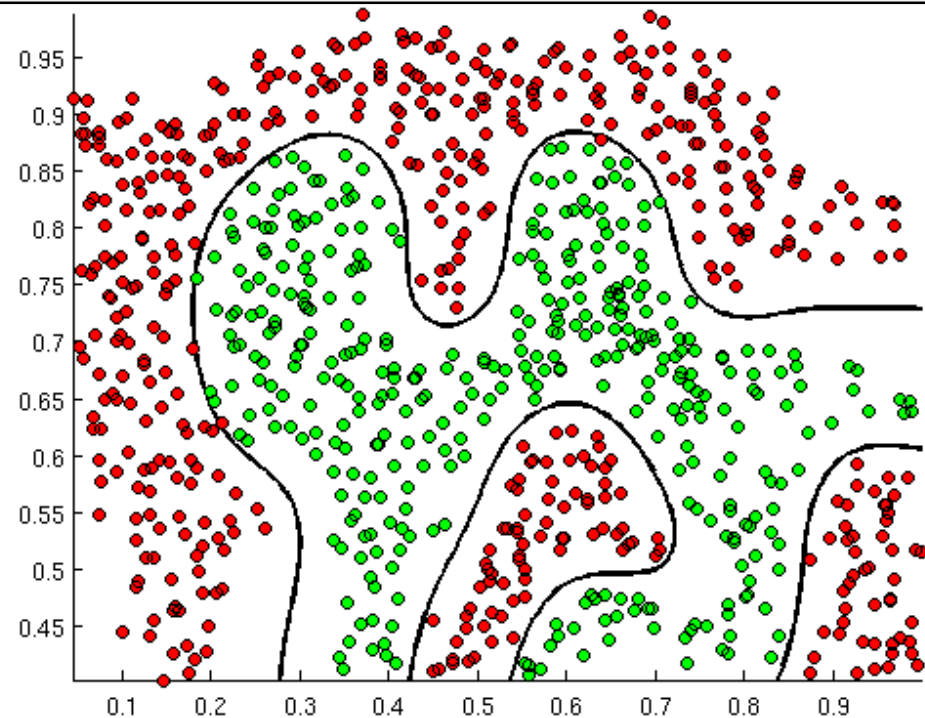
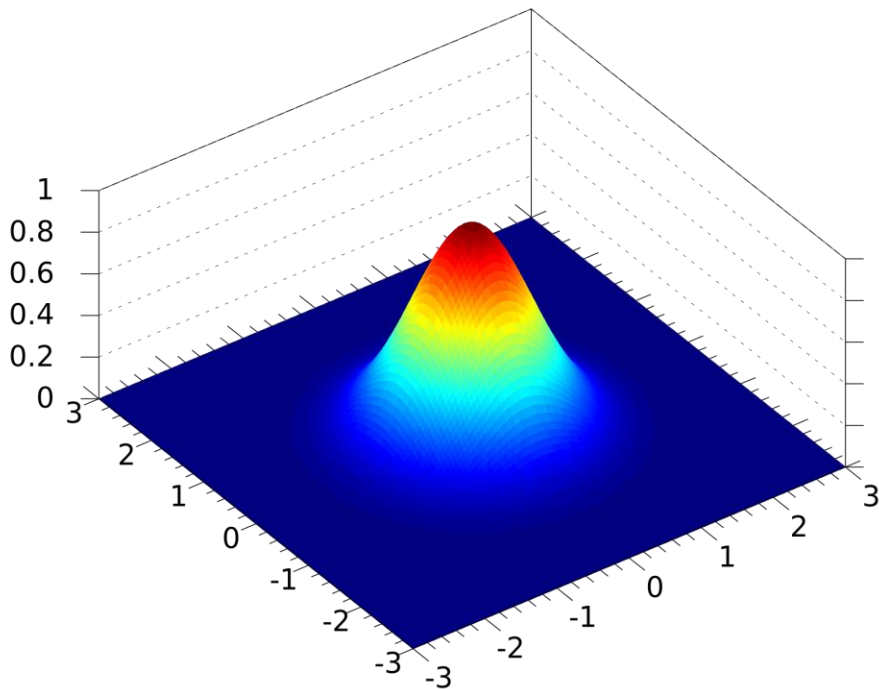
- Model can capture interactions between two features ($x_1 x_2$)
-

