Linear regression

Daniel Hsu (COMS 4771)

Maximum likelihood estimation

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

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

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

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

$$\sum_{i=1}^n \left\{ \ln \frac{1}{\sqrt{2\pi\sigma^2}} - \frac{(y_i - \boldsymbol{x}_i^\intercal \boldsymbol{w})^2}{2\sigma^2} \right\} + T,$$

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

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

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

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

(It is not necessarily uniquely determined.)

Empirical risk minimization

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

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

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

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

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

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

$$\hat{\boldsymbol{w}} \in \operatorname*{arg\,min}_{\boldsymbol{w}\in\mathbb{R}^d} \widehat{\mathcal{R}}(\boldsymbol{w}) = \operatorname*{arg\,min}_{\boldsymbol{w}\in\mathbb{R}^d} \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{x}_i^\mathsf{T}\boldsymbol{w} - y_i)^2.$$

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

Normal equations

Let

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

We can write the empirical risk as

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

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

$$\nabla\widehat{\mathcal{R}}(\boldsymbol{w}) = \nabla\left\{(\boldsymbol{A}\boldsymbol{w} - \boldsymbol{b})^\mathsf{T}(\boldsymbol{A}\boldsymbol{w} - \boldsymbol{b})\right\} = 2\boldsymbol{A}^\mathsf{T}(\boldsymbol{A}\boldsymbol{w} - \boldsymbol{b}), \quad \boldsymbol{w} \in \mathbb{R}^d;$$

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

$$\boldsymbol{A}^\mathsf{T}\boldsymbol{A}\boldsymbol{w} = \boldsymbol{A}^\mathsf{T}\boldsymbol{b}.$$

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

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

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

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

Statistical interpretation

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

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

Which linear functions have smallest risk?

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

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

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

$$\mathbb{E}[\boldsymbol{X}\boldsymbol{X}^{\mathsf{T}}]\boldsymbol{w} = \mathbb{E}[Y\boldsymbol{X}].$$

These linear equations in $\boldsymbol{w}$, 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 $\boldsymbol{w}^\star$ be any solution to the population normal equations. Now consider any other $\boldsymbol{w} \in \mathbb{R}^d$. We write the empirical risk of $\boldsymbol{w}$ as follows:

$$\begin{aligned}
\mathcal{R}(\boldsymbol{w}) &= \mathbb{E}[(\boldsymbol{X}^{\mathsf{T}}\boldsymbol{w} - Y)^2] \\
&= \mathbb{E}[(\boldsymbol{X}^{\mathsf{T}}(\boldsymbol{w} - \boldsymbol{w}^\star) + \boldsymbol{X}^{\mathsf{T}}\boldsymbol{w}^\star - Y)^2] \\
&= \mathbb{E}[(\boldsymbol{X}^{\mathsf{T}}(\boldsymbol{w} - \boldsymbol{w}^\star))^2 + 2(\boldsymbol{X}^{\mathsf{T}}(\boldsymbol{w} - \boldsymbol{w}^\star))(\boldsymbol{X}^{\mathsf{T}}\boldsymbol{w}^\star - Y) + (\boldsymbol{X}^{\mathsf{T}}\boldsymbol{w}^\star - Y)^2] \\
&= \mathbb{E}[(\boldsymbol{X}^{\mathsf{T}}(\boldsymbol{w} - \boldsymbol{w}^\star))^2] + 2(\boldsymbol{w} - \boldsymbol{w}^\star)^{\mathsf{T}}\left(\mathbb{E}[\boldsymbol{X}\boldsymbol{X}^{\mathsf{T}}]\boldsymbol{w}^\star - \mathbb{E}[Y\boldsymbol{X}]\right) + \mathbb{E}[(\boldsymbol{X}^{\mathsf{T}}\boldsymbol{w}^\star - Y)^2] \\
&= \mathbb{E}[(\boldsymbol{X}^{\mathsf{T}}(\boldsymbol{w} - \boldsymbol{w}^\star))^2] + \mathbb{E}[(\boldsymbol{X}^{\mathsf{T}}\boldsymbol{w}^\star - Y)^2] \\
&\geq \mathcal{R}(\boldsymbol{w}^\star).
\end{aligned}$$

The second-to-last step above uses the fact that $\boldsymbol{w}^\star$ is a solution to the population normal equations. Therefore, we conclude that $\mathcal{R}(\boldsymbol{w}) \geq \mathcal{R}(\boldsymbol{w}^\star)$ for all $\boldsymbol{w} \in \mathbb{R}^d$ and all solutions $\boldsymbol{w}^\star$ 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 $(\boldsymbol{X}_1, Y_1), \ldots, (\boldsymbol{X}_n, Y_n)$) are precisely

$$\mathbb{E}[\tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}^{\mathsf{T}}]\boldsymbol{w} = \mathbb{E}[\tilde{Y}\tilde{\boldsymbol{X}}]$$

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

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

Geometric interpretation

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

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

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

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