Linear Regression

compiled by Alvin Wan from Professor Jitendra Malik's lecture

We will begin by a series of assumptions for our training set. The conditional density $p(y|x)$ to be Gaussian.

$$\Pr(y|x) \sim \mathcal{N}(w^T x, \sigma^2)$$

We will also consider the outputs to be a linear combination of the weights with x, with some noise $\epsilon \sim \mathcal{N}(0, \sigma^2)$ applied.

$$y = w^T x + \epsilon$$

We will now consider our training set, which is composed of x_i, a vector in $\mathbb{R}^d$, and y_i, a scalar. Let the matrix of x_i be called X. In this lecture, we will explore two interpretations of our linear regression solution, first with X row-major and then with X column-major.

1 Perspective 1 : Maximum Likelihood Estimation

As it turns out, maximum likelihood estimate of the Gaussian distribution is exactly the solution to least squares. This will give us the interpretation of the solution to least-squares. Consider the probability of our distribution given the data. Since all x_i are i.i.d., we can take the product of the probabilities of distributions given each x_i.

$$\begin{aligned}\Pr(X|\theta) &= \Pi_{i=1}^n \Pr(y_i|x_i, \theta) \\ &= \Pi_{i=1}^n \frac{1}{2\pi\sigma^2} \exp(-\frac{1}{2\sigma^2}(y_i - w^T x_i)^2)\end{aligned}$$

We will now take the log-likelihood.

$$\log \Pr(X|\theta) = \sum_{i=1}^{n} \log(\frac{1}{\sqrt{2\pi\sigma^2}}) - \sum_{i=1}^{n} \frac{1}{2\sigma^2}(y_i - w^T x_i)^2$$

The first term is a constant, so maximizing likelihood is equivalent to minimizing only the second term.

$$\text{minimize}_w \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - w^T x_i)^2$$

Let us express $\sum_{i=1}^{n}(y_i - w^T x_i)^2$ in vector-matrix form.

$$\begin{bmatrix} y_1 - w^T x_1 \\ \vdots \\ y_n - w^T x_n \end{bmatrix} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} - \begin{bmatrix} x_1^T \\ \vdots \\ x_n^T \end{bmatrix} w$$

Consider A, an $n \times d$ matrix that contains all of the x_i. A is sometimes called the "design matrix", and in this case, we see that $y_i - w^T x_i = y - Aw$, so the above expression is equivalent to the below.

$$\text{minimize}_w \frac{1}{2\sigma^2} \sum_{i=1}^{n} \|y - Aw\|^2$$

We can ignore σ^2, since we're minimizing the quantity. Additionally, we will expand $\|y - Aw\|^2$.

$$\begin{aligned} &= \frac{1}{2} \sum_{i=1}^{n} (y - Aw)^T (y - Aw) \\ &= \frac{1}{2} \sum_{i=1}^{n} (\|y\|^2 - 2(Aw)^T y + \|Aw\|^2) \end{aligned}$$

Differentiating with respect to w gives us the gradient, below.

$$\nabla_w \alpha = -A^T y + A^T A w$$

Differentiating again with respect to w gives us the Hessian below.

$$\nabla_w^2 \alpha = A^T A$$

Since the Hessian is positive semidefinite, we have that the solution to $\nabla_w \alpha = 0$ will give us a *minimum*. So, let us now solve for w.

$$\begin{aligned} \nabla_w \alpha &= 0 \\ A^T A w &= A^T y \\ \hat{w} &= (A^T A)^{-1} A^T y \end{aligned}$$

2 Perspective 2 : Linear Algebra

We have another interpretation, using more linear algebra intuition. First, we will rewrite $y \approx Aw$, ignoring noise..

$$\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \approx \begin{bmatrix} | & | & & | \\ a_1 & a_2 & \cdots & a_d \\ | & | & & | \end{bmatrix} \begin{bmatrix} w_1 \\ \vdots \\ w_d \end{bmatrix}$$

On the right hand side, we have that Aw is equivalent to the following.

$$Aw = \sum_{i=1}^{d} a_i w_i$$

This means that Aw is actually a linear combination of the column vectors in A, weighted by the entries of the vector w. Thus, for any choice of w, we find an Aw that exists in the **column space** of A. Recall that the column space of a matrix A is the vector space spanned by its column vectors. This means that if y exists in the column space of A, then there is a solution to w. Otherwise, we pick the closest w. Consider the "error" vector, after picking the best w.

$$e = y - \hat{y} = y - Aw$$

If y exists in the column space, the error vector will be 0. Note that if e has some component that isn't perpendicular to the column space of A, we can perturb w so that e decreases. Thus, if this is true for all components along the column space of A, we see that e must be perpendicular to A. Since e is orthogonal to A, we obtain the following equality.

$$\begin{aligned} A^T e &= 0 \\ A^T(y - Aw) &= 0 \\ A^T y - A^T Aw &= 0 \\ A^T y &= A^T Aw \\ w &= (A^T A)^{-1} A^T y \end{aligned}$$

3 Variants

We can obtain two different variants by adding a **penalty** or **regularization** term to the objective function.

1. Ridge Regression or L2 Regularization

$$\text{minimize} \frac{1}{2} \|y - Aw\|^2 + \lambda \|w\|^2$$

It is easy to compute the gradient vector to find that the optimal solution is the following.

$$\hat{w} = (A^T A + \lambda I)^{-1} A^T y$$

As $\lambda \to 0$, we note that this becomes identical to the least-squares objective function and least-squares solution. However, the addition of λI makes the solution numerically stable, as the positive semidefinite $A^T A$ is not always guaranteed to have an inverse. With that said, adding λI guarantees that $A^T A + \lambda I$ is positive definite and therefore invertible.

2. Lasso or L1 Regularization

$$\text{minimize} \frac{1}{2} \|y - Aw\|^2 + \lambda \|w\|_1$$

We cannot differentiate the lasso objective function, because the 1-norm term is not differentiable at 0. However, the function is convex, so we can apply gradient descent to achieve the global minimum.

We will explore the effects of Regularization in Note 13, which primarily prevents overfitting.