Linear regression

Victor Kitov

v.v.kitov@yandex.ru

Yandex School of Data Analysis

Table of Contents

Linear regression

- Linear model $f(x,\beta)=\langle x,\beta\rangle=\sum_{i=1}^{D}\beta_i x^i$
- Define $X\in\mathbb{R}^{NxD}, \{X\}_{ij}$ defines the j-th feature of i-th object, $Y\in\mathbb{R}^n$, $\{Y\}_i$ - target value for i-th object.
- Ordinary least squares (OLS) method:

$$\sum_{n=1}^{N}(f(x_n,\beta)-y_n)^2=\sum_{n=1}^{N}\left(\sum_{d=1}^{D}\beta_d x_n^d-y_n\right)^2\to\min_{\beta}$$

Solution

Stationarity condition:

$$2\sum_{n=1}^{N} x_n \left(\sum_{d=1}^{D} \beta_d x_n^d - y_n \right) = 0$$

In matrix form:

$$2X^T(X\beta - Y) = 0$$

so

$$\widehat{\beta} = (X^T X)^{-1} X^T Y$$

This is the global minimum, because the optimized criteria is convex.

- Geometric interpretation of linear regression, estimated with OLS.

Linearly dependent features

- Solution $\widehat{\beta} = (X^T X)^{-1} X^T Y$ exists when $X^T X$ is non-degenerate
- Using property $rank(X) = rank(X^T) = rank(X^T X) = rank(XX^T)$
  - problem occurs when one of the features is a linear combination of the other
    - example: constant unity feature c and one-hot-encoding $e_1, e_2, ...e_K$, because $\sum_k e_k \equiv c$
    - interpretation: non-identifiability of $\widehat{\beta}$
  - solved using:
    - feature selection
    - extraction (e.g. PCA)
    - regularization.

Analysis of linear regression

Advantages:

- single optimum, which is global (for the non-singular matrix)
- analytical solution
- interpretability algorithm and solution

Drawbacks:

- too simple model assumptions (may not be satisfied)
- $X^T X$ should be non-degenerate (and well-conditioned)

Table of Contents

Generalization by nonlinear transformations

Nonlinearity by x in linear regression may be achieved by applying non-linear transformations to the features:

$$x \to [\phi_0(x),\ \phi_1(x),\ \phi_2(x),\ \dots\ \phi_M(x)]$$

$$f(x) = \langle \phi(x), \beta \rangle = \sum_{m=0}^{M} \beta_m \phi_m(x)$$

The model remains to be linear in w, so all advantages of linear regression remain.

Typical transformations

$\phi_k(x)$	**comments**
$\exp\left\{-\frac{\lVert x-\mu\rVert^2}{s^2}\right\}$	closeness to point μ in feature space
$x^i x^j$	interaction of features
$\ln x_k$	the alignment of the distribution with heavy tails
$F^{-1}(x_k)$	conversion of atypical continious distribution to uniform[1]

[1] why?

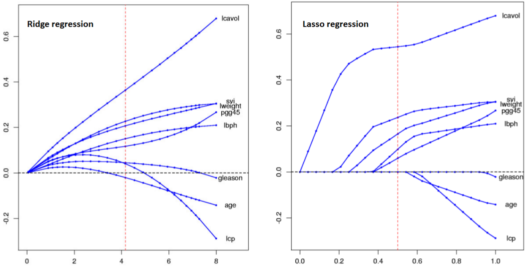
Table of Contents

Regularization

- Variants of target criteria $Q(\beta)$ with regularization[2]:

$$
\begin{array}{ll}
\sum_{n=1}^{N}\left(x_n^T\beta-y_n\right)^2+\lambda||\beta||_1 & \text{Lasso}\\
\sum_{n=1}^{N}\left(x_n^T\beta-y_n\right)^2+\lambda||\beta||_2^2 & \text{Ridge}\\
\sum_{n=1}^{N}\left(x_n^T\beta-y_n\right)^2+\lambda_1||\beta||_1+\lambda_2||\beta||_2^2 & \text{Elastic net}
\end{array}
$$

- Dependency of β from $\frac{1}{\lambda}$:



[2] Derive solution for ridge regression. Will it be uniquely defined for correlated features?

