

Linear regression

Daniel Hsu (COMS 4771)

Maximum likelihood estimation

One of the simplest linear regression models is the following: $(\mathbf{X}_1, Y_1), \dots, (\mathbf{X}_n, Y_n), (\mathbf{X}, Y)$ are iid random pairs taking values in $\mathbb{R}^d \times \mathbb{R}$, and

$$Y \mid \mathbf{X} = \mathbf{x} \sim \mathcal{N}(\mathbf{x}^\top \mathbf{w}, \sigma^2), \quad \mathbf{x} \in \mathbb{R}^d.$$

Here, the vector $\mathbf{w} \in \mathbb{R}^d$ and scalar $\sigma^2 > 0$ are the parameters of the model. (The marginal distribution of \mathbf{X} is unspecified.)

The *log-likelihood* of (\mathbf{w}, σ^2) given $(\mathbf{X}_i, Y_i) = (\mathbf{x}_i, y_i)$ for $i = 1, \dots, n$ is

$$-\frac{n}{2} \log \left(\frac{1}{n} \sum_{i=1}^n \frac{(\mathbf{x}_i^\top \mathbf{w} - y_i)^2}{\sigma^2} \right) + \frac{n}{2} \log \sigma^2$$

where T is some quantity that does not depend on (\mathbf{w}, σ^2) . Therefore, maximizing the log-likelihood over $\mathbf{w} \in \mathbb{R}^d$ (for any $\sigma^2 > 0$) is the same as minimizing

$$\frac{1}{n} \sum_{i=1}^n (\mathbf{x}_i^\top \mathbf{w} - y_i)^2.$$

So, the *maximum likelihood estimator* (MLE) of \mathbf{w} in this model is

$$\hat{\mathbf{w}} \in \arg \min_{\mathbf{w} \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^n (\mathbf{x}_i^\top \mathbf{w} - y_i)^2.$$

(It is not necessarily uniquely determined.)

Empirical risk minimization

Let P_n be the *empirical distribution* on $(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n) \in \mathbb{R}^d \times \mathbb{R}$, i.e., the probability distribution over $\mathbb{R}^d \times \mathbb{R}$ with probability mass function p_n given by

$$p_n((\mathbf{x}, y)) = \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\{(\mathbf{x}, y) = (\mathbf{x}_i, y_i)\}}, \quad (\mathbf{x}, y) \in \mathbb{R}^d \times \mathbb{R}.$$

The distribution assigns probability mass $1/n$ to each (\mathbf{x}_i, y_i) for $i = 1, \dots, n$; no mass is assigned anywhere else. Now consider $(\tilde{\mathbf{X}}, \tilde{Y}) \sim P_n$. The expected squared loss of the linear function $\mathbf{w} \in \mathbb{R}^d$ on $(\tilde{\mathbf{X}}, \tilde{Y})$ is

$$\hat{\mathcal{R}}(\mathbf{w}) := \mathbb{E}[(\tilde{\mathbf{X}}^\top \mathbf{w} - \tilde{Y})^2] = \frac{1}{n} \sum_{i=1}^n (\mathbf{x}_i^\top \mathbf{w} - y_i)^2;$$

we call this the *empirical risk* of \mathbf{w} on the data $(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n)$.

Empirical risk minimization is the method of choosing a function (from some *class of functions*) based on data by choosing a minimizer of the empirical risk on the data. In the case of *linear functions*, the *empirical risk minimizer (ERM)* is

$$\hat{\mathbf{w}} \in \arg \min_{\mathbf{w} \in \mathbb{R}^d} \hat{\mathcal{R}}(\mathbf{w}) = \arg \min_{\mathbf{w} \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^n (\mathbf{x}_i^\top \mathbf{w} - y_i)^2.$$

This is the same as the MLE from above. (It is not necessarily uniquely determined.)

Normal equations

Let

$$\mathbf{A} := \frac{1}{\sqrt{n}} \begin{bmatrix} \leftarrow & \mathbf{x}_1^\top & \rightarrow \\ & \vdots & \\ \leftarrow & \mathbf{x}_n^\top & \rightarrow \end{bmatrix}, \quad \mathbf{b} := \frac{1}{\sqrt{n}} \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}.$$

We can write the empirical risk as

$$\hat{\mathcal{R}}(\mathbf{w}) = \|\mathbf{A}\mathbf{w} - \mathbf{b}\|_2^2, \quad \mathbf{w} \in \mathbb{R}^d.$$

The gradient of $\hat{\mathcal{R}}$ is given by

$$\nabla \hat{\mathcal{R}}(\mathbf{w}) = \nabla \{(\mathbf{A}\mathbf{w} - \mathbf{b})^\top (\mathbf{A}\mathbf{w} - \mathbf{b})\} = 2\mathbf{A}^\top (\mathbf{A}\mathbf{w} - \mathbf{b}), \quad \mathbf{w} \in \mathbb{R}^d.$$

it is equal to zero for $\mathbf{w} \in \mathbb{R}^d$ satisfying

$$\mathbf{A}^\top \mathbf{A} \mathbf{w} = \mathbf{A}^\top \mathbf{b}.$$

These linear equations in \mathbf{w} , which define the *critical points* of $\hat{\mathcal{R}}$, are collectively called the *normal equations*.

