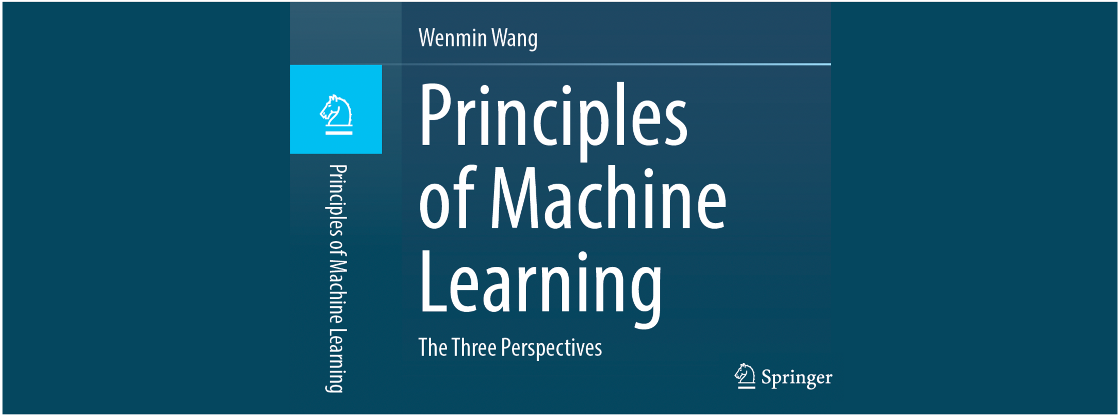


Wenmin Wang

Principles of Machine Learning

The Three Perspectives

Principles of Machine Learning

Springer

https://link.springer.com/book/10.1007/978-981-97-5333-8

Part IV Tasks

12 Classification Task

13 Regression Task

14 Clustering Task

15 Dimensionality Reduction Task

13 Regression Task

13 Regression Task

13.1 Problem and Definition

13.2 Working Principle

13.3 Related Elements

13.4 Typical Algorithms

13.5 Loss Functions

Regression Problem

Regression is to statistically analyze the relationship between sample data and use it to predict unknown input data.

The term “regression” was coined by Sir Francis Galton, first used it in a paper in 1886, titled “Regression Towards Mediocrity in Hereditary Stature”.

Galton’s work has been extended to more general statistical scenarios, making the term gradually widely used in statistics.

The essence of regression is to find a curve that fits the relationship between multiple sample data points, represented as a regression function.

For fitting multiple data points, a certain compromise needs to be made, which is a product of “regression” towards the average value of multiple data points.

Definition

Definition: (Regression)

Regression in machine learning is the process of training a regression algorithm based on samples of corresponding values between input and output data, obtaining an optimal regression hypothesis, and then using it to predict unknown input data. The output is a continuous corresponding value.

- The input is equivalent to observed value.
- The output is a continuous corresponding value.

Regression vs. Classification

Similarity both regression and classification belong to supervised learning paradigm, and need to be trained based on labeled samples.

Difference it lies in the output, that is shown in the following table.

	Regression	Classification
Difference	The output is a continuous value	The output is a discrete category
Example	Sales forecasting, risk analysis	$\{\text{sunny}, \text{cloudy}, \text{rainy}\}$, $\{0, 1, \ldots, 9\}$

Regression Model

Definition: (Regression Model)

A regression model can be formalized as the following equation:

$$y = f(\boldsymbol{x}, \theta) + \varepsilon,$$

where y is dependent variable, $\boldsymbol{x}$ is independent variable, θ is parameter, ε is error term, and $f(\boldsymbol{x}, \theta)$ is referred to as regression function.

The goal of regression model is to find best-fitting regression function based on labeled sample data.

Underlying Assumptions

The underlying assumptions in order for regression to be an effective method:

- The sample is representative of the population at large.
- The independent variables are measured without error.
- The error ε of the model has an expected value of zero, given the independent variables, $\mathbb{E}\left[\varepsilon|\boldsymbol{x}\right] = 0$.
- The variance of error ε is constant across all input data, i.e., homoscedasticity.
- The errors ε are uncorrelated. Mathematically, this satisfies the diagonal property of the variance-covariance matrix of errors.

13 Regression Task

13.1 Problem and Definition

13.2 Working Principle

13.3 Related Elements

13.4 Typical Algorithms

13.5 Loss Functions

Formal Description

Let $\mathcal{X} \subseteq \mathbb{R}^m$ denote input space, and $\mathcal{Y} \subseteq \mathbb{R}$ denote output space.

Through a probability $P(\boldsymbol{x})$, we obtain n independent and identically distributed (i.i.d.) observed data:

$$D = \{\boldsymbol{x}_i \mid \boldsymbol{x}_i \in \mathcal{X} \text{ and } i = 1, \ldots, n\}.$$

Let the target function of regression be: $f : \mathcal{X} \to \mathcal{Y}$. It can be represented as $P(y|\boldsymbol{x})$, where $\boldsymbol{x} \in \mathcal{X}$, $y \in \mathcal{Y}$.

