Classification and Regression: Learning Functions

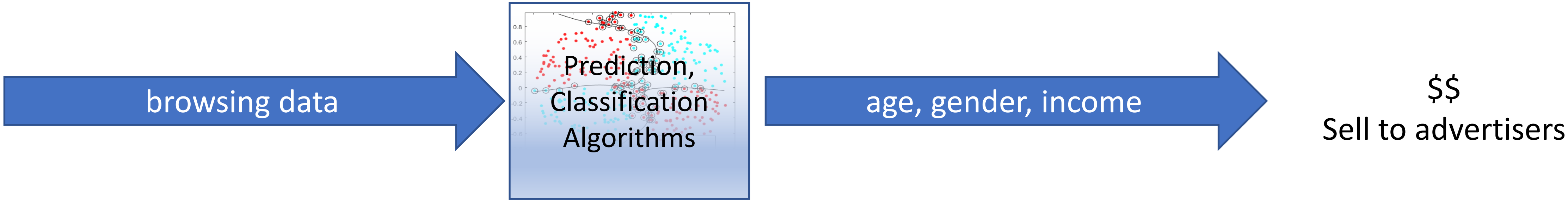
Contents

- Example classification problem
- Classification and regression problems
- Learning functions

Example: digital advertising and tracking data

Example data: *identifier=1234* visited *cars.com, espn.com, hiking.com, stackexchange.com*

cookie



Question: is *identifier=1234* male or female?

How can we design a simple algorithm that takes a list of domains and makes a prediction?

Ambiguity?

Binary Classifier

unlabeled data: *{cars.com, espn.com, hiking.com}*

$$x = \{\text{cars.com}, \text{espn.com}, \text{hiking.com}\}$$

Which is more probable? Male or Female?

$$\mathbb{P}(y = \text{female}|x) \qquad \mathbb{P}(y = \text{male}|x)$$

How can we estimate $\mathbb{P}(y|x)$?

labeled data: *2 x {cars.com, espn.com, hiking.com}, Male*
1 x {cars.com, espn.com, hiking.com}, Female

Histogram Classifier

How can we estimate $\mathbb{P}(y|x)$?

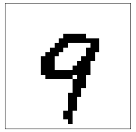
labeled data: *2 x {cars.com, espn.com, hiking.com}, Male*
1 x {cars.com, espn.com, hiking.com}, Female

Classification

- **classification** is the process of assigning items to classes
- a **classifier** is a function that maps inputs to output classes
- **classifier design**: learn a function $f(x)$ that maps inputs to output classes

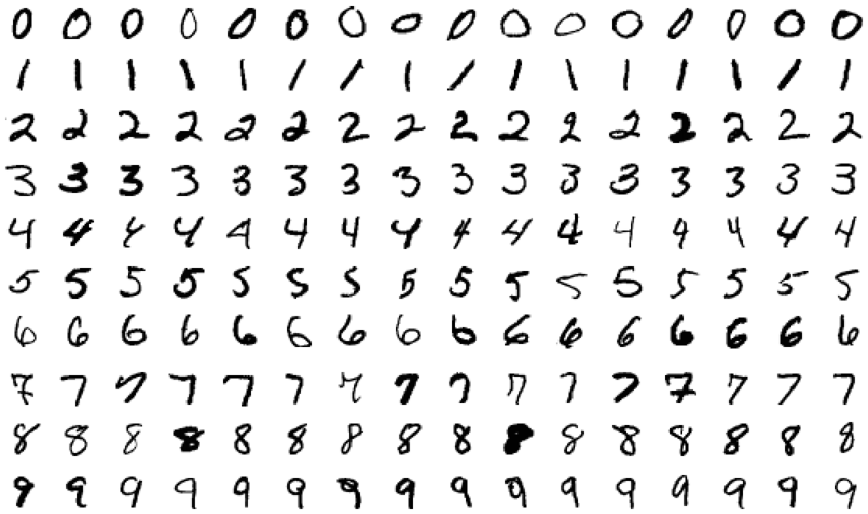
  i.e., learning a mapping from inputs x to outputs y, where $y \in \{1, \ldots, c\}$

  - binary classification $\Rightarrow c = 2$
  - multiclass classification $\Rightarrow c > 2$
  - multilabel classification $\Rightarrow$ multiple mappings

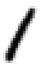
$x =$ 

- example: classifying handwritten digits as $0, 1, \ldots 9$

  goal: build a function so that $f($  $) = 9$



- training data $\mathcal{D}$

  $\mathcal{D} = \{($  $, 1), ($  $, 4), ($  $, 3) \ldots\}$

  $$\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{n}$$

Regression

- **prediction**, **forecasting**, or **regression**

- just like **classification**, regression is
  learning a function that maps inputs to output ~~classes~~

- **regression**: learning a function $f(x)$ that maps inputs to output
  i.e., learning a mapping from inputs x to outputs y, where $y \in \mathbb{R}$

- example: predict the high temperature in Madison based on day of year

  goal: build a function so that $f\,(\text{Jan } 4) = -5.6\ ^{o}\text{C}$

- training data $\mathcal{D}$

  $$\mathcal{D} = \{(\text{April 1 1981}, 3.2), (\text{June 17 1956}, 18.2), \ldots\}$$

prediction, **forecasting**, **regression** and **classification** are about learning functions from data!

