CORNELL TECH

Deep Learning Clinic
(DLC)

Lecture 2

A Brief Introduction to Machine Learning

Jin Sun

9/17/2019

Today

- **Overview**
- Formulation of Learning
- Learning Models
- Loss Function
- Optimization
- Data and Evaluation

Overview

“Any plausible approach to artificial intelligence must involve learning, at some level, … it’s hard to call a system intelligent if it *cannot* learn.”

-- CIML Book

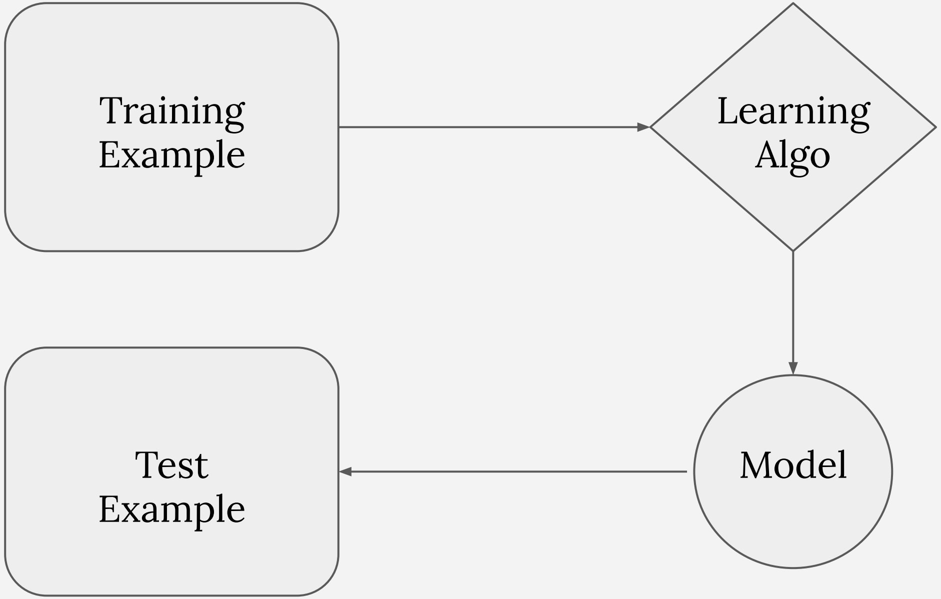
What is *Machine Learning* (ML)?

“ML is about predicting the future based on the past.” (CIML)

Core questions:

What to learn? How to learn? How good is the learning?

Machine Learning Paradigm



Example: Spam Detection

Training Data/Examples

ham	Go until jurong point, crazy.. Available only in bugis n great world la e buffet... Cine there got amore wat...
ham	Ok lar... Joking wif u oni...
spam	Free entry in 2 a wkly comp to win FA Cup final tkts 21st May 2005. Text FA to 87121 to receive entry question(std txt rate)T&C's apply 08452810075over18's
ham	U dun say so early hor... U c already then say...
ham	Nah I don't think he goes to usf, he lives around here though
spam	FreeMsg Hey there darling it's been 3 week's now and no word back! I'd like some fun you up for it still? Tb ok! XxX std chgs to send, å£1.50 to rcv
ham	Even my brother is not like to speak with me. They treat me like aids patent.
ham	As per your request 'Melle Melle (Oru Minnaminunginte Nurungu Vettam)' has been set as your callertune for all Callers. Press *9 to copy your friends Callertune
spam	WINNER!! As a valued network customer you have been selected to receivea å£900 prize reward! To claim call 09061701461. Claim code KL341. Valid 12 hours only.
spam	Had your mobile 11 months or more? U R entitled to Update to the latest colour mobiles with camera for Free! Call The Mobile Update Co FREE on 08002986030

https://towardsdatascience.com/spam-detection-with-logistic-regression-23e3709e522

Example: Spam Detection

Translate data into some easier to manipulate form

ham	Go until jurong point, crazy.. Available only in bugis n great world la e buffet... Cine there got amore wat...
ham	Ok lar... Joking wif u oni...
spam	Free entry in 2 a wkly comp to win FA Cup final tkts 21st May 2005. Text FA to 87121 to receive entry question(std txt rate)T&C's apply 08452810075over18's
ham	U dun say so early hor... U c already then say...
ham	Nah I don't think he goes to usf, he lives around here though

Dictionary

1. email 2163
2. order 1648
3. address 1645
4. language 1534
5. report 1384
6. mail 1364

...

$\longrightarrow$

Bag of Words Features

1 7 1
1 12 2
1 19 2
1 22 1
1 25 1
...

https://towardsdatascience.com/spam-detection-with-logistic-regression-23e3709e522
https://medium.com/analytics-vidhya/building-a-spam-filter-from-scratch-using-machine-learning-fc58b178ea56

Example: Spam Detection

A learning algorithm - Naive Bayes

$$p(C_k \mid x_1, \ldots, x_n)$$

0: Spam
1: Non-spam

The word 'homework' is in the sentence.
True/False

The word 'reward' is in the sentence.
True/False

Example: Spam Detection

A learning algorithm - Naive Bayes

$$p(C_k \mid x_1, \dots, x_n)$$

Likelihood

Bayes rule: $$p(C_k \mid \mathbf{x}) = \frac{p(C_k)\, p(\mathbf{x} \mid C_k)}{p(\mathbf{x})}$$

Posterior

Feature Prior

Class prior

Example: Spam Detection

A learning algorithm - Naive Bayes

$$p(C_k \mid x_1, \dots, x_n)$$

$$p(C_k \mid \mathbf{x}) = \frac{p(C_k)\ p(\mathbf{x} \mid C_k)}{p(\mathbf{x})}$$

$$\begin{aligned} p(C_k \mid x_1, \dots, x_n) &\propto p(C_k, x_1, \dots, x_n) \\ &= p(C_k)\ p(x_1 \mid C_k)\ p(x_2 \mid C_k)\ p(x_3 \mid C_k)\ \cdots \\ &= p(C_k) \prod_{i=1}^{n} p(x_i \mid C_k)\,, \end{aligned}$$

"Naive" assumption

Value comes from training data -> Learning!

Example: Spam Detection

p(spam | ‘reward’=1, ‘homework’=0, ...)

~ p(spam) * p(‘reward’=1 | spam) * p(‘homework’=0 | spam) * ...

= 0.2 * 0.7 * 0.9 * ...

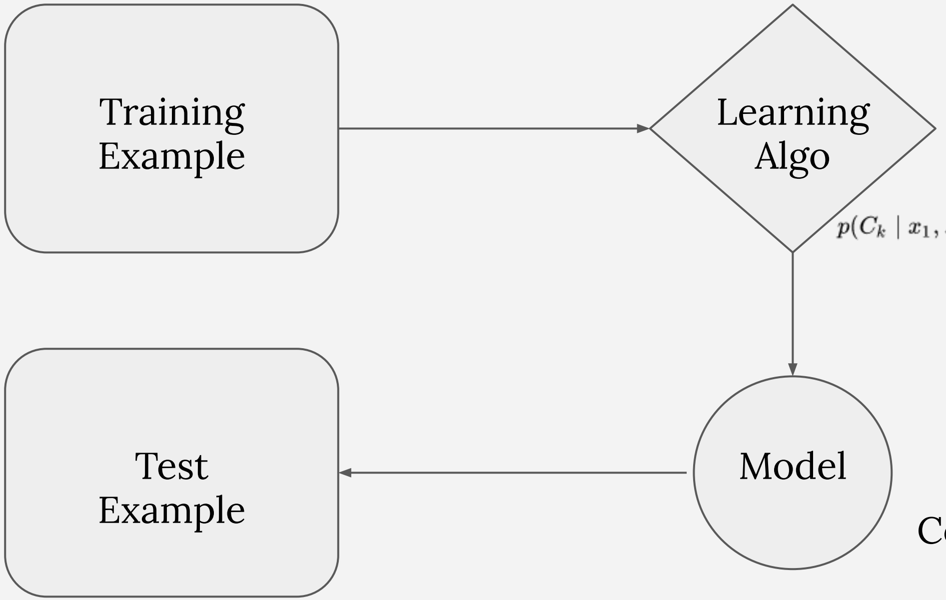
p(spam | ‘reward’=0, ‘homework’=1, ...)

~ p(spam) * p(‘reward’=0 | spam) * p(‘homework’=1 | spam) * ...

= 0.2 * 0.3 * 0.1 * ...

