Modeling Data
Introduction to Machine Learning

Task <-> Methodology

- Explain/Describe data
  - Descriptive Statistics. E.g., what percentage of people are late/on time?
- Use observed data to infer information about a population
  - Inferential Statistics. E.g., what's the support for this candidate?
- Draw a causal connection, explain
  - SCMs, quasi-experiments, experiments, human subjects, etc
- **Predict** characteristics of **out-of-sample data**
  - Decision Theory, Machine Learning. E.g., prediction, forecasting, classification…

High-Level Intuition

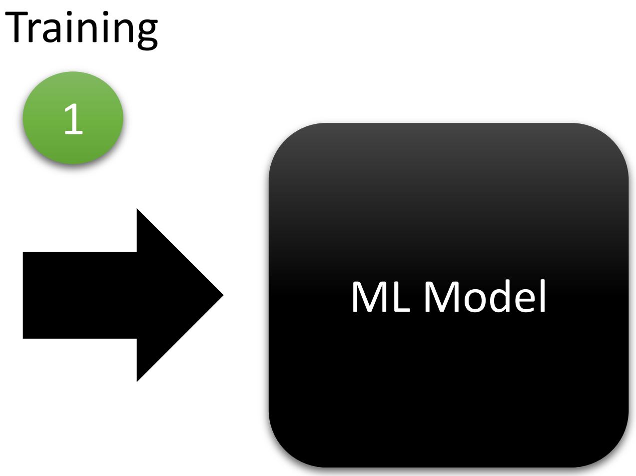
High-Level Intuition

Fox

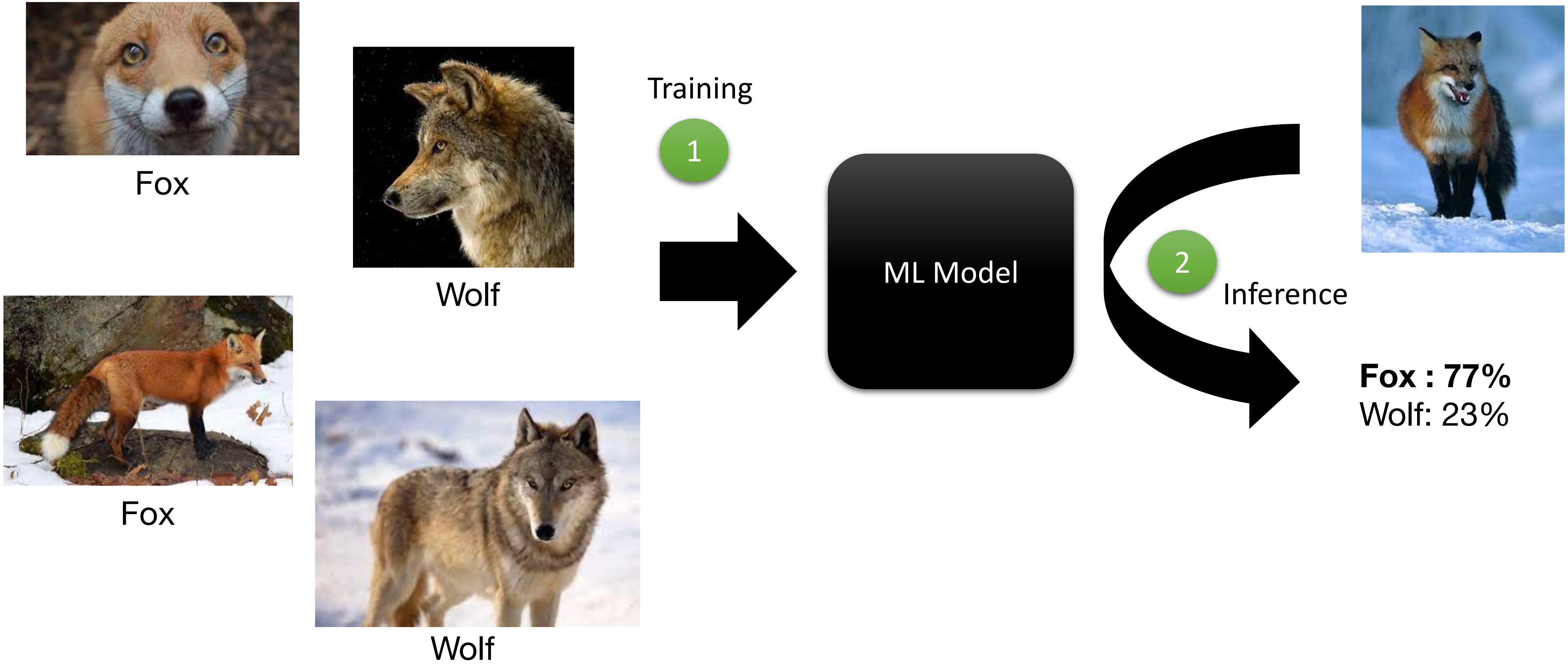
Wolf

Fox

Wolf



High-Level Intuition

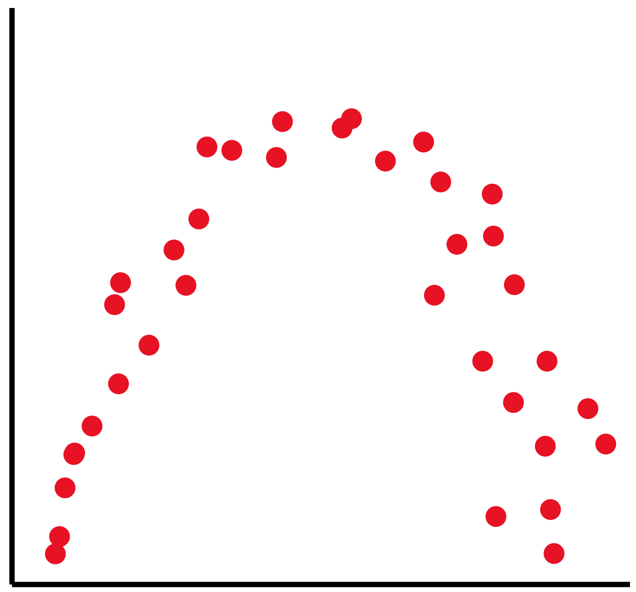


Outline

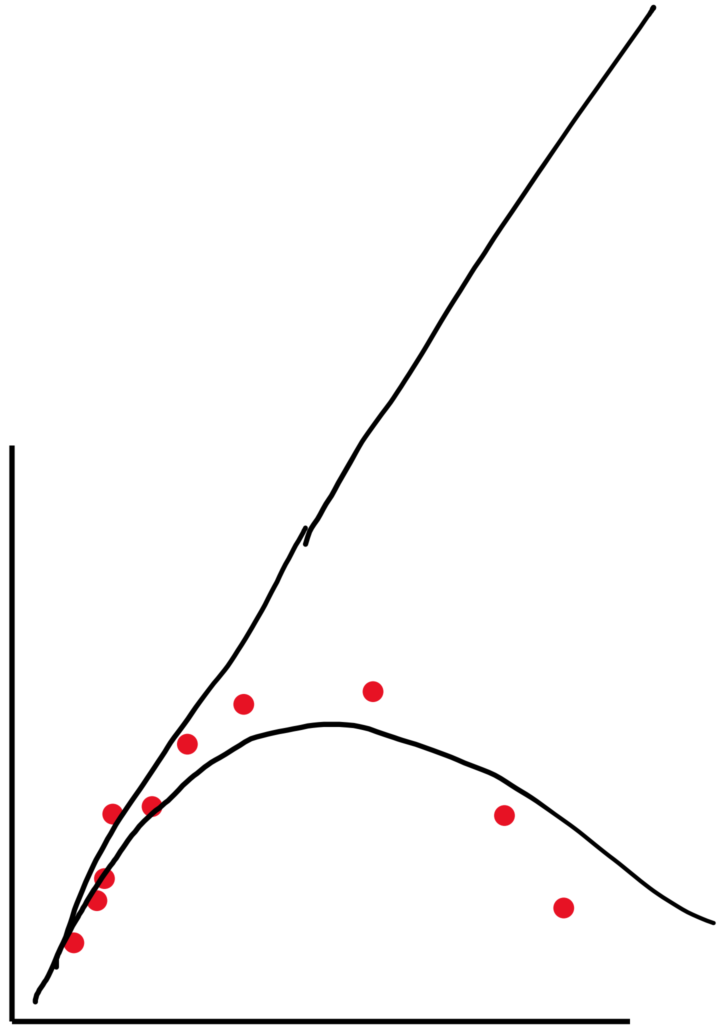
- Intro to Modeling Data
- ML 101
- Regression. Linear Regression
- Classification

