CORNELL TECH

Deep Learning Clinic (DLC)

Lecture 3 - A Brief Introduction to Machine Learning

Jin Sun

10/12/2018

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, if for no other reason than it’s hard to call a system intelligent if it *cannot* learn.” -- [CIML] Book

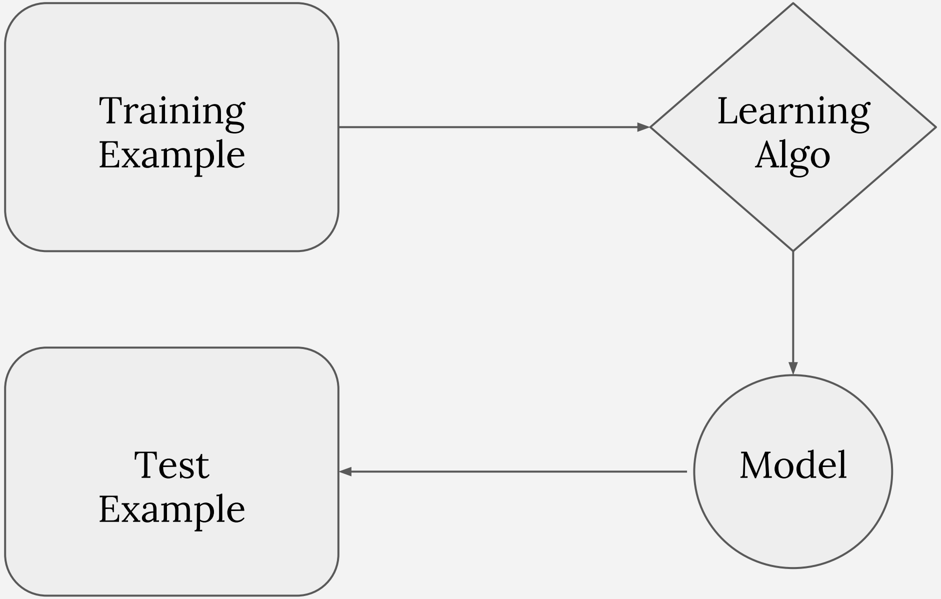
What is Machine Learning (ML)?

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

Two core questions:

How to learn? How good is the learning?

Machine Learning Paradigm



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)

Figure 2.4: A typical design process for a machine learning application.

* 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., 2D vectors of ad size and ad brightness)

y : our ground truth labels (e.g., whether the ad is clicked or not)

$f(\cdot)$: 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
(ad size, ad brightness)

Label
ad is clicked or not

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

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

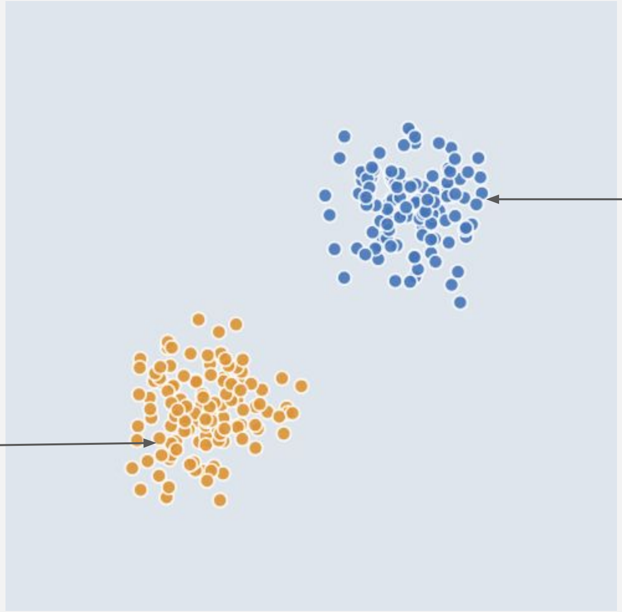
Learning Model

Loss function
How good is our ad click predictor?

A Concrete Example - Binary Classification

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

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



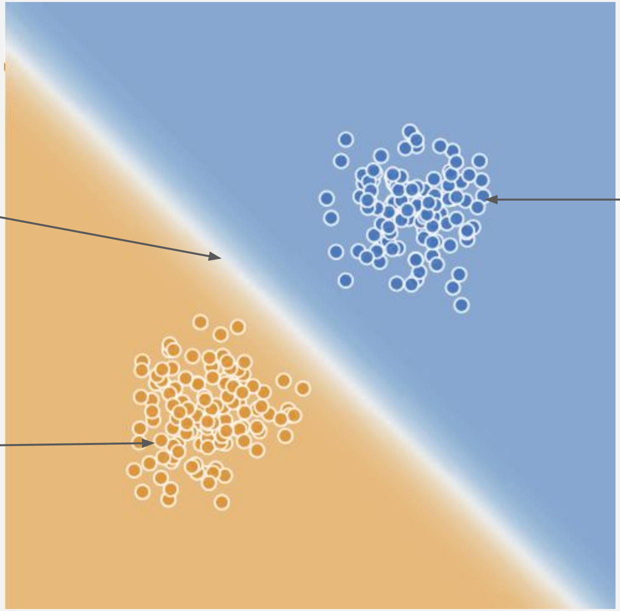
http://playground.tensorflow.org/

Today

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

A Concrete Example - Binary Classification

$$
\begin{aligned}
e &\doteq \mathop{\mathbb{E}}_{(x,y)\ \sim D}[L(y, f(x))] \\
&\doteq \frac{1}{N}\sum_{n=1}^{N} L(y_n, f(x_n))
\end{aligned}
$$



