Introduction to Machine Learning Part 3: k-Nearest Neighbor and Linear Regression

CS 540

Yingyu Liang

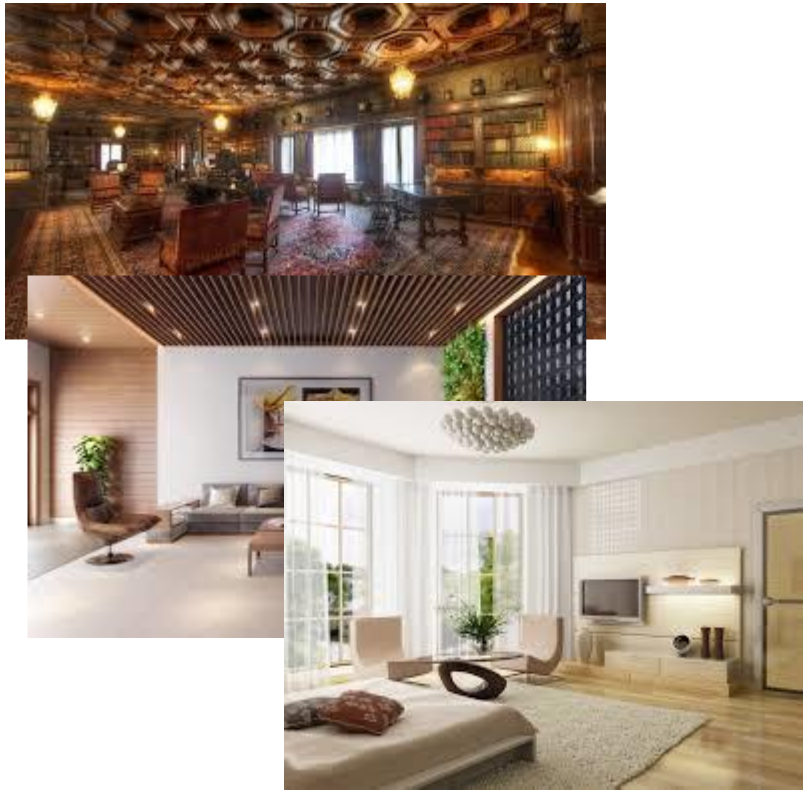
Supervised Learning

Example: image classification

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

Example: image classification

Experience/Data:
images with labels



Indoor

outdoor

Example: image classification

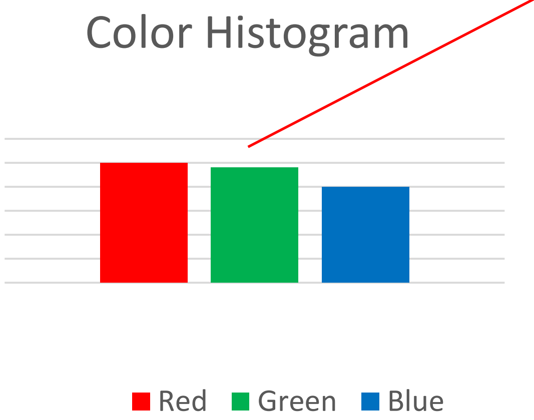
- A few terminologies
  - Training data: the images given for learning
  - Test data: the images to be classified
  - Binary classification: classify into two classes

Example: image classification (multi-class)

ImageNet figure borrowed from vision.standford.edu

Math formulation

Extract **features**



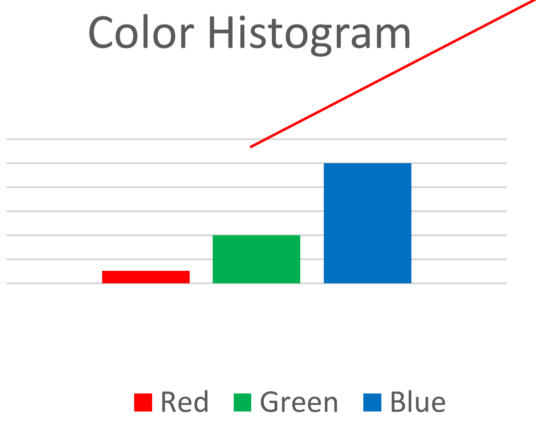
Feature vector: x_i

Indoor → 0

Label: y_i

Math formulation

Extract
features



Feature vector: x_j

Label: y_j

outdoor

1

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 $\{(x_i, y_i): 1 \leq i \leq n\}$
- 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 some unknown 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 some unknown 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