Based on $P(\boldsymbol{x}, y) = P(y|\boldsymbol{x})\, P(\boldsymbol{x})$ to label the actual output value y_i for each input data $\boldsymbol{x}_i$, $(\boldsymbol{x}_i, y_i) \in \mathcal{X} \times \mathcal{Y}$, we obtain a set of training sample S with n elements:

$$S = \{(\boldsymbol{x}_i,\ y_i) \mid i = 1, \ldots, n\}.$$

Formal Description

Let H be a set of hypothesis functions for regression, $H : \mathcal{X} \to \mathcal{Y}$. The goal is to obtain $h \in H$ through the training on S, and $h(\boldsymbol{x}) = \hat{y}$ has smallest expected error with the target function $f(\boldsymbol{x}) = y$:

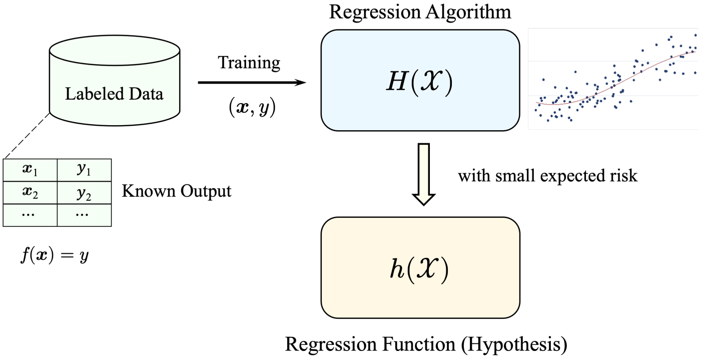
$$\arg\min_{h \in H} R\left(h\right) = \arg\min_{h \in H} \mathbb{E}\left[\mathcal{L}\left(h\left(\boldsymbol{x}\right), f\left(\boldsymbol{x}\right)\right)\right] = \arg\min_{h \in H} \mathbb{E}\left[\mathcal{L}\left(\hat{y}, y\right)\right],$$

where $\hat{y}$ is predicted value, y is target value, $\mathcal{L}\left(\cdot, \cdot\right)$ is loss function of regression.

After obtaining h, use it to predicate regression result for each unknown data $\boldsymbol{x}_i \in \mathcal{X}$. The total output result is:

$$T_{\text{Output}} = \left\{\left(\boldsymbol{x}_i, \hat{y}_i\right) \mid \hat{y}_i = h\left(\boldsymbol{x}_i\right) \text{ and } i = 1, \ldots, n'\right\} \subseteq \mathcal{Y}.$$

Illustrated Description



Regression Algorithm

Labeled Data

Training

$(\boldsymbol{x}, y)$

$H(\mathcal{X})$

$\boldsymbol{x}_1$	y_1
$\boldsymbol{x}_2$	y_2
...	...

Known Output

$f(\boldsymbol{x}) = y$

with small expected risk

$h(\mathcal{X})$

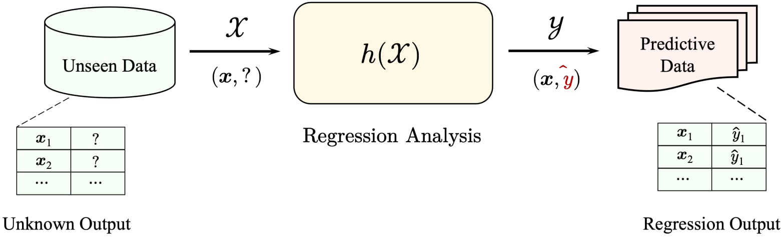
Regression Function (Hypothesis)

Regression function can be regarded as fitting function of training data. This is the key to understanding the regression method.

Illustrated Description

The testing process is also similar to classification.

The difference is that: for a given input $\boldsymbol{x}_i$, through the regression function $h\left(\boldsymbol{x}_i\right)$, what we obtained is an output value $\hat{y}_i$, but not a category.



$\boldsymbol{x}_1$	?
$\boldsymbol{x}_2$	?
···	···

Unknown Output

Regression Analysis

$\boldsymbol{x}_1$	$\hat{y}_1$
$\boldsymbol{x}_2$	$\hat{y}_1$
···	···

Regression Output

13 Regression Task

13.1 Problem and Definition

13.2 Working Principle

13.3 Related Elements

13.4 Typical Algorithms

13.5 Loss Functions

Uni- and Multi-Independent Variables

In the regression model $y = f(\boldsymbol{x}, \theta) + \varepsilon$, the $\boldsymbol{x}$ is independent variable. If $\boldsymbol{x} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_m) \in \mathcal{X}$, m is the number of the independent variables, after substituting the independent variables back into the regression model:

$$y = f(\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_m, \theta) + \varepsilon.$$

Therefore, there are two situations regarding the number of independent variables in the regression model:

1) Uni-independent variable

   Only one independent variable, i.e., $m = 1$.