It turns out the normal equations in fact determine the *minimizers* of $\hat{\mathcal{R}}$. To see this, let $\hat{\mathbf{w}}$ be any solution to the normal equations. Now consider any other $\mathbf{w} \in \mathbb{R}^d$. We write the empirical risk of \mathbf{w} as follows:

$$\begin{aligned} \hat{\mathcal{R}}(\mathbf{w}) &= \|\mathbf{A}\mathbf{w} - \mathbf{b}\|_2^2 \\ &= \|\mathbf{A}(\mathbf{w} - \hat{\mathbf{w}}) + \mathbf{A}\hat{\mathbf{w}} - \mathbf{b}\|_2^2 \\ &= \|\mathbf{A}(\mathbf{w} - \hat{\mathbf{w}})\|_2^2 + 2(\mathbf{A}(\mathbf{w} - \hat{\mathbf{w}}))^\top (\mathbf{A}\hat{\mathbf{w}} - \mathbf{b}) + \|\mathbf{A}\hat{\mathbf{w}} - \mathbf{b}\|_2^2 \\ &= \|\mathbf{A}(\mathbf{w} - \hat{\mathbf{w}})\|_2^2 + 2(\mathbf{w} - \hat{\mathbf{w}})^\top (\mathbf{A}^\top \mathbf{A}\hat{\mathbf{w}} - \mathbf{A}^\top \mathbf{b}) + \|\mathbf{A}\hat{\mathbf{w}} - \mathbf{b}\|_2^2 \\ &= \|\mathbf{A}(\mathbf{w} - \hat{\mathbf{w}})\|_2^2 + \|\mathbf{A}\hat{\mathbf{w}} - \mathbf{b}\|_2^2 \\ &\geq \hat{\mathcal{R}}(\hat{\mathbf{w}}). \end{aligned}$$

The second-to-last step above uses the fact that $\hat{\mathbf{w}}$ is a solution to the normal equations. Therefore, we conclude that $\hat{\mathcal{R}}(\mathbf{w}) \geq \hat{\mathcal{R}}(\hat{\mathbf{w}})$ for all $\mathbf{w} \in \mathbb{R}^d$ and all solutions $\hat{\mathbf{w}}$ to the normal equations. So the solutions to the normal equations are the minimizers of $\hat{\mathcal{R}}$.

Statistical interpretation

Suppose $(\mathbf{X}_1, Y_1), \dots, (\mathbf{X}_n, Y_n), (\mathbf{X}, Y)$ are iid random pairs taking values in $\mathbb{R}^d \times \mathbb{R}$. The *risk* of a linear function $\mathbf{w} \in \mathbb{R}^d$ is

$$\mathcal{R}(\mathbf{w}) := \mathbb{E}[(\mathbf{X}^\top \mathbf{w} - Y)^2].$$

Which linear functions have smallest risk?

The gradient of \mathcal{R} is given by

$$\nabla \mathcal{R}(\mathbf{w}) = \mathbb{E} \left[\nabla \{(\mathbf{X}^\top \mathbf{w} - Y)^2\} \right] = 2\mathbb{E} [\mathbf{X}(\mathbf{X}^\top \mathbf{w} - Y)], \quad \mathbf{w} \in \mathbb{R}^d;$$

it is equal to zero for $\mathbf{w} \in \mathbb{R}^d$ satisfying

$$\mathbb{E}[\mathbf{X}\mathbf{X}^\top]\mathbf{w} = \mathbb{E}[Y\mathbf{X}].$$

These linear equations in \mathbf{w} , which define the *critical points* of \mathcal{R} , are collectively called the *population normal equations*.

It turns out the population normal equations in fact determine the *minimizers* of \mathcal{R} . To see this, let \mathbf{w}^* be any solution to the population normal equations. Now consider any other $\mathbf{w} \in \mathbb{R}^d$. We write the empirical risk of \mathbf{w} as follows:

$$\begin{aligned}\mathcal{R}(\mathbf{w}) &= \mathbb{E}[(\mathbf{X}^\top \mathbf{w} - Y)^2] \\ &= \mathbb{E}[(\mathbf{X}^\top (\mathbf{w} - \mathbf{w}^*) + \mathbf{X}^\top \mathbf{w}^* - Y)^2] \\ &= \mathbb{E}[(\mathbf{X}^\top (\mathbf{w} - \mathbf{w}^*))^2 + 2(\mathbf{X}^\top (\mathbf{w} - \mathbf{w}^*))(\mathbf{X}^\top \mathbf{w}^* - Y) + (\mathbf{X}^\top \mathbf{w}^* - Y)^2] \\ &= \mathbb{E}[(\mathbf{X}^\top (\mathbf{w} - \mathbf{w}^*))^2] + 2(\mathbf{w} - \mathbf{w}^*)^\top (\mathbb{E}[\mathbf{X}\mathbf{X}^\top]\mathbf{w}^* - \mathbb{E}[Y\mathbf{X}]) + \mathbb{E}[(\mathbf{X}^\top \mathbf{w}^* - Y)^2] \\ &= \mathbb{E}[(\mathbf{X}^\top (\mathbf{w} - \mathbf{w}^*))^2] + \mathbb{E}[(\mathbf{X}^\top \mathbf{w}^* - Y)^2] \\ &\geq \mathcal{R}(\mathbf{w}^*).\end{aligned}$$