Popular Kernels: Gaussian Kernel

Polynomial kernel of degree d :

$$k(x, y) = e^{-\gamma \|x - y\|^2}$$

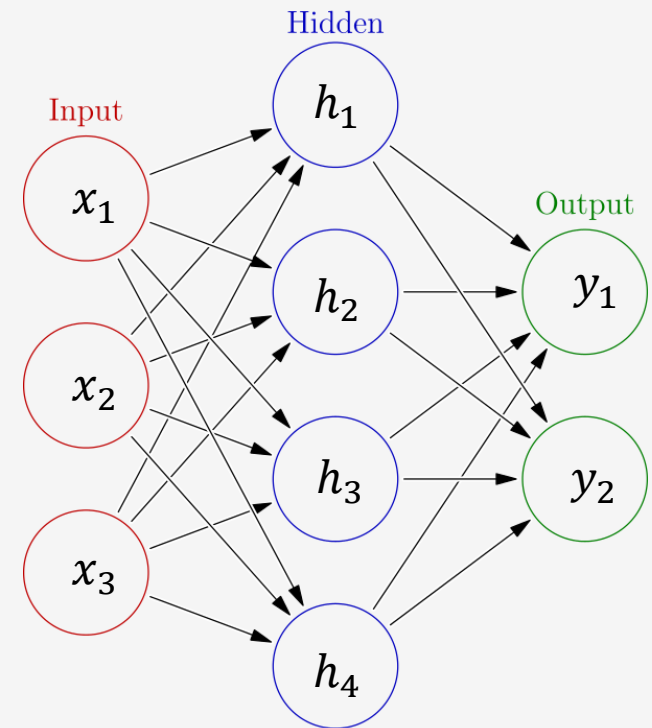
- Projects data points to infinite dimensional space



Artificial Neural Networks (ANN)

Network used to approximate functions

- Nodes in the network are called neurons
- Can depend on large number of inputs
- Architecture: topological structure
- Activation: how are neurons activated

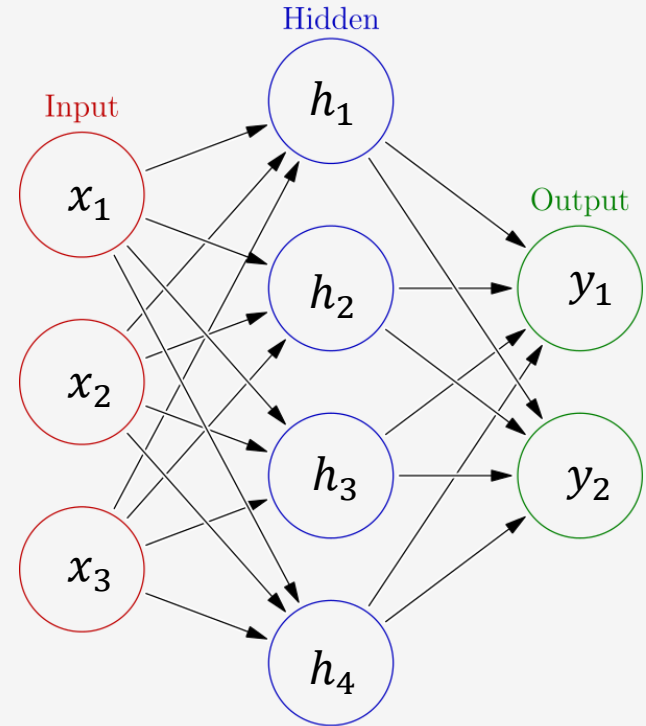


Multilayer Perceptron

Feedforward Artificial Neural Network

- Supervised learning approach
- Consists of at least 3 layers: input, hidden, output
- Activation functions:
 - Linear: $h = Ax, A \in \mathbb{R}^{n,m}$
 - Non-linear: e.g. $g_i = \tanh(h_i)$
- Categorical output for classification can be obtained using **softmax** function:

