MACHINE LEARNING

Linear Models: Linear Regression

Last Update: 9th November 2022

Prof. Dr. Shadi Albarqouni

Director of Computational Imaging Research Lab. (Albarqouni Lab.)
University Hopsital Bonn | University of Bonn | Helmholtz Munich

STRUCTURE

LINEAR REGRESSION

LINEAR REGRESSION

Predicting Manhattan Rent with Linear Regression

An Analysis of StreetEasy Rental Listings

Photo by Florian Wehde on Unsplash

LINEAR REGRESSION -- HYPOTHESIS REPRESENTATION

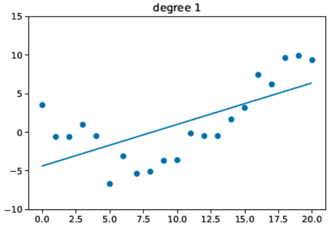
Definition

Linear regression is a widely used regression model $p(y|\mathbf{x};\theta) = \mathcal{N}(y|\mathbf{w}^T\mathbf{x} + b, \sigma^2)$ for predicting a real-valued output $\mathbf{y} \in \mathbb{R}$, given a fixed-dimensional input vector $\mathbf{x} \in \mathbb{R}^D$ (also called independent variables, explanatory variables, or covariates) where $\theta = (\mathbf{w}, b, \sigma^2)$ are the parameters with $\mathbf{w}$ as weights or regression coefficients and b or w_0 as the offset or bias term.

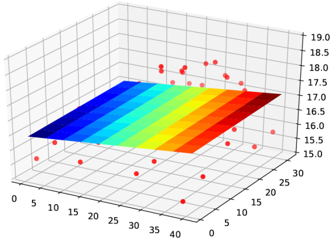
The key property of the model is that the expected value of the output is assumed to be a linear function of the input, $\mathbb{E}[y|x] = w^T x$, which makes the model easy to interpret, and easy to fit to data.

TERMINOLOGY

Simple linear regression: The input is one-dimensional (so $D = 1$), the model has the form $f(x; w) = ax + b$, where $b = w_0$ is the intercept, and $a = w_1$ is the slope.



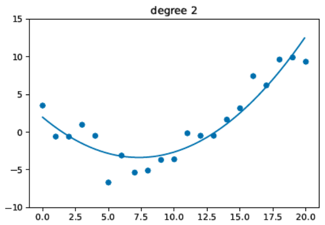
Multiple linear regression: The input is multi-dimensional, $x \in \mathbb{R}^D$ where $D > 1$.

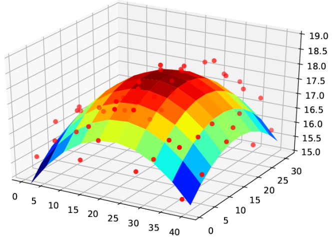


TERMINOLOGY -- CONT.

Multivariate linear regression: The output is multi-dimensional, $y \in \mathbb{R}^J$ where $J > 1$, and the likelihood can be written as $p(y|\mathbf{x};\theta) = \prod_{j=1}^{J} \mathcal{N}(y_j|\mathbf{w}_j^T\mathbf{x}, \sigma_j^2)$

Polynomial linear regression: A non-linear transformation $\phi(\cdot)$, e.g., a polynomial expanision of degree d is applied to the input vector. Consider a one-dimensional input (so $D = 1$), the $\phi(x) = [1, x, x^2, \ldots, x^d]$ and the likelihood can be written as $p(y|\mathbf{x};\theta) = \mathcal{N}(y|\mathbf{w}^T\phi(\mathbf{x}), \sigma^2)$





Polynomial Linear Regression in for 1D and 2D inputs

MAXIMUM LIKELIHOOD ESTIMATION (MLE)

Maximum Likelihood Estimation (MLE)