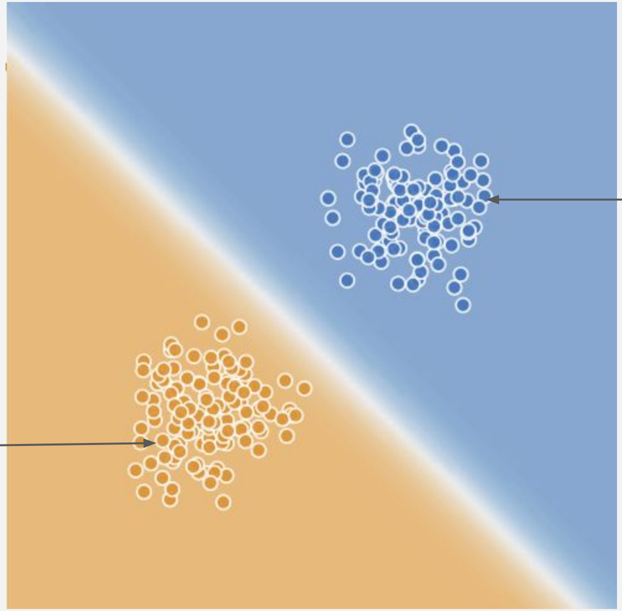
Positive Samples

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

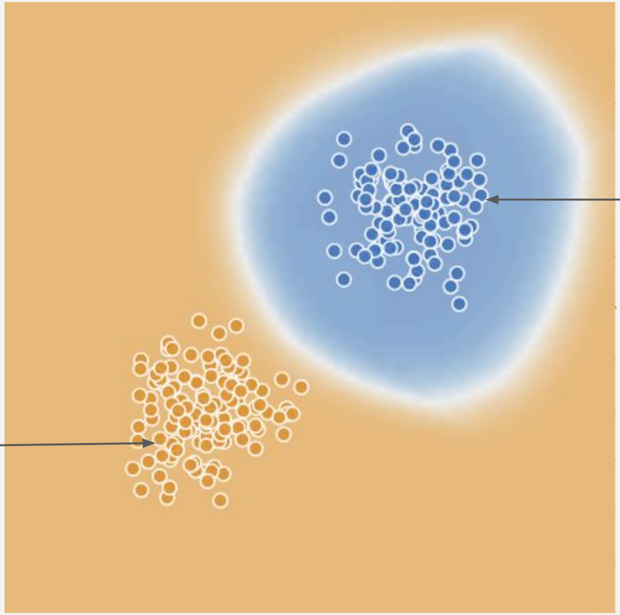
Negative Samples

Linear Function

Choose Your Model

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

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

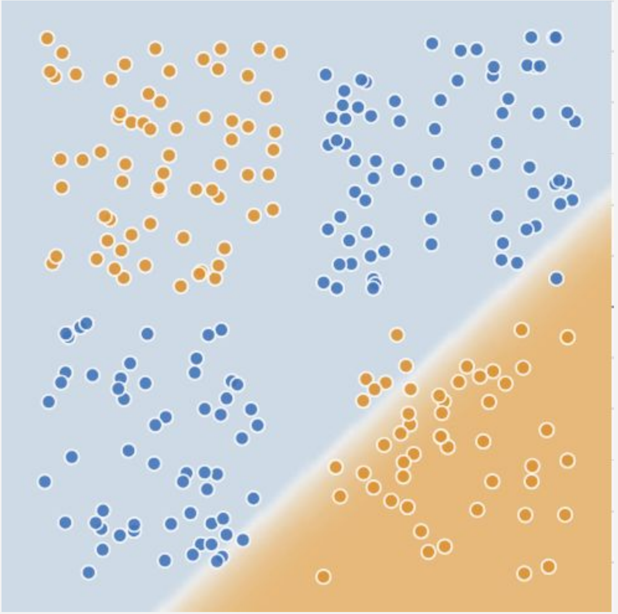


Positive Samples

Negative Samples

Non-linear Function

Pick a Model That Fits the Data Complexity

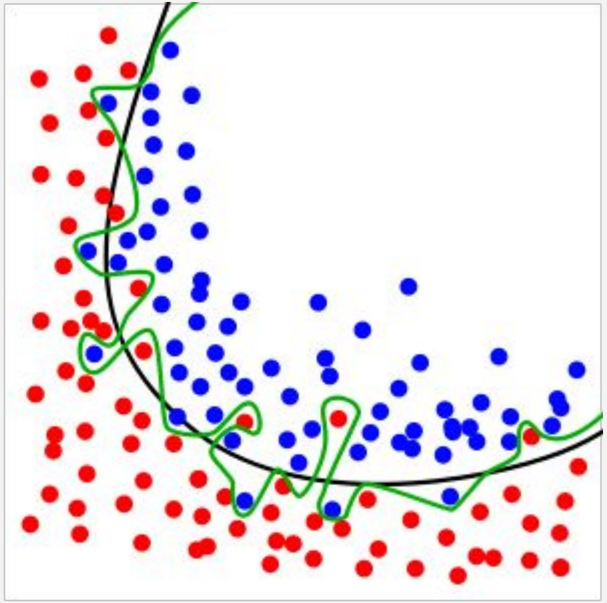


Linear Function
Not Suitable

Generalization

So why not always pick the most complex model?

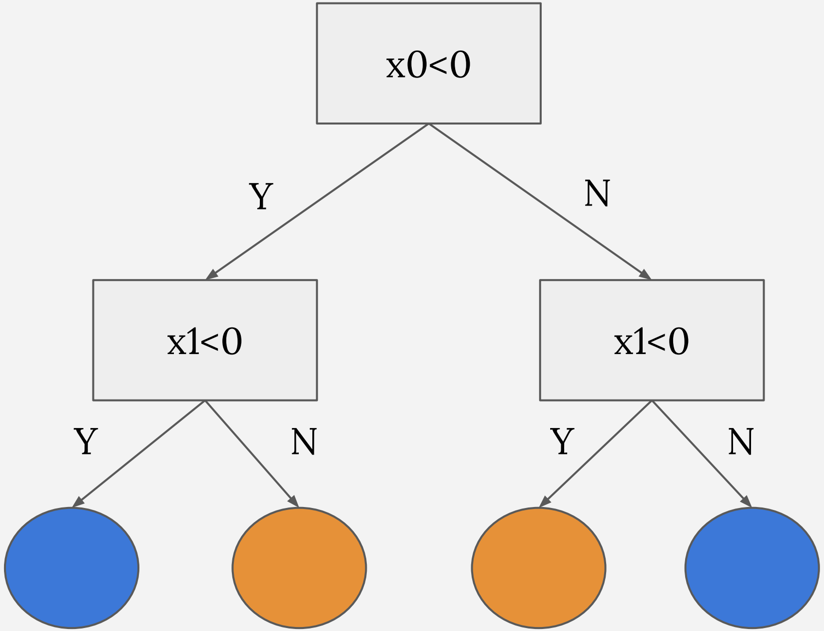
We care about our model's performance on *unseen* test data: the *generalization* ability.

