Chapter 9. Linear Regression

Chapter 9. Linear Regression

FPT Education
FPT UNIVERSITY

9.1 Problem Formulation

Consider a regression problem with the likelihood function

$$p(y|x) = \mathcal{N}(y|f(x), \sigma^2), \tag{1}$$

where $x \in \mathbb{R}^D$ are inputs and $y \in \mathbb{R}$ are noisy function values (targets).

The relation between x and y is given by

$$y = f(x) + \epsilon \text{ with } \epsilon \sim \mathcal{N}(0, \sigma^2), \sigma^2 \text{ is known.} \tag{2}$$

Our object is to find a function that is close to the unknown function f that generated the data and that generalizes well.

We choose a parametrized function and parameters θ that work well for modeling the data.

In the linear regression, we consider the parameter θ appear linearly in our model. An exapmle,

$$p(y|x,\theta) = \mathcal{N}(y|x^T\theta, \sigma^2) \tag{3}$$

$$\Leftrightarrow y = x^T\theta + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2). \tag{4}$$

Example

For $x, \theta \in \mathbb{R}$, the linear regression model in (4) describes straight lines and the parameter θ is the slope of the line.

Consider the linear regression setting in (4) and a training set

$$\mathcal{D} := \{(x_1, y_1), \dots, (x_N, y_N)\}$$

consisting of N inputs $x_n \in \mathbb{R}^D$ and corresponding targets $y_n \in \mathbb{R}$.

The likelihood:

$$\begin{aligned} p(\mathcal{Y}|\mathcal{X}, \theta) &= p(y_1, \dots, y_N | x_1, \dots, x_N, \theta) \\ &= \prod_{n=1}^{N} p(y_n | x_n, \theta) = \prod_{n=1}^{N} \mathcal{N}(y_n | x_n^T \theta, \sigma^2) \end{aligned} \tag{5}$$

where

$$\begin{aligned} \mathcal{X} &:= \{x_1, \dots, x_N\}: \text{ training inputs set} \\ \mathcal{Y} &:= \{y_1, \dots, y_N\}: \text{ corresponding targets set.} \end{aligned}$$

We shall discuss how to find optimal parameter $\theta^\star \in \mathbb{R}^D$ for the model (4):

$$p(y_\star | x_\star, \theta^\star) = \mathcal{N}(y_\star | x_\star^T \theta^\star, \sigma^2).$$

Find the maximum likelihood estimation

$$\theta_{ML} = \arg\max_{\theta} p(\mathcal{Y}|\mathcal{X}, \theta).$$

To find θ_{ML}, we can perform gradient ascent (or gradient descent on the negative likelihood).

However, in practice, we apply the log-transformation to the likelihood function and minimize the negative log-likelihood

$$-\log p(\mathcal{Y}|\mathcal{X}, \theta) = -\sum_{n=1}^{N} \log p(y_n|x_n, \theta). \tag{6}$$

In the model (4), the likelihood is Gaussian

$$\log p(y_n|x_n, \theta) = -\frac{1}{2\sigma^2}(y_n - x_n^T\theta) + \text{const}.$$

Substitute to (6) (ignoring the constant term)

$$\mathcal{L}(\theta) := \frac{1}{2\sigma^2} \sum_{n=1}^{N} (y_n - x_n^T \theta)^2 \tag{7}$$

$$= \frac{1}{2\sigma^2} (y - X\theta)^T (y - X\theta) = \frac{1}{2\sigma^2} \|y - X\theta\|^2$$

where

$$X := \left[x_1 \cdots x_N \right]^T \in \mathbb{R}^{N \times D}$$

$$y := \left[y_1 \dots y_N \right]^T \in \mathbb{R}^{N}.$$

The gradient of $\mathcal{L}$ w.r.t θ

$$
\begin{aligned}
\frac{d\mathcal{L}}{d\theta} &= \frac{1}{2\sigma^2}\frac{d}{d\theta}\left(\|y - X\theta\|^2\right) \\
&= \frac{1}{2\sigma^2}\frac{d}{d\theta}(y^T y - 2y^T X\theta + \theta^T X^T X\theta) \\
&= \frac{1}{\sigma^2}(-y^T X + \theta^T X^T X) \in \mathbb{R}^{1\times D}.
\end{aligned}
\tag{8}
$$

Find θ_{ML} by solving $\frac{d\mathcal{L}}{d\theta} = 0$:

$$
\begin{aligned}
\frac{d\mathcal{L}}{d\theta} = 0 &\Leftrightarrow \theta_{ML}^T X^T X = y^T X \\
&\Leftrightarrow \theta_{ML}^T = y^T X(X^T X)^{-1} && (9) \\
&\Leftrightarrow \theta_{ML} = (X^T X)^{-1} X^T y. && (10)
\end{aligned}
$$

Maximum Likelihood Estimation with Features

Straight lines are not sufficiently expressive when it comes to fitting more interesting data. We can perform a nonlinear transformation $\Phi(x)$ of the inputs x and then linearly combine the components of this transformation. The corresponding linear regression model is

$$p(y|x,\theta) = \mathcal{N}(y|\phi^T(x)\theta, \sigma^2) \tag{11}$$

$$\Leftrightarrow y = \phi^T(x)\theta + \epsilon = \sum_{k=1}^{K} \theta_k \phi_k(x) + \epsilon,$$

where $\phi : \mathbb{R}^D \to \mathbb{R}^K$ is a transformation of inputs x and $\phi_k : \mathbb{R}^D \to \mathbb{R}$ is the k^{th} component of the feature vector ϕ.