It can be obtained by minimizing the Negative Log Likelihood as an objective function

$$\theta_{MLE} = \arg\min_{\theta} NLL(\theta) \qquad \text{where} \qquad \theta = (\mathbf{w}, b, \sigma^2) \triangleq (\mathbf{w}, \sigma^2)^1$$

The Negative Log Likelihood (NLL) for the linear regression is given by

$$NLL(\theta) = -\log \prod_{n=1}^{N} \underbrace{\mathcal{N}(y_n|w^T x_n + b, \sigma^2)}_{p(y_n|x_n;\theta)} \triangleq \frac{1}{2\sigma^2} \sum_{n=1}^{N} (y_n - \hat{y_n})^2 + \frac{N}{2} \log\left(2\pi\sigma^2\right)$$

where $\hat{y_n} = f(\mathbf{x}_n; \mathbf{w}) = \mathbf{w}^T \mathbf{x}_n$ is the prediction with bias $w_0 = b$ and $x_0 = 1$.

The NLL is equal (up to irrelevant constants) to the residual sum of squares,

$$\text{RSS}(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^{N} (y_n - \hat{y_n})^2 = \frac{1}{2} \|\mathbf{X}\mathbf{w} - \mathbf{y}\|_2^2 = \frac{1}{2} (\mathbf{X}\mathbf{w} - \mathbf{y})^T (\mathbf{X}\mathbf{w} - \mathbf{y})$$

[1] b is included in w by simply adding a column with a value of 1 to the feature vector $x_{1:D+1} = [1, x_{1:D}]$

LINEAR REGRESSION AS SYSTEMS OF EQUATIONS

Linear regression problem as systems of equations

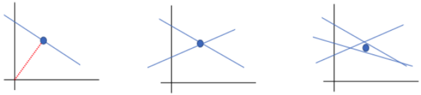
$$
\begin{aligned}
y_1 &= w_0 + w_1 x_{11} + \cdots + w_D x_{1D} \\
y_2 &= w_0 + w_1 x_{21} + \cdots + w_D x_{2D} \\
&\cdots \\
y_N &= w_0 + w_1 x_{N1} + \cdots + w_D x_{ND}
\end{aligned}
$$

The system of equations can be written in a matrix form as $y = Xw$ with $Y \in \mathbb{R}^N$ as targets, $X \in \mathbb{R}^{N \times D}$ as design input matrix, and $w \in \mathbb{R}^D$ as the weight parameters.

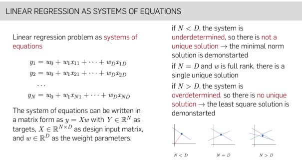
if $N < D$, the system is underdetermined, so there is not a unique solution $\rightarrow$ the minimal norm solution is demonstarted

if $N = D$ and w is full rank, there is a single unique solution

if $N > D$, the system is overdetermined, so there is no unique solution $\rightarrow$ the least square solution is demonstarted



$N < D$ $\quad$ $N = D$ $\quad$ $N > D$



LINEAR REGRESSION AS SYSTEMS OF EQUATIONS

Linear regression problem as systems of equations

$$
\begin{aligned}
y_1 &= w_0 + w_1 x_{11} + \cdots + w_D x_{1D} \\
y_2 &= w_0 + w_1 x_{21} + \cdots + w_D x_{2D} \\
&\cdots \\
y_N &= w_0 + w_1 x_{N1} + \cdots + w_D x_{ND}
\end{aligned}
$$

The system of equations can be written in a matrix form as $y = Xw$ with $Y \in \mathbb{R}^N$ as targets, $X \in \mathbb{R}^{N \times D}$ as design input matrix, and $w \in \mathbb{R}^D$ as the weight parameters.

if $N < D$, the system is underdetermined, so there is not a unique solution $\rightarrow$ the minimal norm solution is demonstarted

if $N = D$ and w is full rank, there is a single unique solution

if $N > D$, the system is overdetermined, so there is no unique solution $\rightarrow$ the least square solution is demonstarted