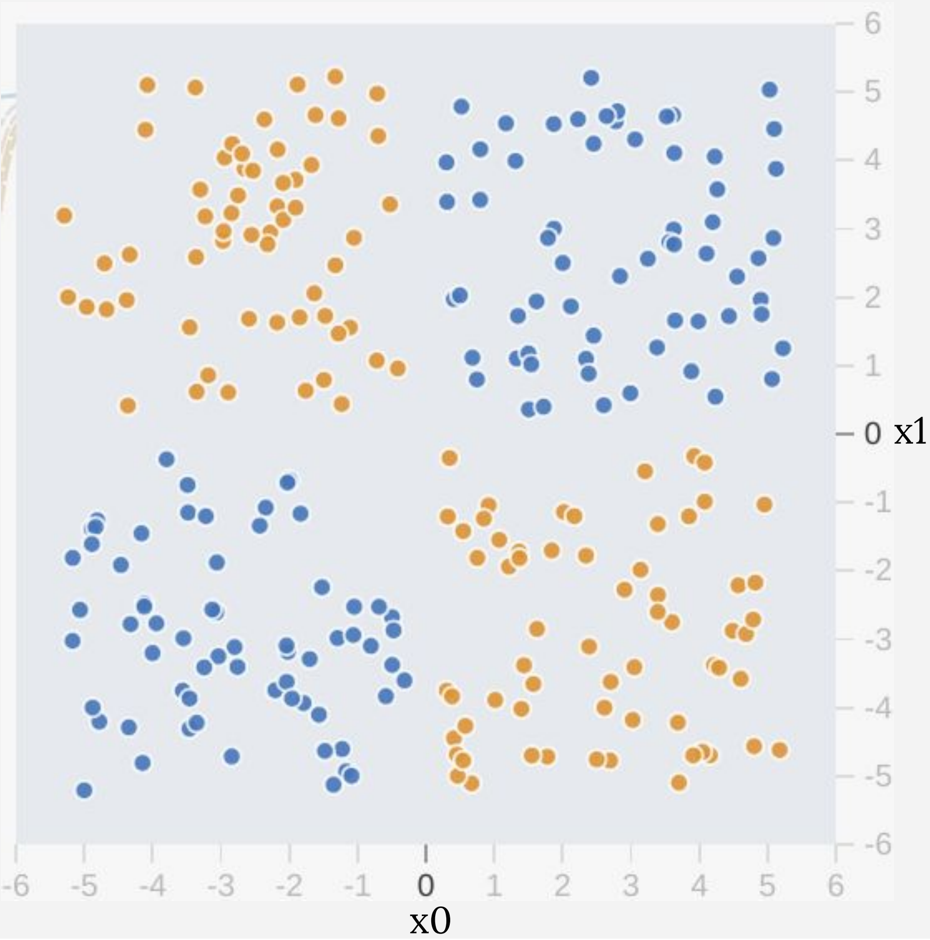


If our model is over-complex, it can be *overfitted* to training and perform poorly on testing data.

