Linear Regression

1. Introduction

Machine Learning is a huge topic in which numerous problems are included, eg. regression, classification, etc. Whatever the problem is, the basic task of machine learning is extracting a *relationship* from data. In regression tasks, the relationship takes the form of $\mathbf{y} = f(\mathbf{x})$. It could be understood as predicting some target $\mathbf{y}$ given some summary of the observations $\mathbf{x}$. We typically consider the target

$\mathbf{y} \in \mathbb{R}$ as some sort of quantity that we'd like to know and $\mathbf{x} \in \mathbb{R}^k$ as some sort of numerical observations. For example, having a basic idea of the house prices of some area is necessary if someone would like to purchase one in this area. Given some observations such as area, number of bedrooms, restaurants around it, etc, it is possible to estimate the price. Here comes the usage of machine learning techniques! It's really cool to have an accurate estimation given some information of the house, isn't it? In a typical machine learning problem, we have no knowledge about this true relationship. All we have is data $\mathcal{D} = \{x_i, y_i\}_{i=1}^n$ and we'd like to infer or approximate this true underlying relationship by designing and performing algorithms. Back to our regression task, our final product would be $\hat{\mathbf{y}} = h(\mathbf{x})$ called **hypothesis** which is our approximation of $f(\cdot)$. Since it is intractable to learn an arbitrary function, we usually restrict our solutions to a **hypothesis class** $\mathcal{H}$ by imposing some assumptions. To be more specific, we usually employ a **parametric model** meaning that there is a finite-dimensional vector $\mathbf{w} \in \mathbb{R}^d$ determines the behavior of the model. The elements of $\mathbf{w}$ is called weights or parameters. We usually use a **Loss Function** to measure the performance of our models. Instructed by the loss function, we could search the best solution to the task. We'll start with **linear regression**, one of the most classic algorithms in Machine Learning and hope you guys enjoy it.

2. Ordinary Least Squares(OLS)

OLS is one of the simplest problems in regression tasks.

Here is the basic background of linear regression: With the basic assumption that the *relationship* between y and x is a linear one, that is we assume: $y_i = w^T x_i$, writing in matrix takes the form:

$$\mathbf{y} = Xw \tag{1}$$
$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \quad X = \begin{pmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_n^T \end{pmatrix}$$

Here $\mathbf{y} \in \mathbb{R}^n$ is a *column vector* where each element is the y_i previously and each row of X is our x_i. Sometimes, $X \in \mathbb{R}^{n \times d}$ is also called *design matrix*. Please note that the convention we use here is usually the implicit assumption and it is redundant to mention it every time. So we can come up with the *loss function* of the problem: $\mathcal{L} = \frac{1}{n}\sum_{i=1}^{n}(y_i - w^T x_i)^2$. It is also called *MSE(Mean Squared Error)* in machine learning. Writing the loss function in a matrix form, we can get:

$$\mathcal{L} = \frac{1}{n}\|\mathbf{y} - Xw\|_2^2 \tag{2}$$

Since the loss measures how bad our model approximates the real data, our goal is to minimize $\mathcal{L}$ in (2).All the remaining is how to solve (2). We introduce 2 ways in the following sections.

3. Vector Calculus

The first way we introduce is using vector calculus to find the optimal solution to (2). I'll briefly introduce vector calculus in Appendix at the end of this post. But first let's get into the loss function together!

I'd like first to clear that $\mathcal{L}$ in (2) is a function about w since all of the others in (2) are given data. Also, I'd like to introduce two vector calculus formulas that we will use later. If you'd like to know the proof, see the appendix!

$$x, w \in \mathbb{R}^k, M \in \mathbb{R}^{k \times k}$$
$$\nabla_x x^T M x = 2Mx$$
$$\nabla_x w^T x = w$$

Then we could easily find out that $\mathcal{L}$ is a *convex function* of w. Therefore, the optimal solution w^* is the one that where the gradient of $\mathcal{L}$ w.r.t. w is 0:

$$\begin{aligned} \arg\min_w \mathcal{L} &= \arg\min_w \|\mathbf{y} - Xw\|_2^2 \\ &= (\mathbf{y} - Xw)^T(\mathbf{y} - Xw) \\ &= w^T X^T X w - 2\mathbf{y}^T X w + \mathbf{y}^T \mathbf{y} \end{aligned} \tag{3}$$

$$\nabla_w \|\mathbf{y} - Xw\|_2^2 = 2X^T X w - 2X^T \mathbf{y} \tag{4}$$

The vector calculus formulas provided above is used in (4). Assume $X^T X$ is invertible, we can get our final solution:

$$\nabla_{w^*} \|\mathbf{y} - Xw\|_2^2 = 0$$
$$2X^T X w^* - 2X^T \mathbf{y} = 0$$

$$w^* = (X^T X)^{-1} X^T \mathbf{y} \tag{5}$$

4. Orthogonal Projection

Another approach for solving (2) is a linear algebraic one. Before we really step into this approach, some theorems of linear algebra must be first understood.

Assume $S, S^\perp \subset \mathbb{R}^m$ where each of them are orthogonal complement of each other. Consider arbitrary vector $v \in \mathbb{R}^m$, we have

$$v = v_S + v_{S^\perp}, v_S \in S, v_{S^\perp} \in S^\perp$$

Let $P_s v$ be the project of v to S, then

$$\|v - P_s v\|_2 \leq \|v - s\|_2, \ \forall s \in S \tag{6}$$

The inequality is tight if and only if $s = P_s v$

Proof:

$$\begin{aligned} \|v - s\| &= \|v - P_s v + P_s v - s\| \\ &= \|v - Ps_v\| + \|P_s v - s\| \\ &\geq \|v - Ps_v\| \end{aligned}$$

The last two step uses Pythagorean theorem since $v - P_s v \in S^\perp, P_s v - s \in S$

Also, knowing Fundamental Theorem of Linear Algebra(FTLA) is necessary:

Here we'll only use

$$range(X)^\perp = null(X^T)$$

Back to our problem: (2), $Xw \in range(X)$ and we could view it as a vector in the space spanned by the column vectors of X (column space). From (6) we could know that the minimum of $\|\mathbf{y} - Xw\|$ is achieved when $Xw = P_X \mathbf{y}$. Also, by FTLA, we have

$$\begin{gathered} \mathbf{y} - Xw^* = \mathbf{y} - P_X \mathbf{y} \\ \mathbf{y} - P_X \mathbf{y} \in range(X)^\perp \\ range(X)^\perp = null(X^T) \\ \mathbf{y} - Xw^* \in null(X^T) \end{gathered}$$

Finally, by the definition of null space, we have

$$X^T(\mathbf{y} - Xw^*) = \mathbf{0}$$

Here we could achieve exactly the same solution as (5)

What's worth mentioning is that if X is a full rank matrix, then Xw should be unique since the column vectors of X is linearly independent from each other. The solution to the problem (2) is thus unique under this condition. However, if X is not full rank, the solution may be multiple and we'll deal with it later.

5. Ridge Regression

6. Appendix: Vector Calculus