$N < D$ $\quad$ $N = D$ $\quad$ $N > D$

Given the following systems of equations,

$$
\begin{aligned}
2 &= 3w_1 + 2w_2 \\
-2 &= 2w_1 - 2w_2
\end{aligned}
$$

it can be written in the matrix form $y = Xw$ as

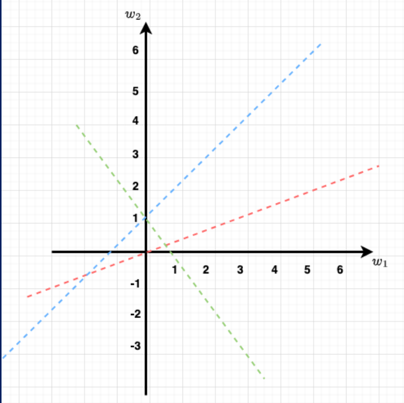
$$
y = \begin{pmatrix} 2 \\ -2 \end{pmatrix}, \quad X = \begin{pmatrix} 3 & 2 \\ 2 & -2 \end{pmatrix}
$$

Since $N = D$, we can simply solve the systems using

$$
w = X^{-1} y = \frac{1}{|detX|} \begin{pmatrix} -2 & -2 \\ -2 & 3 \end{pmatrix} \begin{pmatrix} 2 \\ -2 \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}.
$$

What happens if:

- we have only the first data point (equation)?
- we have an additional data point $0 = -w_1 + 3w_2$?



LEAST NORM ESTIMATION

When $N < D$ (short and fat), the system is underdetermined, so there is not a unique solution $\rightarrow$ Least norm estimation?

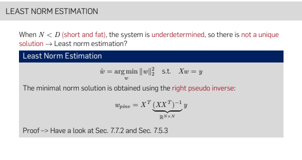
Least Norm Estimation

$$\hat{w} = \arg\min_{w} \|w\|_2^2 \quad \text{s.t.} \quad Xw = y$$

The minimal norm solution is obtained using the right pseudo inverse:

$$w_{pinv} = X^T \underbrace{(XX^T)^{-1}}_{\mathbb{R}^{N \times N}} y$$

Proof –> Have a look at Sec. 7.7.2 and Sec. 7.5.3



LEAST NORM ESTIMATION

When $N < D$ (short and fat), the system is underdetermined, so there is not a unique solution → Least norm estimation?

Least Norm Estimation

$$\hat{w} = \arg\min_w \|w\|_2^2 \quad \text{s.t.} \quad Xw = y$$

The minimal norm solution is obtained using the right pseudo inverse:

$$w_{pinv} = X^T \underbrace{(XX^T)^{-1}}_{\mathbb{R}^{N \times N}} y$$

Proof -> Have a look at Sec. 7.7.2 and Sec. 7.5.3

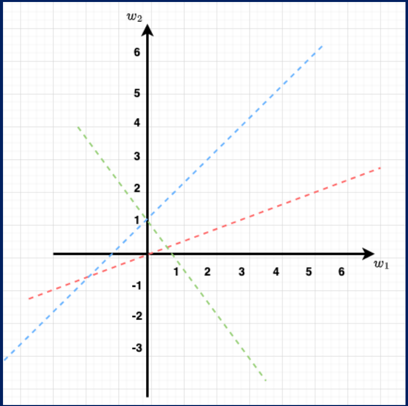
What happens if:

we have only the first data point (equation)?

