

Regression

Victor Kitov

Table of Contents

- 1 Linear regression
- 2 Nonlinear transformations
- 3 Regularization & restrictions.
- 4 Different loss-functions
- 5 Weighted account for observations
- 6 Nadaraya-Watson regression

Linear regression

- Linear model $f(x, \beta) = \langle x, \beta \rangle = \sum_{i=1}^D \beta_i x^i$
- Define $X \in \mathbb{R}^{N \times D}$, $\{X\}_{ij}$ 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:

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

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

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

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

- Geometric interpretation of linear regression, estimated with OLS.

Linearly dependent features

- Solution $\hat{\beta} = (X^T X)^{-1} X^T Y$ exists when $X^T X$ is non-degenerate
- Using property
$$\text{rank}(X) = \text{rank}(X^T) = \text{rank}(X^T X) = \text{rank}(X X^T)$$
 - 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, \dots, e_K , because $\sum_k e_k \equiv c$
 - interpretation: non-identifiability of $\hat{\beta}$
 - 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

- 1 Linear regression
- 2 Nonlinear transformations**
- 3 Regularization & restrictions.
- 4 Different loss-functions
- 5 Weighted account for observations
- 6 Nadaraya-Watson regression

Generalization by nonlinear transformations

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

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

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

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

Typical transformations

$\phi_k(x)$	comments
$\exp \left\{ -\frac{\ x-\mu\ ^2}{s^2} \right\}$	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 continuous distribution to uniform ¹

¹why?

Table of Contents

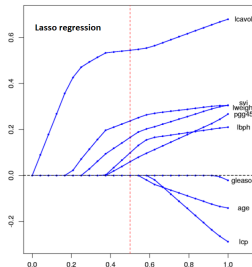
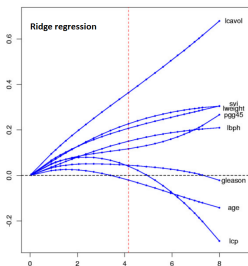
- 1 Linear regression
- 2 Nonlinear transformations
- 3 Regularization & restrictions.**
- 4 Different loss-functions
- 5 Weighted account for observations
- 6 Nadaraya-Watson regression

Regularization

- Variants of target criteria $Q(\beta)$ with regularization²:

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

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



²Derive 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:

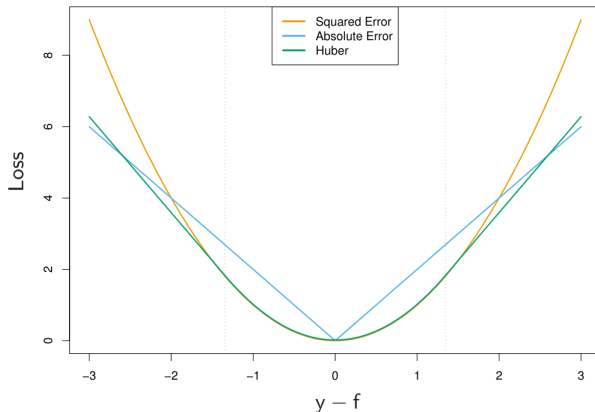
$$\begin{cases} Q(\beta) = \|X\beta - Y\|^2 \rightarrow \min_{\beta} \\ \beta_i \geq 0, \quad i = 1, 2, \dots, D \end{cases}$$

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

Table of Contents

- 1 Linear regression
- 2 Nonlinear transformations
- 3 Regularization & restrictions.
- 4 Different loss-functions**
- 5 Weighted account for observations
- 6 Nadaraya-Watson regression

Non-quadratic loss functions³⁴



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

⁴What 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 :

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

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

Table of Contents

- 1 Linear regression
- 2 Nonlinear transformations
- 3 Regularization & restrictions.
- 4 Different loss-functions
- 5 Weighted account for observations**
- 6 Nadaraya-Watson regression

Weighted account for observations⁵

- 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

⁵Derive solution for weighted regression.

Robust regression

- Initialize $w_1 = \dots = w_N = 1/N$
- Repeat:
 - estimate regression $\hat{y}(x)$ using observations (x_i, y_i) with weights w_i .
 - for each $i = 1, 2, \dots, N$:
 - re-estimate $\varepsilon_i = \hat{y}(x_i) - y_i$
 - 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 $w_1 = \dots = w_N = 1/N$
- Repeat:
 - estimate classifier discriminant functions $\{g_y(\cdot)\}_{y=1,\dots,C}$ using observations (x_i, y_i) with weights w_i .
 - for each $i = 1, 2, \dots, N$:
 - 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.

Table of Contents

- 1 Linear regression
- 2 Nonlinear transformations
- 3 Regularization & restrictions.
- 4 Different loss-functions
- 5 Weighted account for observations
- 6 Nadaraya-Watson regression**

Nadaraya-Watson regression

- Names: local constant regression, kernel regression
- For each x assume $f(x) = \text{const} = \alpha$, $\alpha \in \mathbb{R}$.

$$Q(\alpha, X_{\text{training}}) = \sum_{i=1}^N w_i(x)(\alpha - y_i)^2 \rightarrow \min_{\alpha \in \mathbb{R}}$$

- Weights depend on the proximity of training objects to the predicted object:

$$w_i(x) = K\left(\frac{\rho(x, x_i)}{h}\right)$$

- From stationarity condition $\frac{\partial Q}{\partial \alpha} = 0$ obtain optimal $\hat{\alpha}(x)$:

$$f(x, \alpha) = \hat{\alpha}(x) = \frac{\sum_{i=1}^N y_i w_i(x)}{\sum_{i=1}^N w_i(x)} = \frac{\sum_{i=1}^N y_i K\left(\frac{\rho(x, x_i)}{h}\right)}{\sum_{i=1}^N K\left(\frac{\rho(x, x_i)}{h}\right)}$$

Comments

Under certain regularity conditions $g(x, \alpha) \xrightarrow{P} E[y|x]$

Typically used kernel functions⁶:

$$K_G(r) = e^{-\frac{1}{2}r^2} - \text{Gaussian kernel}$$

$$K_P(r) = (1 - r^2)^2 \mathbb{I}[|r| < 1] - \text{quartic kernel}$$

- The specific form of the kernel function does not affect the accuracy much
- h controls the adaptability of the model to local changes in data
 - *how h affects under/overfitting?*
 - h can be constant or depend on x (if concentration of objects changes significantly)

⁶ Compare them in terms of required computation.

Example