2) Multi-independent variables

   Two or more independent variables, i.e., $m \geq 2$.

Uni- and Multi-Dependent Variables

In the regression model $y = f(\boldsymbol{x}, \theta) + \varepsilon$, the y is dependent variable. If $y = (y_1, \ldots, y_n) \in \mathcal{Y}$, n is the number of the dependent variables, after substituting the dependent variables into the regression model:

$$y = (y_1, \ldots, y_n) = f(\boldsymbol{x}, \theta) + \varepsilon.$$

Therefore, there are two situations for the number of dependent variable in the regression model:

1) Uni-dependent variable

   Only one dependent variable, i.e., $n = 1$.

2) Multi-dependent variables

   Two or more dependent variables, i.e., $n \geq 2$.

Four Submodels

Since each of independent and dependent variables have two situations, so the following four types of regression submodels can be derived.

	Uni-independent variable	Multi-independent variables
Uni-dependent variable	Uni-independent variable, Uni-dependent variable	Multi-independent variables, Uni-dependent variable
Multi-dependent variables	Uni-independent variable, Multi-dependent variables	Multi-independent variables, Multi-dependent variables

The regression problems with multi-independent and uni-dependent variables are common and easy to understand. But the regression problems with multi-independent and multi-dependent variables are also not uncommon.

Linear Combination

Linear combination is an important concept in mathematics.

Definition: (Linear Combination of Parameters)

A linear combination of parameters is an expression composed of a set of terms, each of which is multiplied by parameter and independent variable, and then added together. The value of the first independent variable is usually 1.

Linear Regression

Definition: (Linear Regression Model)

A regression model $y = f(\boldsymbol{x}, \theta) + \varepsilon$ is referred to as a linear model, if where the expression $f(\boldsymbol{x}, \theta)$ is a linear combination of parameters.

Linear regression does not rely on if its fitted line presented is a straight line.

The following is a linear regression model, which appears as a straight line:

$$y = \theta_0 + \theta_1 \boldsymbol{x} + \varepsilon.$$

Following is also a linear regression model, although it it appears as a parabola:

$$y = \theta_0 + \theta_1 \boldsymbol{x} + \theta_2 \boldsymbol{x}^2 + \varepsilon.$$

Nonlinear Regression

Definition: (Nonlinear Regression Model)

A regression model $y = f(\boldsymbol{x}, \theta) + \varepsilon$ is referred to as a nonlinear regression, if where the expression $f(\boldsymbol{x}, \theta)$ is not a linear combination of parameters.

E.g., in the following function, although there is only one independent variable, because the parameters θ_0 and θ_1 are not linearly combined, it is still nonlinear.

$$y = \frac{\theta_1 \boldsymbol{x}}{\theta_0 + \boldsymbol{x}} + \varepsilon.$$

Parametric Regression

Definition: (Parametric Regression Model)

A regression model $y = f(\boldsymbol{x}, \theta) + \varepsilon$ is referred to as a parametric regression model, if the θ is a finite and fixed number of parameters.

In other words, if the relationship between the independent and dependent variables in the regression model is known and can be defined using a finite number of parameters, it is a parametric regression model.

Nonparametric Regression

Definition: (Nonparametric Regression Model)

A regression model $y = f(\boldsymbol{x}, \theta) + \varepsilon$ is referred to as a nonparametric regression model, if the number of parameters of θ is not predetermined but adjusted according to the size of the dataset.

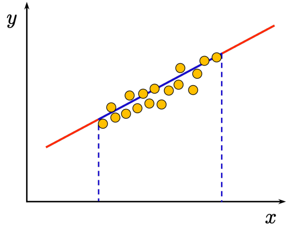
Nonparametric regression models are not parameter-free, but the number of parameters is not fixed.

Interpolation and Extrapolation

Regression models estimate the relationship between independent and dependent variables based on training samples and obtain their best-fit line.

Interpolation:

Based on the best-fit line to estimate the value of dependent variable for a given value of the independent variable, provided it is within the data range (within the blue line).



Extrapolation:

Be the estimation of the value of the dependent variable for a given value of the independent variable outside the data range (outside of blue line).

13 Regression Task

13.1 Problem and Definition

13.2 Working Principle

13.3 Related Elements

13.4 Typical Algorithms

13.5 Loss Functions

Multiple Linear Regression

This algorithm has uni-dependent variable and multi-independent variables, and its expression $f\left(\boldsymbol{x},\theta\right)$ is a linear combination of parameters.

Given training samples $S=\{(\boldsymbol{x}_i,y_i)\mid i=1,\dots,n\}$, and objective function:

$$y_i=f(\boldsymbol{x}_i,\theta)+\varepsilon_i=\theta_0+\theta_1x_{i1}+\cdots+\theta_mx_{im}+\varepsilon_i=\sum_{j=0}^{m}\theta_jx_{ij}+\varepsilon_i,$$

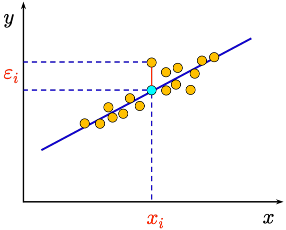
where $m\geq 1$, and $x_{i0}=1$.