Models

Decision Trees



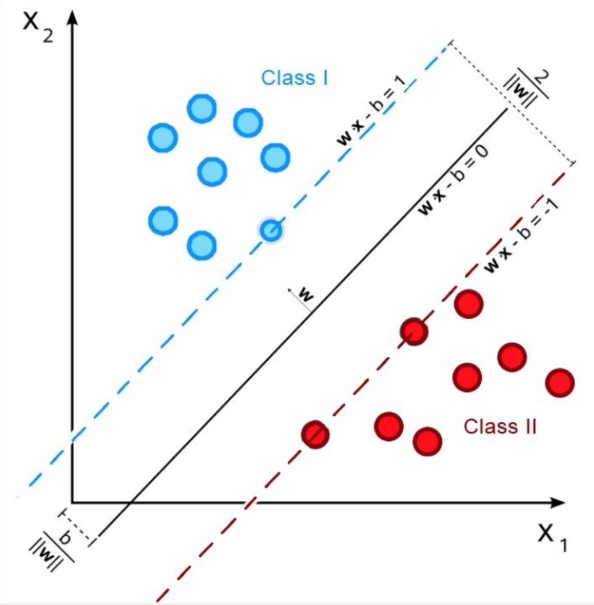


Models

Linear Function

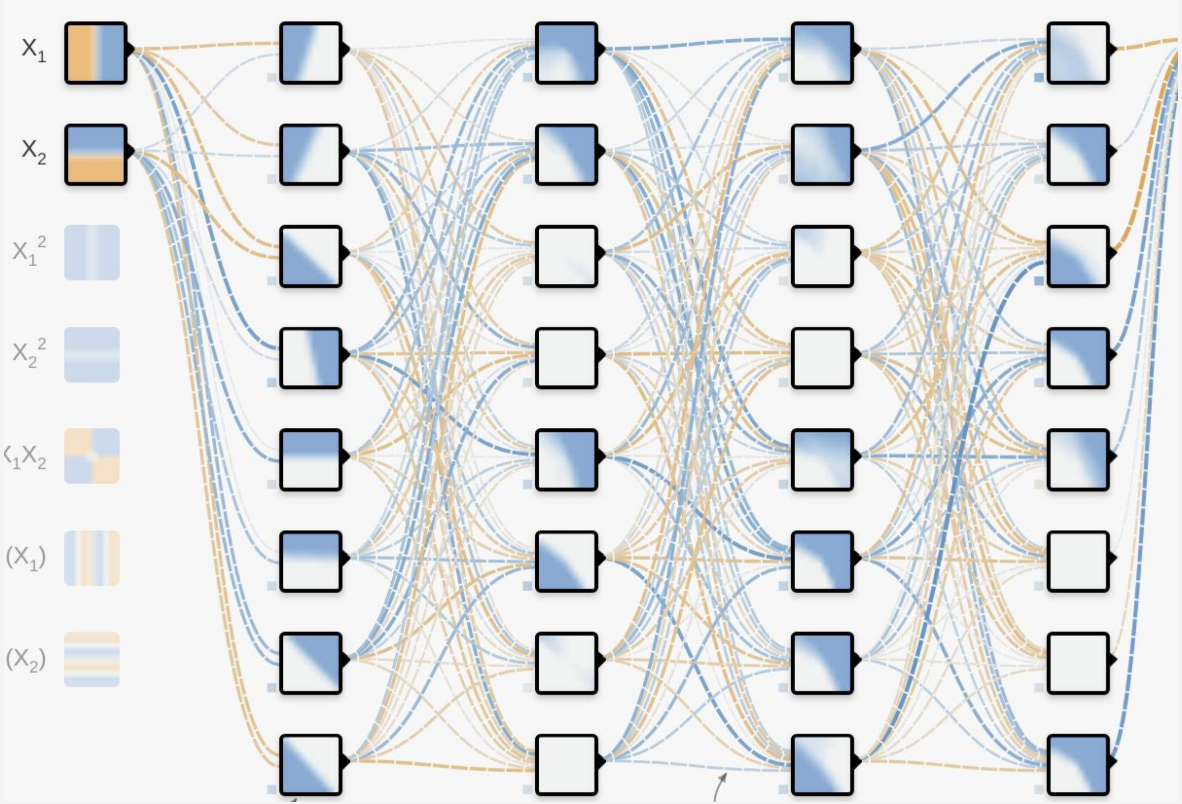
$$f(x) = Wx - b$$

Support Vector Machine (SVM)



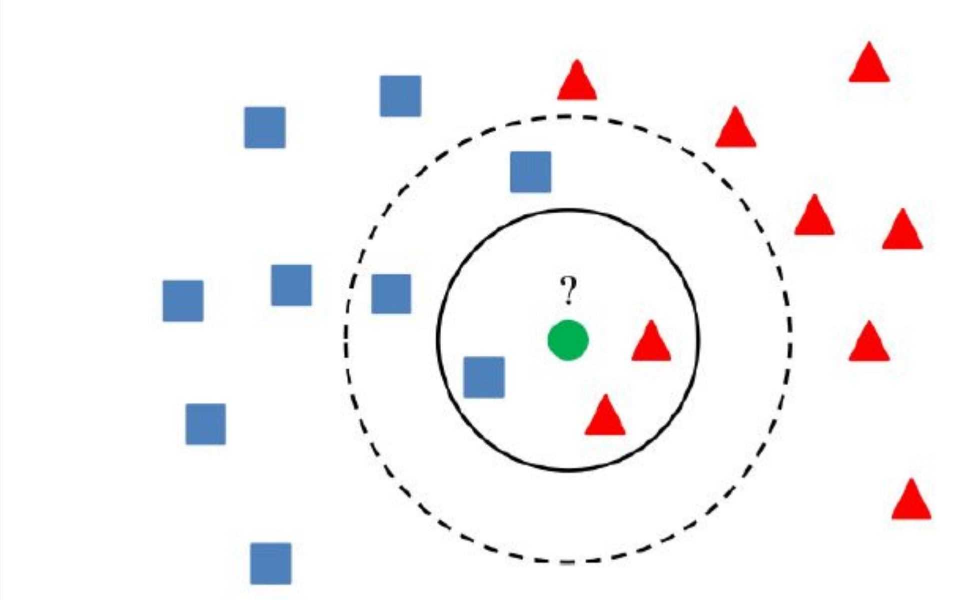
Models

Neural Networks



Non-Parametric Models

Nearest Neighbor



Today

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

Loss Function

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

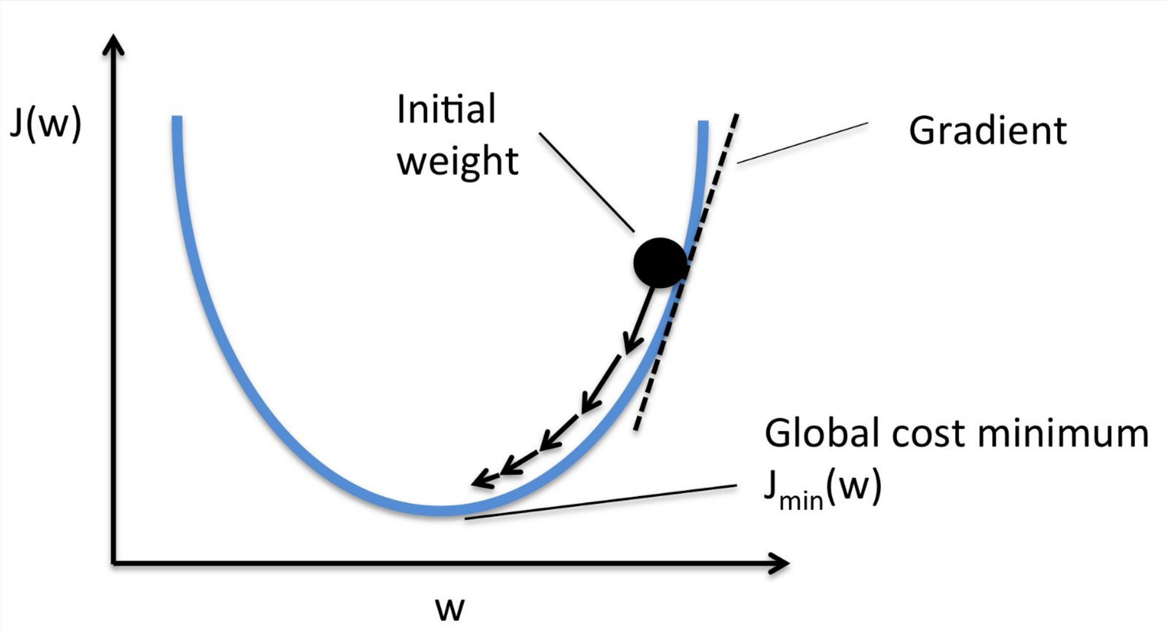
- Overview
- Formulation of Learning
- Learning Models
- Loss Function
- **Optimization**
- Data and Evaluation

