Machine Learning Basics
Lecture 1: Linear Regression

Princeton University COS 495

Instructor: Yingyu Liang

Machine learning basics

What is machine learning?

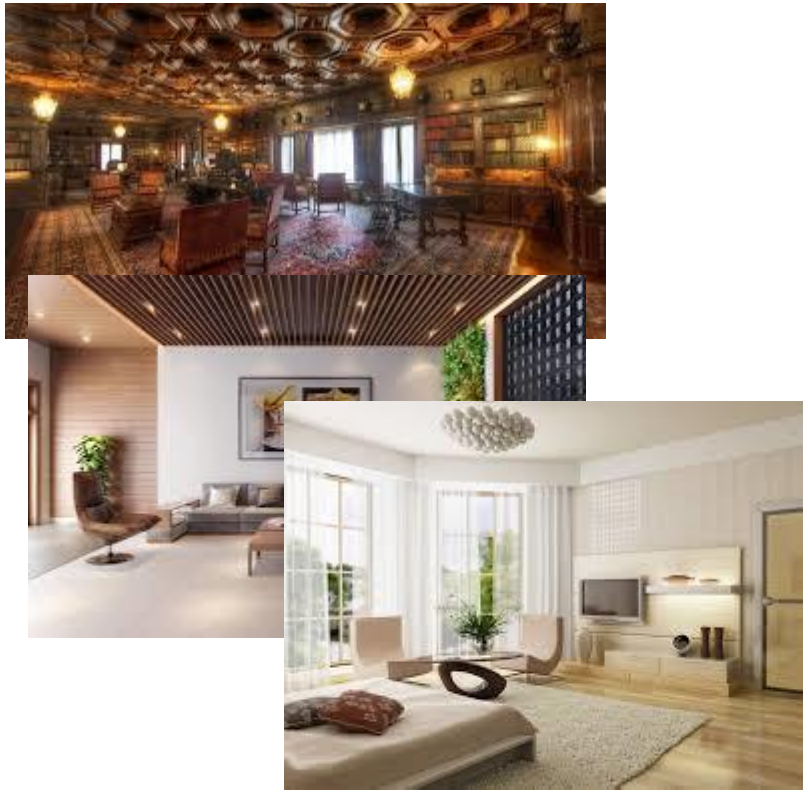
- "A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T as measured by P, improves with experience E."

  ------- *Machine Learning*, Tom Mitchell, 1997

Example 1: image classification

Task: determine if the image is indoor or outdoor
Performance measure: probability of misclassification