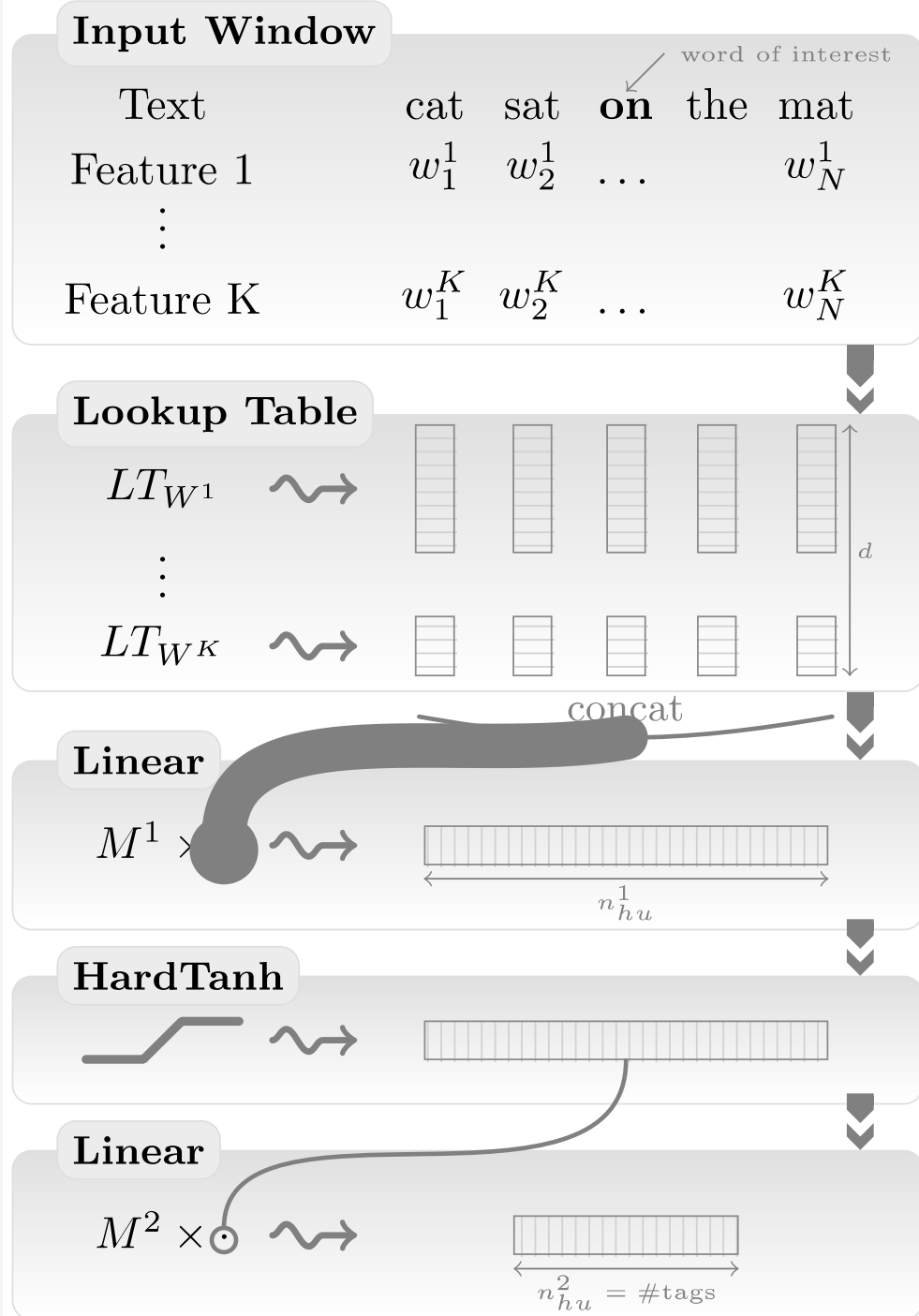
$$P(i|x) = \frac{e^{y_i}}{\sum_j e^{y_j}}$$



NLP Tasks

Natural Language Processing
(Almost) from Scratch:

- Neural Network approach to standard NLP tasks:
- Part-of-Speech tagging
- Chunking
- Named-entity Recognition
- Semantic Role Labeling



<http://playground.tensorflow.org/>

TENSORFLOW

How to Fit a Model

Task: from space of all models (*hypothesis space*) select “the best” model

What do we mean by “best”?

- Model should produce good result on the training set => optimize loss function
- Model should generalize well to unseen data => regularization



How to measure loss?

Compare model outputs with the expected outputs from the training set.

Preferably loss should be easily optimized over.