Machine Learning Paradigm - Spam Detection

ham	Go until jurong point, crazy.. Availa
ham	Ok lar... Joking wif u oni...
spam	Free entry in 2 a wkly comp to win
ham	U dun say so early hor... U c alread
ham	Nah I don't think he goes to usf, he



Naive Bayes

$$p(C_k \mid \mathbf{x}) = \frac{p(C_k)\ p(\mathbf{x} \mid C_k)}{p(\mathbf{x})}$$

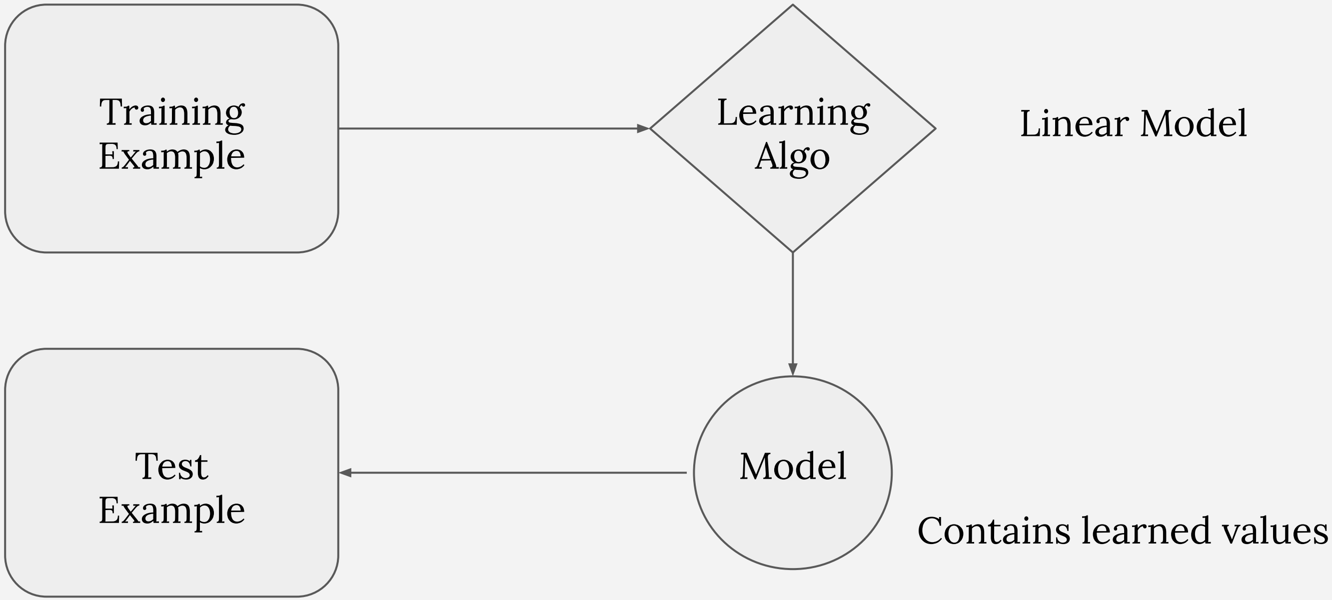
$$\begin{aligned} p(C_k \mid x_1, \dots, x_n) &\propto p(C_k, x_1, \dots, x_n) \\ &= p(C_k)\ p(x_1 \mid C_k)\ p(x_2 \mid C_k)\ p(x_3 \mid C_k) \\ &= p(C_k) \prod_{i=1}^{n} p(x_i \mid C_k)\,, \end{aligned}$$

ham	Even my brother is not like to spea
ham	As per your request 'Melle Melle ((
spam	WINNER!! As a valued network cus
spam	Had your mobile 11 months or mo

Contains learned values

Machine Learning Paradigm - Spam Detection

ham	Go until jurong point, crazy.. Availa
ham	Ok lar... Joking wif u oni...
spam	Free entry in 2 a wkly comp to win
ham	U dun say so early hor... U c alread
ham	Nah I don't think he goes to usf, he



ham	Even my brother is not like to spea
ham	As per your request 'Melle Melle ((
spam	WINNER!! As a valued network cus
spam	Had your mobile 11 months or mo

Example: Spam Detection - A Linear Model

X = [‘reward’=1, ‘homework’=0, …]

```
If    W*X > 0:
      Spam;
Else:
      Ham.
```

How do we know W?

Example: Spam Detection - A Linear Model

X = [‘reward’=1, ‘homework’=0, …]

If W*X > 0:

Spam;

Else:

Ham.

We pick the ‘W’ that has the best spam prediction accuracy in training data.

Example: Spam Detection - A Linear Model

X = [‘reward’=1, ‘homework’=0, …]

If W*X > 0:

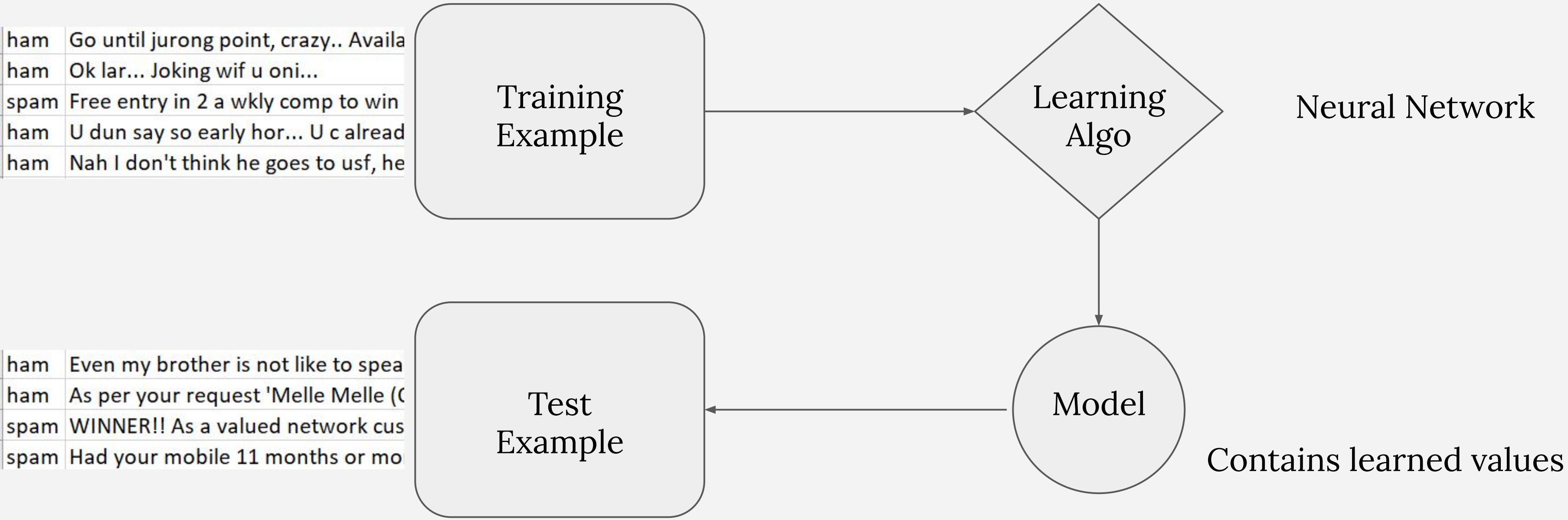
Spam;

Else:

Ham.

$$\mathbf{W} = \underset{W}{\operatorname{argmin}} \| W*X - y \|$$

Machine Learning Paradigm - Spam Detection



Another ML example:

1	real world goal	increase revenue
2	real world mechanism	better ad display
3	learning problem	classify click-through
4	data collection	interaction w/ current system
5	collected data	query, ad, click
6	data representation	bow^2, $\pm$ click
7	select model family	decision trees, depth 20
8	select training data	subset from april'16
9	train model & hyperparams	final decision tree
10	predict on test data	subset from may'16
11	evaluate error	zero/one loss for $\pm$ click
12	deploy!	(hope we achieve our goal)

* CIML Fig 2.4.

Types of Learning Problems

Classification

Predict Yes/No (Binary), or from a set of labels (Multi-class).

Regression

Predict a real value: e.g., tomorrow's stock price.

Structure Learning

Predict a graph, a ranking, etc.

Today

- ❑ Overview
- ❑ **Formulation of Learning**
- ❑ Learning Models
- ❑ Loss Function
- ❑ Optimization
- ❑ Data and Evaluation

Formal Definition of Learning

Notations and their meaning:

x : our input features (e.g., words frequency)

y : our ground truth labels (e.g., whether is a spam)

$f(\cdot)$: the function we are learning to predict y from x

$L(\cdot, \cdot)$: "loss function" -- how good a given function is on the training data

Formal Definition of Learning

Data
(word freq)

Label
spam or not

$$e \doteq \mathop{\mathbb{E}}_{(x,y)\ \sim D}[L(y, f(x))]$$

$$\doteq \frac{1}{N}\sum_{n=1}^{N} L(y_n, f(x_n))$$

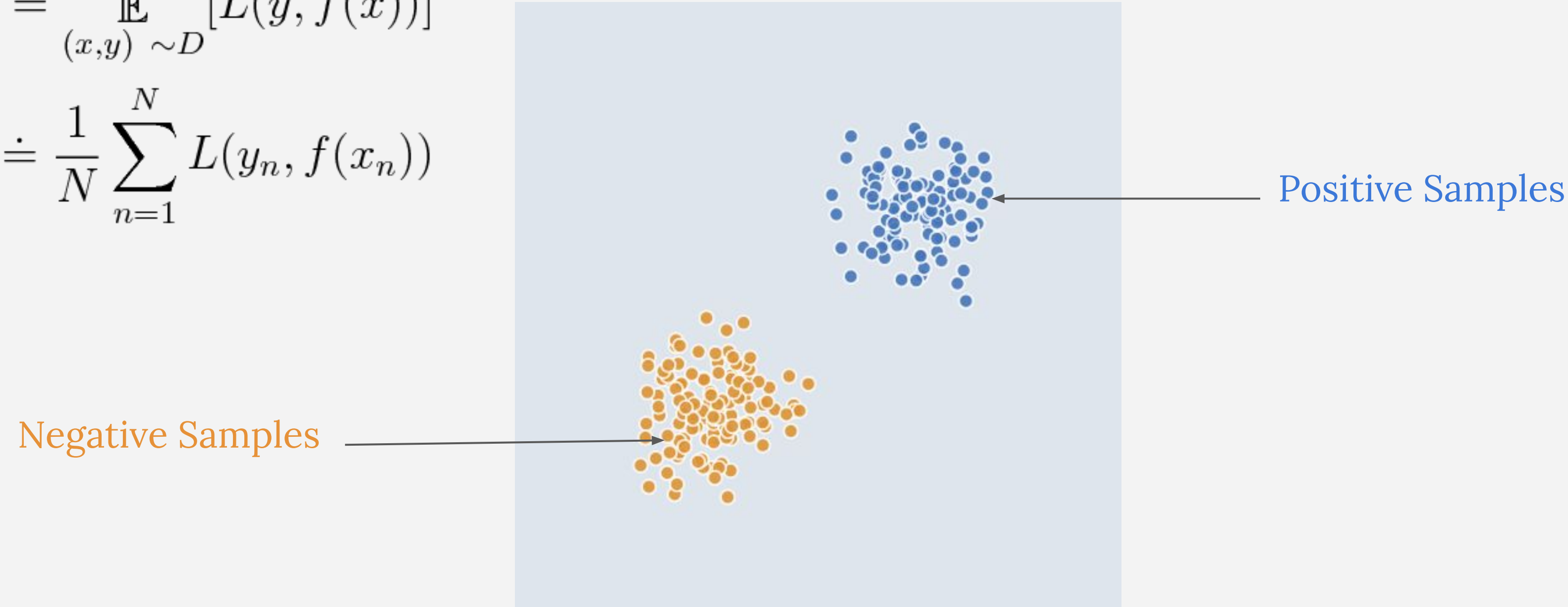
Learning Model

Loss function
How good is our spam predictor?