Obtain prediction model of multiple linear regression using training samples S:

$$\hat{y}_i=\hat{\theta}_0+\hat{\theta}_1x_{i1}+\cdots+\hat{\theta}_mx_{im}=\sum_{j=0}^{m}\hat{\theta}_jx_{ij}.$$

Multiple Linear Regression

Let $\varepsilon_i = y_i - \hat{y}_i$ be the difference between actual value y_i in objective function and the predicted value $\hat{y}_i$ in prediction model, and ε_i is called the residual, depicted in the right figure.



The commonly used parameter estimation is least square method, which calculates its parameters $\hat{\theta}$ by minimizing the residual sum of squares (RSS).

$$\mathrm{RSS}_{\text{multi}}\left(\hat{\theta}\right) = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} \left(y_i - \sum_{j=0}^{m} \hat{\theta}_j x_{ij}\right)^2 .$$

Multiple Linear Regression

To simplify the derivation process, it can be denoted as:

$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \ \boldsymbol{x} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1m} \\ 1 & x_{21} & \cdots & x_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{nm} \end{bmatrix}, \ \theta = \begin{bmatrix} \theta_0 \\ \theta_1 \\ \vdots \\ \theta_m \end{bmatrix}, \ \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix},$$

where $\boldsymbol{x}$ is design matrix, with dimensions $n \times (m+1)$.

Thus, the residual sum of squares can be represented as:

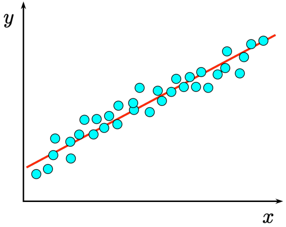
$$\begin{aligned} \mathrm{RSS}_{\text{multi}}\left(\hat{\theta}\right) &= \|\boldsymbol{y} - \hat{\boldsymbol{y}}\|^2 = \|\boldsymbol{y} - \boldsymbol{x}\hat{\theta}\|^2 = (\boldsymbol{y} - \boldsymbol{x}\hat{\theta})^\mathsf{T}(\boldsymbol{y} - \boldsymbol{x}\hat{\theta}) \\ &= \boldsymbol{y}^\mathsf{T}\boldsymbol{y} - \boldsymbol{y}^\mathsf{T}(\boldsymbol{x}\hat{\theta}) - (\boldsymbol{x}\hat{\theta})^\mathsf{T} + (\boldsymbol{x}\hat{\theta})^\mathsf{T}(\boldsymbol{x}\hat{\theta}) = \boldsymbol{y}^\mathsf{T}\boldsymbol{y} - 2\boldsymbol{y}^\mathsf{T}\boldsymbol{x}\hat{\theta} + \hat{\theta}^\mathsf{T}\boldsymbol{x}^\mathsf{T}\boldsymbol{x}\hat{\theta}. \end{aligned}$$

Multiple Linear Regression

Because $\mathrm{RSS}_{\textsf{multi}}\left(\hat{\theta}\right)$ is a convex function, and its optimal solution is at the point where the gradient is zero. By calculating its derivative, we get:

$$\frac{\partial \mathrm{RSS}_{\textsf{multi}}\left(\hat{\theta}\right)}{\partial \hat{\theta}} = \frac{\partial(\boldsymbol{y}^\mathsf{T}\boldsymbol{y} - 2\boldsymbol{y}^\mathsf{T}\boldsymbol{x}\hat{\theta} + \hat{\theta}^\mathsf{T}\boldsymbol{x}^\mathsf{T}\boldsymbol{x}\hat{\theta})}{\partial \hat{\theta}}$$

$$= -2\boldsymbol{x}^\mathsf{T}\boldsymbol{y} + 2\boldsymbol{x}^\mathsf{T}\boldsymbol{x}\hat{\theta} = 0.$$



Thus far, the equation for solving the parameters using the method of least squares (i.e., minimizing the sum of residual squares) is:

$$\hat{\theta}_{\textsf{multi}} = \left(\boldsymbol{x}^\mathsf{T}\boldsymbol{x}\right)^{-1}\boldsymbol{x}^\mathsf{T}\boldsymbol{y}.$$

Polynomial Regression

Given training samples $S = \{(\boldsymbol{x}_i, y_i) \mid i = 1, \ldots, n\}$, and objective function:

$$y_i = f(\boldsymbol{x}_i, \theta) + \varepsilon_i = \theta_0 + \theta_1 x_i + \theta_2 x_i^2 + \cdots + \theta_m x_i^m + \varepsilon_i = \sum_{j=0}^{m} \theta_j x_i^j + \varepsilon_i,$$

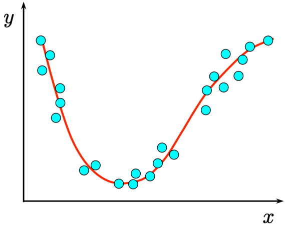
where $m \geq 2$, and $x_i^0 = 1$. The function can be represented in matrix form:

