

Machine Learning

March 24, 2019

(E)xperience, (T)ask, (P)erformance

1 Supervised Learning

With supervised learning, we are given a data set and already know what our correct output should look like, having the idea that there is a relationship between input and output. Supervised learning problems are categorized into "regression" and "Classification" problems. Regression - The goal of predicting a continuous valued output. Classification problem - The goal to predict a discrete valued output. In a regression problem, we are trying to predict results within a continuous output, meaning we are trying to map our input variables into a continuous function. In a classification problem we are instead trying to predict results in a discrete output.

Example 1:

Given data about houses on the market, try to predict their price. Price as a function of size is a continuous output, thus a regression problem.

This could be made into a classification problem if we made the output whether the house sells for more or less than the asking price.

Example 2:

(a) Regression - Given a picture of a person, we have to predict their age on the basis of the given picture

(b) Classification - Given a patient with a tumor, we have to predict whether the tumor is malignant or benign.

2 Unsupervised Learning

Unsupervised Learning deals with a data set with no given information and tries to group it into a pattern. Unsupervised learning allows us to approach problems with little or no idea what our results should look like. We can derive structure from data where we don't necessarily know the effect of the variables.

3 Model Representation

m = Number of training variables. x 's = "input" variable / features y 's = output variable / target variable (x, y) - one training example $(x^{(i)}, y^{(i)})$ i^{th} training example

training set - \mathcal{D} learning algorithm - J hypothesis

3.1 Representing the hypothesis(h)

h maps from x 's to y 's

$$h_{\theta}(x) = \theta_0 + \theta_1 x$$

shorthand $h(x)$

Linear regression with one variable (univariate linear regression) We will also use X to denote the space of input values, and Y to denote the space of output values. In this example, $X = Y = \mathbb{R}$

4 Cost Function

θ_i 's are parameters. How to choose θ_i 's?

Minimize $\theta_0 \theta_1$ Choose $\theta_0 \theta_1$ so that $h_{\theta}(x)$ is close to y for our training examples (x, y)

$$\frac{1}{2m} \sum_{i=1}^m (h_{\theta}(x^{(i)}) - y^{(i)})^2$$

$$\text{Btw... } (\theta_0(x^{(i)})) = \theta_0 + \theta_1 x^{(i)}$$

In english: You want the sum of i to m (size of training set) of the squared difference between prediction of the hypothesis when it is inputted the (house number i) minus the actual price.

"We can measure the accuracy of our hypothesis function by using a cost function. This takes an average difference (actually a fancier version of an average) of all the results of the hypothesis with inputs from x 's and the actual output y 's."