A Concrete Example - Binary Classification

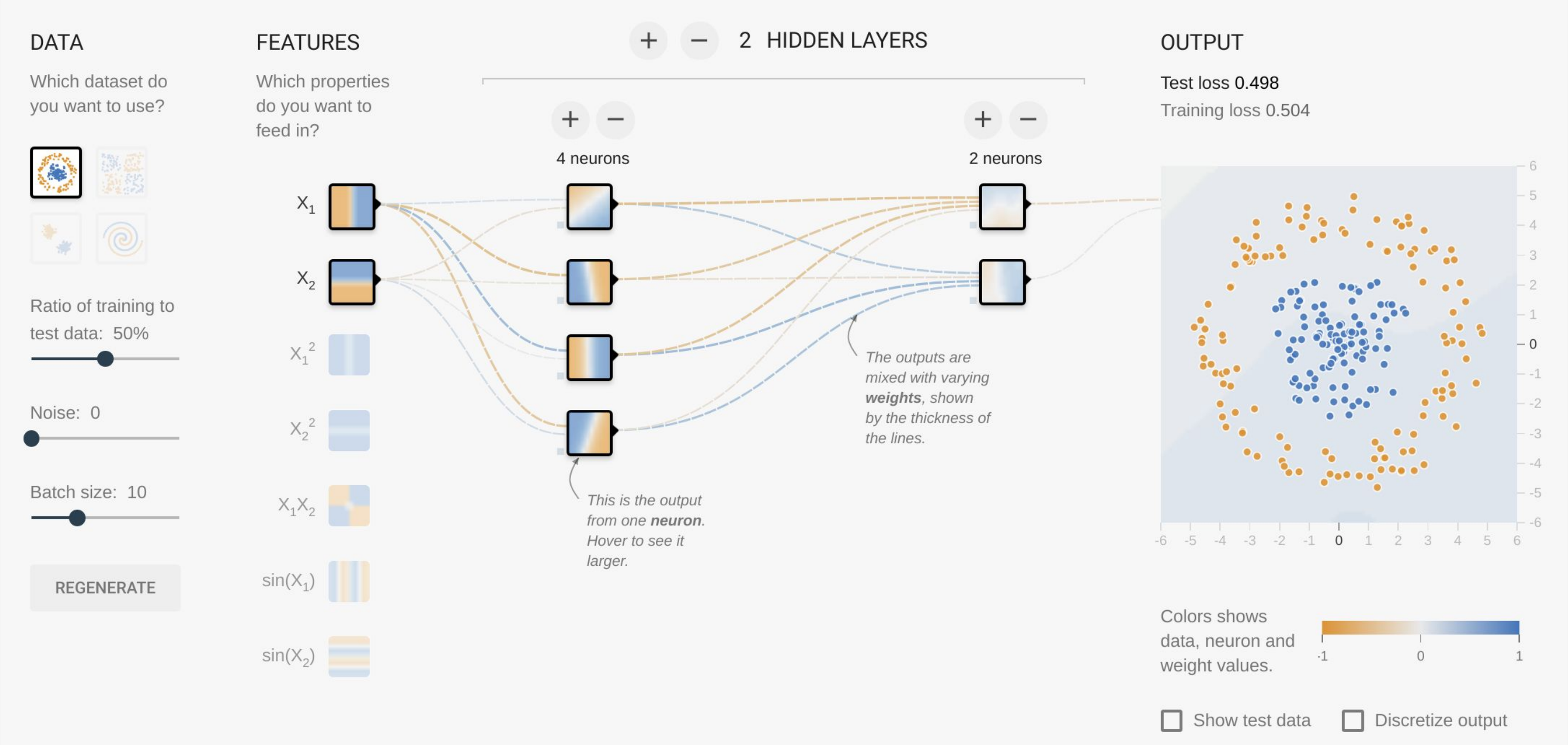
$$e \doteq \mathop{\mathbb{E}}_{(x,y)\ \sim D}[L(y, f(x))]$$

$$\doteq \frac{1}{N}\sum_{n=1}^{N} L(y_n, f(x_n))$$

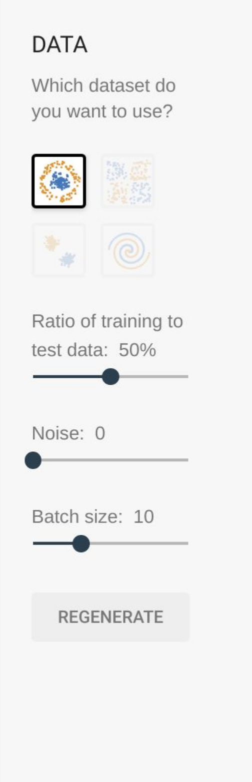


http://playground.tensorflow.org/

A Simple Interactive Machine Learning Example



A Neural Network Playground [Link]



Data:

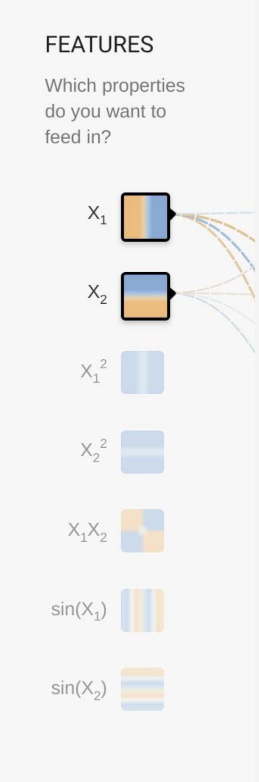
(x,y) 2D Points

Binary Label

Train/Test Split

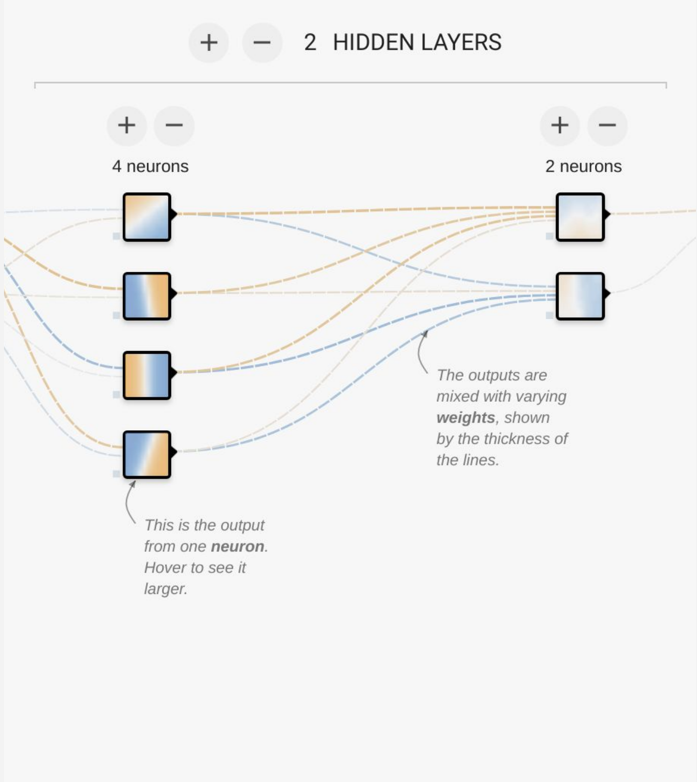
Noise Level

Batch Size



Feature Representation:

Learning problem becomes easier/harder with different feature representations, even with the same data!



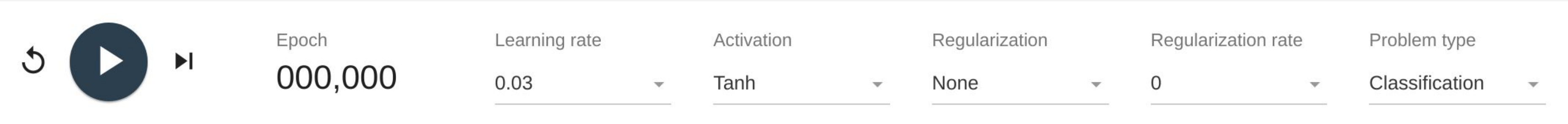
A Learning Model:

Network Structure

Layers

Connectivity

No network works for all the problems!