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 & x_1^2 & \cdots & x_1^m \\ 1 & x_2 & x_2^2 & \cdots & x_2^m \\ 1 & x_3 & x_3^2 & \cdots & x_3^m \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & x_n^2 & \cdots & x_n^m \end{bmatrix} \begin{bmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \\ \vdots \\ \theta_m \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \vdots \\ \varepsilon_n \end{bmatrix}.$$

Polynomial Regression

So, the objective function and prediction model can be represented as:

$$\boldsymbol{y} = \mathbf{X}_{\text{poly}}\theta + \varepsilon, \quad \hat{\boldsymbol{y}} = \mathbf{X}_{\text{poly}}\hat{\theta}.$$



Similar to the method of least squares in multiple linear regression, calculate its residual sum of squares and calculate its gradient to obtain the parameter estimation of polynomial regression:

$$\hat{\theta} = \left(\mathbf{X}_{\text{poly}}{}^{\mathsf{T}}\mathbf{X}_{\text{poly}}\right)^{-1}\mathbf{X}_{\text{poly}}{}^{\mathsf{T}}\boldsymbol{y}.$$

It should be pointed out, polynomial regression is a linear regression model, although its expression $\theta_0 + \theta_1 x_i + \theta_2 x_i^2 + \cdots + \theta_m x_i^m$ seems nonlinear.

Ridge Regression

It is a method for estimating the parameters of a linear regression model when independent variables are highly correlated (called multicollinearity), which adds a regularization term, also known as a shrinkage penalty term:

$$\widehat{\mathrm{RSS}}_{\mathrm{ridge}}\left(\hat{\theta}\right) = \sum_{i=1}^{n}\left(y_i - \sum_{j=0}^{m}\hat{\theta}_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{m}\left(\hat{\theta}_j\right)^2 = \mathrm{RSS}_{\mathrm{multi}}(\hat{\theta}) + \lambda\sum_{j=1}^{m}\left(\hat{\theta}_j\right)^2,$$

where λ is the tuning parameter to be determined.

Minimizing $\widehat{\mathrm{RSS}}_{\mathrm{ridge}}\left(\hat{\theta}\right)$ requires a trade-off: $\mathrm{RSS}_{\mathrm{multi}}(\hat{\theta})$ is minimized using the method of least squares to estimate the parameters, and the second term is minimized by making $\lambda\sum_{j=1}^{m}\left(\hat{\theta}_j\right)^2 \leq c$ for some constant $c > 0$.

Ridge Regression

Similar to multiple linear regression, the residual sum of squares of ridge regression is represented as follows:

$$\widehat{\mathrm{RSS}}_{\mathrm{ridge}}\left(\hat{\theta}\right) = (\boldsymbol{y} - \mathbf{X}\hat{\theta})^{\mathsf{T}}(\boldsymbol{y} - \mathbf{X}\hat{\theta}) + \lambda\hat{\theta}^{\mathsf{T}}\hat{\theta} = \boldsymbol{y}^{\mathsf{T}}\boldsymbol{y} - 2\boldsymbol{y}^{\mathsf{T}}\mathbf{X}\hat{\theta} + \hat{\theta}^{\mathsf{T}}\mathbf{X}^{\mathsf{T}}\mathbf{X}\hat{\theta} + \lambda\hat{\theta}^{\mathsf{T}}\hat{\theta}.$$

$$\frac{\partial\widehat{\mathrm{RSS}}_{\mathrm{ridge}}(\hat{\theta})}{\partial\hat{\theta}} = \frac{\partial(\boldsymbol{y}^{\mathsf{T}}\boldsymbol{y} - 2\boldsymbol{y}^{\mathsf{T}}\mathbf{X}\hat{\theta} + \hat{\theta}^{\mathsf{T}}\mathbf{X}^{\mathsf{T}}\mathbf{X}\hat{\theta} + \lambda\hat{\theta}^{\mathsf{T}}\hat{\theta}\ )}{\partial\hat{\theta}} = -2\mathbf{X}^{\mathsf{T}}\boldsymbol{y} + 2\mathbf{X}^{\mathsf{T}}\mathbf{X}\hat{\theta} + 2\lambda\hat{\theta} = 0.$$

Therefore, the parameter estimation of ridge regression is:

$$\hat{\theta}_{\mathrm{ridge}} = (\mathbf{X}^{\mathsf{T}}\mathbf{X} - \lambda\mathbf{I})^{-1}\mathbf{X}^{\mathsf{T}}\boldsymbol{y},$$

where $\mathbf{I}$ is the identity matrix.

Lasso Regression

“Lasso” is an abbreviation for “least absolute shrinkage and selection operator”.