Consider training inputs $x_n \in \mathbb{R}^D$ and targets $y_n \in \mathbb{R}$, define the feature matrix

$$\Phi := \begin{bmatrix} \phi^T(x_1) \\ \vdots \\ \phi^T(x_N) \end{bmatrix} = \begin{bmatrix} \phi_1(x_1) & \cdots & \phi_K(x_1) \\ \vdots & \ddots & \vdots \\ \phi_1(x_N) & \cdots & \phi_K(x_N) \end{bmatrix}. \tag{12}$$

The negative log-likelihood for the model (11):

$$-\log p(\mathcal{Y}|\mathcal{X},\theta) = \frac{1}{2\sigma^2}(y - \Phi\theta)^T(y - \Phi\theta) + \text{const.} \tag{13}$$

The maximum likelihood estimate:

$$\theta_{ML} = \left(\Phi^T\Phi\right)^{-1}\Phi^T y$$

for the linear regression problem (11).

Estimating the Noise Variance

We assumed that the noise variance σ^2 is known. However, we can obtain the maximum likelihood estimator σ^2_{ML} for the noise variance:

$$\log p(\mathcal{Y}|\mathcal{X},\theta,\sigma^2) = \sum_{n=1}^{N} \log \mathcal{N}(y_n|\phi^T(x_n)\theta,\sigma^2)$$

$$= \sum_{n=1}^{N} \left(-\frac{1}{2}\log(2\pi) - \frac{1}{2}\log\sigma^2 - \frac{1}{2\sigma^2}(y_n - \phi^T(x_n)\theta)^2 \right)$$

$$= -\frac{N}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{n=1}^{N}(y_n - \phi^T(x_n)\theta)^2 + \text{const}.$$

Solving

$$\frac{\partial \log p(\mathcal{Y}|\mathcal{X},\theta,\sigma^2)}{\partial \sigma^2} = 0,$$

we get

$$\sigma^2_{ML} = \frac{1}{N}\sum_{n=1}^{N}(y_n - \phi^T(x_n)\theta)^2.$$

9.2.3 Maximum A Posteriori Estimation

Given a data set $\mathcal{X}, \mathcal{Y}$, the maximum a posteriori (MAD) estimation is a procedure that instead of maximizing the likelihood, we seek parameters that maximize the posterior distribution

$$p(\theta|\mathcal{X}, \mathcal{Y}) = \frac{p(\mathcal{Y}|\mathcal{X}, \theta)p(\theta)}{p(\mathcal{Y}|\mathcal{X})}. \tag{14}$$

We have

$$\log p(\theta|\mathcal{X}, \mathcal{Y}) = \log p(\mathcal{Y}|\mathcal{X}, \theta) + \log p(\theta) + \text{const}, \tag{15}$$

where the constant is independent of θ.

$$-\frac{d \log p(\theta|\mathcal{X}, \mathcal{Y})}{d\theta} = -\frac{d \log p(\mathcal{Y}|\mathcal{X}, \theta)}{d\theta} - \frac{d \log p(\theta)}{d\theta}. \tag{16}$$

With $p(\theta) = \mathcal{N}(0, b^2 I)$

$$-\log p(\theta, \mathcal{X}, \mathcal{Y}) = \frac{1}{2\sigma^2}(y - \Phi\theta)^T(y - \Phi\theta) + \frac{1}{2b^2}\theta^T\theta + \text{const}. \tag{17}$$

Hence

$$-\frac{d\log p(\theta|\mathcal{X},\mathcal{Y})}{d\theta} = \frac{1}{\sigma^2}(\theta^T\Phi^T\Phi - y^T\Phi) + \frac{1}{b^2}\theta^T. \tag{18}$$

Find $\theta_{MAP} \in \arg\min_\theta\{-\log p(\mathcal{Y}|\mathcal{X},\theta) - \log p(\theta)\}$ by solving

$$\begin{aligned}
&-\frac{d\log p(\theta|\mathcal{X},\mathcal{Y})}{d\theta} = 0\\
\Rightarrow&\frac{1}{\sigma^2}(\theta^T\Phi^T\Phi - y^T\Phi) + \frac{1}{b^2}\theta^T = 0\\
\Leftrightarrow&\theta^T\left(\frac{1}{\sigma^2}\Phi^T\Phi + \frac{1}{b^2}I\right) - \frac{1}{\sigma^2}y^T\Phi = 0\\
\Leftrightarrow&\theta^T = y^T\Phi\left(\Phi^T\Phi + \frac{\sigma^2}{b^2}I\right)^{-1}.
\end{aligned}$$

Hence

$$\theta_{MAP} = \left(\Phi^T\Phi + \frac{\sigma^2}{b^2}I\right)^{-1}\Phi^T y.$$

Summary

We have discussed linear regression for

- Gaussian likelihoods;
- conjugate Gaussian priors on the parameters of the model.