6.7900 Fall 2024: Lecture Notes 4

Revision: 9/20/24 3:29PM

1 Linear Regression

Why linear regression? Modern, sophisticated robots do not make hammers obsolete; similarly, linear regression is still a widely used method in modern machine learning. Benefits of linear regression includes its speed, interpretability, and theoretical insights to fields such as deep learning. Given that many models are essentially black boxes, linear regression stands out as an important way to provide direct views of the underlying distribution. It also serves as an important baseline and ablation in machine learning research (Lipton, Steinhardt 2019, Troubling Trends in Machine Learning).

We want to understand the data from a machine learning standpoint rather than relying on a back box, especially when linear regression can do just as good as more complicated and resource-exhausting methods.

Motivating Example Advertising data: "sales, in thousands of units, as a function of TV, radio, and newspaper budgets, in thousands of dollars, for 200 different markets"

- Data point: $y^{(n)} \in \mathbb{R}_+$
- Features: $x^{(n)} = [x_1^{(n)}, x_2^{(n)}, x_3^{(n)}]^T \in \mathbb{R}_+^3$
- We want to predict sales under different advertising budgets

Proposition: Consider regression with $X = \mathbb{R}^D, Y = \mathbb{R}$ and square loss $L(a, g) = (a - g)^2$, then decision rule minimizes risk of a new point $h(x) = \mathbb{E}[Y|X = x]$.

Approach Make a model for $p(y|x)$ then use maximum likelihood estimate of the parameter. Due to the form of $h(x)$, we don't need to worry about a model for $p(x)$. We have the likelihood $y^{(n)} = \theta^T x^{(n)} + \epsilon^{(n)}$, where $\epsilon^{(n)} \sim \mathcal{N}(0, \sigma^2)$.

Let's check the dimensions of our variables: y is 1×1, x is a $D \times 1$ vector, which implies that the dimension of θ is $D \times 1$. ϵ is a 1x1 scalar.

Recall that "all models are wrong"; what about this one here? In fact, we see the noise of the distribution is not the same everywhere, and it grows as the x-axis grows. Additionally, we assumed that the noise in our model is distributed with center at 0, but sales are strictly positive and cannot be negative.

We would like to have an intercept term, which can be done by taking $x_1^{(n)} = 1$ for any n.

An equivalent formulation is

$$y^{(n)} \mid x^{(n)} \sim \mathcal{N}(\theta^T x^{(n)}, \sigma^2)$$

$$p(y \mid x, \theta, \sigma) = \mathcal{N}(y \mid \theta^T x, \sigma^2)$$

Then, we take the MLE approach and we have our optimization problem

$$\hat{\theta}, \hat{\sigma}^2 = argmax_{\theta,\sigma^2} p(\{y^{(n)}\}_1^N | \{x^{(n)}\}_1^N, \theta, \sigma^2)$$

and the prediction we want can be formulated as

$$h(x) = \mathbb{E}[Y|X = x] = \hat{\theta}^T x$$

which depends only on $\hat{\theta}$ but not on $\hat{\sigma}$!

We have chosen a model that led to a predictor, but we still haven't solved the optimization problem that gives $\hat{\theta}$.

2 MLE & ERM

Maximizing likelihood Recall that we want to maximize

$$p(\{y^{(n)}\}_1^N | \{x^{(n)}\}_1^N, \theta, \sigma^2) = \Pi_{n=1}^N p(y^{(n)}|x^{(n)}, \theta, \sigma^2)$$

$$= \Pi_{n=1}^N \mathcal{N}(y^{(n)}|\theta^T x^{(n)}, \sigma^2)$$

$$-\Pi_{n=1}^N \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(y^{(n)} - \theta^T x^{(n)})^2/(2\sigma^2)}$$

Note that maximizing log likelihood is equivalent to minimizing the negative log likelihood:

$$-\log p(\{y^{(n)}\}_1^N | \{x^{(n)}\}_1^N, \theta, \sigma^2)$$

$$= \frac{N}{2} \log \sigma^2 + (2\sigma^2)^{-1} \sum_{n=1}^{N} (y^{(n)} - \theta^T x^{(n)})^2 + c$$

where c is a constant.

Recall that we only need $\hat{\theta}$ but not $\hat{\sigma}^2$. So in order to minimize θ, it is enough to minimize

$$\sum_{n=1}^{N} (y^{(n)} - \theta^T x^{(n)})^2$$

We call the difference between observed value and prediction the residual. A residual is the general concept defined as

$$y^{(n)} - h(x^{(n)})$$

Here, we have the signed residual inside the sum, so we call the entire objective "residual sum of squares", or RSS.

Now, let's do some visualization before diving into the actual optimization (see lecture slides). Looking at the plot, we see that our assumption of noise seems to be off; the noise tend to increase with y rather than staying constant. Analyzing the residual plot of a fitted predictor is a good way to check model assumptions, and here we see the residuals increasing with y. A simple solution to this problem is to describe y in log scale, and the residuals are much closer to linear in this case.

Now look at the case of two feature dimensions, where the prediction lies in a hyperplane. Note that regressing label on the features is not equivalent to regressing one feature on label and remaining features.

Empirical risk minimization Recall if we have squared loss as our empirical risk:

$$\frac{1}{N} \sum_{n=1}^{N} L(y^{(n)}, h(x^{(n)})) = \frac{1}{N} \sum_{n=1}^{N} (Y^{(n)} - h(x^{(n)}))^2$$

This is not that useful if we allow all decision rules h. We can either 1. change the distribution of data, or 2. restrict h. What if we only allow linear $h(x) = \theta^T x$? Then minimizing the empirical risk minimizes

$$\sum_{n=1}^{N} (y^{(n)} - \theta^T x^{(n)})^2$$

which is the exact same objective from maximum likelihood! Note that this similarity is specific to this model and may not always be the case. We can also call this objective as mean squared error, or MSE.

3 Optimization & Closed Form Linear Regression

Notation Let

$$X = [x^{(1)}, x^{(2)}, ..., x^{(n)}]^T$$

which has dimension $N \times D$, and

$$Y = [y^{(1)}, ..., y^{(n)}]^T$$

which has dimension $N \times 1$. Then

$$RSS(\theta) = (X\theta - Y)^T(X\theta - Y)$$

Exercise: Check the dimension of this expression of RSS and confirm that it gives the same result as the previous formulation.

Optimization We want to minimize

$$RSS(\theta) = (X\theta - Y)^T(X\theta - Y)$$

If the function is convex and the derivative is 0 at some point, then it is the minimum we want.

Notation: for $f(\theta)$, $\nabla_\theta f$ is the $D \times 1$ vector with dth element $\frac{\partial f}{\partial \theta_d}$. the Hessian $\nabla^2_\theta f$ is $D \times D$ matrix with i, jth element $\frac{\partial^2 f}{\partial \theta_i \partial \theta_j}$. First order condition gives

$$\nabla RSS(\theta) = 2X^T(X\theta - Y) = 0$$

Exercise: Check that you can derive the above.

Note that the leftmost term $\nabla RSS(\theta)$ is a $D \times 1$ vector, and we are asking for the whole vector to be 0.

To ensure that we get a minimum rather than a maximum, we need to check the second order condition

$$\nabla^2 RSS(\theta) = 2X^T X > 0$$

In other words, the Hessian needs to be positive definite. For now, suppose that $N > D$ and X is full rank. Then, $X^T X$ is positive definite and invertible. Let's solve the first order conditions:

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

We call this the "ordinary least squares", or OLS for short. Note that a closed form solution like this is not always desirable; matrix inversion can get expensive when working with large dimensions.

Visualization in next lecture.