Given the following systems of equations,

$$2 = 3w_1 + 2w_2$$

it can be written in the matrix form $y = Xw$ as
$y = 2, \quad X = \begin{pmatrix} 3 & 2 \end{pmatrix}$
Since $N < D$, we can simply solve the systems using
$w_{pinv} = X^T(XX^T)^{-1}y =$
$$\begin{pmatrix} 3 \\ 2 \end{pmatrix} \left(\begin{pmatrix} 3 & 2 \end{pmatrix} \begin{pmatrix} 3 \\ 2 \end{pmatrix} \right)^{-1} (2) = \begin{pmatrix} 0.46 \\ 0.31 \end{pmatrix}.$$



LEAST SQUARES ESTIMATION

When $N > D$ (tall and skinny), the system is overdetermined, so there is no unique solution $\rightarrow$ Least square estimation?

Least Squares Estimation (LSE)

To find the solution that gets as close as possible to satisfying all of the constraints specified by $y = Xw$, we need to minimize the following cost function, known as the least squares objective

$$\hat{w} = \arg\min_{w} \frac{1}{2} \|Xw - y\|_2^2$$

The corresponding solution known as ordinary least squares (OLS) is obtained using the left pseudo inverse or by taking the derivative w.r.t w, $\nabla_w RSS(w) = 0$,

$$X^T(Xw - y) = 0 \rightarrow w_{OLE} = \underbrace{(X^TX)^{-1}}_{\mathbb{R}^{D\times D}} X^T y$$

LEAST SQUARES ESTIMATION

When $N > D$ (tall and skinny), the system is overdetermined, so there is no unique solution $\rightarrow$ Least square estimation?

Least Squares Estimation (LSE)

To find the solution that gets as close as possible to satisfying all of the constraints specified by $y = Xw$, we need to minimize the following cost function, known as the least squares objective

$$\hat{w} = \arg\min_{w} \frac{1}{2}\|Xw - y\|_2^2$$

The corresponding solution known as ordinary least squares (OLS) is obtained using the left pseudo inverse or by taking the derivative w.r.t w, $\nabla_w RSS(w) = 0$,

$$X^T(Xw - y) = 0 \rightarrow w_{OLE} = \underbrace{(X^TX)^{-1}}_{\mathbb{R}^{D\times D}} X^Ty$$

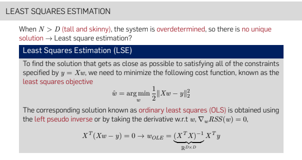
Let $\nabla_w RSS(w) = 0$

$$\nabla_w \frac{1}{2}(Xw - y)^T(Xw - y) = 0$$

$$\frac{1}{2}(2)X^T(Xw - y) = 0$$

$$X^TXw - X^Ty = 0$$

$$w = (X^TX)^{-1}X^Ty$$



LEAST SQUARES ESTIMATION

When $N > D$ (tall and skinny), the system is overdetermined, so there is no unique solution → Least square estimation?

Least Squares Estimation (LSE)

To find the solution that gets as close as possible to satisfying all of the constraints specified by $y = Xw$, we need to minimize the following cost function, known as the least squares objective

$$\hat{w} = \arg\min_{w} \frac{1}{2}\|Xw - y\|_2^2$$

The corresponding solution known as ordinary least squares (OLS) is obtained using the left pseudo inverse or by taking the derivative w.r.t w, $\nabla_w RSS(w) = 0$,

$$X^T(Xw - y) = 0 \rightarrow w_{OLE} = \underbrace{(X^TX)^{-1}}_{\mathbb{R}^{D\times D}} X^T y$$

What happens if:

we have an additional data point $0 = -w_1 + 3w_2$?

Given the following systems of equations,

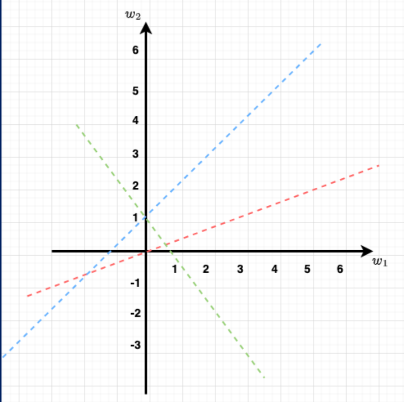
$$\begin{aligned} 2 &= 3w_1 + 2w_2 \\ -2 &= 2w_1 - 2w_2 \\ 0 &= -w_1 + 3w_2 \end{aligned}$$

it can be written in the matrix form $y = Xw$ as

$$y = \begin{pmatrix} 2 \\ -2 \\ 0 \end{pmatrix}, \quad X = \begin{pmatrix} 3 & 2 \\ 2 & -2 \\ -1 & 3 \end{pmatrix}$$