Training:

Train/Test Loss

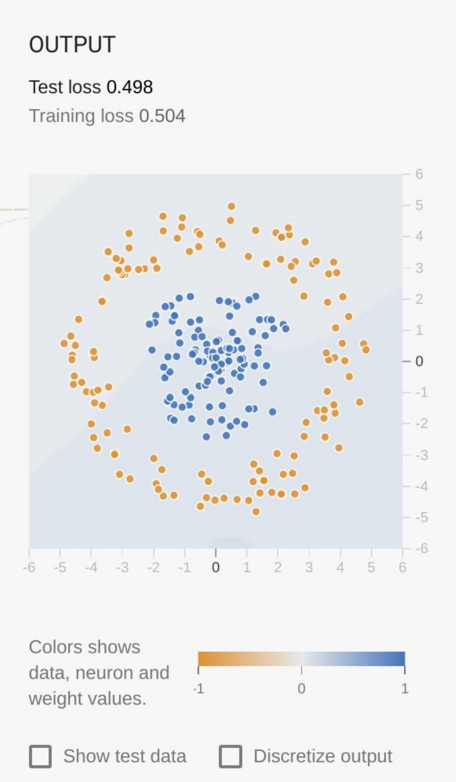
Epochs

Optimization Algorithm

Evaluation:

Metric (accuracy, distance, …)

Cross-validation



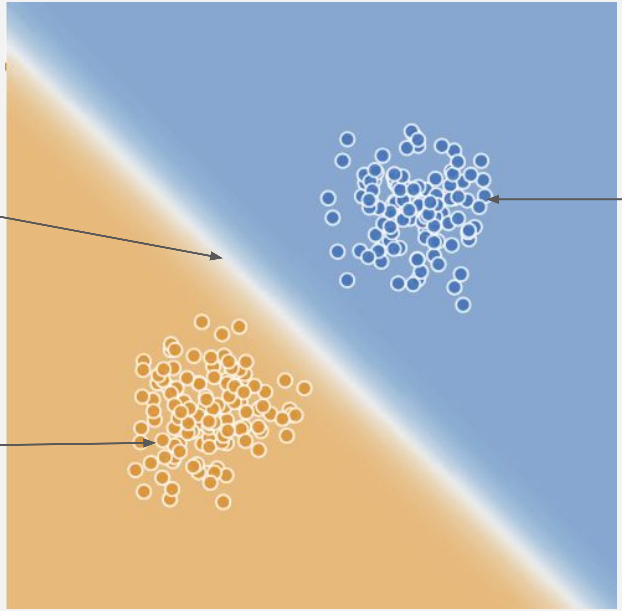
Today

- ❑ Overview
- ❑ Formulation of Learning
- ❑ **Learning Models**
- ❑ Loss Function
- ❑ Optimization
- ❑ Data and Evaluation

A Concrete Example - Binary Classification

$$e \doteq \mathop{\mathbb{E}}_{(x,y)\sim D}[L(y, f(x))]$$

$$\doteq \frac{1}{N}\sum_{n=1}^{N} L(y_n, f(x_n))$$



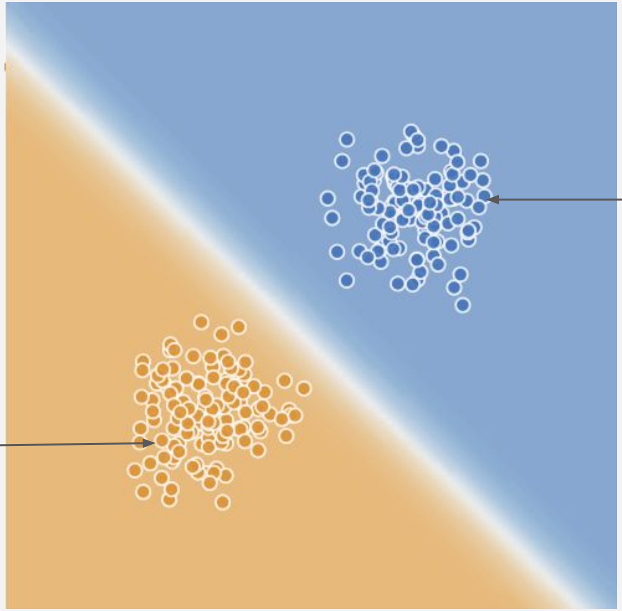
Positive Samples

Negative Samples

Choose Your Model

$$e \doteq \mathop{\mathbb{E}}_{(x,y)\sim D}[L(y, f(x))]$$

$$\doteq \frac{1}{N}\sum_{n=1}^{N} L(y_n, f(x_n))$$



Positive Samples

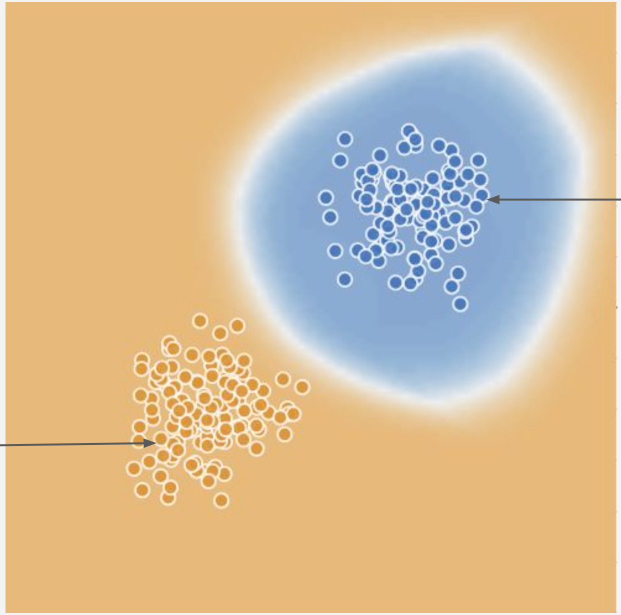
Negative Samples

Linear Function

Choose Your Model

$$e \doteq \mathop{\mathbb{E}}_{(x,y)\sim D}[L(y, f(x))]$$

$$\doteq \frac{1}{N}\sum_{n=1}^{N} L(y_n, f(x_n))$$

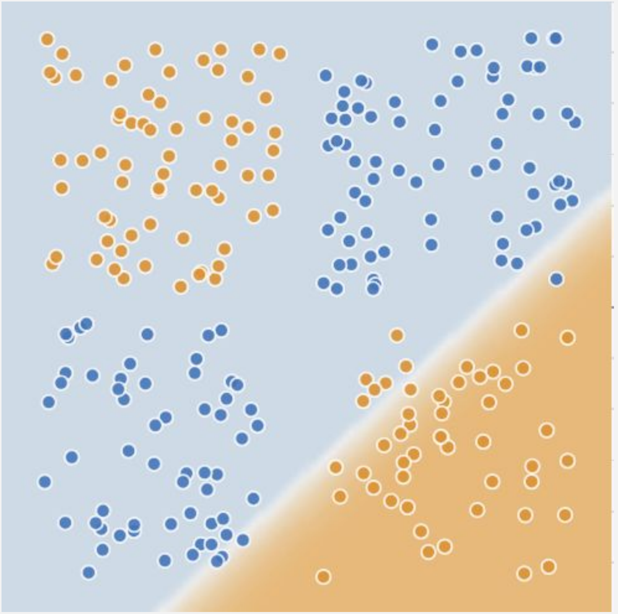


Positive Samples

Negative Samples

Non-linear Function

Pick a Model That Fits the Data Complexity



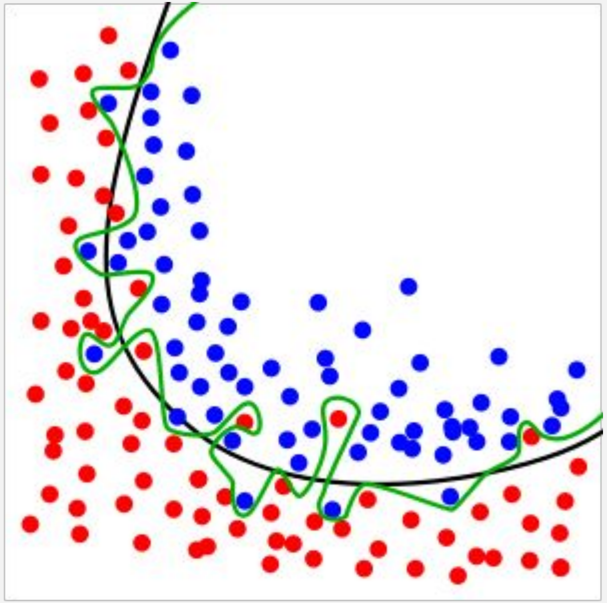
Linear Function
Not Suitable

Generalization

So why not always pick the most complex model?

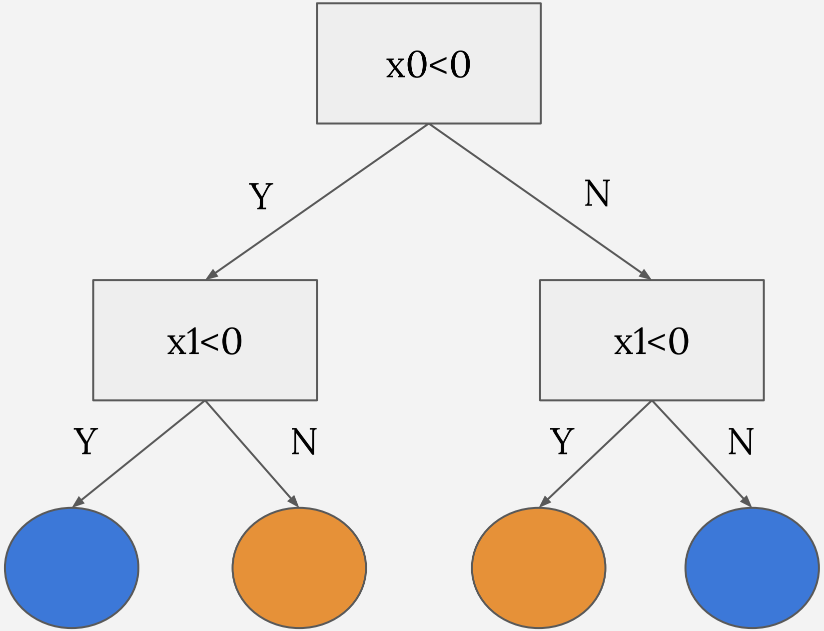
We care about our model's performance on *unseen* test data: the *generalization* ability.