Why we build models…

- To understand data
  - We often wish to understand the data generating process
- To make predictions about *out-of-sample* data



Ground truth



All Models are Wrong

- "All models are wrong, but some are useful"
  - George Box
- "Modelling in science remains, partly at least, an art. Some principles do exist, however, to guide the modeler. The first is that *all models are wrong*; some, though, are better than others and we can search for the better ones. At the same time we must recognize that eternal truth is not within our grasp"
  - *McCullagh, P.; Nelder, J. A. (1983), Generalized Linear Models, Chapman & Hall, §1.1.4*.

Let’s Build a Model To Understand Data

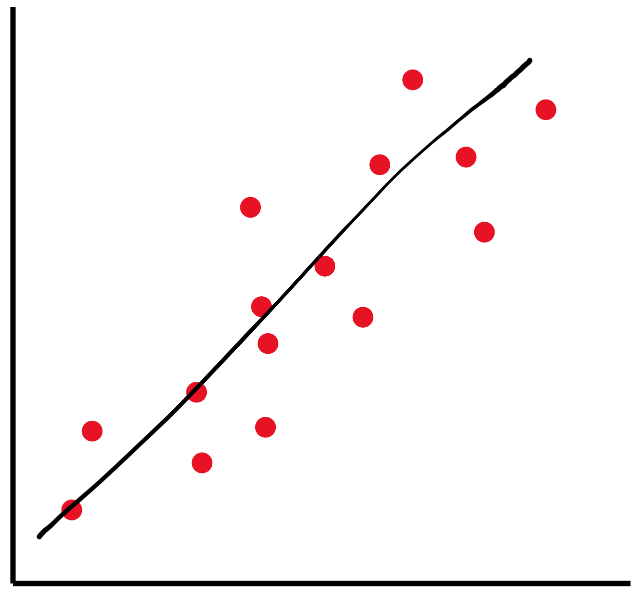
- Running example: a regression problem
- Example:

Name	Age	Department	Gender	Title	Salary
Jack	55	CS	M	Professor	??
Jane	27	Stats	F	Assistant Professor	??

Given these input vectors…

…predict this input variable

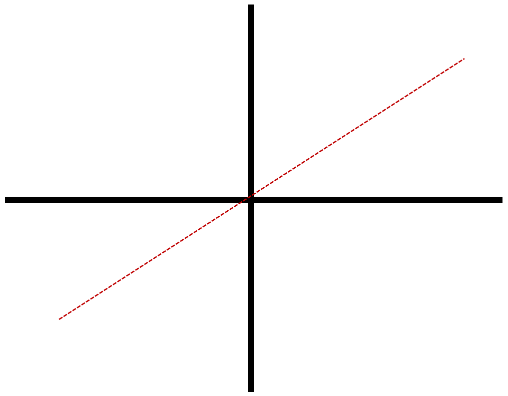
Building Intuition: Fitting a Line



Given an input vector x, predict y

- We need to choose a model to do that

$$\hat{y} = 0.3x$$



Parameters / weights

Output value / Explanatory

Input vector / predictor

$$\hat{y} = w^T x$$

$$x \in \mathbb{R}^n$$

$$y \in \mathbb{R}$$

$$w \in \mathbb{R}^n$$

Let's Build a Model To Understand Data

- Running example: a regression problem
- Example:

Name	Age	Department	Gender	Title	**Salary**
Jack	55	CS	M	Professor	??
Jane	27	Stats	F	Assistant Professor	??

x_1, x_2, x_3, x_4, x_5 $\hat{y}$

Variables/Attributes/Columns become 'features' of the input vector

Linear Regression Model

- ‘Linear’ because of the relationship between x and y

$$\hat{y} = w^T x + b$$

Linear Regression Model

- ‘Linear’ because of the relationship between x and y

- A model is an assumption…
  - …of what function represents data *well*

$$\hat{y} = w^T x + b$$

- Once we’ve fixed a model:
  - We <u>find</u> the parameters/weights *w* that make the model <u>perform well</u>

Linear Regression Model

- 'Linear' because of the relationship between x and y
- A model is an assumption…
  - …of what function represents data *well*

$$\hat{y} = w^T x + b$$

- Once we've fixed a model:
  - We find the parameters/weights *w* that make the model perform well

We need a method to find those parameters

This suggests we need a performance metric

The data we have…

- A dataset becomes a matrix
  - Each row is an input vector

Name	Age	Department	Gender	Title	Salary
Jack	55	CS	M	Professor	33000
Jill	23	Econ	F	Professor	32000
Josh	32	Bio	M	Staff	28000
Jenn	44	Bio	F	Associate Professor	24000
Jane	27	Stats	F	Assistant Professor	25000

Input vectors

- A dataset becomes a matrix
  - Each row is an input vector

Dataset contains the target variable / label

	Name	Age	Department	Gender	Title	Salary
Training dataset	Jack	55	CS	M	Professor	33000
	Jill	23	Econ	F	Professor	32000
	Josh	32	Bio	M	Staff	28000
Test dataset	Jenn	44	Bio	F	Associate Professor	24000
	Jane	27	Stats	F	Assistant Professor	25000

Performance Metric

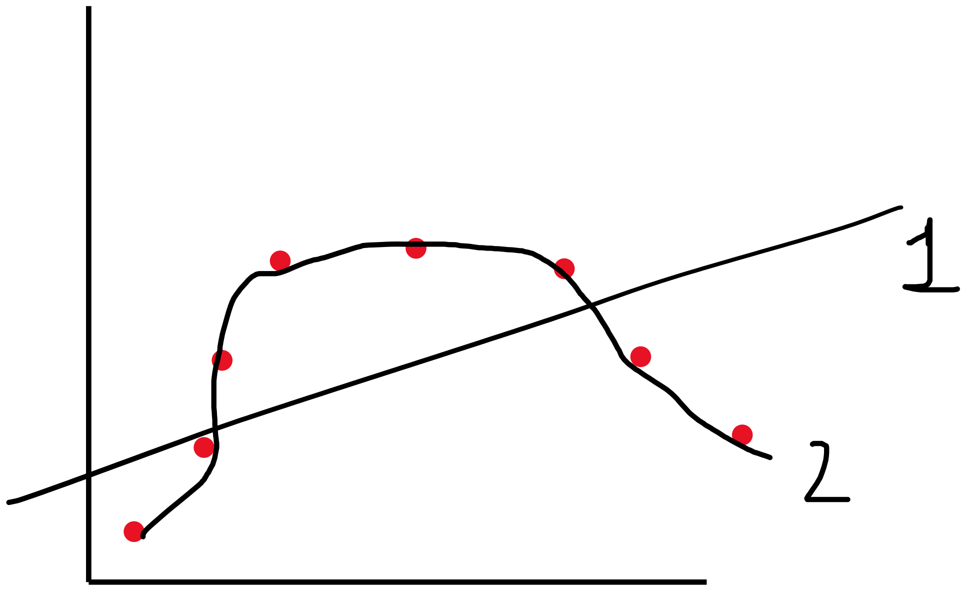
- Mean Squared Error (MSE)
  - Error decreases to 0 when *predicted y = ground-truth y*

