Regression

Victor Kitov

v.v.kitov@yandex.ru

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:

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

Solution

Stationarity condition:

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

In vector 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.

Restriction of the solution

- Restriction: matrix $X^T X$ should be non-degenerate
  - occurs when one of the features is a linear combination of the other
    - interpretation: non-identifiability of $\widehat{\beta}$
  - solved using feature selection, extraction (e.g. PCA) or regularization.
  - example: constant feature $c = [1, 1, ...1]^T$ and one-hot-encoding $e_1, e_2, ...e_K$, because $\sum_k e_k \equiv c$

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)

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

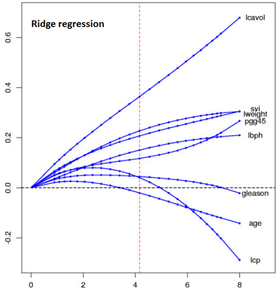
$\phi_k(x)$	**comments**
$\exp\left\{-\frac{\lVert x-\mu\rVert^2}{s^2}\right\}$	closeness to point μ in feature space
$x^i x^j$	interaction of features
$\ln x_k$	the alignment of the distribution with heavy tails
$F^{-1}(x_k)$	conversion of atypical distribution to uniform

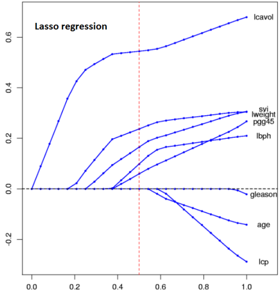
Regularization

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

$$
\begin{array}{ll}
||X\beta - Y||^2 + \lambda||\beta||_1 & \text{Lasso} \\
||X\beta - Y||^2 + \lambda||\beta||_2 & \text{Ridge} \\
||X\beta - Y||^2 + \lambda_1||\beta||_1 + \lambda_2||\beta||_2 & \text{Elastic net}
\end{array}
$$

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





Different account for different features

- Optimization task with regularization:

$$\sum_{n=1}^{N}\mathcal{L}(\widehat{y}_n, y_n|w)+\lambda R(w)\to\min_{w}$$

- Here λ controls complexity of the model:

Different account for different features

- Optimization task with regularization:

$$\sum_{n=1}^{N} \mathcal{L}(\widehat{y}_n, y_n|w) + \lambda R(w) \to \min_w$$

- Here λ controls complexity of the model:$\uparrow \lambda \Leftrightarrow$ complexity$\downarrow$.
- Suppose we have K groups of features with indices:

$$I_1, I_2, ...I_K$$

- We may control the impact of each group on the model:

$$\sum_{n=1}^{N} \mathcal{L}(\widehat{y}_n, y_n|w) + \lambda_1 R(\{w_i|i \in I_1\}) + ... + \lambda_K R(\{w_i|i \in I_K\}) \to \min_w$$

- $\lambda_1, \lambda_2, ...\lambda_K$ can be set using cross-validation

Weighted account for observations

- 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
- In probabilistic models different weights represent different variances.

Solution for weighted regression

$$\sum_{n=1}^{N} w_n \left(x_n^T \beta - y_n\right)^2 \to \min_{\beta \in \mathbb{R}}$$

Stationarity condition:

$$\sum_{n=1}^{N} w_n x_n^d \left(x_n^T \beta - y_n\right) = 0$$

Define $\{X\}_{n,d} = x_n^d$, $W = diag\{w_1, ...w_N\}$. Then

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

$$\beta = \left(X^T W X\right)^{-1} X^T W Y$$

Robust regression

- Robust means it is not affected much by outliers.
- Initialize $w_1 = ... = w_N = 1$
  - repeat until convergence of ε_i:
    - estimate regression $\widehat{y}(x)$ using observations (x_i, y_i) with weights w_i.
    - re-estimate $\varepsilon_i = \widehat{y}(x_i) - y_i$, $i = 1, 2, ...N$.
    - recalculate $w_i = w(|\varepsilon_i|)$ with $\varepsilon_1, ...\varepsilon_N$ where $w(\cdot)$ is some decreasing function.
    - normalize weights $w_i = \frac{w_i}{\sum_{n=1}^{N} w_n}$

Non-quadratic loss functions