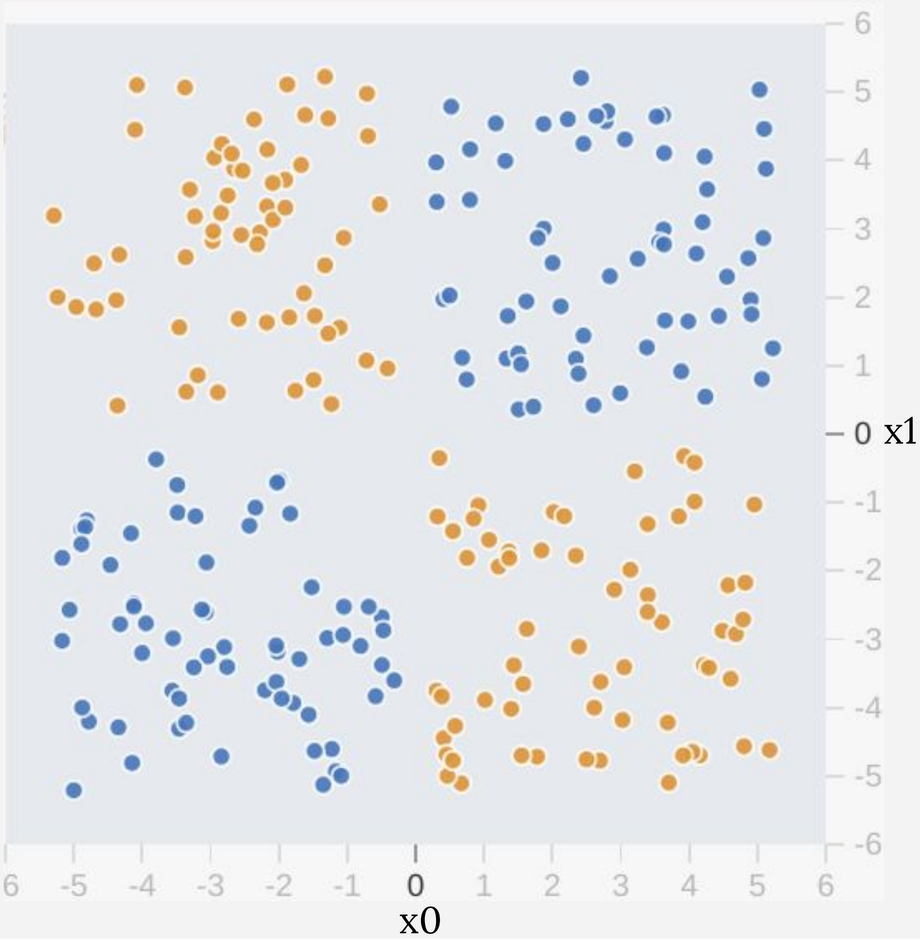
If our model is over-complex, it can be *overfitted* to training and perform poorly on testing data.



Models

Decision Trees



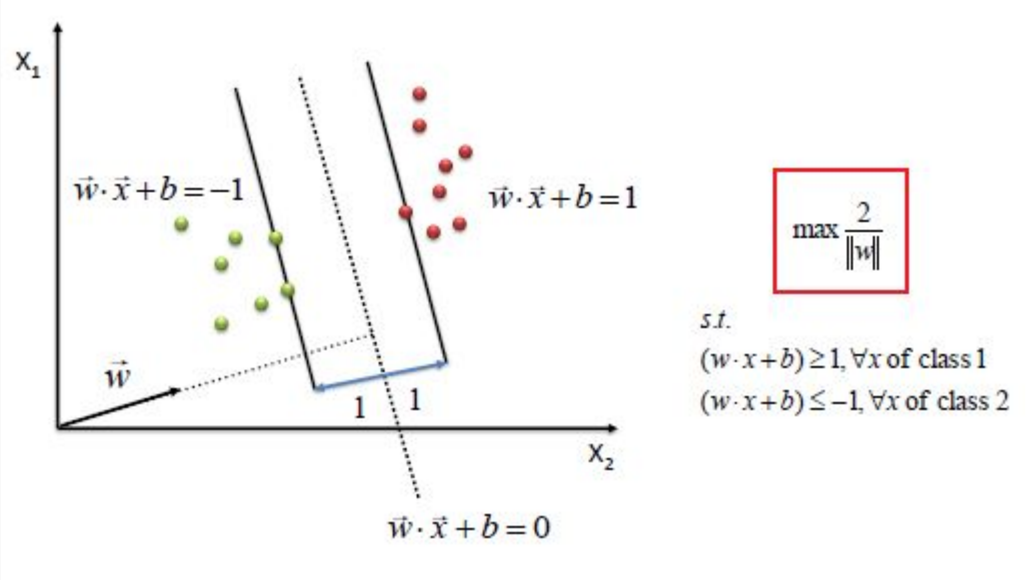


Models

Linear Function

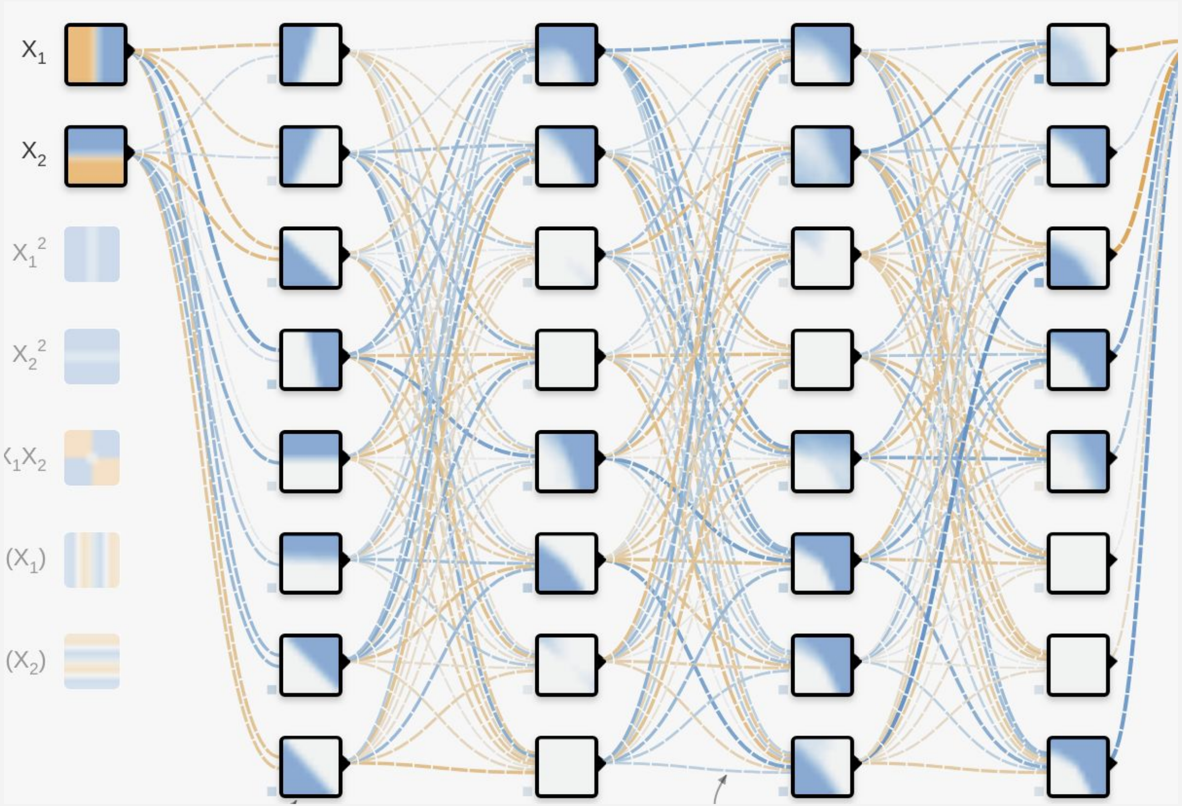
$$f(x) = Wx - b$$

Support Vector Machine (SVM)



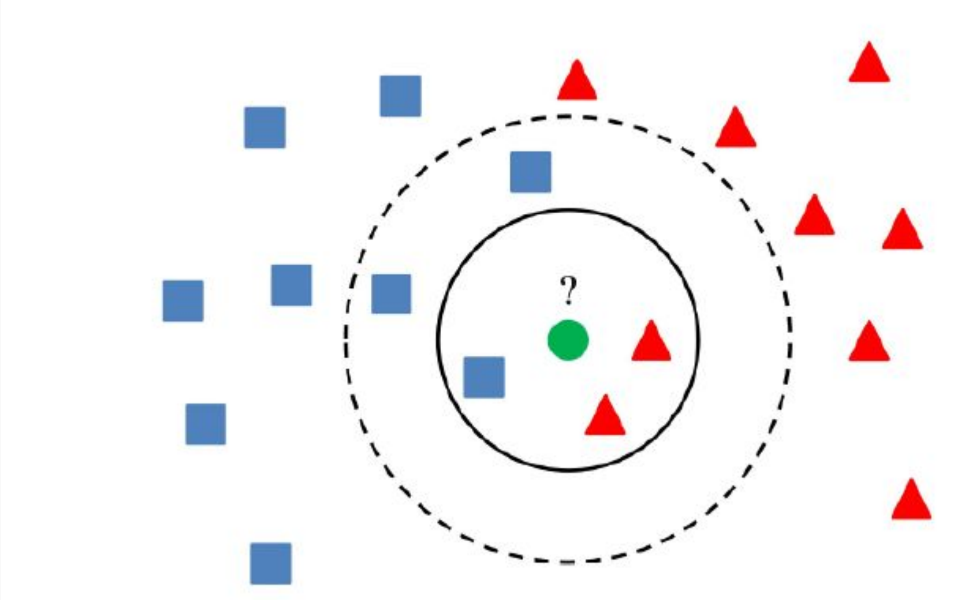
Models

Neural Networks



Non-Parametric Models

Nearest Neighbor



Today

- ❑ Overview
- ❑ Formulation of Learning
- ❑ Learning Models
- ❑ **Loss Function**
- ❑ Optimization
- ❑ Data and Evaluation