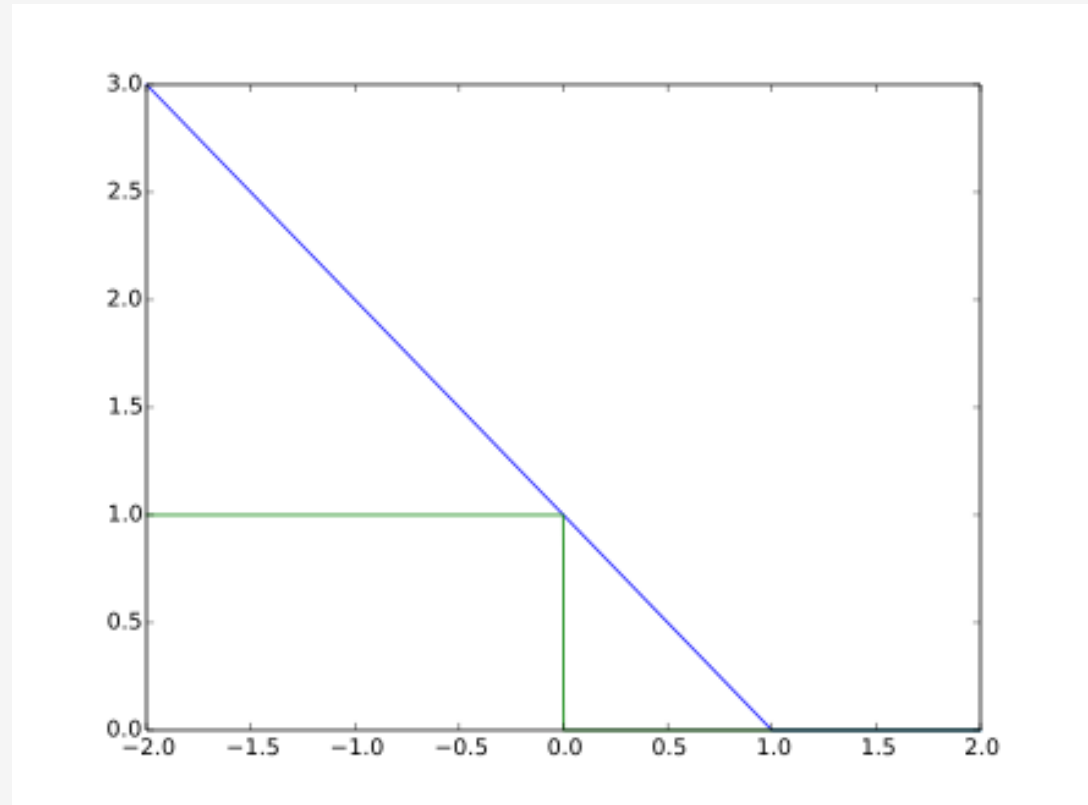
Classification

o-1 loss:

$$L(f) = \sum_i I(y_i f(x_i) > 0)$$

Hinge loss:

$$L(f) = \sum_i \max(0, 1 - y_i f(x))$$



Maximum likelihood estimation

Training data: $\{(x_i, y_i); x_i \in \mathbb{R}^n, y_i \in \{0,1\}\}$

- x_i are IID (Independent and identically distributed)

We can compute likelihood of the training data given a model θ :

$$\prod_i p(y_i|x_i, \theta)$$

In case of logistic regression:

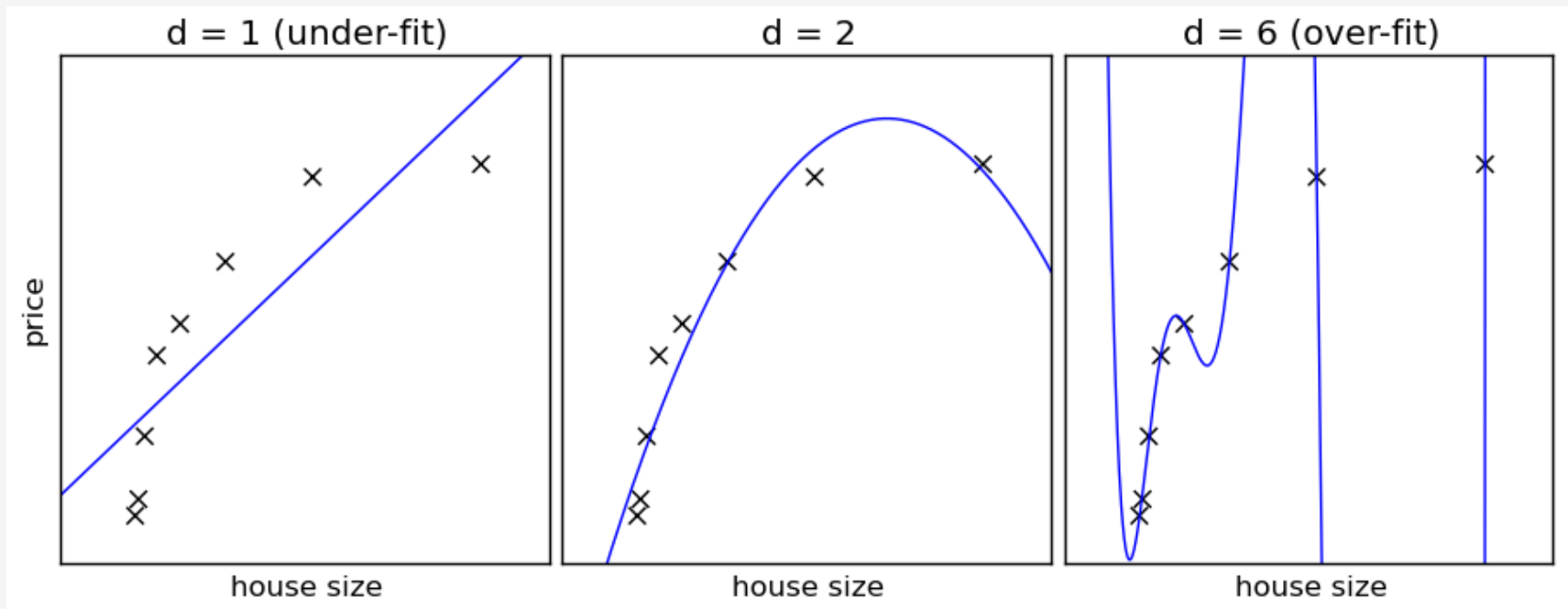
$$p(y_i|x_i, \theta) = \frac{e^{x_i^t \theta}}{e^{x_i^t \theta} + 1}$$

In practice we work with log-likelihood:

$$\sum_i \log p(y_i|x_i, \theta)$$

How to ensure we generalize well?

