1. INTRODUCTION

CSC 462[THPMLB-1]

WHAT IS MACHINE LEARNING?

- Term “Machine Learning” coined by Arthur Samuel in 1959.
- Machine learning is a field that grew out from AI
- Machine learning gives new capabilities for computers
- It is a multidisciplinary field (vision, computational biology, www, statistics, industry …)

WHAT IS MACHINE LEARNING?

- It is a set methods of data analysis that automates analytical model building.
- Algorithms that iteratively learn from data by finding the hidden insights without being explicitly programmed where to look.

EXAMPLES

- Self-driving Google car
- Online recommendation offers like those from Amazon, Netflix, and IMDB
- Knowing what customers are saying about you on Twitter
- Fraud detection
- Web search engines: Google uses ranking algorithms to rank websites in search results
- Facebook: when you tag your friend, Facebook learn to recognize your friends' faces
- Spam filtering

EXAMPLES

- Database mining
  - Large datasets from growth of automation/web
  - E.g. Web click data, medical records, biology, engineering
- Applications you can't program by hand
  - E.g., Autonomous helicopter, handwriting recognition, most of Natural Language Processing (NLP), Computer Vision.

MACHINE LEARNING DEFINITION

- Arthur Samuel (1959). Machine Learning: Field of study that gives computers the ability to learn without being explicitly programmed.

MACHINE LEARNING DEFINITION

- Tom Mitchell (1998) Well-posed Learning Problem:

A computer program is said to *learn* from experience E with respect to some task T and some performance measure P, if its performance on T, as measured by P, improves with experience E.

SPAM FILTERING

- Suppose your email program watches you marking emails as spams and other as not spams. According to the previous definition:
  - What is the task T performed ?
  - What is the performance measure P ?
  - What is the experience E ?

