170S Week 5 Discussion Notes

Colin Ni

February 3, 2025

The lectures and textbook discuss linear regression for samples living in one dimension. We will discuss the nicer and more general theory for multiple dimensions, which you will learn in Math 156 (machine learning), should you choose to take it (you should). The hope is that this will elucidate some of the gross formulas you may have seen so far. In particular, we will not discuss sufficient statistics or Bayesian statistics this week.

Linear regression

Consider the following model:

$$y(\mathbf{x}) = \alpha + \mathbf{x}\beta,$$

where the input $\mathbf{x} \in \mathbb{R}^{1\times d}$ is a length d row vector, the parameter $\alpha \in \mathbb{R}$ is a scalar, and the parameter $\beta \in \mathbb{R}^d$ is a length d column vector. This model expresses a linear relationship. In particular, the graph of y is a d-dimensional (affine hyper-) plane living in $\mathbb{R}^{d+1}$.

Warmup. Plot the points

$$(2.0, 1.3), (3.3, 3.3), (3.7, 3.3), (2.0, 2.0), (2.3, 1.7),$$
$$(2.7, 3.0), (4.0, 4.0), (3.7, 3.0), (3.0, 2.7), (2.3, 3.0).$$

Graph the model y for $d = 1$ for the parameters $\alpha = 2$ and $\beta = \frac{1}{2}$, and discuss why these parameters are not good for this dataset. In particular, find an example of a point that the model predicts poorly and one that it predicts well, and approximately compute the residuals (difference between the true value and the predicted value) for each. Make a guess for what the best parameters are. Finally, discuss what things would look like for $d = 2$.

Proposition. Suppose we are given a noisy dataset of input $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^d$ values and their output $y_1, \ldots, y_n$ values, and suppose we want to find the parameters $\hat{\alpha}$ and $\hat{\beta}$ whose plane best fits the dataset. The best such parameters are given by the following recipe:

(1) Append a one to the LHS of each input $\mathbf{x}_i$ so that now $\mathbf{x}_i$ is a length $d+1$ row vector.

(2) Combine the inputs $\mathbf{x}_i$ into an $n \times (d+1)$ matrix $\mathbf{X}$ and the outputs y_i into a length n column vector $\mathbf{y}$.

(3) Compute the length $d+1$ column vector $(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$.

(4) Take the scalar $\hat{\alpha}$ to be the top-most entry, and the length d column vector $\hat{\beta}$ to consist of the remaining entires.

Remark. By "noisy," we mean that the y values are distributed as

$$y(\mathbf{x}) \sim \alpha + \mathbf{x}\beta + N(0, \sigma^2),$$

and by "best fit" we mean that the $\hat{\alpha}$ and $\hat{\beta}$ parameters minimize the mean squared error

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y(\mathbf{x}_i) - y_i)^2,$$

which is equivalent to being the maximum likelihood estimators for α and β.

Proof of Proposition. Observe that the purpose of (1) is to absorb the α into the β, in the sense that now

$$y(\mathbf{x}_i) = \mathbf{x}_i \begin{pmatrix} \alpha \\ \beta \end{pmatrix} = \mathbf{x}_i\gamma,$$

where we defined γ to be this length $d+1$ column vector. The purpose of (2) is so that the $y(\mathbf{x}_i)$ values can be expressed together as the length n column vector

$$\begin{pmatrix} y(\mathbf{x}_1) \\ \vdots \\ y(\mathbf{x}_n) \end{pmatrix} = \mathbf{X}\gamma.$$

The mean squared error can now be written nicely as

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y(\mathbf{x}_i) - y_i)^2 = \frac{1}{n}(\mathbf{X}\gamma - \mathbf{y})^T(\mathbf{X}\gamma - \mathbf{y}).$$

It remains to explain why (3) and (4) describe the γ that minimizes MSE. The derivative (*i.e.* the gradient) is given by

$$\frac{\mathrm{d}}{\mathrm{d}\gamma}\mathrm{MSE} = \frac{1}{n}(2\mathbf{X}^T\mathbf{X}\gamma - 2\mathbf{X}^T\mathbf{y}),$$

so the only critical point of MSE is at

$$\hat{\gamma} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y},$$

where we can assume $\mathbf{X}$ has full rank. This is a minimum because by the second partial derivative test:

$$\frac{\mathrm{d}^2}{\mathrm{d}\gamma^2}\mathrm{MSE} = \frac{2}{n}\mathbf{X}^T\mathbf{X}$$

is positive definite, again using that $\mathbf{X}$ has full rank. We then unpack $\hat{\gamma}$ to get $\hat{\alpha}$ and $\hat{\beta}$. □

Remark. Following the previous remark, the maximum likelihood estimator of σ^2 is given by

$$\hat{\sigma^2} = \frac{1}{n}(\mathbf{y} - \mathbf{X}\hat{\gamma})^T(\mathbf{y} - \mathbf{X}\hat{\gamma}).$$

This is a biased estimator for σ^2, but

$$\frac{1}{n-d-1}(\mathbf{y} - \mathbf{X}\hat{\gamma})^T(\mathbf{y} - \mathbf{X}\hat{\gamma}) = \frac{n}{n-d-1}\hat{\sigma^2}$$

is an unbiased estimator for σ^2.

Example. Exercise 6.5-4, which is one of the homework problems, uses the points from our warmup. Let us use our recipe to find the best $\hat{\alpha}$ and $\hat{\beta}$. We have

$$\mathbf{X} = \begin{pmatrix} 1 & 2.0 \\ 1 & 3.3 \\ 1 & 3.7 \\ 1 & 2.0 \\ 1 & 2.3 \\ 1 & 2.7 \\ 1 & 4.0 \\ 1 & 3.7 \\ 1 & 3.0 \\ 1 & 2.3 \end{pmatrix} \qquad \text{and} \qquad \mathbf{y} = \begin{pmatrix} 1.3 \\ 3.3 \\ 3.3 \\ 2.0 \\ 1.7 \\ 3.0 \\ 4.0 \\ 3.0 \\ 2.7 \\ 3.0 \end{pmatrix}.$$

So

$$\mathbf{X}^T\mathbf{X} = \begin{pmatrix} 10 & 29 \\ 29 & 89.14 \end{pmatrix} \qquad \text{and} \qquad (\mathbf{X}^T\mathbf{X})^{-1} = \frac{1}{50.4}\begin{pmatrix} 89.14 & -29 \\ -29 & 10 \end{pmatrix},$$

where we note that the entires of $\mathbf{X}^T\mathbf{X}$ consist of the number, the sum, and the sum of the squares of the samples and where to compute the inverse we use the classic formula for the inverse of a 2×2 matrix. Thus

$$(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \approx \begin{pmatrix} 0.06 \\ 0.92 \end{pmatrix}.$$