Loss Function

Measure how good a model is on the training data.

$$e \doteq \mathop{\mathbb{E}}_{(x,y)\ \sim D}[L(y, f(x))]$$

$$\doteq \frac{1}{N}\sum_{n=1}^{N} L(y_n, f(x_n))$$

Loss function

Loss/Cost/Objective Function

Choose a Loss Function

Classification:

Hinge Loss	$\max(0, 1 - f(x) \cdot y)$
Cross Entropy	$-(y \ln(f(x)) + (1 - y) \ln(1 - f(x)))$

Regression:

MSE Loss	$(f(x) - y)^2$
L1 Loss	$\lvert f(x) - y \rvert$
KL Divergence	$\sum f(x) \ln \frac{f(x)}{y}$

Today

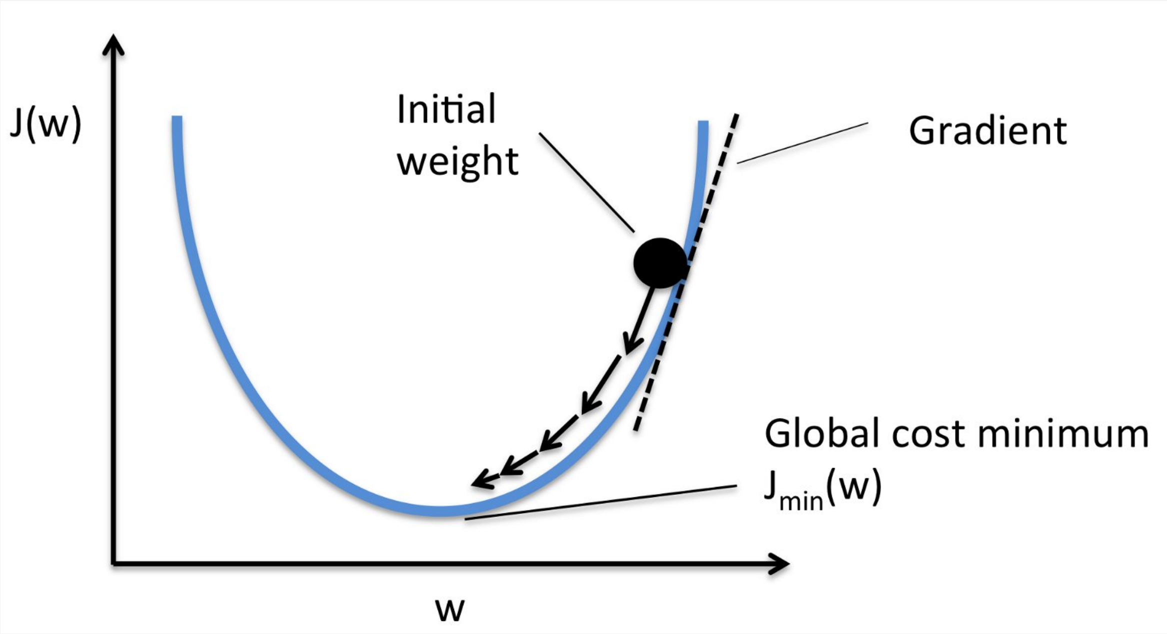
- ❑ Overview
- ❑ Formulation of Learning
- ❑ Learning Models
- ❑ Loss Function
- ❑ **Optimization**
- ❑ Data and Evaluation

Get Training Started - Optimization

$$\text{minimize}_\theta \ e \doteq \mathbb{E}_{(x,y)\sim D}[L(y, f(x;\theta))]$$

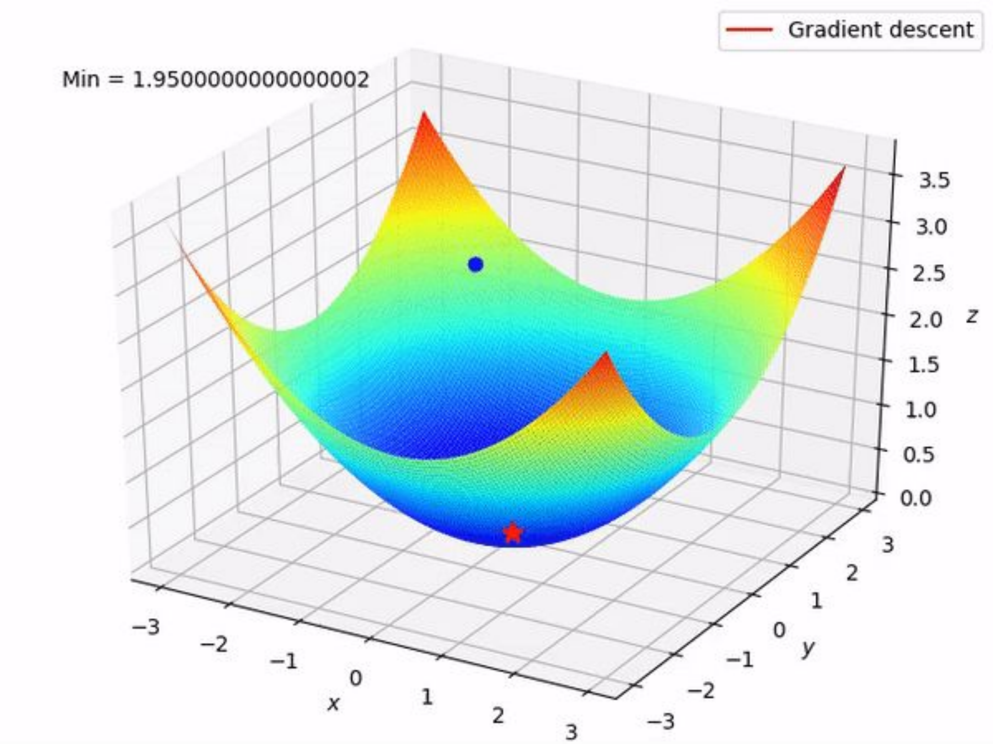
Find the best θ that minimizing the expected loss.

Gradient Descent



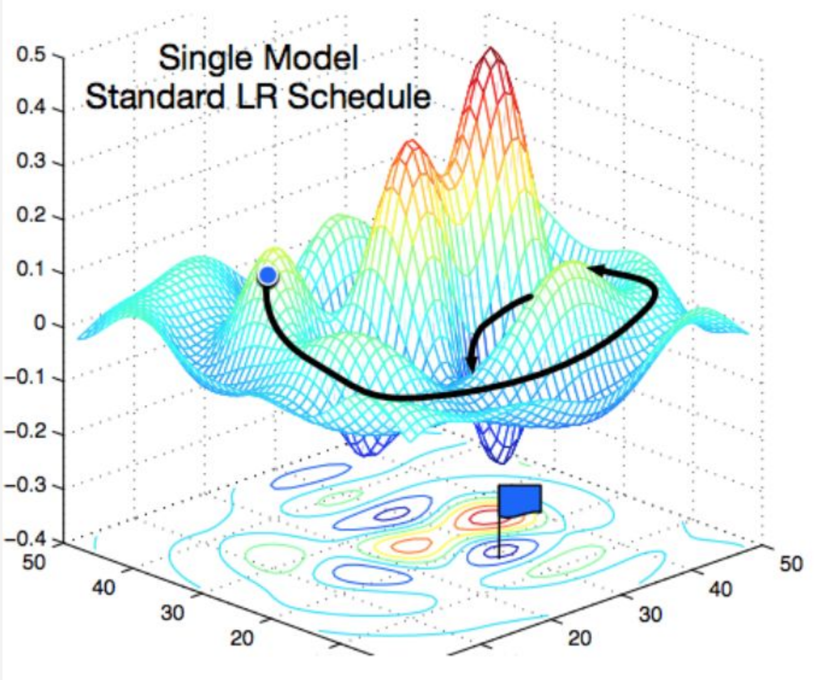
1D Loss Function

Gradient Descent



2D Loss Surface

Gradient Descent



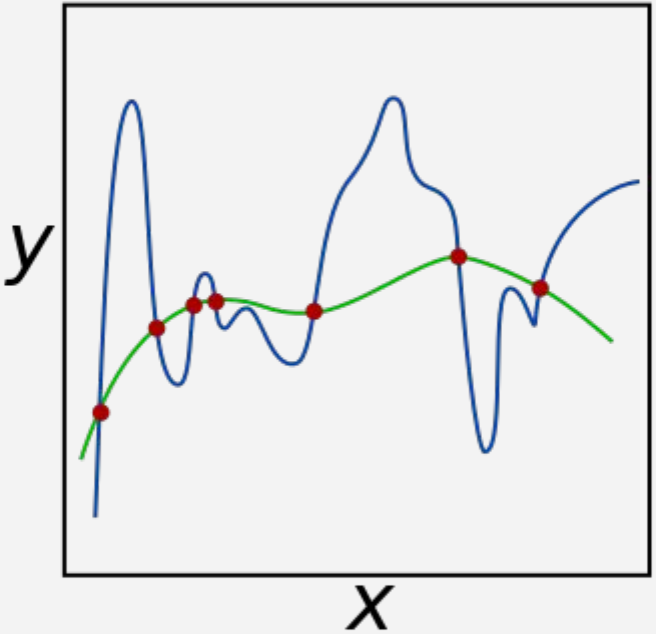
Non-Convex Loss Surface

Optimization Solvers

Dlib	Optimization library in C++
SciPy	Numeric package for Python
MATLAB	[Commercial]
Gurobi	[Commercial]
Deep Learning Frameworks (PyTorch, Tensorflow, and etc)	Built-in GD solvers