Example 1: image classification



Experience/Data:
images with labels

Indoor

outdoor

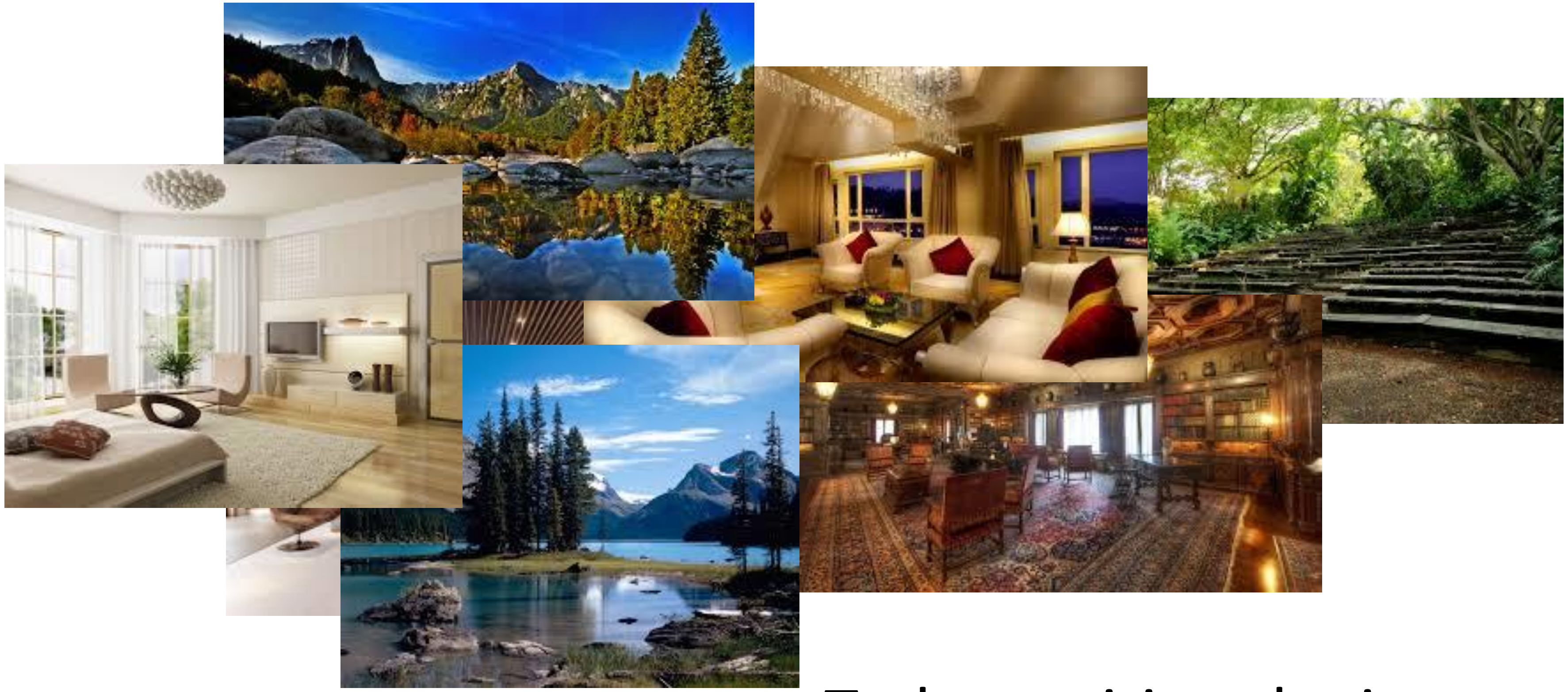
Example 1: image classification

- A few terminologies
  - Training data: the images given for learning
  - Test data: the images to be classified
  - Binary classification: classify into two classes

Example 1: image classification (multi-class)

ImageNet figure borrowed from vision.standford.edu

Example 2: clustering images



Task: partition the images into 2 groups

Performance: similarities within groups

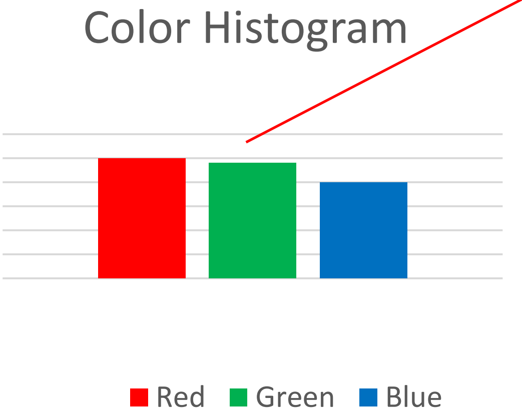
Data: a set of images

Example 2: clustering images

- A few terminologies
  - Unlabeled data vs labeled data
  - Supervised learning vs unsupervised learning

Math formulation

Extract
features



Feature vector: x_i

Indoor

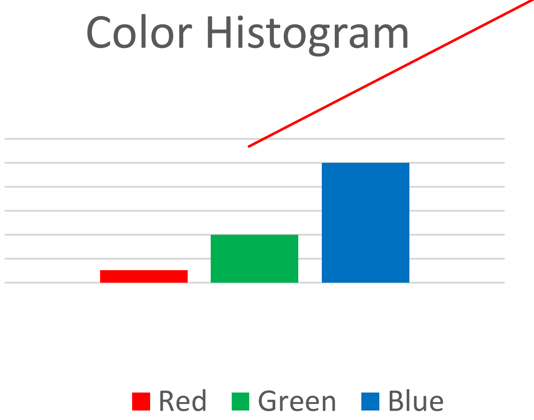
0

Label: y_i

Math formulation

Extract

features



Feature vector: x_j

outdoor

1

Label: y_j

Math formulation

- Given training data $\{(x_i, y_i): 1 \leq i \leq n\}$
- Find $y = f(x)$ using training data
- s.t. f correct on test data

What kind of functions?

