Linear regression

Daniel Hsu (COMS 4771)

Maximum likelihood estimation

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

$$Y \mid \vec{X} = \vec{x} \sim \mathrm{N}(\vec{x} \cdot \vec{w}, \sigma^2), \quad \text{for all } \vec{x} \in \mathbb{R}^d.$$

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

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

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

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

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

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

$$\vec{w}_{\mathrm{mle}} \in \operatorname*{arg\,min}_{\vec{w} \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^n (\vec{x}_i \cdot \vec{w} - y_i)^2.$$

(It is not necessarily uniquely determined.)

Empirical risk minimization

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

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

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

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

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

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

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

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

Normal equations

Let

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

We can write the empirical risk as

$$\widehat{\mathcal{R}}(\vec{w}) = \frac{1}{n} \|A\vec{w} - \vec{b}\|_2^2, \quad \text{for all } \vec{w} \in \mathbb{R}^d.$$

The gradient of $\widehat{\mathcal{R}}$ (up to a factor of n) is given by

$$n \nabla \widehat{\mathcal{R}}(\vec{w}) = \nabla \{ (A\vec{w} - \vec{b})^{\mathsf{T}} (A\vec{w} - \vec{b}) \} = 2A^{\mathsf{T}}(A\vec{w} - \vec{b}), \quad \text{for all } \vec{w} \in \mathbb{R}^d;$$

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

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

These linear equations in $\vec{w}$, 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 $\vec{w}_{\mathrm{sol}}$ be any solution to the normal equations. Now consider any other $\vec{w} \in \mathbb{R}^d$. We write the (scaled) empirical risk of $\vec{w}$ as follows:

$$\begin{aligned} n\widehat{\mathcal{R}}(\vec{w}) &= \|A\vec{w} - \vec{b}\|_2^2 \\ &= \|A(\vec{w} - \vec{w}_{\mathrm{sol}}) + A\vec{w}_{\mathrm{sol}} - \vec{b}\|_2^2 \\ &= \|A(\vec{w} - \vec{w}_{\mathrm{sol}})\|_2^2 + 2(A(\vec{w} - \vec{w}_{\mathrm{sol}}))^{\mathsf{T}}(A\vec{w}_{\mathrm{sol}} - \vec{b}) + \|A\vec{w}_{\mathrm{sol}} - \vec{b}\|_2^2 \\ &= \|A(\vec{w} - \vec{w}_{\mathrm{sol}})\|_2^2 + 2(\vec{w} - \vec{w}_{\mathrm{sol}})^{\mathsf{T}}(A^{\mathsf{T}}A\vec{w}_{\mathrm{sol}} - A^{\mathsf{T}}\vec{b}) + \|A\vec{w}_{\mathrm{sol}} - \vec{b}\|_2^2 \\ &= \|A(\vec{w} - \vec{w}_{\mathrm{sol}})\|_2^2 + \|A\vec{w}_{\mathrm{sol}} - \vec{b}\|_2^2 \\ &\geq n\widehat{\mathcal{R}}(\vec{w}_{\mathrm{sol}}). \end{aligned}$$

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

Statistical interpretation

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

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

Which linear functions have smallest risk?

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

$$\nabla \mathcal{R}(\vec{w}) = \mathbb{E}\left[\nabla\{(\vec{X} \cdot \vec{w} - Y)^2\}\right] = 2\mathbb{E}\left[\vec{X}(\vec{X} \cdot \vec{w} - Y)\right], \quad \text{for all } \vec{w} \in \mathbb{R}^d;$$

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

$$\mathbb{E}[\vec{X}\vec{X}^\intercal]\vec{w} = \mathbb{E}[Y\vec{X}].$$

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

$$\begin{aligned}
\mathcal{R}(\vec{w}) &= \mathbb{E}[(\vec{X} \cdot \vec{w} - Y)^2] \\
&= \mathbb{E}[(\vec{X} \cdot (\vec{w} - \vec{w}_{\text{opt}}) + \vec{X} \cdot \vec{w}_{\text{opt}} - Y)^2] \\
&= \mathbb{E}[(\vec{X} \cdot (\vec{w} - \vec{w}_{\text{opt}}))^2 + 2(\vec{X} \cdot (\vec{w} - \vec{w}_{\text{opt}}))(\vec{X} \cdot \vec{w}_{\text{opt}} - Y) + (\vec{X} \cdot \vec{w}_{\text{opt}} - Y)^2] \\
&= \mathbb{E}[(\vec{X} \cdot (\vec{w} - \vec{w}_{\text{opt}}))^2] + 2(\vec{w} - \vec{w}_{\text{opt}}) \cdot \Big(\mathbb{E}[\vec{X}\vec{X}^\intercal]\vec{w}_{\text{opt}} - \mathbb{E}[Y\vec{X}]\Big) + \mathbb{E}[(\vec{X} \cdot \vec{w}_{\text{opt}} - Y)^2] \\
&= \mathbb{E}[(\vec{X} \cdot (\vec{w} - \vec{w}_{\text{opt}}))^2] + \mathbb{E}[(\vec{X} \cdot \vec{w}_{\text{opt}} - Y)^2] \\
&\geq \mathcal{R}(\vec{w}_{\text{opt}}).
\end{aligned}$$

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

$$\mathbb{E}[(\vec{X}')(\vec{X}')^\intercal]\vec{w} = \mathbb{E}[Y'\vec{X}']$$

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

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

Geometric interpretation

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

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

Since $\operatorname{range}(A) = \{A\vec{w} : \vec{w} \in \mathbb{R}^d\}$, minimizing $\|A\vec{w} - \vec{b}\|_2^2$ is the same as finding the vector $\vec{p} \in \operatorname{range}(A)$ closest to $\vec{b}$ (in Euclidean distance), and then specifying the linear combination of $\vec{a}_1, \ldots, \vec{a}_d$ that is equal to $\vec{p}$, i.e., specifying $\vec{w} = (w_1, \ldots, w_d)$ such that $w_1\vec{a}_1 + \cdots + w_d\vec{a}_d = \vec{p}$. The solution $\vec{p}$ is the *orthogonal projection* of $\vec{b}$ to $\operatorname{range}(A)$. This vector $\vec{p}$ is uniquely determined; however, the coefficients $\vec{w}$ are uniquely determined if and only if $\vec{a}_1, \ldots, \vec{a}_d$ are linearly independent. The vectors $\vec{a}_1, \ldots, \vec{a}_d$ are linearly independent exactly when the rank of A is equal to d.

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