Regularization

$$\text{minimize}_\theta \; e \doteq \mathbb{E}_{(x,y)\sim D}[L(y, f(x;\theta))] \boxed{+ \lambda R(\theta)}$$

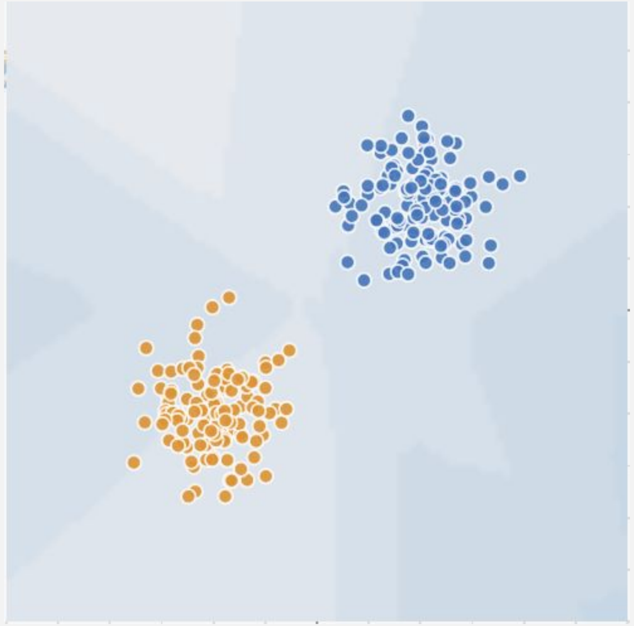


E.g., L1, L2 norm

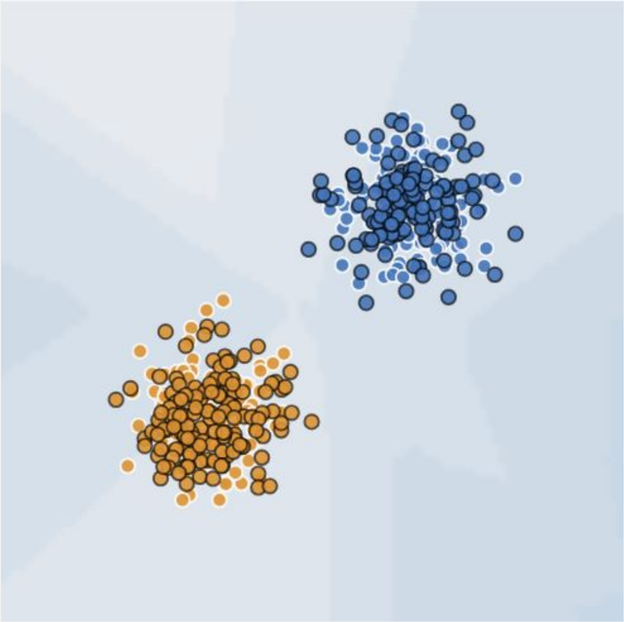
Today

- ❑ Overview
- ❑ Formulation of Learning
- ❑ Learning Models
- ❑ Loss Function
- ❑ Optimization
- ❑ **Data and Evaluation**

Data



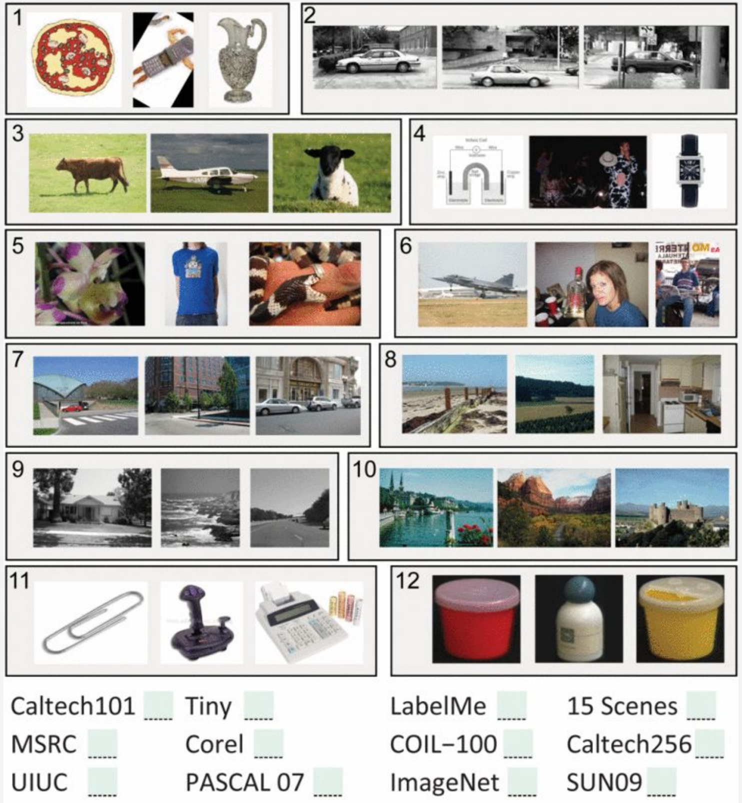
Training Set



Testing Set

Both sets need to come from the same distribution.

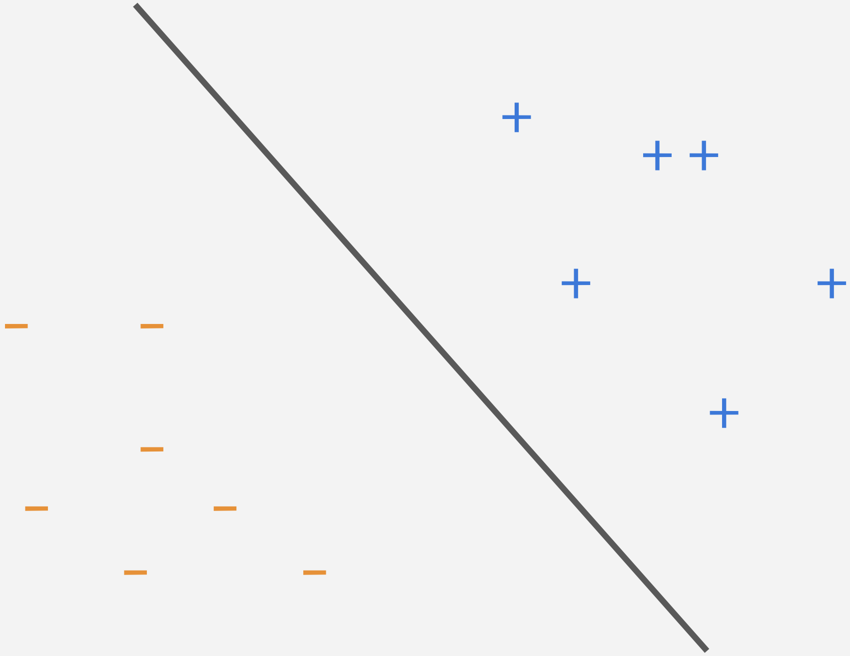
Data Bias



Torralba, Antonio, and Alexei A. Efros. "Unbiased look at dataset bias." CVPR, 2011.