$$\mathrm{MSE}_{\mathrm{test}} = \frac{1}{m} \sum_i (\hat{\boldsymbol{y}}^{(\mathrm{test})} - \boldsymbol{y}^{(\mathrm{test})})_i^2.$$

m test examples

- Goal: We want the model to perform well on the test data which has never seen before
  - Out-of-sample data

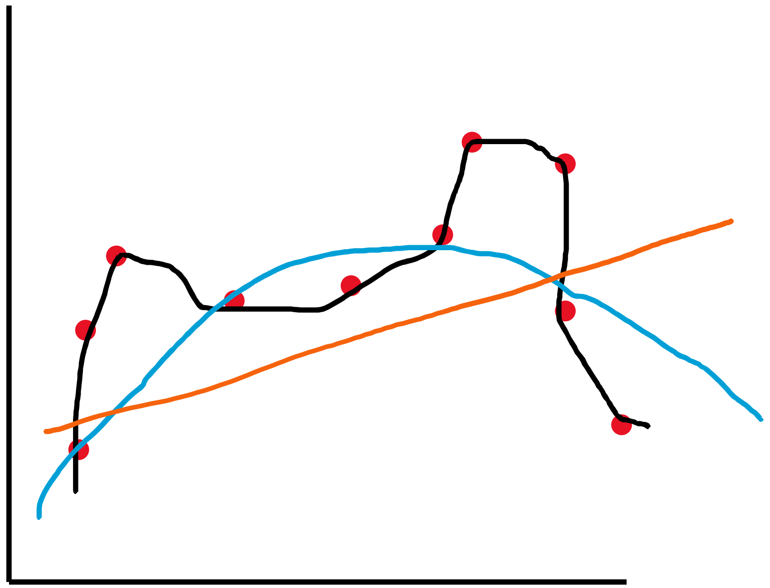
Building some intuition…



- “*With four parameters I can fit an elephant, and with five I can make him wiggle his trunk*.”
    - Von Neumann

Higher Capacity Models

- We can increase the capacity of the model by adding more parameters
  - This will help with obtaining a ‘better’ fit.



