Notes on Linear Regression

Liangjie Hong

May 19, 2010

1 Basic Model

Let us assume the input is a set of vectors $x_i = \{x_{i1}, x_{i1}, ..., x_{in}\}$ where each x_{ij} is a item in feature vector (For convenience, $x_{i1} = 1$). For each input vector, we have a real-valued output variable y_i associated to the input. We assume that each output variable is "generated" by the input through the following linear equations:

$$y_i = f(\vec{x_i}) = \sum_{j=0}^{n} \theta_j x_{ij} \tag{1}$$

Our problem is how to determine the value of each coefficient θ_j. In order to estimate those values, we usually use a cost function which aims to minimize the residuals:

$$C = \arg\min_{\theta} \ \frac{1}{2} \sum_{i=0}^{m} (y_i - f(\vec{x_i}))^2 = \arg\min_{\theta} \ \frac{1}{2} \sum_{i=0}^{m} \left(y_i - \theta_0 - \sum_{j=1}^{n} \theta_j x_{ij} \right)^2 \tag{2}$$

2 Iterative Algorithms

In order to solve Equation 2, in this section, we introduce two iterative algorithms. Both algorithms are based on the simple updating rule:

$$\theta_j \Leftarrow \theta_j - \alpha \frac{\partial C}{\partial \theta_j}$$

, which requires the gradients as follows:

$$\begin{aligned} \frac{\partial C}{\partial \theta_j} &= [y_i - f(\vec{x_i})](-x_{ij}) \\ &= (f(\vec{x_i}) - y_i)x_{ij} \end{aligned}$$

Therefore, the first algorithm called **batch gradient descent** is shown below:

```
Repeat until convergence {
    $\theta_j \Leftarrow \theta_j + \alpha \sum_{i=0}^{m} (y_i - f(\vec{x_i}))x_{ij}$   (for every $j$)
}
```

The second algorithm is called **stochastic gradient descent**, shown as follows:

```
Loop {
    for $i = 1$ to $m$ {
    $\theta_j \Leftarrow \theta_j + \alpha(y_i - f(\vec{x_i}))x_{ij}$   (for every $j$)
    }
}
```

3 L2 Regularization

For some problems, we would like to impose a penalty on the size of the coefficients. One popular method is to use L2 norm penalty. L2 regularized linear regression is also called Ridge regression. We can write the cost function as follows:

$$C = \arg\min_{\theta} \ \frac{1}{2}\sum_{i=0}^{m}\left(y_i - \theta_0 - \sum_{j=1}^{n}\theta_j x_{ij})\right)^2 + \frac{\lambda}{2}\sum_{j=1}^{n}\theta_j^2 \tag{3}$$

Note, we do not shrink θ_0 and instead we use the following pre-process steps:

- $x_{ij} \Leftarrow x_{ij} - \overline{x_j}$
- $\theta_0 \Leftarrow \sum_{i=0}^{m} y_i/M$

In other words, we estimate all coefficients without intercept. We can also use both algorithms introduced above to obtain the solutions. Here, we show the gradients as follows:

$$\begin{aligned}\frac{\partial C}{\partial \theta_j} &= [y_i - f(\vec{x_i})](-x_{ij}) + \lambda\theta_j \\ &= (f(\vec{x_i}) - y_i)x_{ij} + \lambda\theta_j\end{aligned}$$

Therefore, the basic updating rule is:

$$\theta_j \Leftarrow \theta_j + \alpha[(y_i - f(\vec{x_i}))x_{ij} - \lambda\theta_j]$$

4 Matrix Representation

We can re-write Equation 2 in matrix form as follows:

$$C = \arg\min_{\theta} \ \frac{1}{2}(\mathbf{y} - \mathbf{X}\theta)^T(\mathbf{y} - \mathbf{X}\theta)$$

Rather than using iterative algorithms, we can indeed obtain the closed form solutions as follows:

$$\begin{aligned}\frac{\partial C}{\partial \theta} &= \mathbf{X}^T(y - \mathbf{X}\theta) = 0 \\ &\Rightarrow \hat{\theta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}\end{aligned}$$

Similarly, we can have the matrix form for Ridge regression as follows:

$$C = \arg\min_{\theta} \ \frac{1}{2}(\mathbf{y} - \mathbf{X}\theta)^T(\mathbf{y} - \mathbf{X}\theta) + \frac{\lambda}{2}\theta^T\theta$$

We also obtain the closed form solution for Ridge regression as follows:

$$\begin{aligned}\frac{\partial C}{\partial \theta} &= \mathbf{X}^T(y - \mathbf{X}\theta) + \lambda\theta = 0 \\ &\Rightarrow \hat{\theta} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y}\end{aligned}$$

5 Probabilistic Interpretation

Before we derive the probabilistic interpretation of linear regression, we first look at one property of Normal Distributions. If a random variable X has a normal distribution $N(\theta, \sigma^2)$, a new random variable $aX + b$ has a normal distribution $N(a\theta + b, a^2\sigma^2)$. Now, let us focus on our assumption, the response y_i is a linear function of a set of features $\vec{x}$:

$$y_i = \sum_{j=0}^{n} \theta_j x_{ij} + \epsilon_i \tag{4}$$

This Equation is slightly different from our original Equation 1 that we have an additional term ϵ_i to represent the error between our estimation and the true value. Now, let us assume that all errors has a Normal Distribution $N(0, \sigma^2)$ where the variance is a fixed value. Therefore, we can immediately know that y_i will also follow a Normal Distribution $N(\theta^T x, \sigma^2)$ by using the property we introduced at the beginning of the section:

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

This is only for one sample. The probability for the whole data set (likelihood) is as follows:

$$L(\theta) = \prod_{i=0}^{m} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y_i - \theta^T x_i)^2}{2\sigma^2}\right)$$

We want to maximize this probability and the estimator obtained is usually called Maximum Likelihood Estimator (MLE). Here, we work with log of the likelihood function:

$$\log L(\theta) = m \log \frac{1}{\sqrt{2\pi}\sigma} - \frac{1}{\sigma^2}\frac{1}{2}\sum_{i=0}^{m}(y_i - \theta^T x_i)^2$$

Note, the first term is a constant if we treat σ^2 as a fixed value and therefore maximizing $\log L(\theta)$ is equivalence to minimizing:

$$\sum_{i=0}^{m}(y_i - \theta^T x_i)^2$$

which we recognize to be our basic model (Equation 2).