Chapter 1 What is Machine Learning?

1.1 Introduction

Definition 1.1 (Machine Learning) Field of study that gives computers thr ability to learn without being explicitly programmed. (older, informal)

Improve performance from task experience. (more modern)

- A task
- Some experience
- Some way to measure

Two board type of ML algorithms:

- Supervised Learning
- Unsupervised Learning

Others:

- Reinforcement Learning
- Recommender Systems

1.1.1 Supervised Learning

Supervised learning has some known relationship between input and output.

Supervised learning problems are categorized into regression and classification problems.

Fitting a straight line or quadratic or 2nd-order curve to a function is a regression. The main idea is to give a continuous solution based on data we have.

Example 1.1 Estimate weekly income of a company.

Estimating some possbility to be certain type or classification is also a supervised learning.

Example 1.2 Classify tumor type based on its size.

When the scale of classes turns to infinity, we need Support Vector Machine.

1.1.2 Unsupervised Learning

Unsupervised learning has no specific relationship between input and output. Data has no lables, but the data could be devided into severval clusters.

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.

Example 1.3 Two recordings of audio are mixed, and computer should get them departed.

1.2 Linear Regression with One Variable

1.2.1 Model and Cost Function

For a training set, Notation:

- m: number of training example
- x: input variable / features
- y: output variable / target variable

Learning algorithm applis to training set to find a hypothesis that could gives back a estimated answer.

The goal of supervised learing is to learn a function (hypothesis) from the training set.

1.2.2 Cost Function

In a univariate problems, we have:

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

And the θ_i is so called Parameters.

Regression is to minimize the difference between $h(x)$ and y. The cost function defined as:

$$J(\theta_0, \theta_1) = \frac{1}{2m} \sum_{i=0}^{m} \left(h_\theta(x^{(i)}) - y^{(i)} \right)^2$$

Of course, for each parameters's vector we can compute its J value. Then we can get the global minimum to get the best fit curve. Later, we will learn an efficient algorithm to find the minimum cost.

1.3 Parameter Learning

Have some function $J(\theta_0, \theta_1)$ and want to minimize it. Here is the outline:

- start with some θ_0, θ_1
- change the parameters until the loss has been minimized

We need a way to minimize loss function.

1.3.1 Gradient Descent

At a parameters' place, take a little baby step to change. And choose the step direction that has the quickest reduction speed.

And the process is to repeat the equantion until convergence:

$$\theta_j := \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta_0, \theta_1), \text{where} j = 0, 1$$

And α is called learning rate, which controlls the step of learning. The partical or derivative part controlls the direction.

If α is too small, gradient descent can be really slow. But if it's too large, gradient descent may not converge or even diverge.

With this algorithm and appropriate learning rate, the parameters are always towards an (local) optimal valley.

1.3.2 Gradient Descent for Linaer Regression

With gradient descent, we can handle with the cost function provided in previous sections.

$$J(\theta_0, \theta_1) = \frac{1}{2m} \sum_{i=0}^{m} \left(h_\theta(x^{(i)}) - y^{(i)} \right)^2$$

$$\frac{\partial}{\partial \theta_j} J(\theta_0, \theta_1) = \frac{1}{m} \sum_{i=0}^{m} \frac{\partial}{\partial \theta_i} h_\theta(x^{(i)}) \left(h_\theta(x^{(i)}) - y^{(i)} \right)$$

Informally, convex function has no local optimum but only one global optimum.

Batch gradient descent: each step takes all the training examples.

1.4 Linear Algebra Review

1.4.1 Matrices and Vectors

Definition 1.2 (Matrix) Rectangular array of numbers, dimension of matrix is # of rows times # of columns. For a matrix A, its (i, j) entry is writen as A_{ij}.

Definition 1.3 (Vector) Vector can be treated as a $n \times 1$ matrix in $\mathbb{R}^n$.

1.4.2 Addition and Scalar Multiplication

Matricx addition is to add elements at the same place of their matrix, of course, the matrices should have the same size. (element wise)

Scalar multiplication is to multiply the scalar number to every element in the matrix.

1.4.3 Matrix Vector Multiplication

For a matrix and a vector to be multiplied, they should respectively has size of $m \times n$, $n \times 1$. This relationship of size accords to that the answer is the linear combination of columns of matrix with coefficient of respected vector elements.

1.4.4 Matrix Matrix Multiplication

For 2 matrices to be multiplied, they should respectively has size of $m \times n$, $n \times p$.

1.4.5 Matrix Multiplication Properties

For matrix multiplication, it's a way to pack a series of hypothesis which has the same structure and different parameters.

1.4.6 Inverse and Transpose

Not all numbers have an inverse. So do matrices.

Only square matrices may have inverse. Else matrices could have pseudo inverse.

Words

quadratic 二次的
ellipse 椭圆
contour 等高线，轮廓
by convention 按照惯例