Linear regression

Tuan Nguyen

Ngày 28 tháng 7 năm 2021

Overview

What is Machine Learning?

Hình 1: Machine Learning

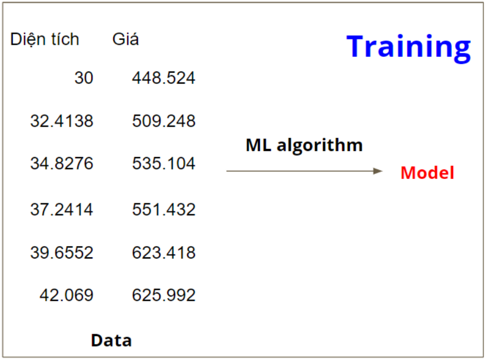
What is Machine Learning? (cont.)

Hình 2: Machine Learning

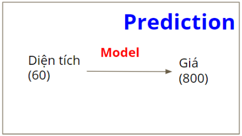
There are two main steps in Machine Learning task

- Training: Data -> Model
- Prediction: Model -> Predict

House price prediction

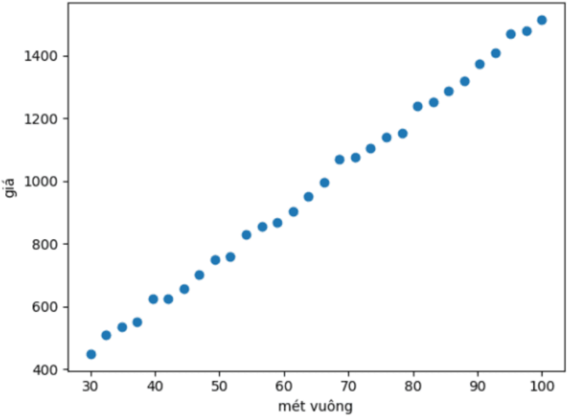


Hình 3: Training

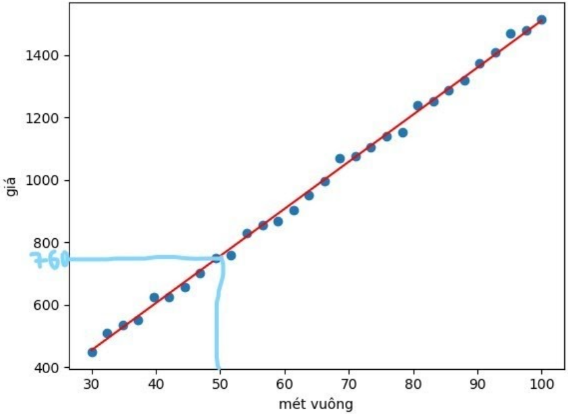


Hình 4: Prediction

Visualize data

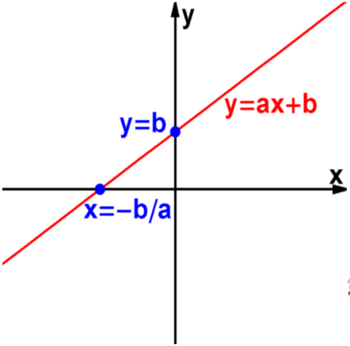


Hình 5: Correlation between square and price



Hình 6: 2 steps in machine learning

Model



$$y = w_1 * x + w_0$$

Hình 7: Model and its parameters

Maximum likelihood

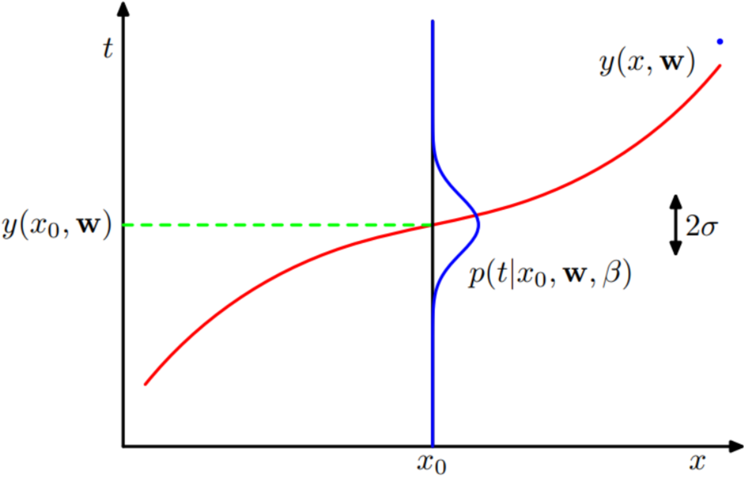
We have a data set of observations $\mathbf{x} = (x_1, x_2, ..., x_N)^T$, representing N observations of the scalar variable x and their corresponding target values $\mathbf{t} = (t_1, t_2, ..., t_N)^T \Rightarrow$ make predictions for some new value of the input variable x.