The second-to-last step above uses the fact that \mathbf{w}^* is a solution to the population normal equations. Therefore, we conclude that $\mathcal{R}(\mathbf{w}) \geq \mathcal{R}(\mathbf{w}^*)$ for all $\mathbf{w} \in \mathbb{R}^d$ and all solutions \mathbf{w}^* to the population normal equations. So the solutions to the population normal equations are the minimizers of \mathcal{R} .

The similarity to the previous section is no accident. The normal equations (based on $(\tilde{X}_1, \tilde{Y}_1), \dots, (\tilde{X}_n, \tilde{Y}_n)$) are precisely

$$\mathbb{E}[\tilde{\mathbf{X}}\tilde{\mathbf{X}}^\top]\mathbf{w} = \mathbb{E}[\tilde{Y}\tilde{\mathbf{X}}]$$

for $(\tilde{X}, \tilde{Y}) \sim P_n$, where P_n is the empirical distribution on $(\mathbf{X}_1, Y_1), \dots, (\mathbf{X}_n, Y_n)$. By the Law of Large Numbers, the left-hand side $\mathbb{E}[\tilde{\mathbf{X}}\tilde{\mathbf{X}}^\top]$ converges to $\mathbb{E}[\mathbf{X}\mathbf{X}^\top]$ and the right-hand side $\mathbb{E}[\tilde{Y}\tilde{\mathbf{X}}]$ converges to $\mathbb{E}[Y\mathbf{X}]$ as $n \rightarrow \infty$. In other words, the normal equations converge to the population normal equations as $n \rightarrow \infty$. Thus, ERM can be regarded as a *plug-in estimator* for \mathbf{w}^* .

Using classical arguments from asymptotic statistics, one can prove that the distribution of $\sqrt{n}(\hat{\mathbf{w}} - \mathbf{w}^*)$ converges (as $n \rightarrow \infty$) to a multivariate normal with mean zero and covariance $\mathbb{E}[\mathbf{X}\mathbf{X}^\top]^{-1} \text{cov}(\varepsilon\mathbf{X}) \mathbb{E}[\mathbf{X}\mathbf{X}^\top]^{-1}$, where $\varepsilon := Y - \mathbf{X}^\top \mathbf{w}^*$. (This assumes, along with some standard moment conditions, that $\mathbb{E}[\mathbf{X}\mathbf{X}^\top]$ is invertible so that \mathbf{w}^* is uniquely defined. But it does *not* require the conditional distribution of $Y \mid \mathbf{X}$ to be normal.)

Geometric interpretation

Let $\mathbf{a}_j \in \mathbb{R}^n$ be the vector in the j -th column of \mathbf{A} , so

$$\mathbf{A} = \begin{bmatrix} \uparrow & & \uparrow \\ \mathbf{a}_1 & \cdots & \mathbf{a}_d \\ \downarrow & & \downarrow \end{bmatrix}.$$

Since $\text{range}(\mathbf{A}) = \{\mathbf{A}\mathbf{w} : \mathbf{w} \in \mathbb{R}^d\}$, minimizing $\|\mathbf{A}\mathbf{w} - \mathbf{b}\|_2^2$ is the same as finding the vector $\hat{\mathbf{b}} \in \text{range}(\mathbf{A})$ closest to \mathbf{b} (in Euclidean distance), and then specifying the linear combination of $\mathbf{a}_1, \dots, \mathbf{a}_d$ that is equal to $\hat{\mathbf{b}}$, i.e., specifying $\hat{\mathbf{w}} = (\hat{w}_1, \dots, \hat{w}_d)$ such that $\hat{w}_1 \mathbf{a}_1 + \dots + \hat{w}_d \mathbf{a}_d = \hat{\mathbf{b}}$. The solution $\hat{\mathbf{b}}$ is the *orthogonal projection* of \mathbf{b} to $\text{range}(\mathbf{A})$. This vector $\hat{\mathbf{b}}$ is uniquely determined; however, the coefficients $\hat{\mathbf{w}}$ are uniquely determined if and only if $\mathbf{a}_1, \dots, \mathbf{a}_d$ are linearly independent. The vectors $\mathbf{a}_1, \dots, \mathbf{a}_d$ are linearly independent exactly when the rank of \mathbf{A} is equal to d .

We conclude that the empirical risk has a unique minimizer exactly when \mathbf{A} has rank d .