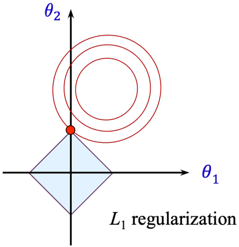
$$\widehat{\mathrm{RSS}}_{\mathrm{lasso}}(\hat{\theta}) = \sum_{i=1}^{n}\left(y_i - \sum_{j=0}^{m}\hat{\theta}_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{m}|\hat{\theta}_j| = \mathrm{RSS}_{\mathrm{multi}}(\hat{\theta}) + \lambda\sum_{j=1}^{m}|\hat{\theta}_j|.$$

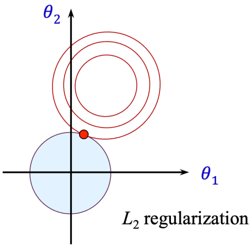
Similarly, lasso regression is also for a certain constant $c > 0$, making its regularization term $\sum_{j=1}^{m}\left|\hat{\theta}_j\right| \leq c$. However, when the parameter λ is sufficiently large, the L_1 regularization of lasso regression has the effect of forcing some parameter estimates to be exactly zero.

Therefore, lasso regression is usually easier for variable selection and exclusion of irrelevant variables than ridge regression.

Lasso Regression

For the regularization term, ridge regression $\sum_{j=1}^{m}\left(\hat{\theta}_j\right)^2$ is L_2 regularization, and lasso regression $\sum_{j=1}^{p}\left|\hat{\theta}_j\right|$ is L_1 regularization, $\|\hat{\theta}\|_1 = \sum|\hat{\theta}_j|$.





This is a subtle but important change, as some parameters of the lasso regression can be precisely reduced to zero.

Lasso Regression

Similar to multiple linear regression, the residual sum of squares of lasso regression is represented as follows:

$$\widehat{\mathrm{RSS}}_{\mathrm{lasso}}(\hat{\theta}) = (\boldsymbol{y} - \mathbf{X}\hat{\theta})^{\intercal}(\boldsymbol{y} - \mathbf{X}\hat{\theta}) + \lambda\hat{\theta} = \boldsymbol{y}^{\intercal}\boldsymbol{y} - 2\boldsymbol{y}^{\intercal}\mathbf{X}\hat{\theta} + \hat{\theta}^{\intercal}\mathbf{X}^{\intercal}\mathbf{X}\hat{\theta} + \lambda\hat{\theta}.$$

$$\frac{\partial\widehat{\mathrm{RSS}}_{\mathrm{lasso}}(\hat{\theta})}{\partial\hat{\theta}} = \frac{\partial(\boldsymbol{y}^{\intercal}\boldsymbol{y} - 2\boldsymbol{y}^{\intercal}\mathbf{X}\hat{\theta} + \hat{\theta}^{\intercal}\mathbf{X}^{\intercal}\mathbf{X}\hat{\theta} + \lambda\hat{\theta}\ )}{\partial\hat{\theta}} = -2\mathbf{X}^{\intercal}\boldsymbol{y} + 2\mathbf{X}^{\intercal}\mathbf{X}\hat{\theta} + \lambda = 0.$$

Therefore, the parameter estimation for lasso regression is:

$$\hat{\theta}_{\mathrm{lasso}} = (2\mathbf{X}^{\intercal}\mathbf{X})^{-1}(2\mathbf{X}^{\intercal}\boldsymbol{y} - \lambda).$$

Bayesian Regression

The multiple linear regression, polynomial regression, ridge regression, and lasso regression are all based on the method of least squares.

Bayesian regression is a linear regression model based on Bayesian inference.

Given training samples $S = \{(\boldsymbol{x}_i, y_i) \mid i = 1, \ldots, n\}$, the objective function $\boldsymbol{y} = \mathbf{X}\theta + \varepsilon$. Based on Gaussian distribution, θ and ε can be represented as:

$$\theta \sim \mathcal{N}\left(0, \Sigma\right), \quad \varepsilon \sim \mathcal{N}(0, \sigma^2 \mathbf{I}),$$

where Σ, σ^2 are known, and $\mathbf{I}$ is an identity matrix. Based on Bayes' theorem:

$$p\left(\theta | S\right) = p\left(\theta | \mathbf{X}, \boldsymbol{y}\right) = \frac{p(\boldsymbol{y} | \mathbf{X}, \theta) p(\theta)}{p(\boldsymbol{y} | \mathbf{X})}.$$