- If label y discrete: classification
- If label y continuous: regression

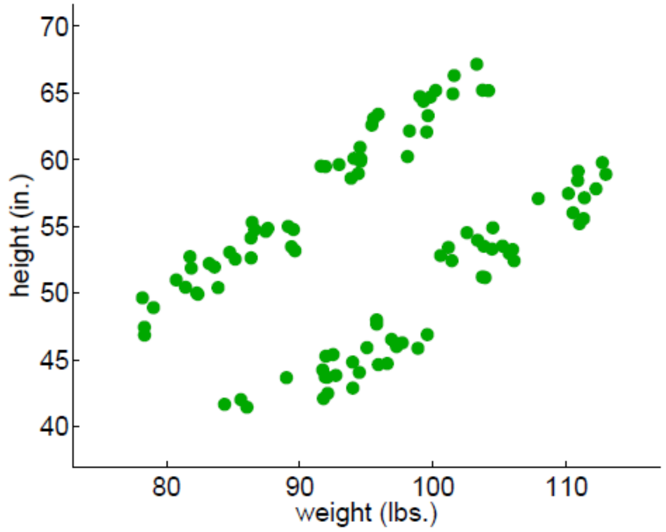
K-Nearest Neighbors

K-nearest neighbors

- Given training data $\{(x_i, y_i): 1 \le i \le n\}$ i.i.d. from distribution D
- Store the training data
- Given a new data point x, predict its label based on its neighbors

Little Green Man

- Little green men:
  - Predict gender (M,F) from weight, height?
  - Predict adult, juvenile from weight, height?



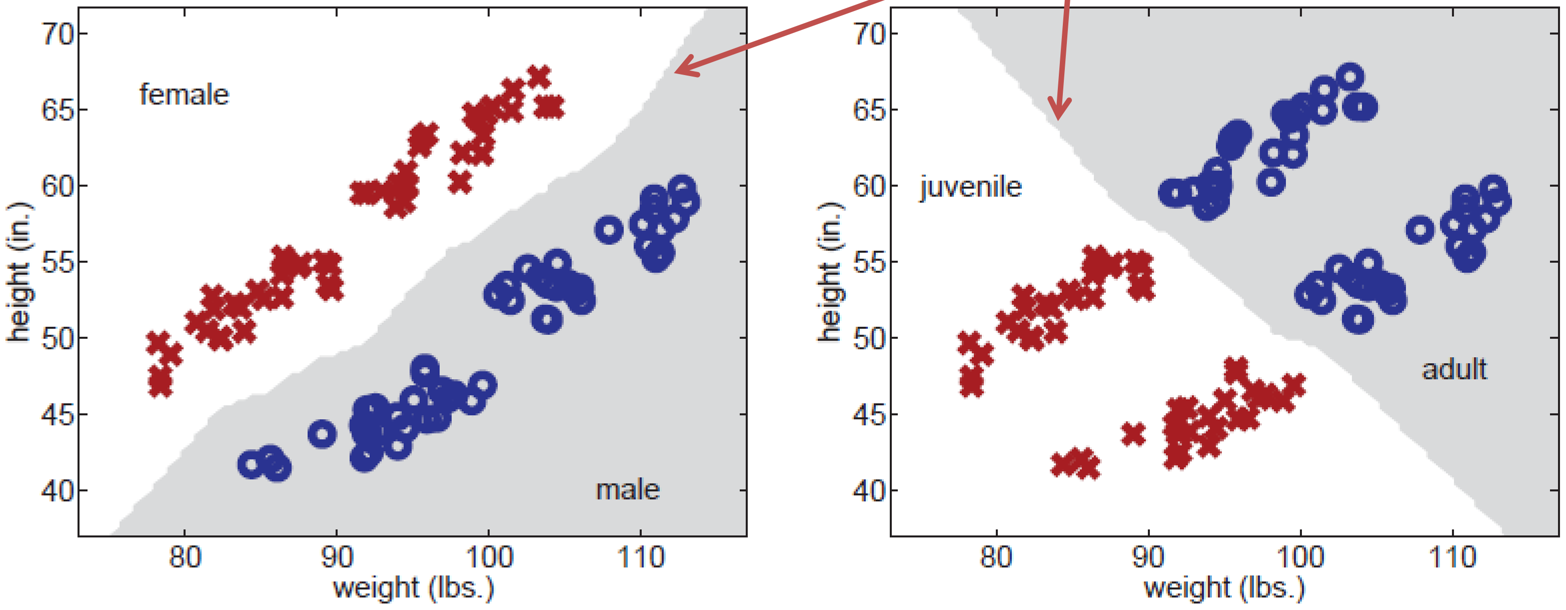
k-nearest-neighbor (kNN)