Since $N > D$, we can simply solve the systems using

$$w_{OLE} = (X^TX)^{-1}X^Ty = \begin{pmatrix} 0.18 & 0.48 \end{pmatrix}^T$$

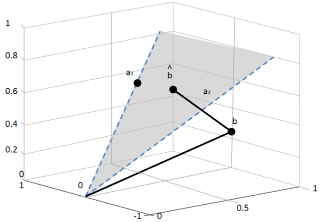


GEOMETRIC INTERPRETATION OF LEAST SQUARES

Our objective is to find the optimal parameters that minimizes the following objective function:

$$\underset{\hat{y}\in \text{span}([x_{:,1},\dots,x_{:,d}])}{\arg\min} \|y-\hat{y}\|_2$$

where $x_{:,d}$ is the $d^{th}-$column of matrix X and $\hat{y} = Xw$ is the prediction which belongs to the span(X).



Geometric representation[2]

It tunrs out that the shortest path with minimal distance (residuals) is the orthogonal projection of y into the supspace span(X), i.e., $x_{:,D} \perp (y-\hat{y})$. This is translated to: $x_{:,D}^T(y-\hat{y}) = 0 \rightarrow X^T(y-Xw) \triangleq X^Ty - X^TXw = 0$, thus $w_{opt} = (X^TX)^{-1}X^Ty \rightarrow$ ordinary least squares (OLS).

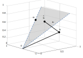
[2]The figure considers $b = Ax$

GEOMETRIC INTERPRETATION OF LEAST SQUARES

Our objective is to find the optimal parameters that minimizes the following objective function:

$$\underset{\hat{y} \in \text{span}([x_{:,1}, \ldots, x_{:,d}])}{\arg\min} \|y - \hat{y}\|_2$$

where $x_{:,d}$ is the d^{th}−column of matrix X and $\hat{y} = Xw$ is the prediction which belongs to the span(X).



Geometric representation[2]

It tunrs out that the shortest path with minimal distance (residuals) is the orthogonal projection of y into the supspace span(X), i.e., $x_{:,D} \perp (y - \hat{y})$. This is translated to: $x_{:,D}^T(y - \hat{y}) = 0 \rightarrow X^T(y - Xw) \triangleq X^T y - X^T X w = 0$, thus $w_{opt} = (X^T X)^{-1} X^T y \rightarrow$ ordinary least squares (OLS).

[2] The figure considers $b = Ax$

Given the following systems of equations $y = Xw$ as

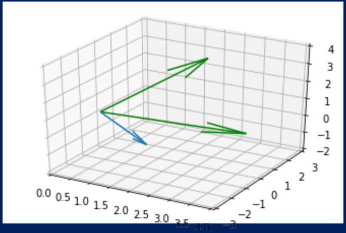
$$y = \begin{pmatrix} 2 \\ -2 \\ 0 \end{pmatrix}, \quad X = \begin{pmatrix} 3 & 2 \\ 2 & -2 \\ -1 & 3 \end{pmatrix},$$