Better fitting to training data does not always translate to better model!



Ockham (Occam)'s Razor

William of Ockham (1295-1349) was a Franciscan friar who applied the criteria to theology:

- "Entities should not be multiplied beyond necessity" (Classical version but not an actual quote)
- "The supreme goal of all theory is to make the irreducible basic elements as simple and as few as possible without having to surrender the adequate representation of a single datum of experience." (Einstein)

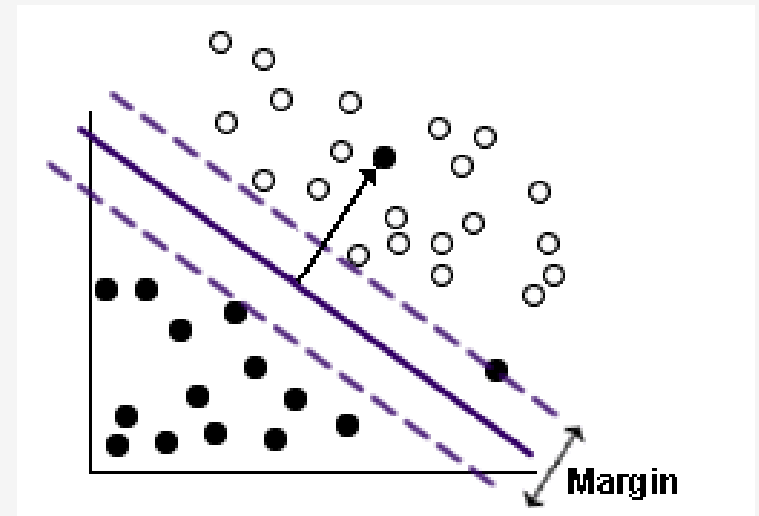
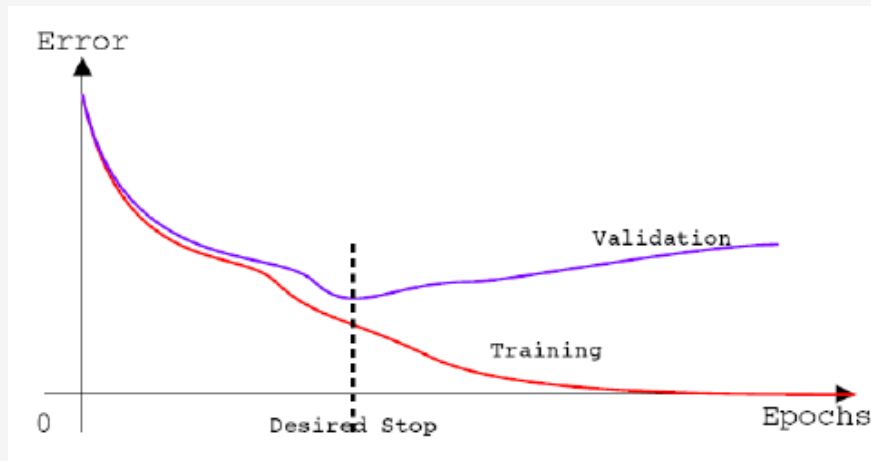
Requires a precise definition of simplicity

Assumes that nature itself is simple.