Math formulation

- Given training data $\{(x_i, y_i): 1 \leq i \leq n\}$
- Find $y = f(x) \in \mathcal{H}$ using training data
- s.t. f correct on test data

Hypothesis class

Math formulation

- Given training data $\{(x_i, y_i): 1 \le i \le n\}$
- Find $y = f(x) \in \mathcal{H}$ using training data
- s.t. f correct on test data

Connection between training data and test data?

Math formulation

- Given training data $\{(x_i, y_i): 1 \le i \le n\}$ i.i.d. from distribution D
- Find $y = f(x) \in \mathcal{H}$ using training data
- s.t. f correct on test data i.i.d. from distribution D

They have the same distribution

i.i.d.: independently identically distributed

Math formulation

- Given training data $\{(x_i, y_i): 1 \leq i \leq n\}$ i.i.d. from distribution D
- Find $y = f(x) \in \mathcal{H}$ using training data
- s.t. f correct on test data i.i.d. from distribution D

What kind of performance measure?

Math formulation

- Given training data $\{(x_i, y_i): 1 \leq i \leq n\}$ i.i.d. from distribution D
- Find $y = f(x) \in \mathcal{H}$ using training data
- s.t. the expected loss is small

$$L(f) = \mathbb{E}_{(x,y)\sim D}[l(f, x, y)]$$

Various loss functions

Math formulation

- Given training data $\{(x_i, y_i): 1 \leq i \leq n\}$ i.i.d. from distribution D
- Find $y = f(x) \in \boldsymbol{\mathcal{H}}$ using training data
- s.t. the expected loss is small

$$L(f) = \mathbb{E}_{(x,y)\sim D}[l(f, x, y)]$$

- Examples of loss functions:
  - 0-1 loss: $l(f, x, y) = \mathbb{I}[f(x) \neq y]$ and $L(f) = \Pr[f(x) \neq y]$
  - l_2 loss: $l(f, x, y) = [f(x) - y]^2$ and $L(f) = \mathbb{E}[f(x) - y]^2$

Math formulation

- Given training data $\{(x_i, y_i): 1 \leq i \leq n\}$ i.i.d. from distribution D
- Find $y = f(x) \in \mathcal{H}$ using training data
- s.t. the expected loss is small

$$L(f) = \mathbb{E}_{(x,y)\sim D}[l(f, x, y)]$$

How to use?

Math formulation

- Given training data $\{(x_i, y_i): 1 \leq i \leq n\}$ i.i.d. from distribution D
- Find $y = f(x) \in \boldsymbol{\mathcal{H}}$ that minimizes $\hat{L}(f) = \frac{1}{n}\sum_{i=1}^{n} l(f, x_i, y_i)$
- s.t. the expected loss is small

$$L(f) = \mathbb{E}_{(x,y)\sim D}[l(f, x, y)]$$

Empirical loss

Machine learning 1-2-3

- Collect data and extract features
- Build model: choose hypothesis class $\mathcal{H}$ and loss function l
- Optimization: minimize the empirical loss

