ONLINE MASTERS IN **DATA SCIENCE**

DSC 255 - MACHINE LEARNING FUNDAMENTALS

THE LANDSCAPE OF MACHINE LEARNING

SANJOY DASGUPTA, PROFESSOR

UC San Diego

COMPUTER SCIENCE & ENGINEERING

HALICIOĞLU DATA SCIENCE INSTITUTE

Three learning modalities

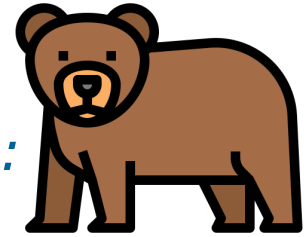
1. **Supervised learning**
   For solving **prediction problems**
2. **Unsupervised learning**
   For finding **good representations**
3. **Learning through interaction**
   E.g., reinforcement learning

Prediction problems: Inputs and outputs

Basic terminology:

- The input space, $\mathcal{X}$.
  E.g., 32 x 32 RGB images of animals.
- The output space, $\mathcal{Y}$.
  E.g., Names of 100 animals.

$\mathcal{X}$:



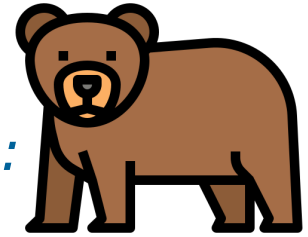
$\mathcal{Y}$: "bear"

Prediction problems: Inputs and outputs

Basic terminology:

- The input space, $\mathcal{X}$.
  E.g., 32 x 32 RGB images of animals.
- The output space, $\mathcal{Y}$.
  E.g., Names of 100 animals.

$\mathcal{X}$:



$\mathcal{Y}$: "bear"

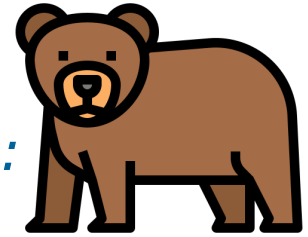
After seeing a bunch of examples (x, y), pick a mapping that accurately recovers the input-output pattern of the examples.

$$f: \mathcal{X} \rightarrow \mathcal{Y}$$

Prediction problems: Inputs and outputs

Basic terminology:

- The input space, $\mathcal{X}$.
  E.g., 32 x 32 RGB images of animals.
- The output space, $\mathcal{Y}$.
  E.g., Names of 100 animals.

x: 

y: "bear"

After seeing a bunch of examples (x, y), pick a mapping that accurately recovers the input-output pattern of the examples.

$$f: \mathcal{X} \rightarrow \mathcal{Y}$$

Prediction problems can be categorized by the type of **output space:** (1) discrete, (2) continuous, or (3) probability values.

Binary classification

E.g., Spam detection

$\mathcal{X}$ = {email messages}

$\mathcal{Y}$ = {spam, not spam}

Discrete output space: classification

Binary classification

E.g., Spam detection

$\mathcal{X}$ = {email messages}

$\mathcal{Y}$ = {spam, not spam}

Multiclass

E.g., News article classification

$\mathcal{X}$ = {news articles}

$\mathcal{Y}$ = {politics; business; sports,...}

Discrete output space: classification

Binary classification

E.g., Spam detection

$\mathcal{X}$ = {email messages}

$\mathcal{Y}$ = {spam, not spam}

Multiclass

E.g., News article classification

$\mathcal{X}$ = {news articles}

$\mathcal{Y}$ = {politics; business; sports,...}

Structured outputs

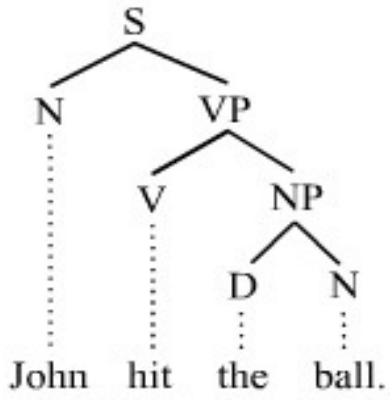
E.g., Parsing

$\mathcal{X}$ = {sentences}

$\mathcal{Y}$ = {parse trees}

$\mathcal{X}$: "John hit the ball"

$\mathcal{Y}$:



Continuous output space: regression

Pollution level prediction

- Predict tomorrow's air quality index in my neighborhood
  $\mathcal{Y} = [0, \infty)$ (< 100: okay, > 200: dangerous)

Insurance company calculations

- What is the expected life expectancy of this person?
  $\mathcal{Y} = [0, 120]$

What are suitable predictor variables $(\mathcal{X})$ in each case?

Probability estimation

$\mathcal{Y} = [0, 1]$ represents **probabilities**

Example: Credit card transactions:

- $\mathcal{X}$ = details of a transaction
- $\mathcal{Y}$ = probability this transaction is fraudulent

Probability estimation

$\mathcal{Y} = [0, 1]$ represents **probabilities**

Example: Credit card transactions:

- $\mathcal{X}$ = details of a transaction
- $\mathcal{Y}$ = probability this transaction is fraudulent

Why not just treat this as a binary classification problem?

Three learning modalities

1. **Supervised learning**
   Nearest neighbor, generative models for prediction, linear regression, logistic regression, support vector machines, kernel methods, decision trees, boosting, random forests, neural nets
2. **Unsupervised learning**
   Clustering, projection, manifold learning, embedding, generative modeling
3. **Learning through interaction**