Bayesian Regression

$$p\left(\boldsymbol{y}\middle|\mathbf{X},\theta\right)=\mathcal{N}\left(\boldsymbol{y}\middle|\mathbf{X}\theta,\sigma^2\mathbf{I}\right),\quad p\left(\theta\right)=\mathcal{N}\left(\theta|0,\Sigma\right).$$

$$\begin{aligned}p\left(\theta|\mathbf{X},\boldsymbol{y}\right)&\propto p(\boldsymbol{y}|\mathbf{X},\theta)p(\theta)=\mathcal{N}\left(\boldsymbol{y}\middle|\mathbf{X}\theta,\sigma^2\mathbf{I}\right)\mathcal{N}\left(\theta|0,\Sigma\right)\\&\propto\exp\left(-\frac{1}{2\sigma^2}\left(\boldsymbol{y}-\mathbf{X}\theta\right)^\intercal\left(\boldsymbol{y}-\mathbf{X}\theta\right)\right)\exp\left(-\frac{1}{2}\theta^\intercal\Sigma^{-1}\theta\right)\\&=\exp\left(-\frac{1}{2}\left(\frac{1}{\sigma^2}\left(\boldsymbol{y}-\mathbf{X}\theta\right)^\intercal\left(\boldsymbol{y}-\mathbf{X}\theta\right)+\theta^\intercal\Sigma^{-1}\theta\right)\right).\end{aligned}$$

$$\begin{aligned}\frac{1}{\sigma^2}\left(\boldsymbol{y}-\mathbf{X}\theta\right)^\intercal\left(\boldsymbol{y}-\mathbf{X}\theta\right)+\theta^\intercal\Sigma^{-1}\theta&=\frac{1}{\sigma^2}\left(\theta^\intercal\mathbf{X}^\intercal\mathbf{X}\theta-2\theta^\intercal\mathbf{X}^\intercal\boldsymbol{y}+\boldsymbol{y}^\intercal\boldsymbol{y}\right)+\theta^\intercal\Sigma^{-1}\theta\\&=\theta^\intercal\left(\frac{1}{\sigma^2}\mathbf{X}^\intercal\mathbf{X}+\Sigma^{-1}\right)\theta-2\frac{1}{\sigma^2}\boldsymbol{y}^\intercal\mathbf{X}\theta.\end{aligned}$$

Bayesian Regression

Let $M = \frac{1}{\sigma^2}\mathbf{X}^\intercal\mathbf{X} + \Sigma^{-1}$, $b = \frac{1}{\sigma^2}\mathbf{X}^\intercal\mathbf{y}$, then the above expression becomes:

$$\frac{1}{\sigma^2}\left(\boldsymbol{y} - \mathbf{X}\theta\right)^\intercal\left(\boldsymbol{y} - \mathbf{X}\theta\right) + \theta^\intercal\Sigma^{-1}\theta = \theta^\intercal M\theta - 2b^\intercal\theta = \left(\theta - M^{-1}b\right)^\intercal M(\theta - M^{-1}b) - b^\intercal M^{-1}b.$$

Substitute the above exponential part, so we have:

$$\begin{aligned} p\left(\theta|\mathbf{X}, \boldsymbol{y}\right) &\propto \exp\left(-\frac{1}{2}\left(\left(\theta - M^{-1}b\right)^\intercal M\left(\theta - M^{-1}b\right) - b^\intercal M^{-1}b\right)\right) \\ &= \exp\left(-\frac{1}{2}\left(\left(\theta - M^{-1}b\right)^\intercal M\left(\theta - M^{-1}b\right)\right)\right)\exp\left(-\frac{1}{2}(-b^\intercal M^{-1}b)\right) \\ &\propto \exp\left(-\frac{1}{2}\left(\left(\theta - M^{-1}b\right)^\intercal M\left(\theta - M^{-1}b\right)\right)\right). \end{aligned}$$

Bayesian Regression

$$\mathcal{N}\left(\theta|\mu,\Sigma\right) = |2\pi\Sigma|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}\left(\theta-\mu\right)^{\intercal}\Sigma^{-1}\left(\theta-\mu\right)\right).$$

Therefore, we have: $p\left(\theta|\mathbf{X},\boldsymbol{y}\right) \propto \mathcal{N}\left(\theta|M^{-1}b, M^{-1}\right)$.

Substituting $M = \frac{1}{\sigma^2}\mathbf{X}^{\intercal}\mathbf{X} + \Sigma^{-1}$ and $b = \frac{1}{\sigma^2}\mathbf{X}^{\intercal}\mathbf{y}$, the posterior mean $M^{-1}b$ is:

$$M^{-1}b = \left(\frac{1}{\sigma^2}\mathbf{X}^{\intercal}\mathbf{X} + \Sigma^{-1}\right)^{-1}\frac{1}{\sigma^2}\mathbf{X}^{\intercal}\boldsymbol{y} = \left(\mathbf{X}^{\intercal}\mathbf{X} + \sigma^2\Sigma^{-1}\right)^{-1}\mathbf{X}^{\intercal}\boldsymbol{y} = \left(\mathbf{X}^{\intercal}\mathbf{X} + \sigma^2\Lambda\right)^{-1}\mathbf{X}^{\intercal}\boldsymbol{y},$$

where $\Lambda = \Sigma^{-1}$. Therefore, M^{-1} is:

$$M^{-1} = \left(\frac{1}{\sigma^2}\mathbf{X}^{\intercal}\mathbf{X} + \Sigma^{-1}\right)^{-1} = \sigma^2\left(\mathbf{X}^{\intercal}\mathbf{X} + \sigma^2\Lambda\right)^{-1}.$$

13 Regression Task

13.1 Problem and Definition

13.2 Working Principle

13.3 Related Elements

13.4 Typical Algorithms

13.5 Loss Functions

About Loss Functions for Regression Model

For a regression model, since its label's values are real numbers, it is unrealistic to expect the model to accurately predict the actual values.