Wait…

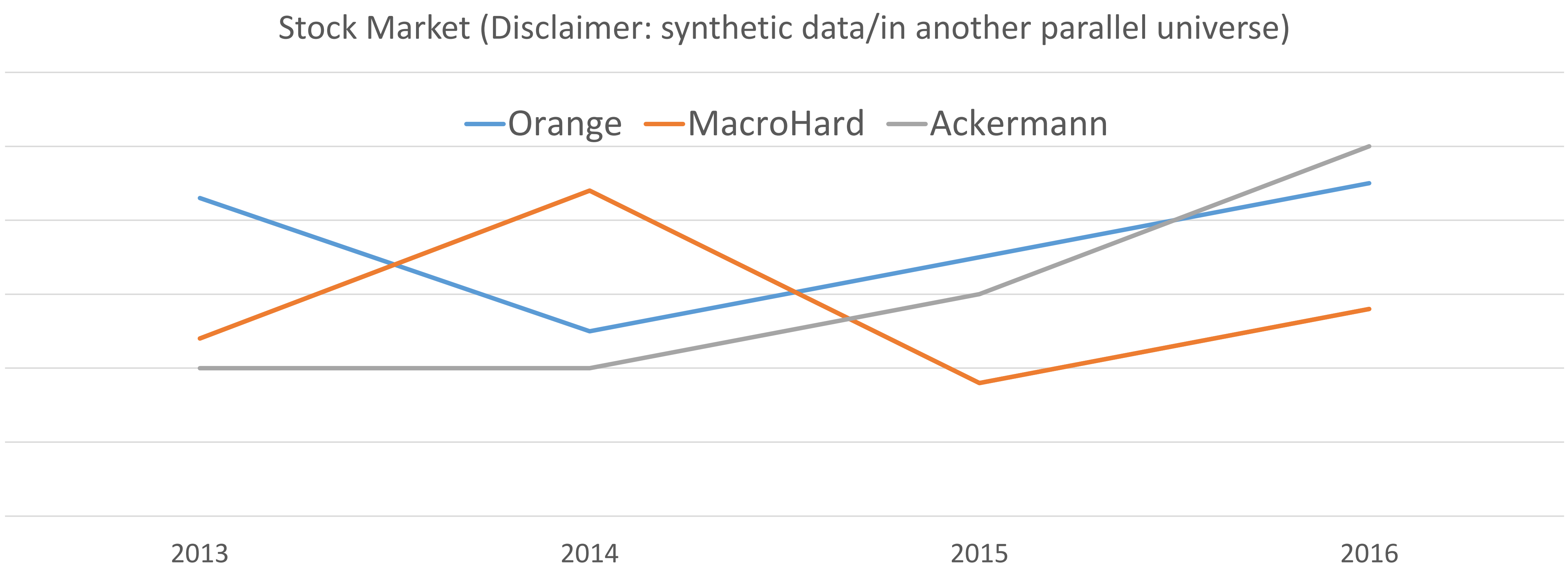
- Why handcraft the feature vectors x, y?
  - Can use prior knowledge to design suitable features

- Can computer learn the features on the raw images?
  - Learn features directly on the raw images: Representation Learning
  - Deep Learning $\subseteq$ Representation Learning $\subseteq$ Machine Learning $\subseteq$ Artificial Intelligence