Get Training Started - Optimization

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

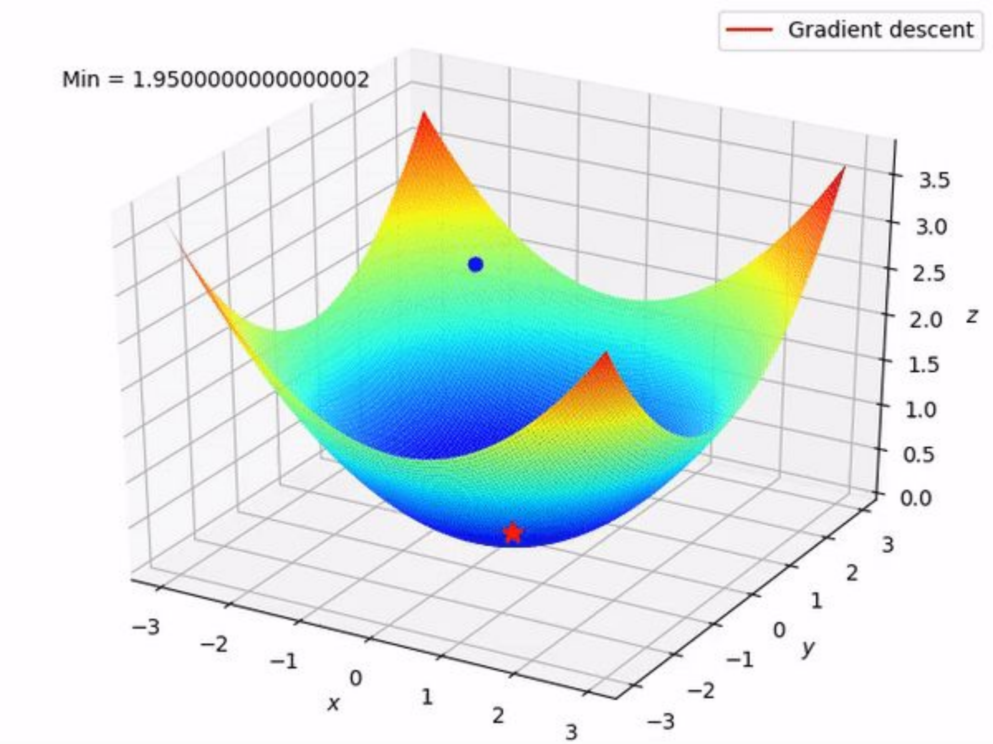
Find the best θ that minimizing the expected loss.