Model Overfitting

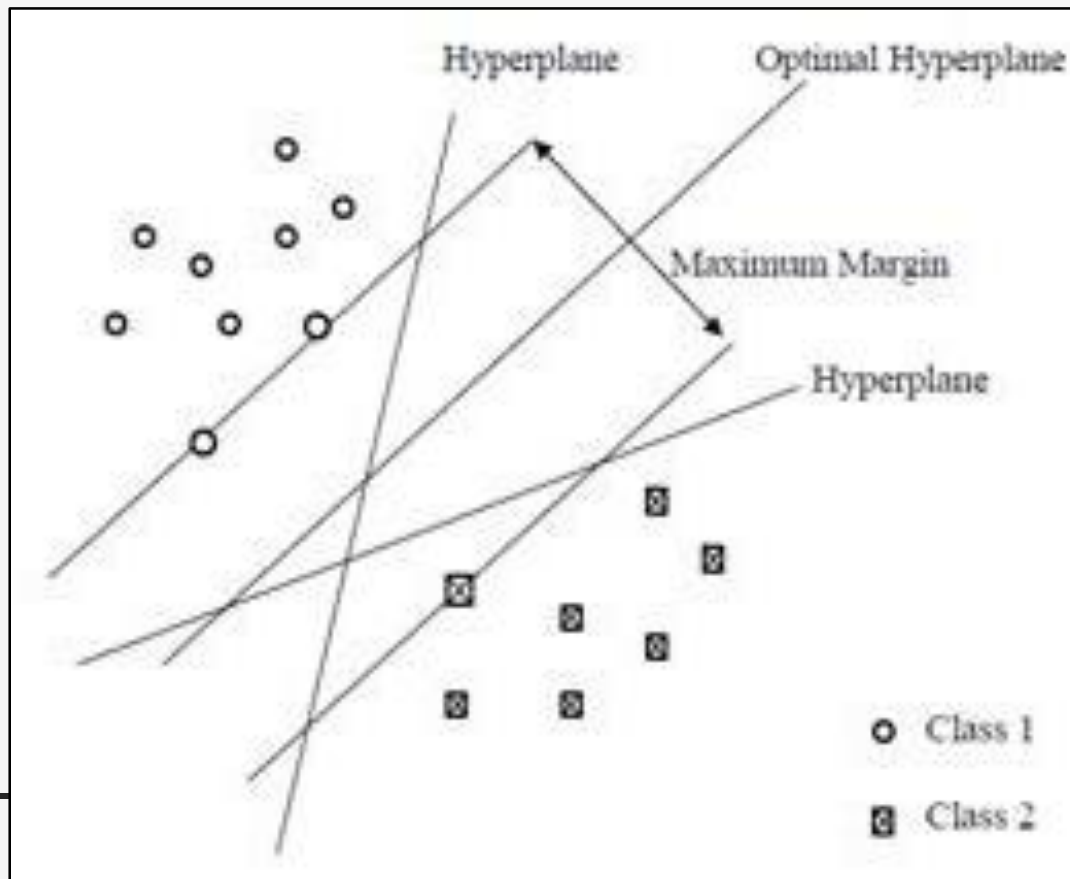
Typically handled through Regularization – favor simpler models

- When model is a linear functional we can add its norm to the criterion
- Early stopping for Neural Networks
- Tree pruning for Decision Tress



Support Vector Machine

- Maximal Margin criterion in Support Vector Machine
- Minimize norm of the model



Support Vector Machine

Hard margin:

$$\begin{aligned} & \min_w \|w\| \\ \text{subject to } & y_i(w^T x_i + b) \geq 1 \end{aligned}$$

Soft margin:

$$\min_w \left(\frac{1}{n} \sum_i \max(0, 1 - y_i w^T x_i) + \|w\|^2 \right)$$

Validation

How to evaluate machine learning model:

- Q1: Can we quantify how well it works?
- Q2: How well does it generalize?
- Q3: Is it better than some other model?



Q1: Quantifying Model Performance

Done via evaluation metrics computed on test set

Evaluation metrics:

- Depend on the problem type
 - E.g. classification vs. regression
- Depend on the domain: what is the cost of a wrong or bad answer
 - E.g. we do not want to miss positive examples
 - E.g. some wrong answers cost more money than other wrong answers?

Difficulties:

- Supervised problems easier to evaluate since we have ground truth
 - Solutions to unsupervised problems can be subjective
 - E.g. how to quantify good clustering vs. bad clustering?
-

Binary Classification

Model $f(x)$ evaluated on test set $\{(x_i, y_i); y_i \in \{1, -1\}\}$

Each element from test set is assigned to one of the following sets:

	y_i	$f(x_i)$
True Positive (TP)	1	1
False Positive (FP)	-1	1
True Negative (TN)	-1	-1
False Negative (FN)	1	-1