Input: Training data $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$*; distance function* $d()$*;*
number of neighbors k*; test instance* $\mathbf{x}^*$

1. *Find the* k *training instances* $\mathbf{x}_{i_1}, \ldots, \mathbf{x}_{i_k}$ *closest to* $\mathbf{x}^*$ *under distance* $d()$.
2. *Output* y^* *as the majority class of* $y_{i_1}, \ldots, y_{i_k}$. *Break ties randomly.*

- 1NN for little green men:

Decision boundary



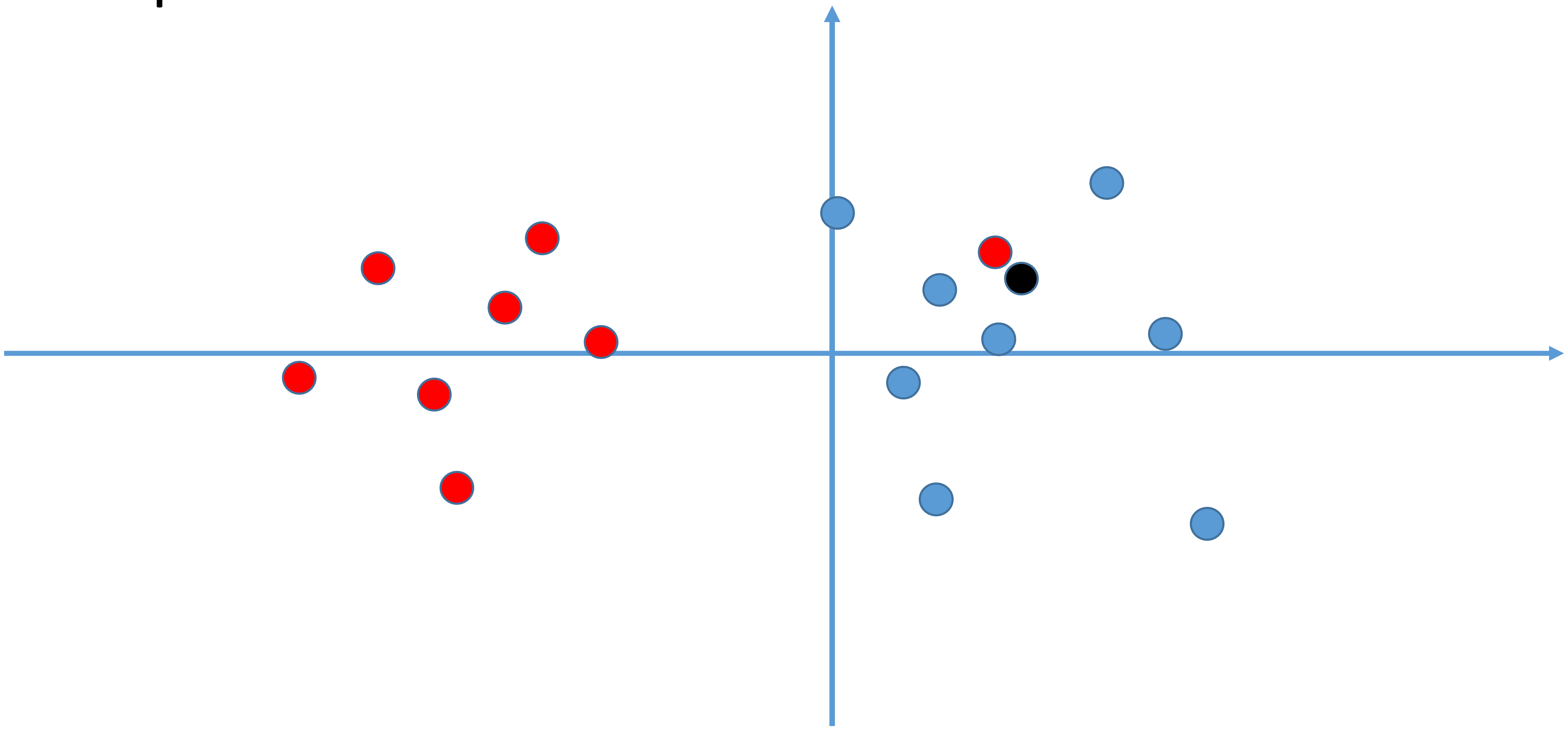
(a) classification by gender

(b) classification by age

kNN

- What if we want regression?
  - Instead of majority vote, take average of neighbors' y
- How to pick *k*?
  - Split data into training and tuning sets
  - Classify tuning set with different *k*
  - Pick *k* that produces least tuning-set error

Example



What's the predicted label for the black dot using 1 neighbor? 2 neighbors? 3 neighbors?

Linear regression

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 $y = f(x) \in \mathcal{H}$ 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 $y = f(x) \in \mathcal{H}$ 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