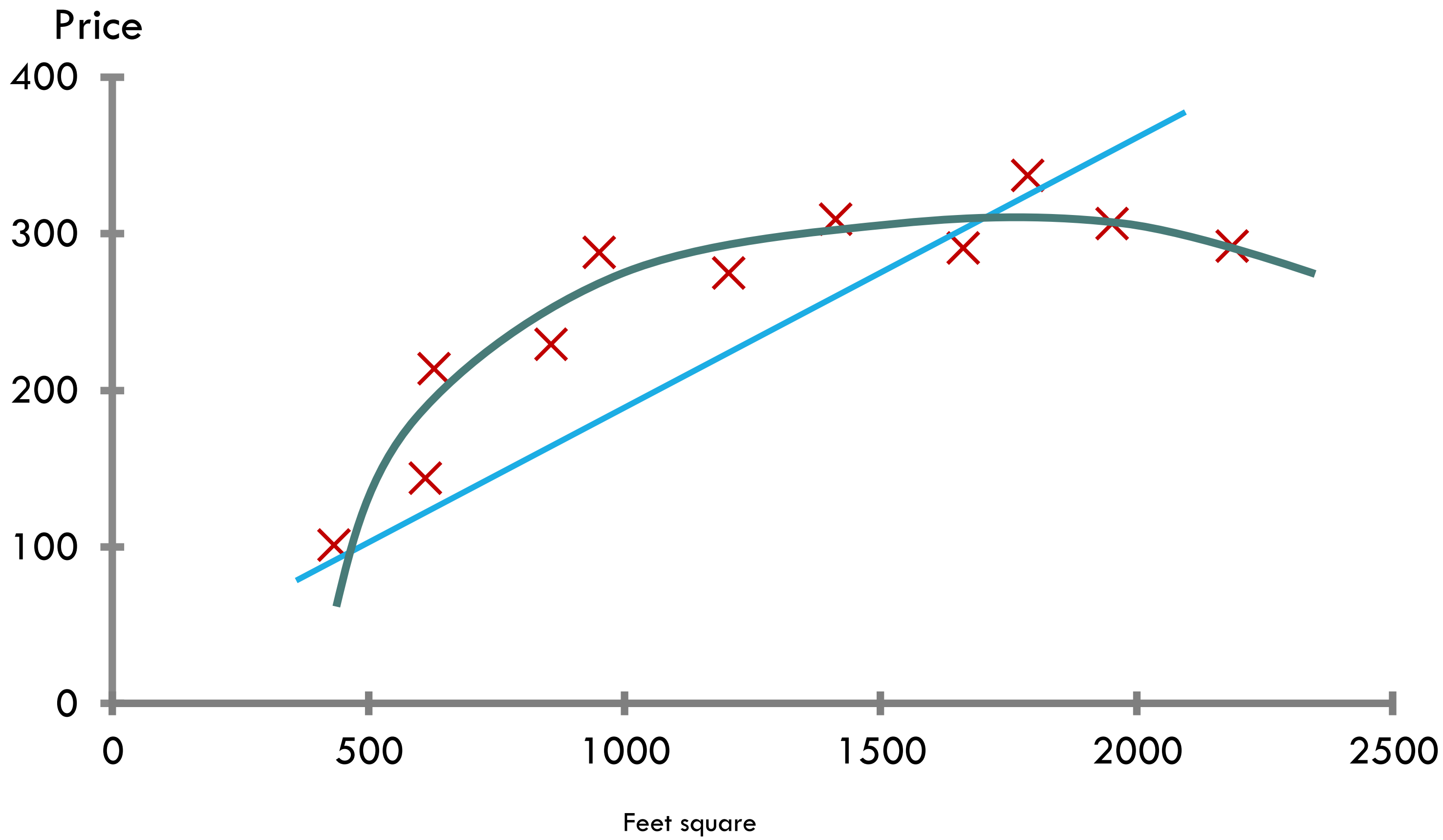
TYPES OF LEARNING

1. **Supervised learning:** The agent observes a set of input-output examples (**labeled examples**) and learns a map from inputs to outputs.
   - **Classification**: output is discrete (e.g., spam email)
   - **Regression**: output is real-valued (e.g., stock market)
2. **Unsupervised learning:** No explicit feedback is given, only the inputs (**unlabeled examples**). The agent learns patterns in the input.
   - **Clustering**: grouping data into K groups. (e.g. clustering images of fish into different species)
3. **Semi-supervised learning:** The agent is given some labeled examples (generally a few) and some unlabelled examples and tries to learn a mapping. (e.g. constrained clustering)
4. **Reinforcement learning**: The agent learns from a series of rewards and punishments, and based on these adapts its behavior.

1. SUPERVISED LEARNING: HOUSING PRICE PREDICTION

Supervised learning: right answers are given

Regression: Predict continuous valued output (price)



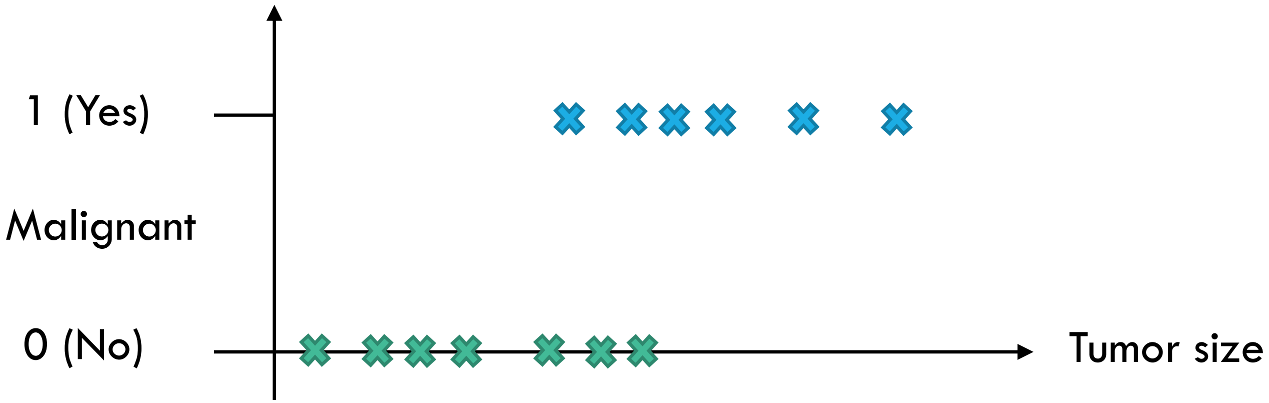
1. SUPERVISED LEARNING: HOUSING PRICE PREDICTION

- What approaches can we use to solve this? Straight line through data or second order polynomial
  - How to chose straight or curved line? (later)
- Each of these approaches represent a way of doing supervised learning
- *What does this mean?* We gave the algorithm a data set where a "right answer" was provided
- So we know actual prices for houses
  - The idea is we can learn what makes the price a certain value from the **training data**
  - The algorithm should then produce more right answers based on new training data where we don't know the price already (i.e. predict the price)