So, we hope its predicted values are as close as possible to the actual ones.

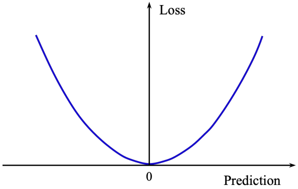
This is the essential difference between regression and classification:

- the loss function of a regression model measures the error between the predicted value and the actual value,
- the loss function of a classification model measures whether it is same between the predicted value and the actual value.

Mean Square Error

Mean square error (MSE), also known as quadratic loss, or L_2 loss, is one of the most common regression loss functions. Its expression is as follows:

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2.$$



In the figure, horizontal axis is predicted value, and vertical axis is mean square error (MSE).

A good regression model will have an MSE value closer to zero.

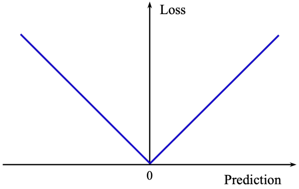
MSE is most commonly used loss function in regression because it is easy to differentiate, making model optimization easy, and it has stable properties.

Mean Absolute Error

Mean absolute error (MAE), also known as L_1 loss, is one of the simplest loss functions used to evaluate the effectiveness of a regression model.

The definition of mean absolute error is as follows:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i|.$$



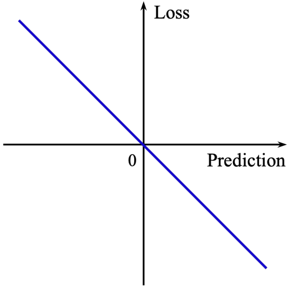
MAE is the average of absolute errors, measuring the size of the error, but not the direction of error.

MAE is one of the simplest loss functions. The lower the MAE, the higher the accuracy of the model.

Mean Bias Error

The definition of mean bias error (MBE) is as follows:

$$\mathrm{MBE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i).$$



The bias in MBE is a tendency to overestimate or underestimate the parameter value.

Positive bias means the error is overestimated, and negative bias means the error is underestimated.

MBE is similar to MAE, the only difference being that MBE does not take the absolute value.

Relative Absolute Error and Relative Squared Error

Relative absolute error (RAE) is dividing the total absolute error by the sum of the absolute differences between actual values and average value.

Relative squared error (RSE) is dividing the mean squared error (MSE) by the square of the difference between actual value and average value.

$$\text{RAE} = \frac{\sum_{i=1}^{n} |y_i - \hat{y}_i|}{\sum_{i=1}^{n} |y_i - \bar{y}|}, \qquad \text{RSE} = \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2},$$

where $\bar{y}$ is the average of n actual values y_i, and its equation is:

$$\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i.$$

Huber Loss

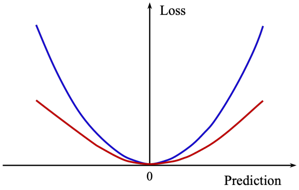
Huber loss is also known as smooth mean absolute error (SMAE).

It is a combination of linear and quadratic scoring methods, balances and combines the characteristics of MAE and MSE.

$$\mathcal{L}_\delta(y,\hat{y}) = \begin{cases} \frac{1}{2}\left(y_i - \hat{y}_i\right)^2, & \text{for } |y_i - \hat{y}_i| \leq \delta \\ \delta\left(|y_i - \hat{y}_i| - \frac{1}{2}\delta\right) & \text{otherwise.} \end{cases}$$

It has a hyperparameter δ, which can be adjusted according to the data.

For values greater than δ, the loss will be linear (L_1 loss), otherwise the loss will be quadratic (L_2 loss).



Log-Cosh Loss

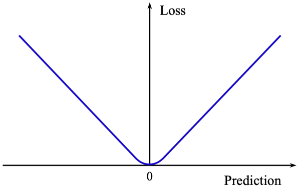
Log-Cosh loss calculates the logarithm of the hyperbolic cosine of the error.

It works similarly to MSE, but is not affected by larger prediction errors.

$$\mathcal{L}(y, \hat{y}) = \sum_{i=1}^{n} \log\left(\cosh(y_i - \hat{y}_i)\right).$$

This function is smoother than the quadratic loss function.

The Log-Cosh loss is very similar to the Huber loss, as it is a combination of linear and quadratic scoring methods.



Coefficient of Determination

The coefficient of determination, also known as R-squared (R^2), represents the proportion of the variance in the dependent variable.

$$\mathrm{R}^2 = 1 - \mathrm{RSE} = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}.$$

It can take any value in 0 to 1, but meaning varies with different values:

- If the value is 0, it means that the dependent variable cannot be explained by the independent variable.
- If the value is 1, it means that the independent variable can perfectly explain the dependent variable, with no error.

Thank You