What is a function?

- functions (in highschool)

  $\Rightarrow$ takes number as input, and outputs another number

  $f(\cdot) : \mathbb{R} \mapsto \mathbb{R}$

- functions (in second/third semester calculus)

  $\Rightarrow$ takes several numbers as input, and outputs another number

  $f(\cdot) : \mathbb{R}^3 \mapsto \mathbb{R}$

- functions (in general, and machine learning)

  $\Rightarrow$ takes an element from a set $\mathcal{X}$ as input
  outputs an element from another set $\mathcal{Y}$

  $f(\cdot) : \mathcal{X} \mapsto \mathcal{Y}$

Examples of functions

- example: $f(x) = 10 - 15\cos(\frac{2\pi x}{365})$

  domain: set of possible inputs $\mathcal{X} = \mathbb{R}$

  output space: set of possible outputs $\mathcal{Y} = [-5, 25]$

  $f(\cdot) : \mathbb{R} \mapsto [-5, 25]$

- example: $f(x)$ is an NMIST image classifier

  domain: $\mathcal{X}$ is the set of all 28×28 black and white images

  $$\mathcal{X} = \{0, 1\}^{28\times 28} \qquad \Rightarrow 2^{784} \text{ possibilities}$$

  output space: set of possible classes $\mathcal{Y} = \{0, 1, \ldots, 9\}$

  $f(\cdot) : \{0, 1\}^{784} \mapsto \{0, 1, \ldots, 9\}$

Example: MNIST Classifier

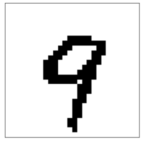
- example: $f(x)$ is an NMIST image classifier

Features, Labels, and Output

- machine learning is about **learning functions from data**

$$f(x)$$

- the inputs to the functions are called **features**

$x =$  or salient features of the image

- the true output is called a label

$$y \in \{0, 1, 2, \ldots, 9\}$$

- labeled examples of features make up *training data*

$$\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{n}$$

- the outputs are called **classes, predictions, forecasts,** etc.

$$\widehat{y} = f(x)$$

Notation

$x \in \mathcal{X}$

$y \in \mathcal{Y}$

$f(\cdot) : \mathcal{X} \mapsto \mathcal{Y}$

$\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$

$\mathbb{R}$

$\mathbb{R} \times \mathbb{R}$

$\mathbb{R}^d$

$\{0, 1\}$

$[a, b]$

$|\mathcal{X}|$

Training: estimating a function from data

- 1-d example: learning a function $f(\cdot) : \mathbb{R} \mapsto \mathbb{R}$

- start with a **model** - *limited set of possible functions*

  example: $f(x) = a - b\cos(\frac{2\pi x}{T})$

- plot some points, find paramaters (a, b, T) that fit data

$$\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{n} = \{(\text{April 1 1981}, 3.2), (\text{June 17 1956}, 18.2), \ldots\}$$

1. $f(x) \approx y$ for (most) of the training data
2. $f(x)$ should work for brand new x

Finding efficient ways to learn functions from data is the fundamental challenges of machine learning.

Probabilistic models

- probabilistic models treat the features $\boldsymbol{x}$ and labels $\boldsymbol{y}$ as random variables
- probabilistic models use probability to help design $f(x)$

$x =$ 

What is the probability each label given the data?

$\mathbb{P}(y = 0|x =$  $)$

$\vdots$

$\mathbb{P}(y = 4|x =$  $)$

$\vdots$

$\mathbb{P}(y = 9|x =$  $)$

$f(x) = \arg \max_{i=0,\ldots,9} \mathbb{P}(y = i|x =$  $)$

MAP (Maximum a posteriori) estimate

Memorization and Histogram Classifier

$$f(x) = \arg\max_{i=0,\ldots,9} \mathbb{P}(y = i \mid x =$$  $$)$$

for $x =$ , count how many examples in $\mathcal{D}$ correspond to $y = 1$, $y = 2$, $\ldots$

$\mathbb{P}(y = 0 \mid x =$  $) \approx$ $\dfrac{\text{count of } x = \text{[image 4]} \text{ with } y = 0 \text{ in } \mathcal{D}}{\text{count of } x = \text{[image 5]} \text{ in } \mathcal{D}}$

$\vdots$

$\mathbb{P}(y = 9 \mid x =$  $)$

Why won't this work?

k-nearest neighbor classifier (knn)

$\mathbb{P}(y = 0|x =$  $) \approx \dfrac{\text{count of examples with } y = 0 \text{ out } k \textit{ closest} \text{ to } x = \text{}}{k}$ 

$\vdots$

$\mathbb{P}(y = 9|x =$  $)$

$f(x) = \arg \max_{i=0,\ldots,9} \mathbb{P}(y = i|x =$  $)$

Discriminative and generative classifiers

$$f(x) = \arg\max_{i=0,\ldots,9} \mathbb{P}(y = i | x =$$  $)$

- learning $f(x)$ requires learning $\mathbb{P}(y|x)$
- learning $\mathbb{P}(y|x)$ might require learning $\mathbb{P}(y, x)$
- if we know $\mathbb{P}(x, y)$, we can often sample from it