Suppose that the observations are drawn independently from a Gaussian distribution. Data points that are drawn independently from the same distribution are said to be independent and identically distributed (i.i.d)

$$t = y(x, \mathbf{w}) + \mathcal{N}(0, \beta^{-1}) t = \mathcal{N}(y(x, \mathbf{w}), \beta^{-1})$$

Precision paramter $\beta = \dfrac{1}{\sigma^2}$

$$p(t|x, \mathbf{w}, \beta) = \mathcal{N}(t|y(x, \mathbf{w}), \beta^{-1})$$



Hình 8: $p(t|x, \mathbf{w}, \beta) = \mathcal{N}(t|y(x, \mathbf{w}), \beta^{-1})$

Maximum likelihood (cont.)

We now use the training data x, t to determine the values of the unknown parameters w and by maximum likelihood. If the data are assumed to be drawn independently from the distribution then the likelihood function:

$$p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta) = \prod_{n=1}^{N} \mathcal{N}(t_n|y(x_n, \mathrm{w}), \beta^{-1})$$

It is convenient to maximize the logarithm of the likelihood function

$$\log p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta) = \sum_{n=1}^{N} \log\big(\mathcal{N}(t_n|y(x_n, \mathrm{w}), \beta^{-1})\big)$$

$$= -\frac{\beta}{2} \sum_{n=1}^{n} \{y(x_n, \mathbf{w}) - t_n\}^2 + \frac{N}{2} \log \beta - \frac{N}{2} \log(2\pi)$$

$$\max_{\mathbf{w}} \log p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta) = -\max_{\mathbf{w}} \frac{\beta}{2} \sum_{n=1}^{n} \{y(x_n, \mathbf{w}) - t_n\}^2$$

$$= \min_{\mathbf{w}} \frac{1}{2} \sum_{n=1}^{n} \{y(x_n, \mathbf{w}) - t_n\}^2.$$

We minimize $P = \frac{1}{2} \sum_{n=1}^{n} \{y(x_n, \mathbf{w}) - t_n\}^2$ to find $\mathbf{w}$. Suppose

$$X = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}, \mathbf{w} = \begin{bmatrix} w_0 \\ w_1 \end{bmatrix}$$

$$\Rightarrow P = \|X\mathbf{w} - \mathbf{t}\|_2^2$$

By minimizing P, we can find $\mathbf{w} = (X^T X)^{-1} X^T t$. P is called Mean Squared Error loss (MSE).

Posterior

Bayes theorem

$$p(A|B) = \frac{p(B|A)p(A)}{p(B)}$$
$$\Leftrightarrow \text{posterior} = \frac{\text{likelihood} \times \text{prior}}{\text{evidence}}$$
$$\Rightarrow p(\mathbf{w}|\mathbf{x}, \mathbf{t}, \alpha, \beta) = \frac{p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta)p(\mathbf{w}|\alpha)}{p(\mathbf{x}, \mathbf{t}, \alpha, \beta)}$$

$p(\mathbf{w}|\mathbf{x}, \mathbf{t}, \alpha, \beta)$ is a posterior. While likelihood is given the parameter how the parameter fit the data, posterior is given the data, what is the probability of parameter. In the posterior, we also includef our belief.

We expect to maximinze the posterior to find $\mathbf{w}$.

$$p(\mathbf{w}|\mathbf{x}, \mathbf{t}, \alpha, \beta) \propto p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta)p(\mathbf{w}|\alpha)$$

Because $p(\mathbf{x}, \mathbf{t}, \alpha, \beta)$ is dependent of $\mathbf{w}$

Posterior (cont.)

Suppose $p(\mathbf{w}|\alpha)$ is a normal distribution. We have

$$p(\mathbf{w}|\alpha) = \mathcal{N}(\mathbf{w}|0, \alpha^{-1}I) = (\frac{\alpha}{2\pi})^{(M+1)/2} \exp\{-\frac{\alpha}{2}\mathbf{w}^T\mathbf{w}\}$$

So

$$\begin{aligned} & p(\mathbf{w}|\mathbf{x}, \mathbf{t}, \alpha, \beta) \\ & \quad \propto p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta)p(\mathbf{w}|\alpha) \\ & \quad \propto \exp\{-\frac{\beta}{2}\sum_{n=1}^{n}\{y(x_n, \mathbf{w}) - t_n\}^2 - \frac{\alpha}{2}\mathbf{w}^T\mathbf{w}\} \end{aligned}$$

we find that the maximum of the posterior is given by the minimum of

$$\frac{\beta}{2}\sum_{n=1}^{n}\{y(x_n, \mathbf{w}) - t_n\}^2 + \frac{\alpha}{2}\mathbf{w}^T\mathbf{w}$$

Posterior (cont.)

or we minimize

$$Q = \|X\mathbf{w} - \mathbf{t}\|_2^2 + \lambda \mathbf{w}^T \mathbf{w}$$

Q is MSE loss with L2 regularization.

By minimizing Q, we can find $\mathbf{w} = (X^T X + \lambda I)^{-1} X^T t$

Gaussian prior is called conjugate prior because the posterior is also Gaussian distribution. So conjugate prior is the distribution that makes the likelihood and posterior have the same distribution.