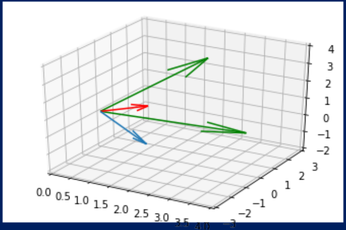
the geometric interpretation of the residual sum of squares is presented on thr right hand side where

$$\begin{pmatrix} 3 \\ 2 \\ -1 \end{pmatrix} w_1 + \begin{pmatrix} 2 \\ -2 \\ 3 \end{pmatrix} w_2 = \begin{pmatrix} 2 \\ -2 \\ 0 \end{pmatrix}$$

$$w_{OLE} = (X^T X)^{-1} X^T y = \begin{pmatrix} 0.18 & 0.48 \end{pmatrix}^T$$

$$\hat{y} = Proj(x) y = \begin{pmatrix} 1.49 & -0.61 & 1.27 \end{pmatrix}^T$$

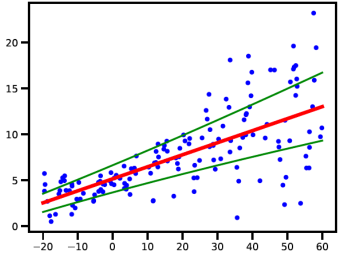




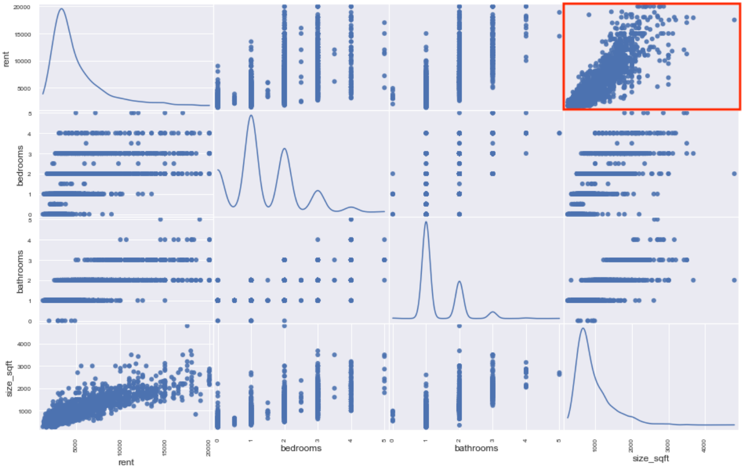
Weighted Least Squares

In some cases, we want to associate a weight with each example. For example, in heteroskedastic regression, the variance depends on the input, so the model has the form $p(y|\mathbf{x};\theta) = \mathcal{N}(y|X\mathbf{w},\Lambda^{-1})$ where $\Lambda = \text{diag}(1/\sigma^2(x))$

$$\hat{w}_{wLSE} = (X^T\Lambda X)^{-1}X^T\Lambda y$$



Scatter Matrix of Features Correlated with Rent



Source:**https://towardsdatascience.com/predicting-manhattan-rent-with-linear-regression-27766041d2d9**

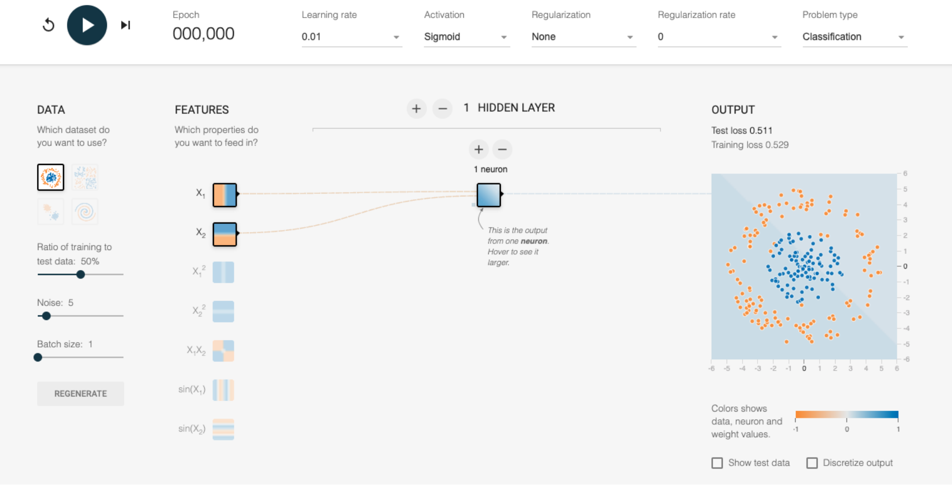
ALGORITHMIC ISSUES

When $N >> D$ (tall and skinny), the system is overdetermined, so there is no unique solution $\rightarrow$

