Linear regression

Lesson No. 3

João Victor da Silva Guerra

March 9, 2020

1 Introduction

Linear regression is a fundamental tool of statistics and supervised machine learning, which models the relationship between two variables by fitting a linear equation ($\vec{\mathbf{y}} = A \times \vec{\mathbf{x}} + B$). In the example, $\vec{\mathbf{x}}$ is the explanatory variable and $\vec{\mathbf{y}}$ is the dependent variable.

2 Model description

In the simplest case, the linear regression model is described as:

$$p(y|\vec{x},\vec{\theta}) = \mathcal{N}(y|\vec{w}^T \times \vec{x}, \sigma^2) \tag{1}$$

In addition, the linear regression can be made to model non-linear relationships by replacing $\vec{x}$ with some non-linear function of $\vec{x}$ (denoted $\phi(\vec{x})$), also known as **basis function expansion**. In this scenario, the model is described as:

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

Note that the model remains linear in the parameters $\vec{w}$.

3 Least-squares regression

The parameters of a statistical model is estimated by the maximum likelihood estimation (MLE), which is defined as follows:

$$\hat{\theta} = \arg\max_{\theta} \log p(\mathcal{D}|\theta) \tag{3}$$

Assuming the training examples as independent and identically distributed, we can define the **log-likelihood (ℓ)** and its counterpart, **negative log likelihood (NLL)**, as follows:

$$\ell(\theta) = \log p(\mathscr{D}|\theta) = \sum_{i=1}^{N} \log p(y_{\mathrm{i}}|\vec{x}_{\mathrm{i}}, \theta) \tag{4}$$

$$NLL(\theta) = -\sum_{i=1}^{N} \log p(y_{\mathrm{i}}|\vec{x}_{\mathrm{i}}, \theta) \tag{5}$$

Note that it is equivalent to maximize (4) and minimize (5); however, the last is usually more convenient, because optimization software packages are commonly designed to find the minima functions, rather than maxima.

Applying MLE's method to the linear equation model with the Gaussian definition inserted as well, the log-likelihood is defined as follows:

$$\begin{aligned} \ell(\theta) &= \sum_{i=1}^{N} \log \left[\left(\frac{1}{2\pi\sigma^2} \right)^{\frac{1}{2}} exp\left(-\frac{1}{2\sigma^2} (y_i - \vec{w}^T \vec{x}_i)^2 \right) \right] \\ &= -\frac{1}{2\sigma^2} RSS(\vec{w}) - \frac{N}{2} \log(2\pi\sigma^2) \end{aligned} \tag{6}$$

where RSS is the **residual sum of squares**, also known as **sum of squared erros** and is defined as follows:

$$RSS(\vec{w}) = \sum_{i=1}^{N} (y_i - \vec{w}^T \vec{x}_i)^2 \tag{7}$$

Based on the above equations, the MLE for $\vec{w}$ is the one that minimizes the RSS.

3.1 Ordinary least squares (???) - incomplete

The corresponding solution of $\hat{w}$ to this linear system of equations is called the ordinary least squares (OLS), which is defined as:

$$\hat{w}_{OLS} = (X^T X)^{-1} X^T y \tag{8}$$

3.2 Convexity

Firstly, we define a **convex set** (S) as follows:

$$\lambda x + (1-\lambda) y, \begin{cases} \forall \lambda \in [0,1] \\ \forall x, y \in S \end{cases} \tag{9}$$

A function $f(x)$ is convex if its epigraph (set of points above the function) defines a convex set. Equivalently, a function $f(x)$ is convex if it is defined on a convex set as follows:

$$f(\lambda x + (1-\lambda) y) \leq \lambda f(x) + (1-\lambda) f(y), \begin{cases} \forall \lambda \in [0,1] \\ \forall x, y \in S \end{cases} \tag{10}$$

3.3 Study information - can be removed

* **trace (often** tr**)** of a square matrix A is defined by the sum of elements on the main diagonal of A.

$$x^T Ax = tr(x^T Ax) \tag{11}$$

So, for derivation of MLE (p. 220 of [1])

References

[1] K. P. Murphy, *Machine Learning: A Probabilistic Perspective*. The MIT Press, 2012, pp. 27–33.