$$precision = \frac{TP}{TP + FP}$$

$$recall = \frac{TP}{TP + FN}$$

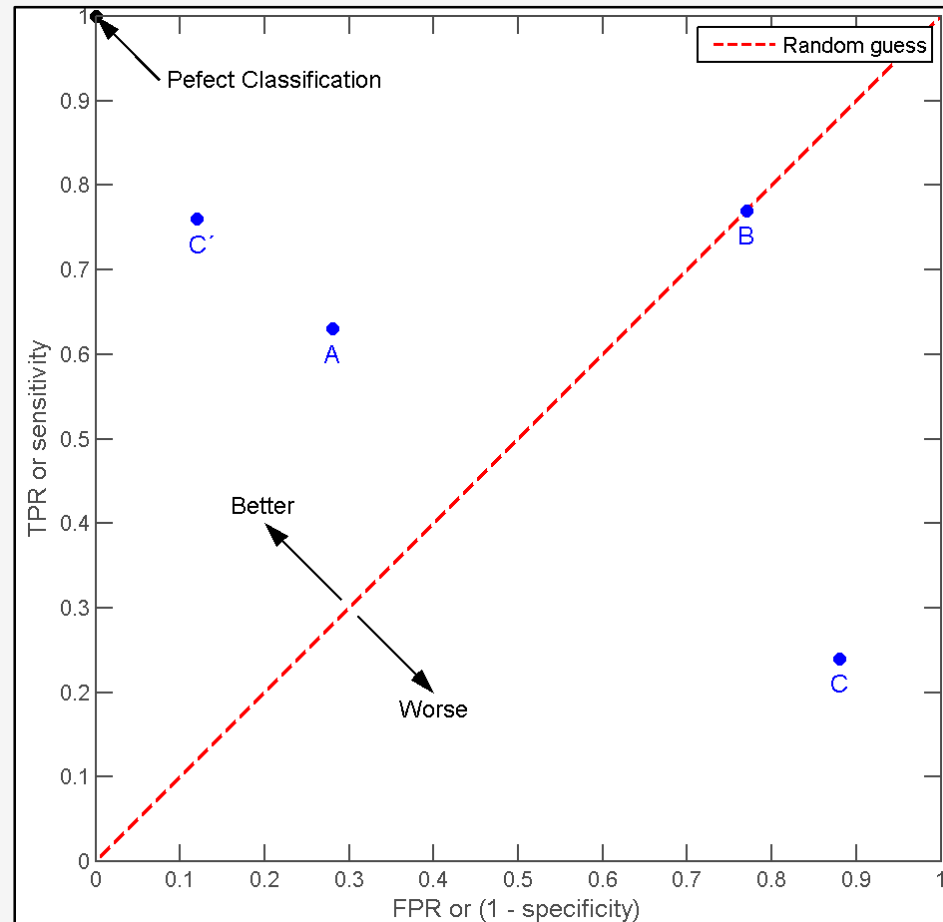
$$F_1 = \frac{2 \text{ precision recall}}{\text{precision} + \text{recall}}$$

ROC Curve

ROC = Receiver operating characteristic

Plot models by two characteristics computed from the test set:

- True positive rate: $P(TP) = \frac{TP}{TP+FN}$
- False positive rate: $P(FP) = \frac{FP}{FP+TN}$

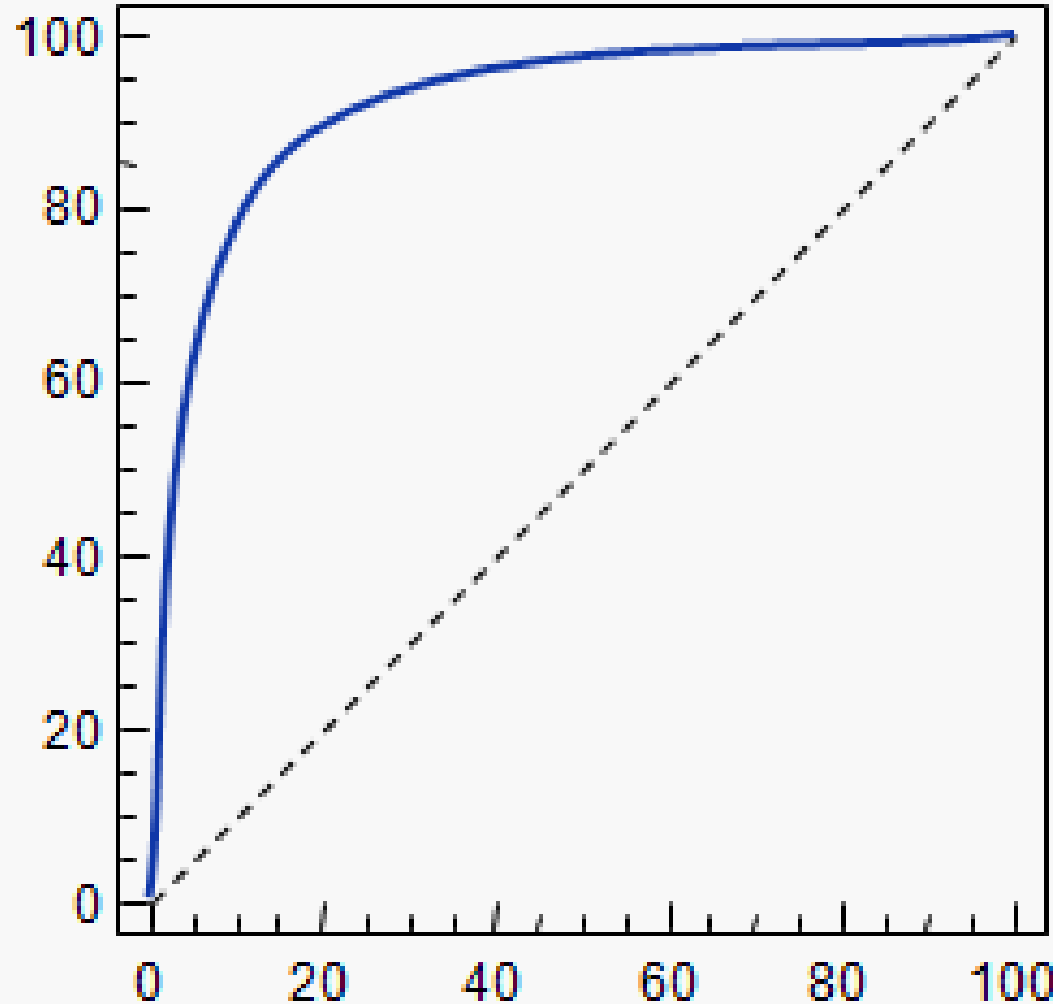


Area Under the Curve (AUC)