Gradient Descent



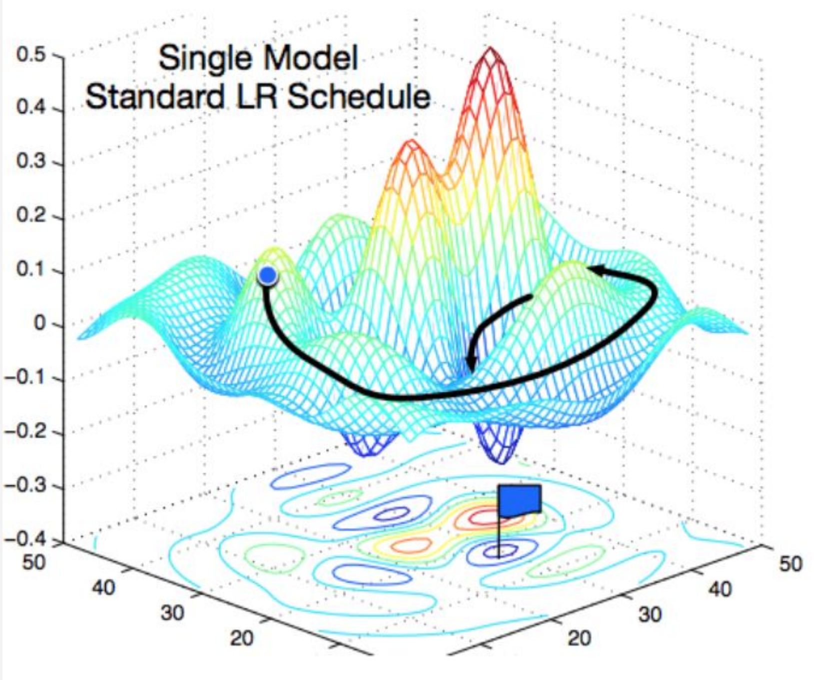
1D Loss Function

Gradient Descent



2D Loss Surface

Gradient Descent



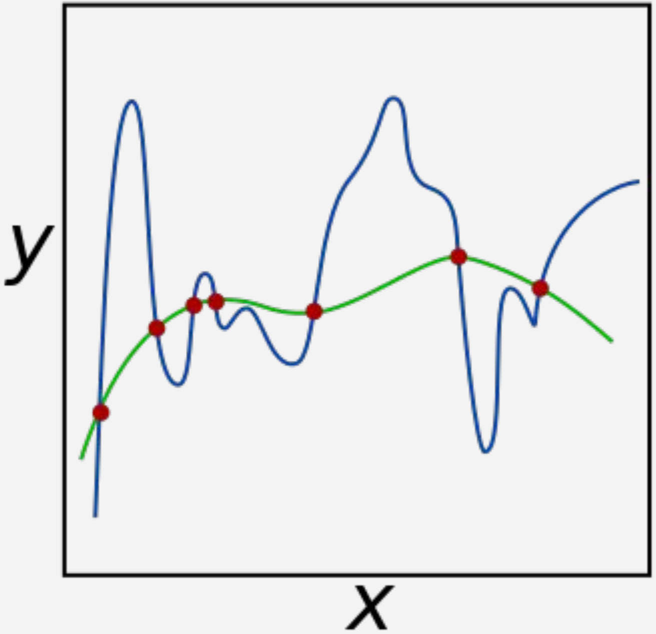
Non-Convex Loss Surface

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

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

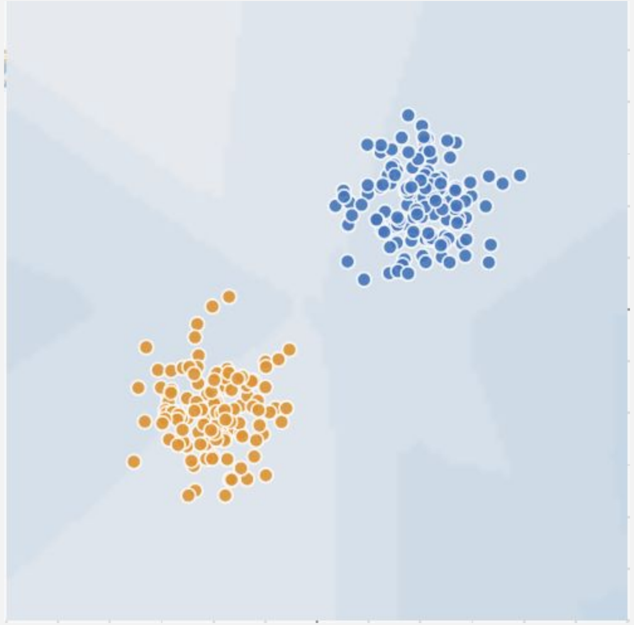


E.g., L1, L2 norm

Today

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

Data

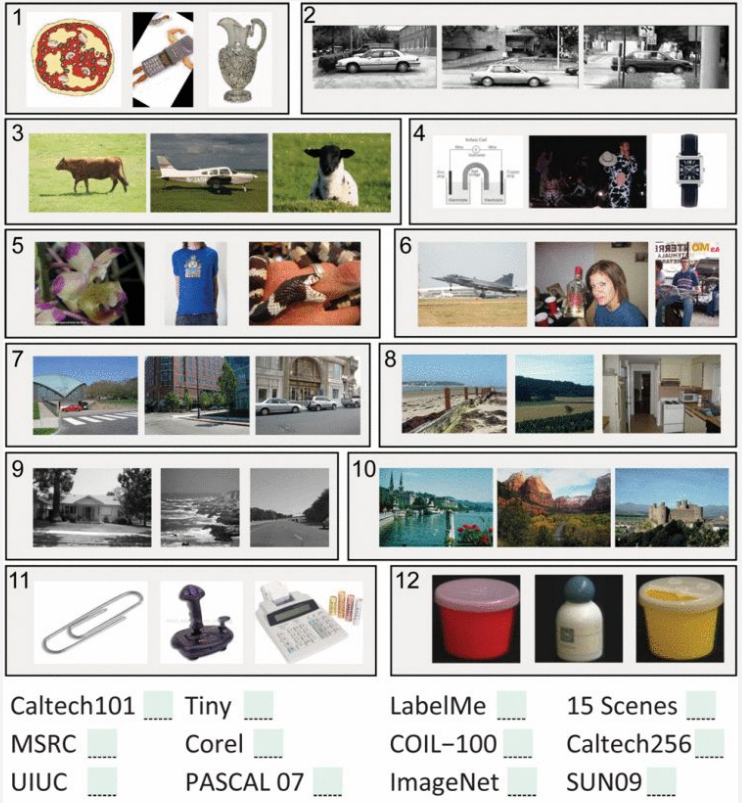


Training Set

Testing Set

Both sets need to come from the same distribution.