1. SUPERVISED LEARNING: BREAST CANCER (MALIGNANT/BENIGN)

Classification: Discrete valued output (0 or 1)

Uses one attribute (size)



1. SUPERVISED LEARNING: BREAST CANCER (MALIGNANT/BENIGN)

- In other problems we may have multiple attributes
- We may also, for example, know the age and tumor size

1. SUPERVISED LEARNING

- Based on that data, you can try and define separate classes by
  - Drawing a straight line between the two groups
  - Using a more complex function to define the two groups (which we'll discuss later)
  - Then, when you have an individual with a specific tumor size and who is a specific age, you can hopefully use that information to place them into one of your classes
- You might have many features to consider
  - Clump thickness
  - Uniformity of cell size
  - Uniformity of cell shape
- The most exciting algorithms can deal with an infinite number of features

QUESTION

- You’re running a company, and you want to develop learning algorithms to address each of two problems.

- Problem 1: You have a large inventory of identical items. You want to predict how many of these items will sell over the next 3 months.
- Problem 2: You’d like software to examine individual customer accounts, and for each account decide if it has been hacked/compromised.

- Should you treat these as classification or as regression problems?

2. UNSUPERVISED LEARNING

- Second major problem type
- In unsupervised learning, we get unlabeled data
  - Just told - here is a data set, can you structure it
- One way of doing this would be to cluster data into groups
  - This is a **clustering algorithm**

2. UNSUPERVISED LEARNING: EXAMPLE OF CLUSTERING ALGORITHM

- Google news: groups news stories into cohesive groups
- Organize computer clusters
  - Identify potential weak spots or distribute workload effectively
- Social network analysis
  - Customer data
- Astronomical data analysis
  - Algorithms give amazing results/theories about how the galaxy is formed.

WHICH WOULD YOU ADDRESS USING AN UNSUPERVISED LEARNING ALGORITHM?

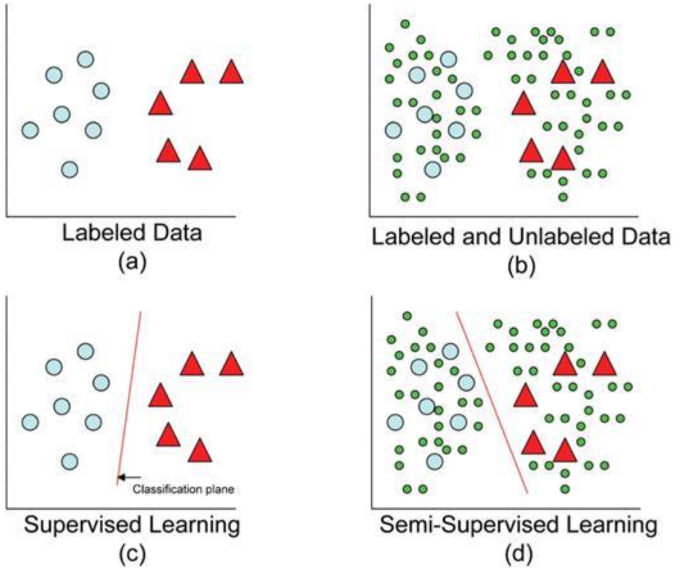
- Given email labeled as spam/not spam, learn a spam filter.
- Given a set of news articles found on the web, group them into set of articles about the same story.
- Given a database of customer data, automatically discover market segments and group customers into different market segments.
- Given a dataset of patients diagnosed as either having diabetes or not, learn to classify new patients as having diabetes or not.

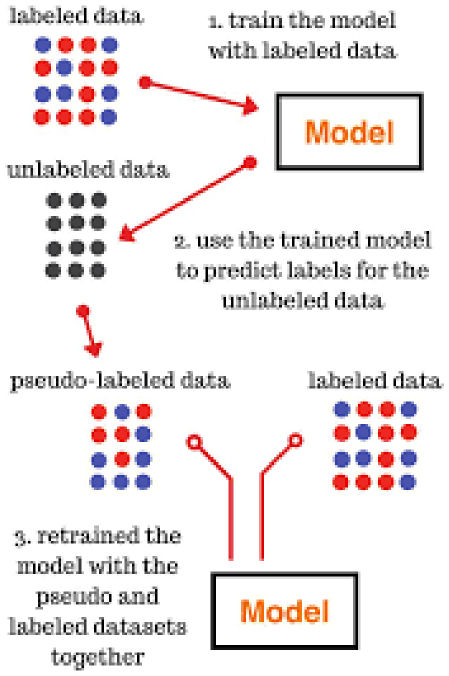
LABELED VS. UNLABELED

- Many applications have unlabeled examples
- Labeled examples are expensive → require human effort
  - NLP: Penn Chinese Treebank → 2 years for 4000 sentences
  - Speech analysis: 400 hours annotation time for 1 hour of speech
  - Medical applications require doctors' opinions, might not be unique!