Classification model returns
some confidence score

Sort classification of test
examples by this score

Compute ROC plot for different
positive classification threshold



Regression

Model $f(x)$ evaluated on test set $\{(x_i, y_i); y_i \in \mathbb{R}\}$

- Residual: $f(x_i) - y_i$

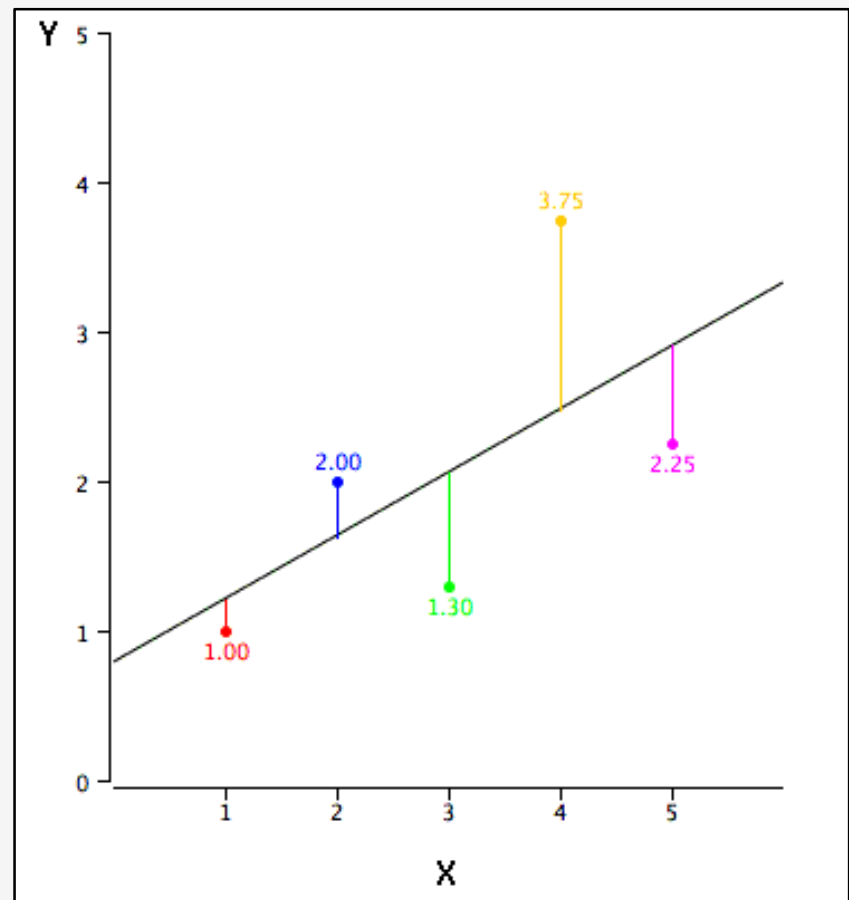
Standard error metrics:

- Mean Square Error

$$\frac{1}{n} \sum_{i=1}^n (f(x_i) - y_i)^2$$

- Coefficient of determination

$$R^2 = 1 - \frac{\sum_{i=1}^n (f(x_i) - y_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$



Q2: How well does it generalize?

Golden rule: never evaluate on the training data

- Split data into training and testing parts
- Train on the training part
- Compute evaluation metrics on the testing parts

Avoid leaks from training to test data:

- Time-specific datasets like news articles
- Feature extraction and representation learning should not see test data



Confidence

Way of obtaining multiple metrics from same dataset:

- k -fold cross-validation
 - Generate k random train-test splits and compute evaluation metrics from each
- Leave-one-out cross validation
- Edge case of k -fold cross-validation



Q3: Is it better than some other model?

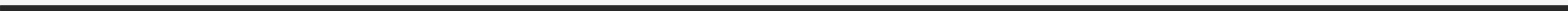
We have two models: f_1 and f_2

How can we tell which is better:

- Compute evaluation metrics on several train-test splits for each model
- Result: two sets of metrics
- Test hypothesis:
 - They have same mean: Permutation test
 - Two sets are drawn from the same distribution: Kolmogorov-Smirnoff test



HANDS-ON



Hands-on

Tools (all on web, no installation necessary)

- Tonic: <https://tonicdev.com/>
- QMiner: <http://qminer.ijs.si>

Example 1: Sentiment Classifier

- <https://tonicdev.com/rupnikj/sentiment-classifier-hands-on>

Example 2: Hangover Regression:

- <https://tonicdev.com/blazf/hangover-regression>
-