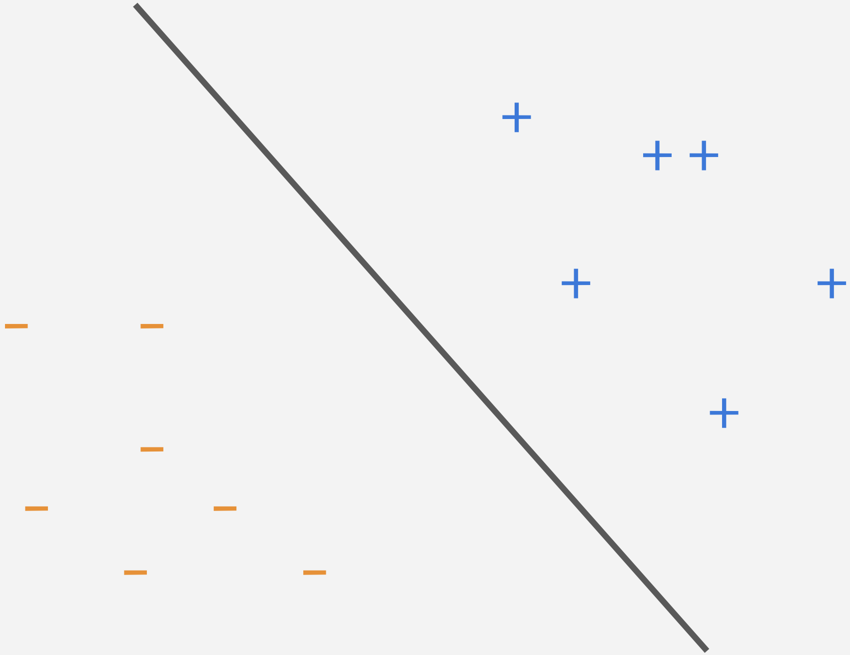
Data Bias



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

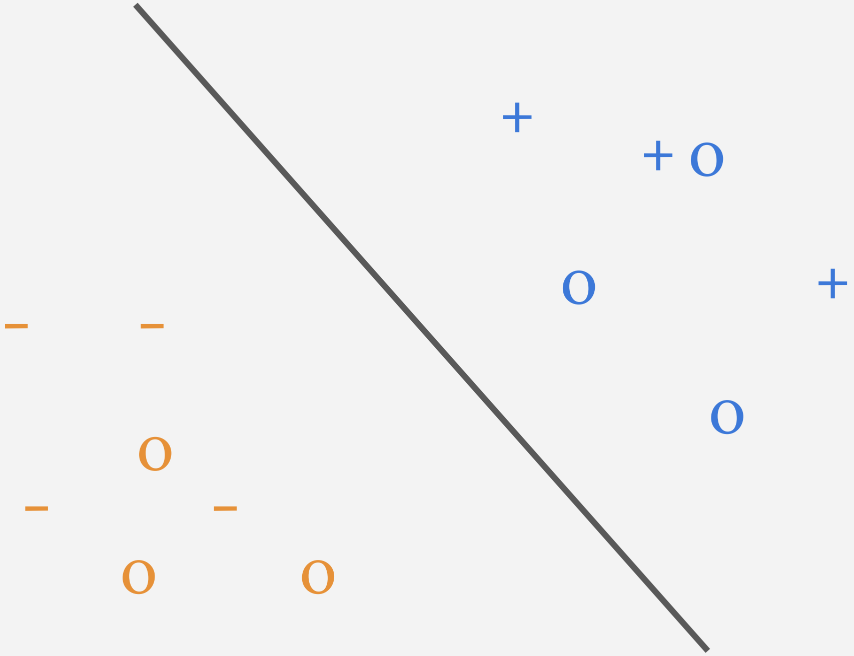
Different Types of Supervision

Fully Supervised



Different Types of Supervision

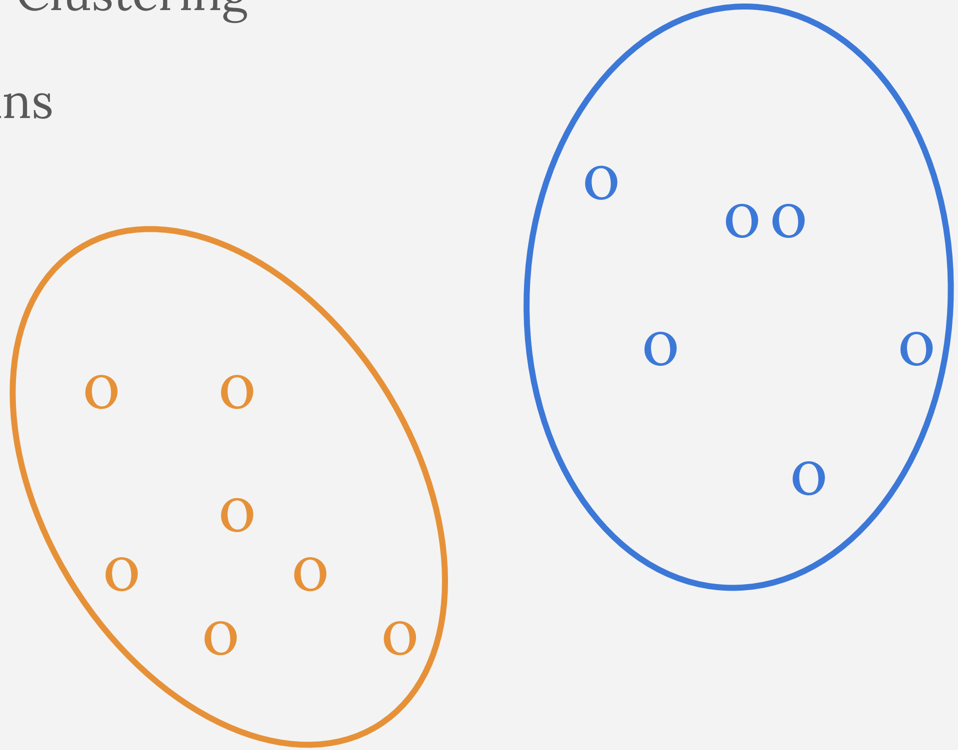
Semi-Supervised



Different Types of Supervision

Unsupervised / Clustering

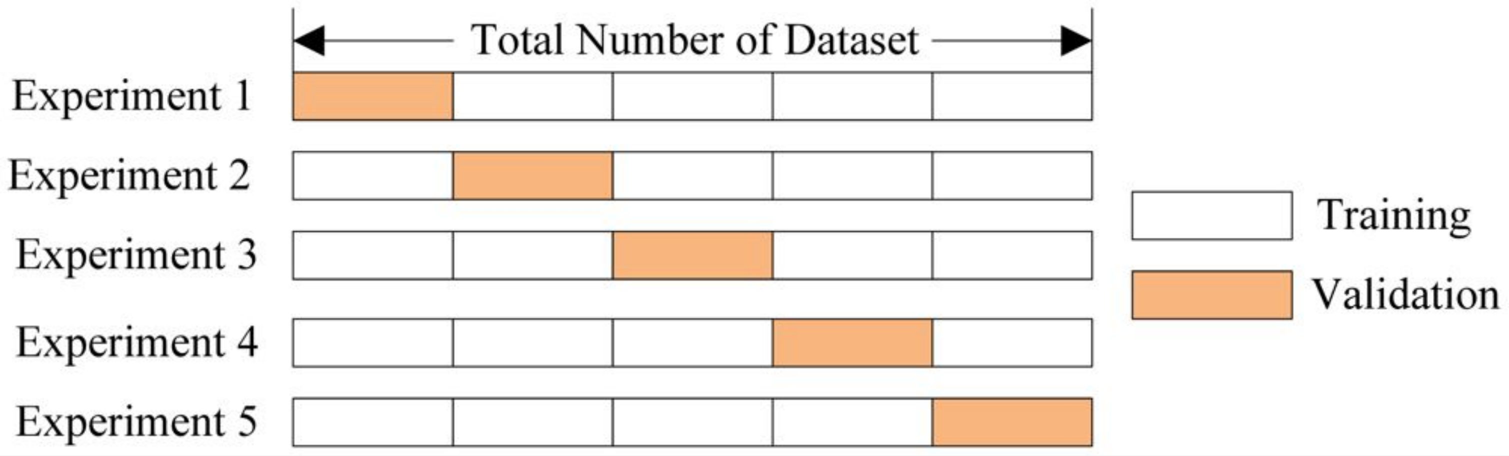
E.g., K-means



Evaluation of A Model