$$\hat{y} = w^T x$$

$$x \in \mathbb{R}^n$$

$$y \in \mathbb{R}$$

$$w \in \mathbb{R}^n$$

We have a goal, let's find *w*

- We want to find parameters *w* using the training dataset

We want to achieve
a low training error

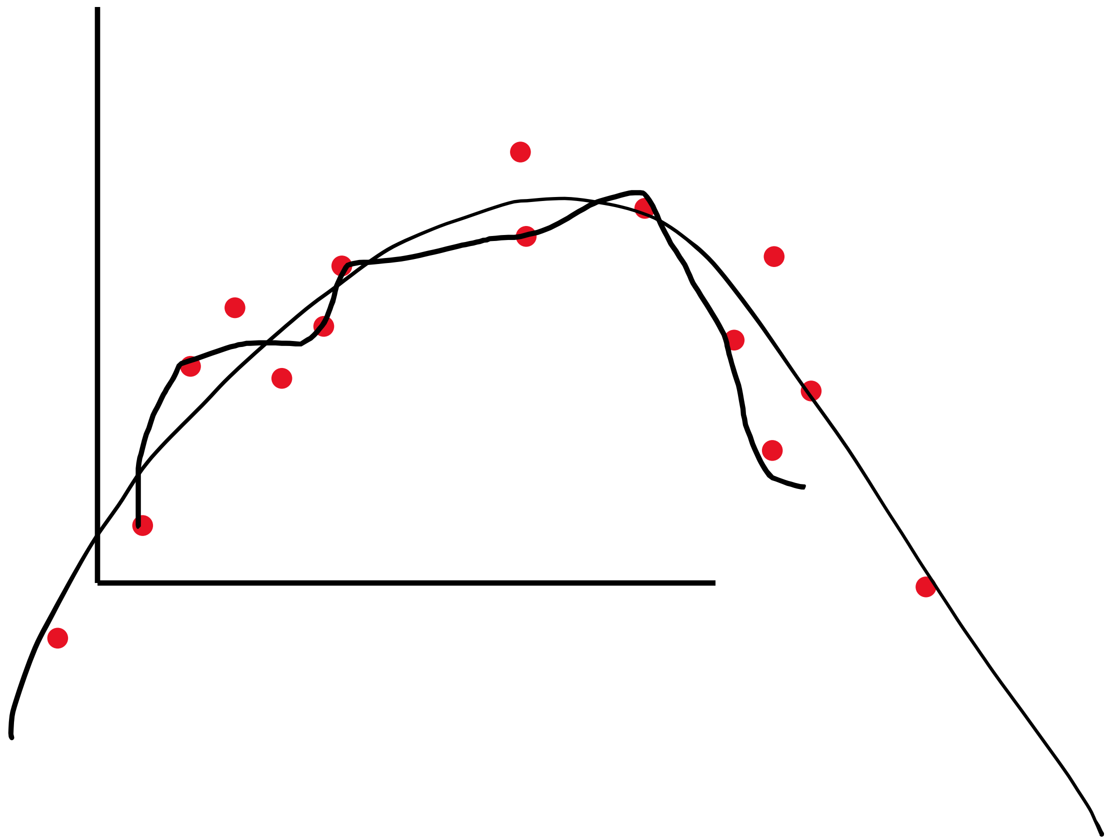
We have a goal, let's find *w*

- We want to find parameters *w* using the training dataset

$$\nabla_{\boldsymbol{w}} \mathrm{MSE}_{\mathrm{train}} = 0$$

- This is an optimization problem that we know how to solve well
  - We can find the minimum MSE
- Consider we run this optimization with the training data. What will happen when we run on test data?