$$w_{OLE} = \underbrace{(X^T X)^{-1}}_{\mathbb{R}^{D\times D}} X^T y$$

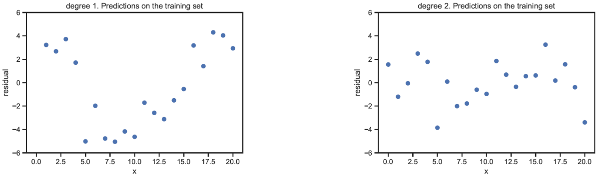
- numerical reasons – $X^T X$ may be ill conditioned or singular (look at this example)
- alternative and less expesnive solutions are SVD and QR decompositions
- alternative to direct methods based on matrix decomposition is iterative solvers
- standarize the data (see Sec. 10.2.8)

DEMO


Epoch
000,000
Learning rate
0.01
Activation
Sigmoid
Regularization
None
Regularization rate
0
Problem type
Classification
DATA
Which dataset do you want to use?
Ratio of training to test data: 50%
Noise: 5
Batch size: 1
REGENERATE
FEATURES
Which properties do you want to feed in?
X_1
X_2
X_1^2
X_2^2
X_1X_2
$\sin(X_1)$
$\sin(X_2)$
1 HIDDEN LAYER
1 neuron
This is the output from one **neuron**. Hover to see it larger.
OUTPUT
Test loss 0.511
Training loss 0.529
Colors shows data, neuron and weight values.
Show test data
Discretize output

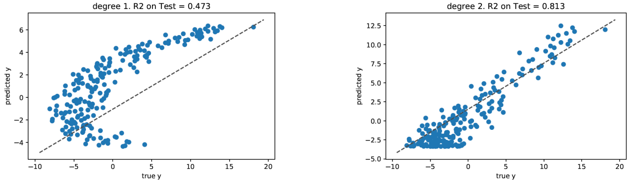
MEASURING GOODNESS OF FIT -- RESIDUAL PLOT

Residual plot for 1D feature



(a) (b)

Residual plot for Multi-dimensional feature



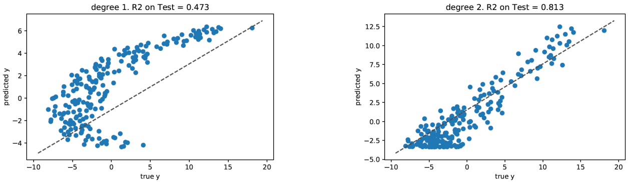
(a) (b)

MEASURING GOODNESS OF FIT -- PREDICTION ACCURACY AND R^2

Residual Sum of Square (RSS): $\frac{1}{2}\sum_{n=1}^{N}(y_n - \hat{y_n})^2$

Root Mean Square Error (RMSE): $\sqrt{\frac{1}{N}\text{RSS}}$

Coefficient of determination: $R^2 = 1 - \frac{RSS}{TSS} = 1 - \frac{\sum_{n=1}^{N}(y_n - \hat{y_n})^2}{\sum_{n=1}^{N}(y_n - \bar{y})^2}$



(a) (b)

RIDGE REGRESSION

Maximum likelihood estimation can result in overfitting. A simple solution to this is to use MAP estimation with a zero-mean Gaussian prior on the weights, $p(w) = \mathcal{N}(w|0, \lambda^{-1}I)$. This is called ridge regression.