Cross-Validation:

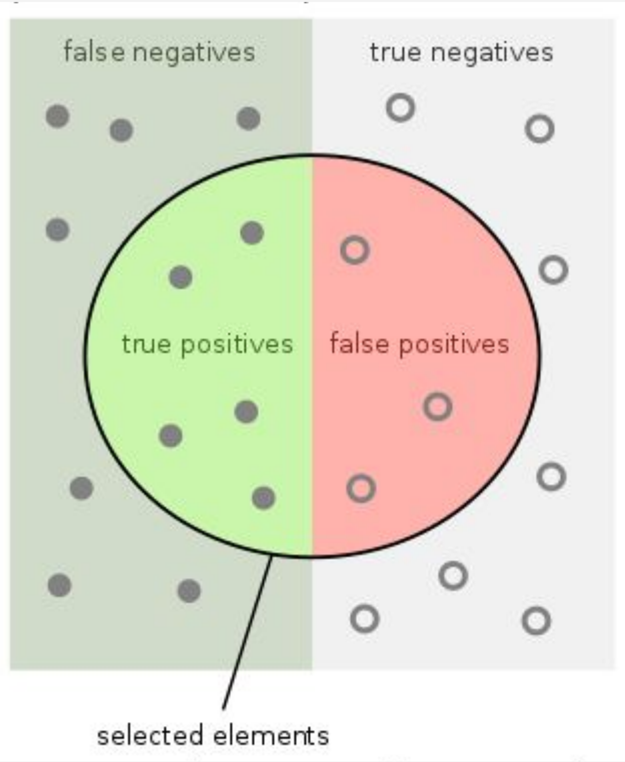
Keep a hold-out set from the collected data to simulate the model's performance on unseen data.



Performance Metrics - Classification

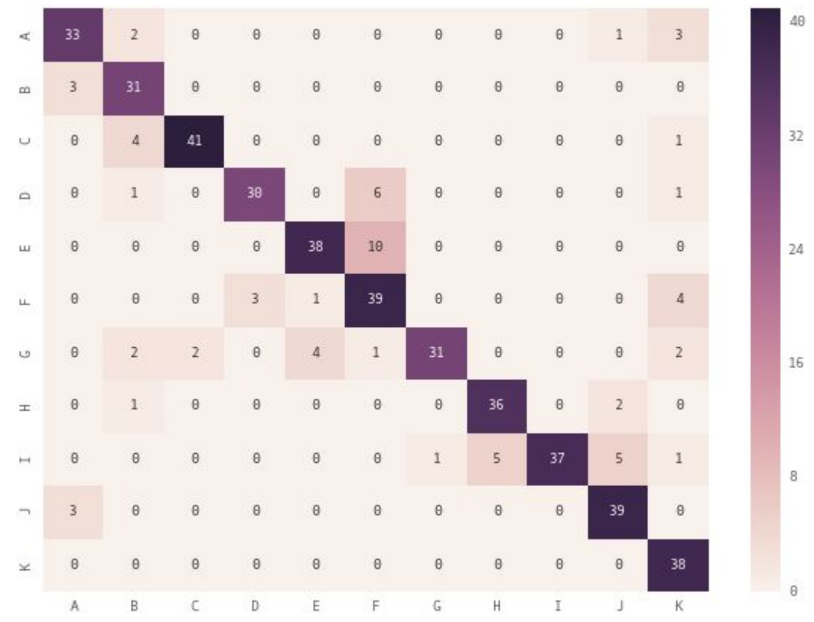
Precision = TP / (TP+FP)

Recall = TP / (TP+FN)

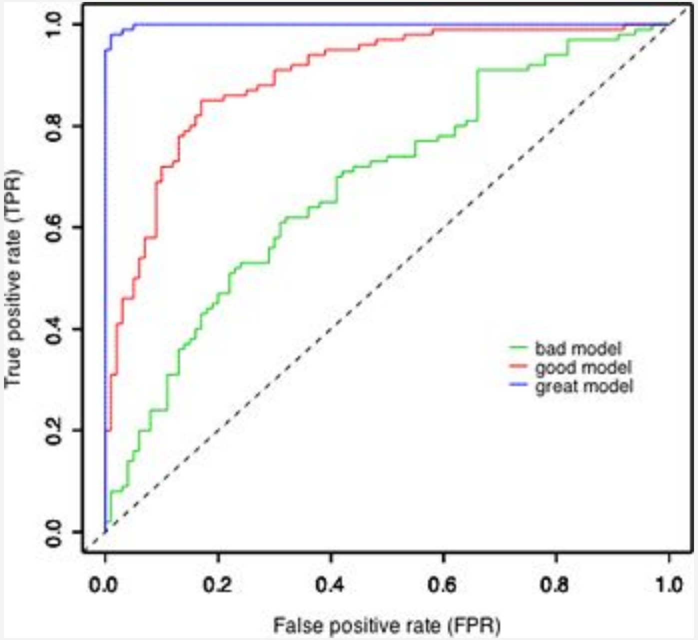


Performance Metrics - Classification

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