Week 1

Nico Limacher

July 24, 2019

1 Introduction

What is Machine Learning

No well accepted definition, but here's Arthur Samuel back in the day:

> The field of study that gives computers the ability to learn without being explicitly learned.

More recently defined by Tom Mitchell of Carnegie Mellon:

> 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.

In the case of an email spam filter:

T: Classifying emails as spam

P: The proportion of emails correctly marked as spam

E: Observing the user mark certain emails themselves

Some goals:

- Discuss Supervised Learning
- Discuss Unsupervised Learning
- Get practical advice for implementation of such methods

Supervised Learning

Formal Definition The task of learning a function that maps an input to an output based on example input-output pairs. It infers a function from labeled training data consisting of a set of training examples.

Essentially, for every example data point in our set we are told the correct answer.

Example 1: Housing Data in Portland, Oregon

Given data in Figure 1. if you own a house with 750 sq. ft., how much can you expect the house to be worth?

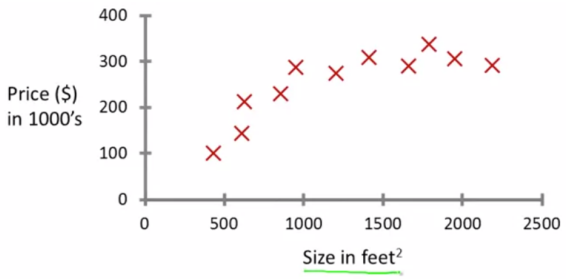

Figure 1: Price vs. Size in Feet

An Idea: Provided we can plot a straight line through the data, we could use that to get a Y-axis value given some value on the X-axis

Perhaps a Better Idea: Rather than a straight line what if we included a quadratic function

How do we decide which to use?

This is an example of a *Regression Problem with Continuous Output*

Example 2: Malignant Tumors

Given data about the size of a tumor can we predict whether it will be benign or malignant This is an example of a *Classification Problem*

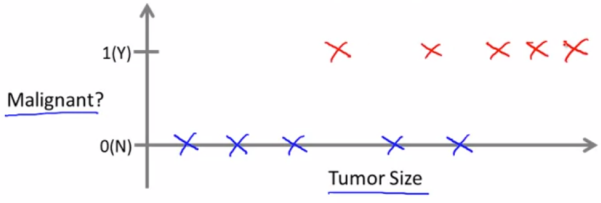

Figure 2: Tumor Size vs. Malignant or Benign

Features In above examples, the features Size in Feet^2 and Tumor Size were used as Inputs to the Machine Learning Algorithm. Mature Machine Learning approaches use many, many features. For example, in the Malignant Tumors exercise: Clump Thickness, Uniformity of Cell Size, and Uniformity of Cell Shape were all considered.

Support Vector Machine A method to support including infinite amounts of features in a model. To discuss in detail later.

Unsupervised Learning

Given data with no classification or labels, what do we do with it? How do we find structure?

Formal Definition

Clustering Algorithm Group data points with no prior knowledge of relationship. Real world example: Google News finding new "stories"

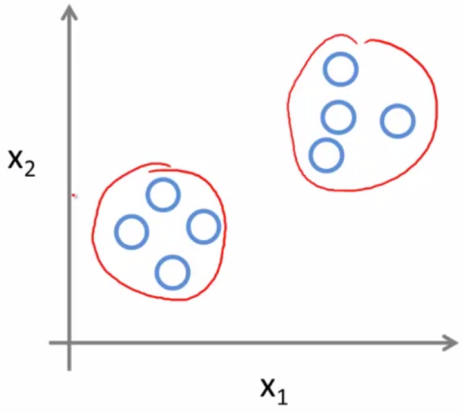

Figure 3: Similar to the Malignant Tumor example, but here we cluster data points on our own with no guidance