Building some intuition…



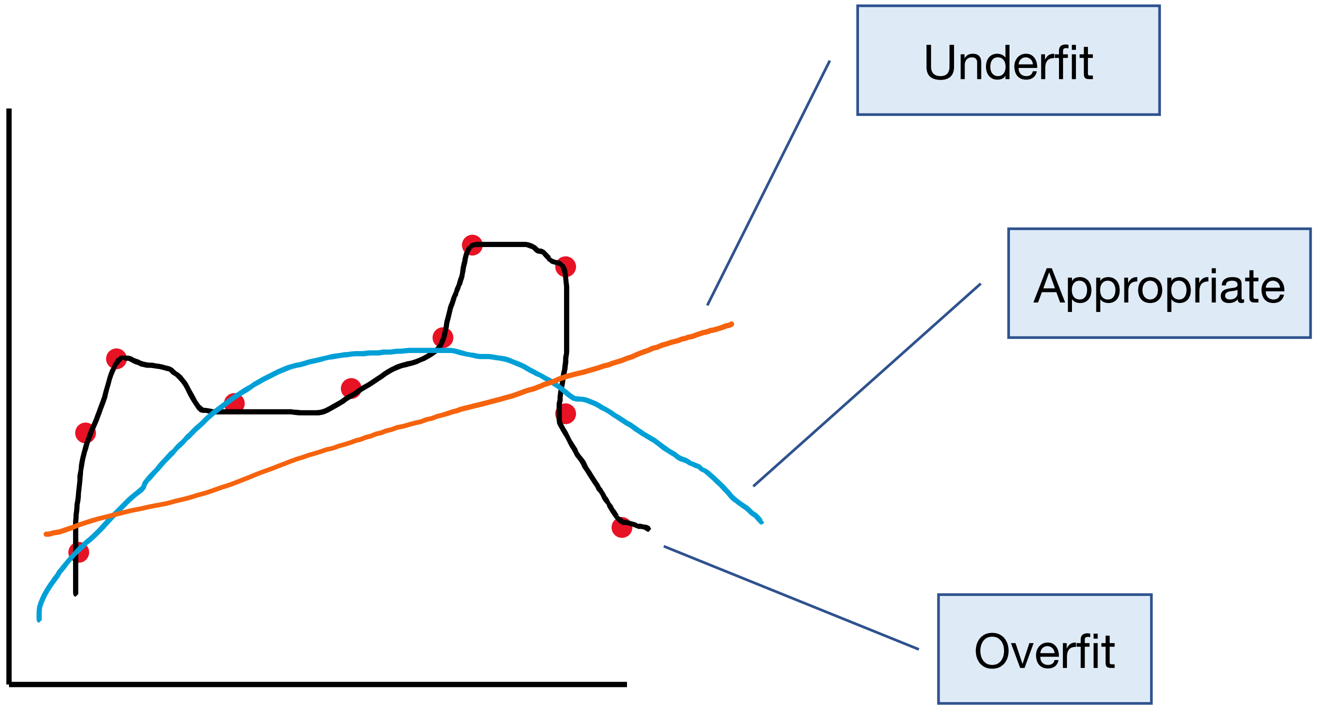
Central challenge of Machine Learning

- Learn parameters so the model performs well on unseen data
    - *Generalize* to unseen data
- As opposed to the optimization problem of doing well on training data
- Remember why we build models:
    - To understand the process that generated the data
    - To make predictions about out-of-sample data
- Do you think minimizing the MSE on the training data helps us achieve any of those two goals?

Underfitting, Overfitting

- Underfitting
  - When a model cannot reduce the *training error*
- Overfitting
  - A model achieves low *training error* but high *test error*
- Ideally, we want low training error and small gap between training and test error
  - That’s a model that explains the data generating process
  - That’s a model that helps us predict out-of-sample data

Underfitting, Overfitting…



Outline

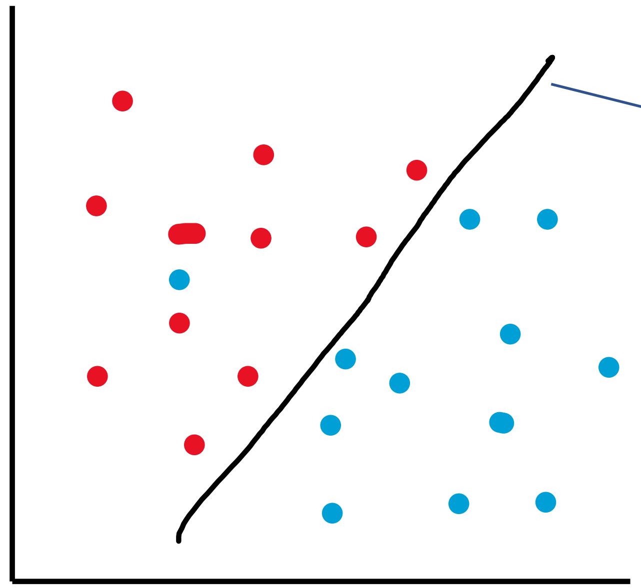
- Intro to Modeling Data
- Walk through Machine Learning
- Regression. Linear Regression
- **Classification**

Classification Problem

- Given an input vector, x, predict a class, c
  - Binary classification problems
    - Spam vs not spam
    - Give loan vs don't
    - Admit student vs don't
    - Will reoffend vs won't