Maximum A Posterior

$$\hat{w}_{MAP} = \arg\min_{w} \frac{1}{2}\|Xw - y\|_2^2 + \lambda\|w\|_2^2$$

The corresponding solution known as Maximum A Posterior (MAP) is obtained by taking the derivative w.r.t w, e.g., $\nabla_w RSS(w) + \lambda\|w\|_2^2 = 0$

$$w_{MAP} = (X^T X + \lambda I)^{-1} X^T y$$

RIDGE REGRESSION

Maximum likelihood estimation can result in overfitting. A simple solution to this is to use MAP estimation with a zero-mean Gaussian prior on the weights, $p(w) = \mathcal{N}(w|0, \lambda^{-1}I)$. This is called ridge regression.

Maximum A Posterior

$$\hat{w}_{MAP} = \arg\min_{w} \frac{1}{2}\|Xw - y\|_2^2 + \lambda\|w\|_2^2$$

The corresponding solution known as Maximum A Posterior (MAP) is obtained by taking the derivative w.r.t w, e.g., $\nabla_w RSS(w) + \lambda\|w\|_2^2 = 0$

$$w_{MAP} = (X^T X + \lambda I)^{-1} X^T y$$

$$\nabla_w RSS(w) + \lambda\|w\|_2^2 = 0$$

$$\nabla_w (Xw - y)^T(Xw - y) + \lambda w^T w \triangleq X^T(Xw - y) + \lambda I w = 0$$

$$X^T X w - X^T y + \lambda I w \triangleq (X^T X + \lambda I)w - X^T y = 0$$

$$w_{MAP} = (X^T X + \lambda I)^{-1} X^T y$$

LASSO REGRESSION

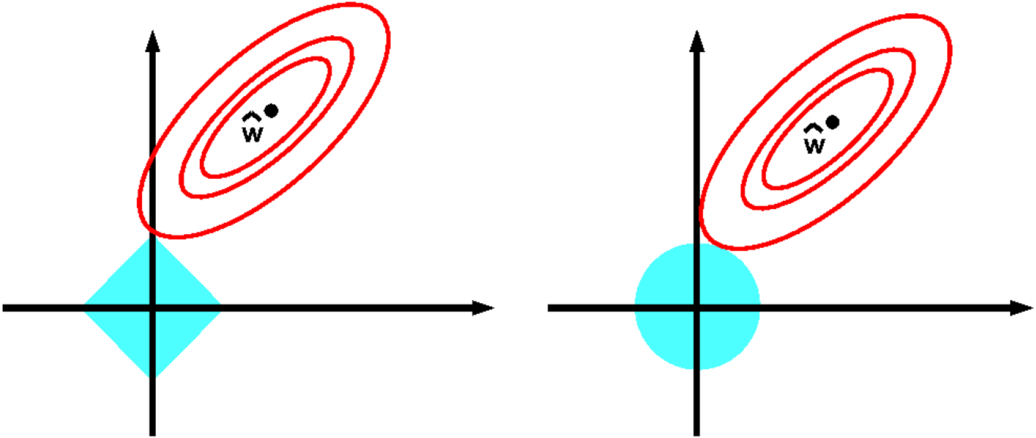
Sometimes we want the parameters to not just be small, but to be exactly zero (compression), i.e., we want w to be sparse, so that we maximize the likelihood $p(w) = \text{Laplace}(w|0, \lambda^{-1})$

Maximum A Posterior

$$\hat{w}_{MAP} = \arg\min_{w} \frac{1}{2}\|Xw - y\|_2^2 + \lambda\|w\|_1$$

where $\|w\|_1 = \sum_{d=1}^{D} |w|$ is the ℓ_1-norm of w.
The corresponding solution known as Maximum A Posterior (MAP). This is mainly used to perform feature selection

LASSO VS. RIDGE REGRESSION



Lasso Ridge

Questions