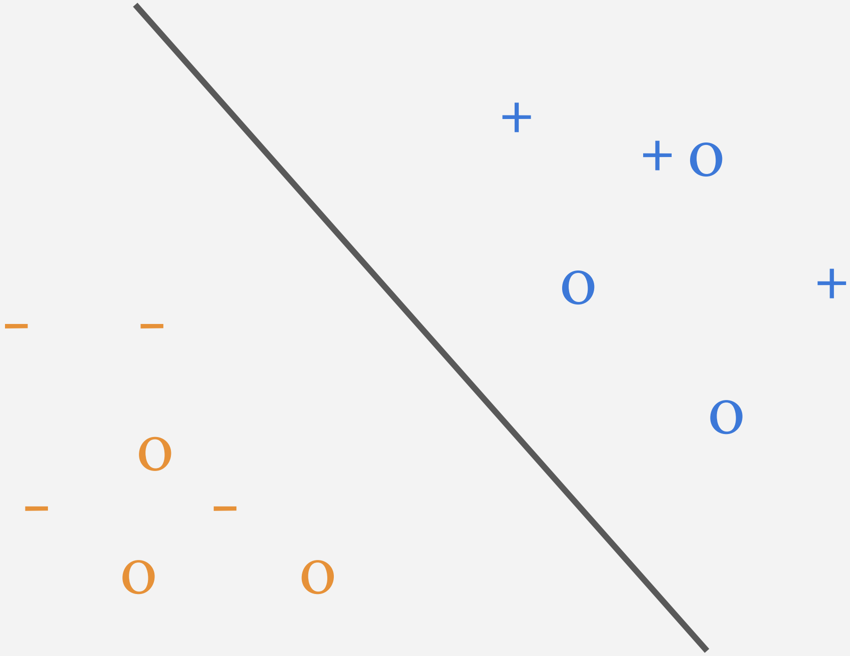
Different Types of Supervision

Fully Supervised



Different Types of Supervision

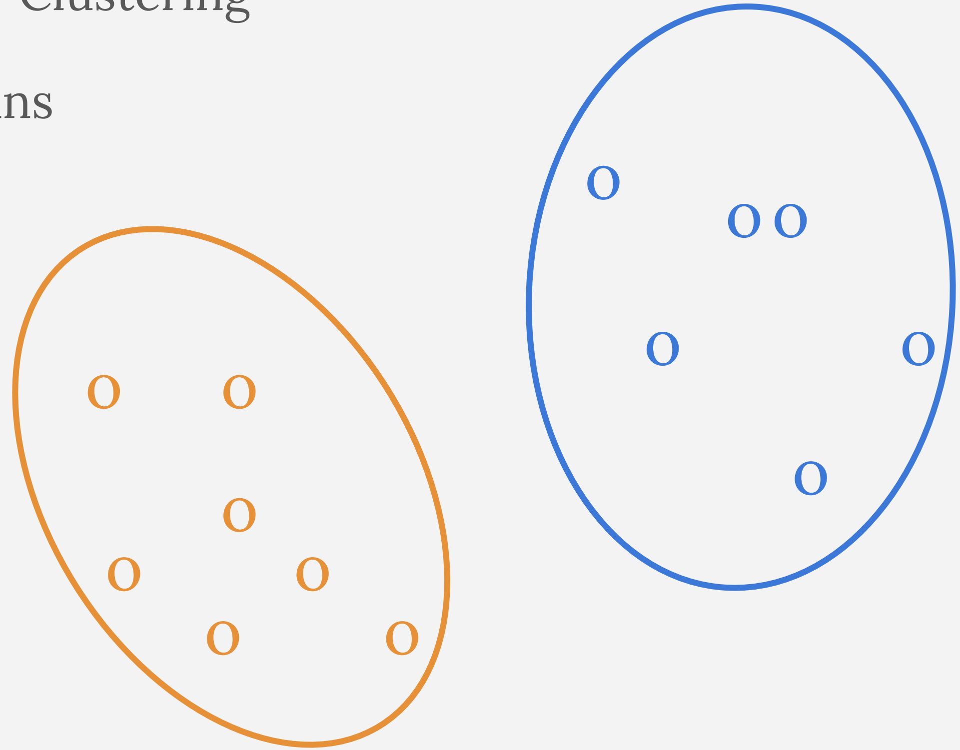
Semi-Supervised



Different Types of Supervision

Unsupervised / Clustering

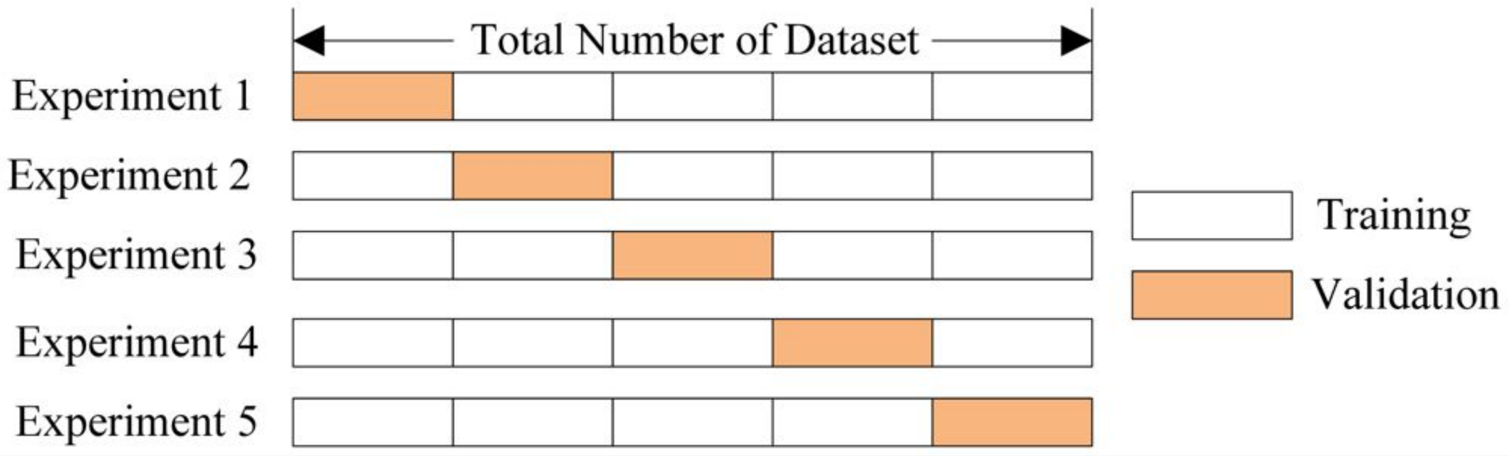
E.g., K-means



Evaluation of A Model

Cross-Validation:

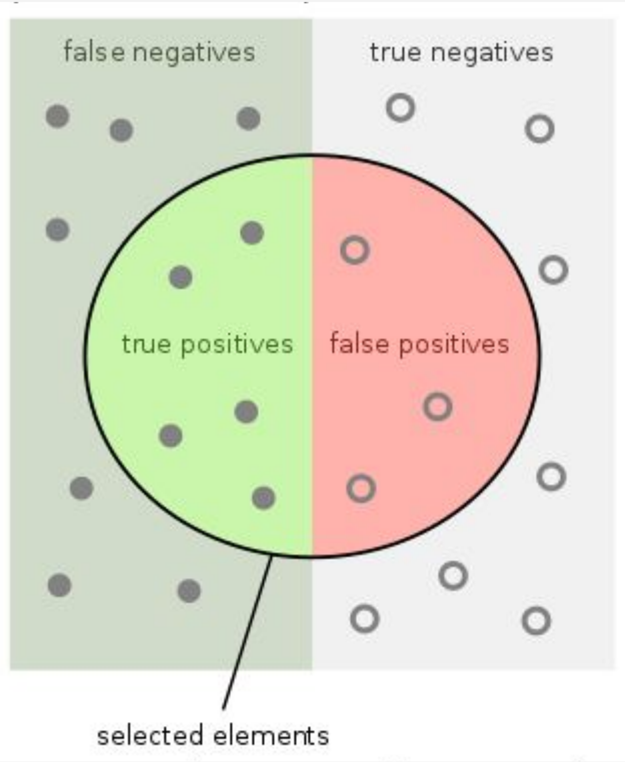
Keep a hold-out set from the collected data to simulate the model's performance on unseen data.



Performance Metrics - Classification

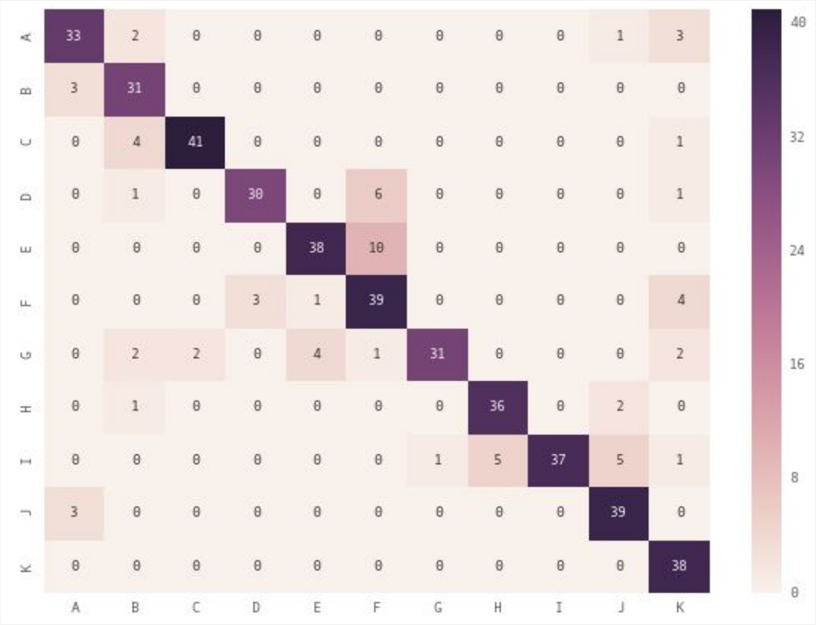
Precision = TP / (TP+FP)

Recall = TP / (TP+FN)

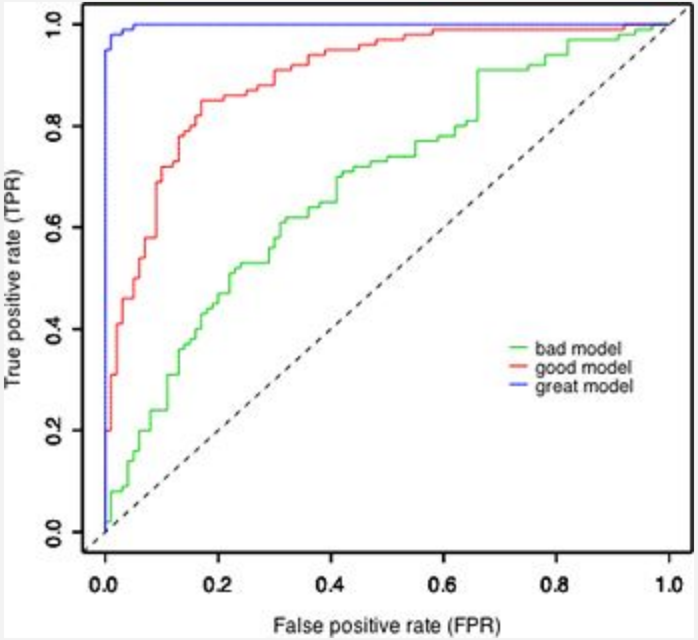


Performance Metrics - Classification

Confusion Matrix



ROC Curve



Summary

- ❑ Overview
- ❑ Formulation of Learning
- ❑ Learning Models
- ❑ Loss Function
- ❑ Optimization
- ❑ Data and Evaluation

Further Readings:

A Course in Machine Learning by Hal Daume III link
Introduction to Machine Learning by Alex Smola et al link
Pattern Classification by Richard O. Duda et al link
Pattern Recognition and Machine Learning by Christopher Bishop link