Binary Classification

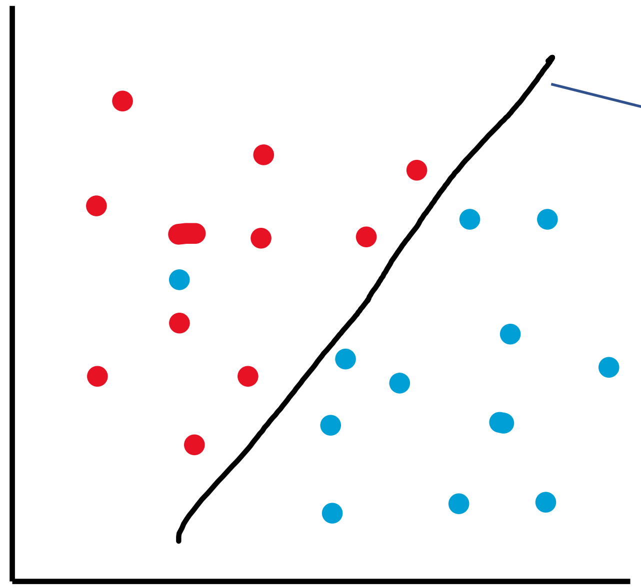
- Given an input vector, x, choose a label 0 or 1
- Intuitively:



Find a hyperplane that separates the space of positive and negative samples

Binary Classification

- Given an input vector, x, choose a label 0 or 1
- Intuitively:



Find a hyperplane that separates the space of positive and negative samples

- How do you evaluate this one?
  - False positives/negatives, accuracy, …

So, what's machine learning?

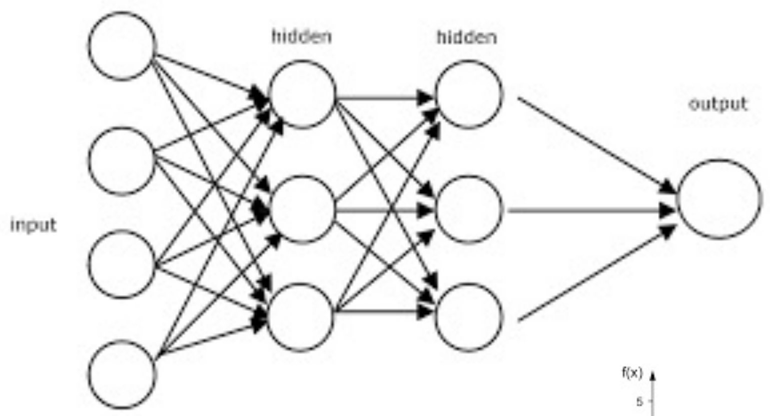
- A model
  - Linear regression, logistic regression, …
- Parameters
- A performance metric
  - MSE
- A training objective
  - Loss function
- A strategy to learn/fit the model parameters

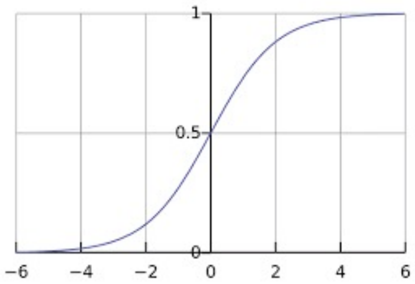
Machine learning models

- Regression models
- Decision trees
- Support vector machines
- Deep networks
- Many, many others:
  - PGM, genetic algorithms…

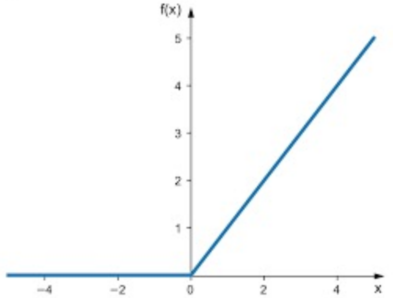
Neural Networks at 10K feet

- Y = f(X)
  - F may be constructed by combining different functions
    - $\mathbf{h^1} = g^1 (W^1 \mathbf{x} + b^1)$
    - $h^2 = g^2 (W^2 \mathbf{h^1} + b^2)$
    - ...
- Activation functions
  - Softmax
  - Relu
  - And many many more…
- Optimizers





Softmax



Relu