

# **INTRO TO DATA SCIENCE**

## **LECTURE 2: MACHINE LEARNING**

---

## **AGENDA**

---

**I. DATAEXPLOR01 REVIEW**

**II. WHAT IS MACHINE LEARNING?**

**III. MACHINE LEARNING ALGORITHMS**

**IV. PYTHON TOOLS FOR ML**

**V. LAB: PRACTICE**

---

## INTRO TO DATA SCIENCE

---

# I. WHAT IS MACHINE LEARNING?

---

# WHAT IS MACHINE LEARNING?

---

from Wikipedia:

“Machine learning, a branch of artificial intelligence, is about the construction and study of systems that *can learn from data.*”

source: [http://en.wikipedia.org/wiki/Machine\\_learning](http://en.wikipedia.org/wiki/Machine_learning)

---

# WHAT IS MACHINE LEARNING?

---

from Wikipedia:

“Machine learning, a branch of artificial intelligence, is about the construction and study of systems that *can learn from data.*”

- "Field of study that gives computers the ability to learn without being explicitly programmed" ~*Arthur Samuel*
- “Improve on task, **T**, with respect to performance metric, **P**, based on experience, **E**” ~*Tom Mitchell*

---

## QUESTION

---

WHAT IS  
MACHINE LEARNING  
USED FOR?

---

# **WHAT IS MACHINE LEARNING USED FOR?**

---

**Prediction**

**Pattern Recognition**

**Search Engines**

**Diagnostics**

**Bioinformatics**

**Summarization**

**Machine Translation**

---

## MACHINE LEARNING STRENGTHS

---

- finding patterns in large data sets
- scaling out decision making that is time-consuming or repetitive for humans



---

## MACHINE LEARNING WEAKNESSES

---

- algorithms vary in ability to generalize over patterns
- possibility of over-generalizing
- limited by available data

---

## INTRO TO DATA SCIENCE

---

# II. MACHINE LEARNING ALGORITHMS

---

## TYPES OF MACHINE LEARNING ALGORITHMS

---

1. supervised
2. unsupervised

---

## TYPES OF MACHINE LEARNING ALGORITHMS

---

1. supervised

➤ making predictions

2. unsupervised

---

## TYPES OF MACHINE LEARNING ALGORITHMS

---

1. supervised

➤ making predictions

2. unsupervised

➤ extracting structure

---

# SUPERVISED LEARNING

---

- Outcome measurement  $Y$ , (also called dependent variable, response, target)
- Vector of  $p$  predictor measurements  $X$  (also called inputs, regressors, covariates, features, independent variables)
- In **regression**,  $Y$  is quantitative (e.g. price, temperature)
- In **classification**,  $Y$  has values in a finite, unordered set (survived/died, cancer class of tissue sample, category of document)

---

## SUPERVISED LEARNING

---

On the basis of training data  $(x_1, y_1), \dots, (x_N, y_N)$  we would like to:

- accurately predict unseen test cases
- understand which inputs affect the outcome, and how
- asses the quality of our predictions and inferences

---

## UNSUPERVISED LEARNING

---

- No outcome variable, just a set of predictors (features) measured on a set of samples
- objective is less clear-- find features/groups of samples that behave similarly, find linear combinations of features with the most variation
- difficult to know how well you are doing
- can be useful as a pre-processing step for supervised learning



---

## **SUPERVISED OR UNSUPERVISED?**

---

- spam filtering

---

## **SUPERVISED OR UNSUPERVISED?**

---

- spam filtering
- character recognition

---

## **SUPERVISED OR UNSUPERVISED?**

---

- spam filtering
- character recognition
- document clustering

---

## SUPERVISED OR UNSUPERVISED?

---

- spam filtering
- character recognition
- document clustering
- fraud detection

---

## SUPERVISED OR UNSUPERVISED?

---

- spam filtering
- character recognition
- document clustering
- fraud detection
- dimensionality reduction

---

## INTRO TO DATA SCIENCE

---

**III. PYTHON TOOLS**

**IV. LAB PRACTICE**