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 $(y_1, \dots, y_n)^T$

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

Linear regression: optimization

- Set the gradient to 0 to get the minimizer

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

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

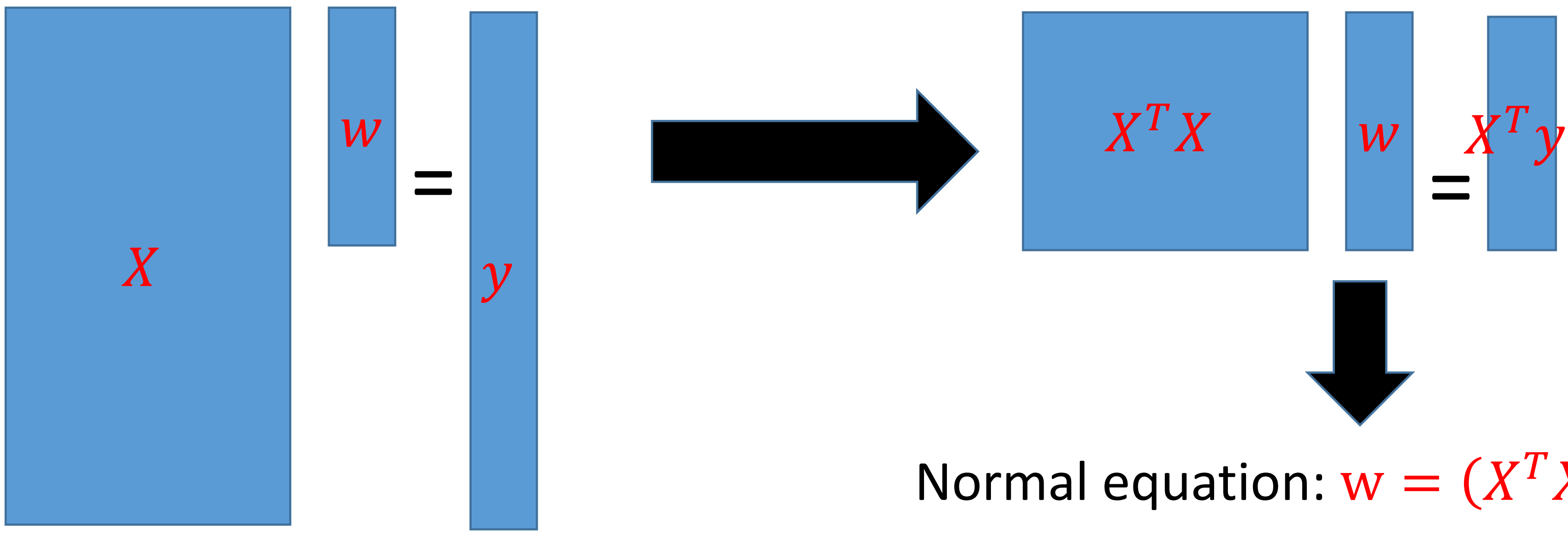
$$\nabla_w [\, w^T X^T X w - 2 w^T X^T y + y^T y] = 0$$

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

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

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

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 $f_{w,b}(x) = w^T x + b = (w')^T (x')$

Linear regression with regularization: Ridge 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 $\widehat{L_R}(f_w) = \frac{1}{n}||Xw - y||_2^2 + \lambda||w||_2^2$

- By setting the gradient to be zero, we have

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