Wait…

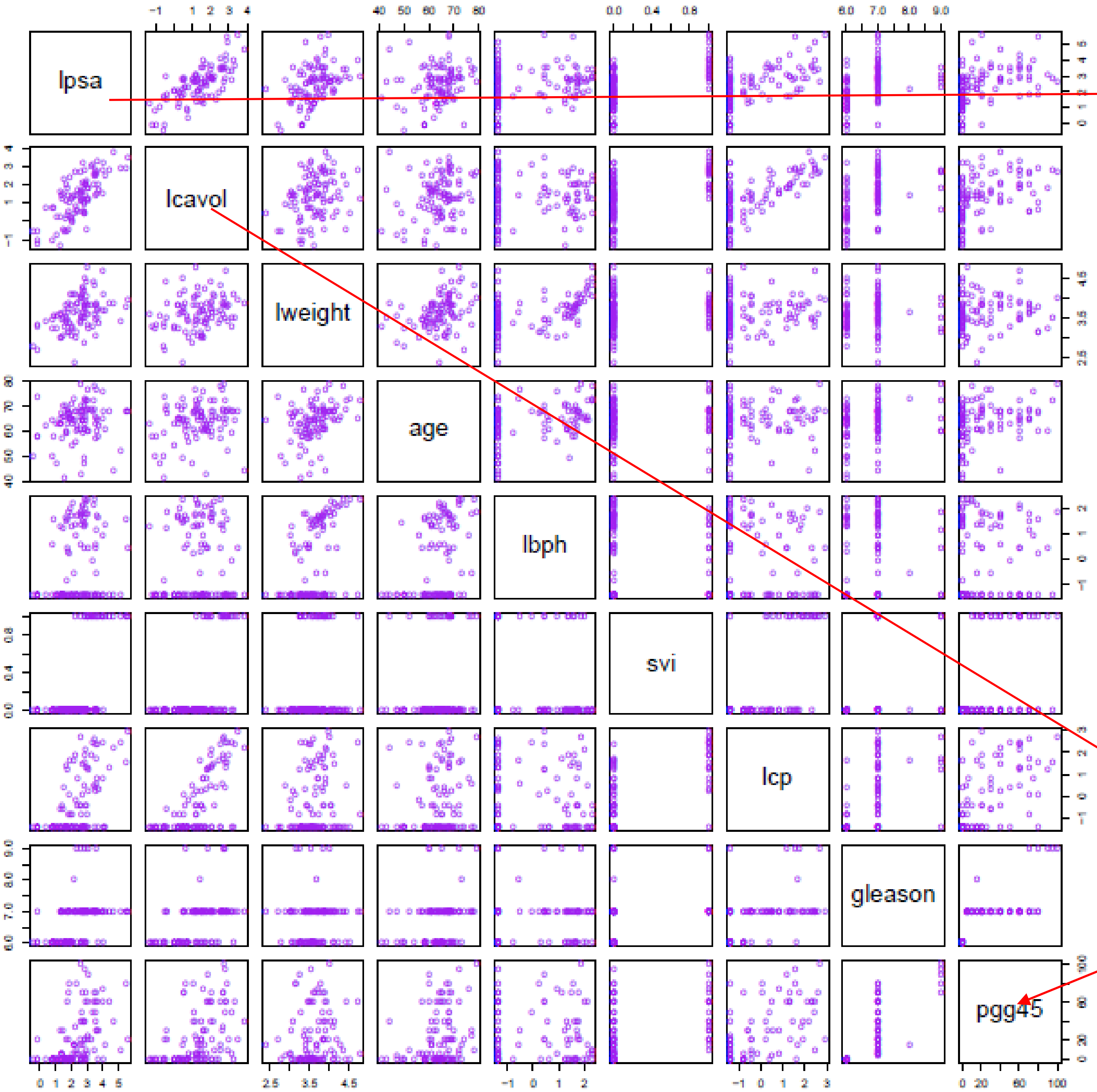
- Does MachineLearning-1-2-3 include all approaches?
    - Include many but not all
    - Our current focus will be MachineLearning-1-2-3

Example: Stock Market Prediction



Sliding window over time: serve as input x; non-i.i.d.

Linear regression



y: prostate specific antigen

$(x_1, \dots, x_8)$: clinical measures

Real data: Prostate Cancer by Stamey et al. (1989)

Figure borrowed from
The Elements of Statistical Learning

Linear regression

- Given training data $\{(x_i, y_i): 1 \leq i \leq n\}$ i.i.d. from distribution D
- Find $f_w(x) = w^T x$ that minimizes $\hat{L}(f_w) = \frac{1}{n} \sum_{i=1}^{n} (w^T x_i - y_i)^2$

l_2 loss; also called mean square error

Hypothesis class $\boldsymbol{\mathcal{H}}$

Linear regression: optimization

- Given training data $\{(x_i, y_i): 1 \leq i \leq n\}$ i.i.d. from distribution D
- Find $f_w(x) = w^T x$ that minimizes $\hat{L}(f_w) = \frac{1}{n}\sum_{i=1}^{n}(w^T x_i - y_i)^2$

- Let X be a matrix whose i-th row is x_i^T, y be the vector $(y_1, \ldots, y_n)^T$

$$\hat{L}(f_w) = \frac{1}{n}\sum_{i=1}^{n}(w^T x_i - y_i)^2 = \frac{1}{n}\|Xw - y\|_2^2$$

