Regression

Debrup Chakraborty

To be covered today

- Linear Regression
- Probabilistic Interpretation of Linear Regression
- Logistic Regression

Material covered is mostly from course notes of Prof. Andrew Ng on regression.
Can be found at:
http://www.stanford.edu/class/cs229/notes/cs229-notes1.pdf

Supervised Learning

- We are given a training set $L = \{(\boldsymbol{x}_i, y_i) : i = 1 \ldots n, \boldsymbol{x}_i \in \mathcal{R}^p, y_i \in \mathcal{R}\}$
- Our goal is to find a good hypothesis h such that $h(\boldsymbol{x})$ is a good predictor for the corresponding value of y.
- When the target variable y takes continuous values as in the above case, we call the learning problem a function approximation problem.
- If y takes discrete values then we call the problem a classification problem.

The structure of h

- To begin with, we need to decide a structure of h
- To start with we assume that h is a linear function of $\boldsymbol{x}$, i.e.,

$$
\begin{aligned}
h_\theta(\boldsymbol{x}_i) &= \theta_0 + \theta_1 x_{i,1} + \theta_p x_{i,2} + \dots \theta_p x_{i,p} \\
&= \sum_{j=0}^{p} \theta_j x_{i,j} \\
&= \boldsymbol{\theta}^T \boldsymbol{x}_i
\end{aligned}
$$

Assuming that $\boldsymbol{x}_{i,0} = 1, \forall i$

The Structure of h (contd.)

- This structure of h

$$h_\theta(\boldsymbol{x}_i) = \boldsymbol{\theta}^T \boldsymbol{x}_i \tag{1}$$

  depends on the parameter vector $\boldsymbol{\theta}$.
- Such a representation of h is called a parametric representation.
- Now the problem boils down to finding the parameter vector $\boldsymbol{\theta}$ such that the function h fits the data the best.

Linear Regression

- Given the training set L, how do we learn the parameters $\boldsymbol{\theta}$.
- One of the reasonable methods would be to make $h(\boldsymbol{x})$ close to y for at least the training set.
- An intuitive cost function for this purpose would be:

$$J(\boldsymbol{\theta}) = \frac{1}{2}\sum_{i=1}^{n}(h_\theta(\boldsymbol{x}_i) - y_i)^2$$

- This function is called the least squares function.
- Our task is to find that $\boldsymbol{\theta}$ which minimizes J.

Gradient Descent Algorithm

- To begin with, we start with an iterative algorithm.
- We start with an initial guess of $\boldsymbol{\theta}$ and in each step change theta to make $J(\boldsymbol{\theta})$ smaller.
- This can be done by the **gradient descent** algorithm which gives the update rule as

$$\theta_j)_{new} \leftarrow \theta_j)_{old} - \alpha \frac{\partial}{\partial \theta_j} J(\boldsymbol{\theta})$$

Here α is called the learning rate.

Gradient Descent (contd.)

- For our specific cost function J, and for a single training example $(\boldsymbol{x}_i, y_i)$ the update rule becomes

$$\theta_j)_{new} \leftarrow \theta_j)_{old} + \alpha \left(y_i - h(\boldsymbol{x})\right) x_{i,j}$$

  How?

- This update rule is called
  - **Least Mean Squares** (LMS) update rule
  - **Widrow-Hoff** learning rule

Gradient Descent (contd.)

This rule can be extended for the case of multiple training data in two *obvious ways*:

- The batch gradient descent
- Stochastic gradient descent

Batch Gradient Descent

Algorithm:

repeat until convergence

{

$$\theta_j)_{new} \leftarrow \theta_j)_{old} + \alpha \sum_{i=1}^{n} (y_i - h(\boldsymbol{x}))\, x_{i,j}, \forall j$$

}

Stochastic Gradient Descent

Algorithm:

repeat until convergence
for $i = 1$ to n
for $j = 0$ to p
$\theta_j)_{new} \leftarrow \theta_j)_{old} + \alpha\,(y_i - h(\boldsymbol{x}))\,x_{i,j}$
end for
end for
end repeat

A Closed Form Solution

- The gradient descent is not a specific method to solve the linear regression problem but can be applied to other problems also.
- The linear regression problem has a closed form solution, which we shall state without proof.
- Let X be the design matrix and Y the responses. Then the value of $\boldsymbol{\theta}$ that minimizes J is given by

$$\boldsymbol{\theta} = (X^T X)^{-1} X^T Y$$

A Probabilistic Interpretation

- Here we take another view of the linear regression problem.
- We find an answer to the question:
  Why the least-squares cost function is a reasonable one
- We will show that under certain reasonable probabilistic assumptions the least squares method has a natural interpretation.

A Prob. Interpretation (contd.)

- We assume that the target variables and the inputs are related via the equation

$$y_i = \boldsymbol{\theta}^T \boldsymbol{x}_i + \epsilon_i$$

- ϵ_i is an error term which takes care of:
  - Unmodeled effects
  - Random Noise
- We assume that the ϵ_i are distributed IID (independent and identically distributed) according to the Gaussian distribution with zero mean and a variance σ^2.

A Prob. Interpretation (contd.)

- Thus we can write,

$$\epsilon_i \sim \mathcal{N}(0, \sigma^2)$$

- Hence, the probability density of ϵ_i will be

$$p(\epsilon_i) = \frac{1}{\sqrt{2\pi}\sigma} exp\left(-\frac{\epsilon_i^2}{2\sigma^2}\right).$$

- This implies that

$$p(y_i|\boldsymbol{x}_i;\boldsymbol{\theta}) = \frac{1}{\sqrt{2\pi}\sigma} exp\left(-\frac{(y_i - \boldsymbol{\theta}^T\boldsymbol{x}_i)^2}{2\sigma^2}\right).$$

A Prob. Interpretation (contd.)

- Given X (all $\boldsymbol{x}_i$) and $\boldsymbol{\theta}$ the probability of the data is given by $p(Y|X;\boldsymbol{\theta})$
- This quantity when viewed as a function of $\boldsymbol{\theta}$ is called the likelihood function.
- By the independence assumption of ϵ_i we can write the likelihood function as

$$\begin{aligned} L(\boldsymbol{\theta}) &= \prod_{i=1}^{n} p(y_i|\boldsymbol{x}_i;\boldsymbol{\theta}) \\ &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} exp\left(-\frac{(y_i - \boldsymbol{\theta}^T\boldsymbol{x}_i)^2}{2\sigma^2}\right) \end{aligned}$$

A Prob. Interpretation (contd.)

- Given this probabilistic model what is the best way to choose $\boldsymbol{\theta}$?
- According to the principle of **maximum likelihood**, we should choose $\boldsymbol{\theta}$ so as to make the data most likely. Thus, we should choose that $\boldsymbol{\theta}$ which maximizes $L(\boldsymbol{\theta})$.
- Maximizing $L(\boldsymbol{\theta})$ is same as minimizing $J(\boldsymbol{\theta})$. Why??