3. SEMI-SUPERVISED LEARNING

- Falls between unsupervised learning and supervised learning
- Uses labeled and unlabeled data (labeled << unlabeled)
  - Can produce considerable improvement in learning accuracy over supervised/unsupervised learning alone.





labeled data

1. train the model with labeled data

Model

unlabeled data

2. use the trained model to predict labels for the unlabeled data

pseudo-labeled data

labeled data

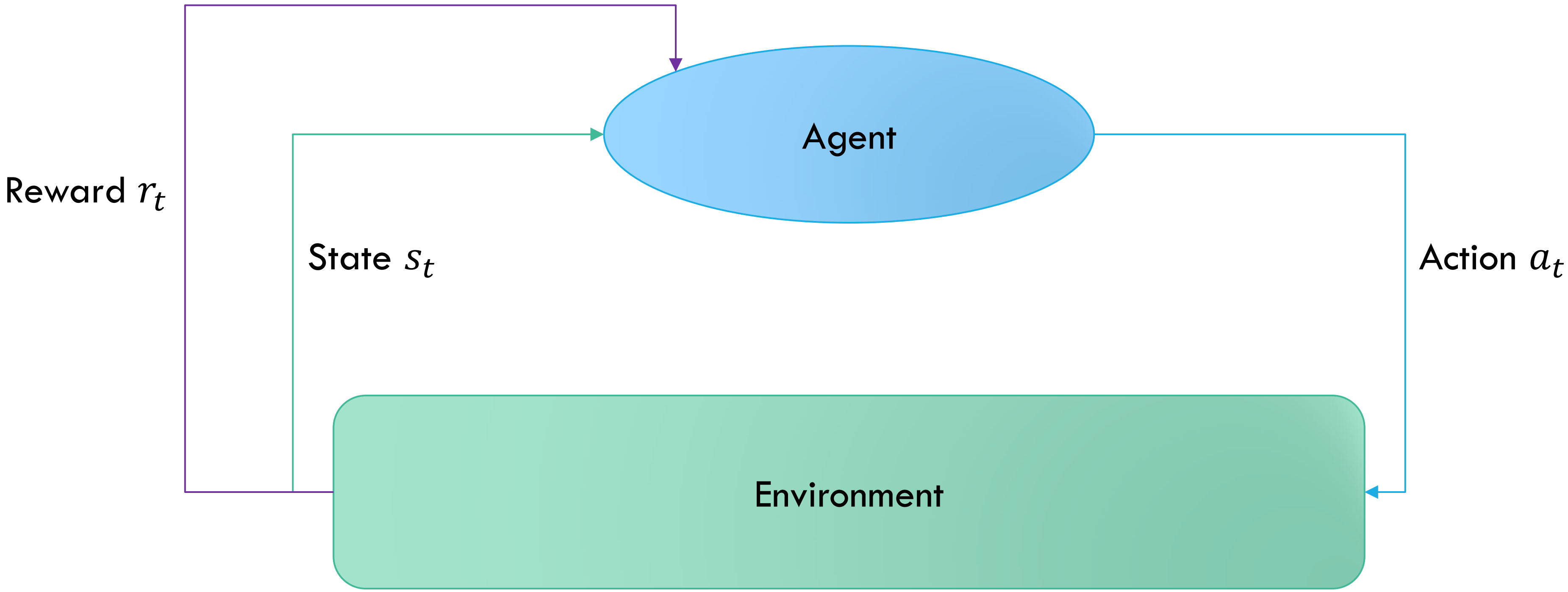
3. retrained the model with the pseudo and labeled datasets together

Model

We can train a classifier on the small amount of labeled data, and then use the classifier to make predictions on the unlabeled data. Since these predictions are likely better than random guessing, the unlabeled data predictions can be adopted as ‘pseudo-labels’ in subsequent iterations of the classifier. While there are many flavors of semi-supervised learning, this specific technique is called **self-training**.

4. REINFORCEMENT LEARNING

- **Goal:** Learn how to take actions to maximize a reward


Agent
Reward r_t
State s_t
Action a_t
Environment