Linear monotonic regression

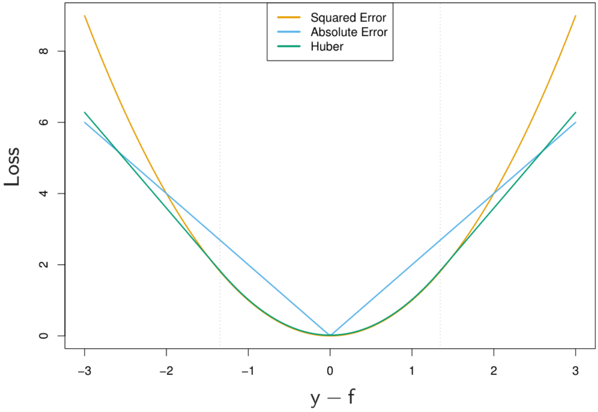
- We can impose restrictions on coefficients such as non-negativity:

$$\begin{cases} Q(\beta) = ||X\beta - Y||^2 \to \min_\beta \\ \beta_i \geq 0, \quad i = 1, 2, ...D \end{cases}$$

- Example: avaraging of forecasts of different prediction algorithms
- $\beta_i = 0$ means, that i-th component does not improve accuracy of forecasting.

Table of Contents

Non-quadratic loss functions[3][4]



[3] What is the value of constant prediction, minimizing sum of squared errors?

[4] What is the value of constant prediction, minimizing sum of absolute errors?

Conditional non-constant optimization

- For $x, y \sim P(x,y)$ and prediction being made for fixed x:

$$\arg\min_{f(x)} \mathbb{E}\left\{ (f(x)-y)^2 \middle| x \right\} = \mathbb{E}[y|x]$$

$$\arg\min_{f(x)} \mathbb{E}\left\{ |f(x)-y| \,|\, x \right\} = \text{median}[y|x]$$

Minimization of expected squared error

- Let $x, y \sim P(x, y)$ and $\mathbb{E}[y|x]$ exist. Then

$$\arg\min_{f(x)} \mathbb{E}\left\{ (f(x) - y)^2 \middle| x \right\} = \mathbb{E}[y|x]$$

$$\begin{aligned}
\mathbb{E}\left\{ (f(x)-y)^2 \middle| x\right\} &= \mathbb{E}\left\{ (f(x)-\mathbb{E}[y|x]+\mathbb{E}[y|x]-y)^2 \middle| x\right\} \\
&= \mathbb{E}\left\{ (f(x)-\mathbb{E}[y|x])^2 \middle| x\right\} + \mathbb{E}\left\{ (\mathbb{E}[y|x]-y)^2 \middle| x\right\} \\
&\quad +2\mathbb{E}\left\{ (f(x)-\mathbb{E}[y|x])(\mathbb{E}[y|x]-y) \middle| x\right\} = \\
&= (f(x)-\mathbb{E}[y|x])^2 + \mathbb{E}\left\{ (\mathbb{E}[y|x]-y)^2 \middle| x\right\} \qquad (1)
\end{aligned}$$

Minimization of expected squared error

We used

$$\mathbb{E}\left\{(f(x)-\mathbb{E}[y|x])\left(\mathbb{E}[y|x]-y\right)|\, x\right\} =$$
$$(f(x)-\mathbb{E}[y|x])\,\mathbb{E}\left\{\mathbb{E}[y|x]-y|\, x\right\} \equiv 0$$

Minimum of (1) is achieved at $f(x)=\mathbb{E}[y|x]$.

Table of Contents

Weighted account for observations[5]

- Weighted account for observations

$$\sum_{n=1}^{N} w_n (x_n^T \beta - y_n)^2$$

- Weights may be:
  - increased for incorrectly predicted objects
    - algorithm becomes more oriented on error correction
  - decreased for incorrectly predicted objects
    - they may be considered outliers that break our model

[5] Derive solution for weighted regression.

Robust regression

- Initialize $w_1 = ... = w_N = 1/N$
- Repeat:
  - estimate regression $\widehat{y}(x)$ using observations (x_i, y_i) with weights w_i.
  - for each $i = 1, 2, ...N$:
    - re-estimate $\varepsilon_i = \widehat{y}(x_i) - y_i$
    - recalculate $w_i = K(|\varepsilon_i|)$
  - normalize weights $w_i = \frac{w_i}{\sum_{n=1}^{N} w_n}$

Comments: $K(\cdot)$ is some *decreasing* function, repetition may be

- predefined number of times
- until convergence of model parameters.

Robust classification

- Initialize $w_1 = ... = w_N = 1/N$
- Repeat:
  - estimate classifier disriminant functions $\{g_y(\cdot)\}_{y=1,...C}$ using observations (x_i, y_i) with weights w_i.
  - for each $i = 1, 2, ...N$:
    - re-estimate $M_i = g_{y_i}(x_i) - \max_{y \neq y_i} g_y(x_i)$
    - recalculate $w_i = K(M_i)$
  - normalize weights $w_i = \frac{w_i}{\sum_{n=1}^{N} w_n}$

Comments: $K(\cdot)$ is some *increasing* function, repetition may be

- predefined number of times
- until convergence of model parameters.