Linear regression: optimization

- Set the gradient to 0 to get the minimizer

$$\nabla_w \hat{L}(f_w) = \nabla_w \frac{1}{n} \|Xw - y\|_2^2 = 0$$

$$\nabla_w [(Xw - y)^T (Xw - y)] = 0$$

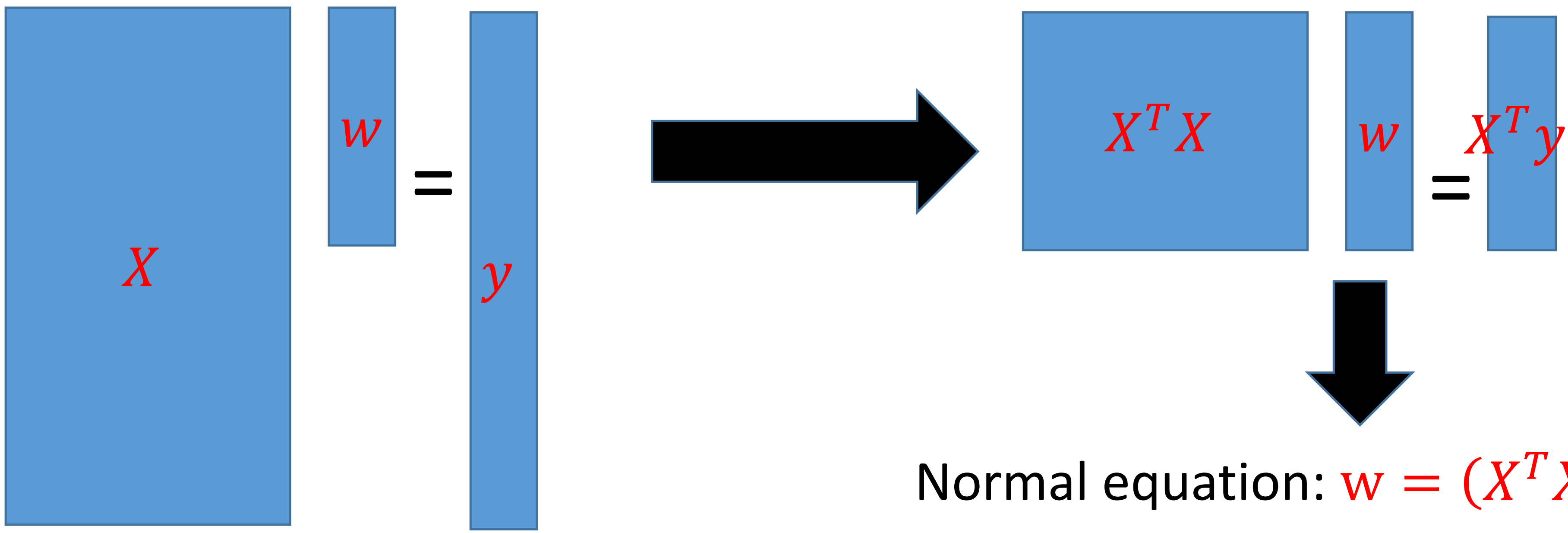
$$\nabla_w [\, w^T X^T X w - 2 w^T X^T y + y^T y] = 0$$

$$2X^T X w - 2X^T y = 0$$

$$\mathrm{w} = (X^T X)^{-1} X^T y$$

Linear regression: optimization

- Algebraic view of the minimizer
  - If X is invertible, just solve $Xw = y$ and get $w = X^{-1}y$
  - But typically X is a tall matrix



Normal equation: $w = (X^TX)^{-1}X^Ty$

Linear regression with bias

Bias term

- Given training data $\{(x_i, y_i): 1 \leq i \leq n\}$ i.i.d. from distribution D
- Find $f_{w,b}(x) = w^T x + b$ to minimize the loss

- Reduce to the case without bias:
  - Let $w' = [w;\ b], x' = [x; 1]$
  - Then $f_{w,b}(x) = w^T x + b = (w')^T (x')$