Regression

Victor Kitov

Yandex School of Data Analysis

February 2, 2016

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:

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

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 non-singular $X^T X$)
- analytical solution
- interpretability of 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**
$\exp\left\{-\frac{\|x-\mu\|^2}{s^2}\right\}$	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 distribution to uniform

Table of contents

Regularization

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

$\|\|X\beta - Y\|\|^2 + \lambda\|\|\beta\|\|_1$	Lasso
$\|\|X\beta - Y\|\|^2 + \lambda\|\|\beta\|\|_2$	Ridge (*analytic solution?*)
$\|\|X\beta - Y\|\|^2 + \lambda_1\|\|\beta\|\|_1 + \lambda_2\|\|\beta\|\|_2$	Elastic net

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

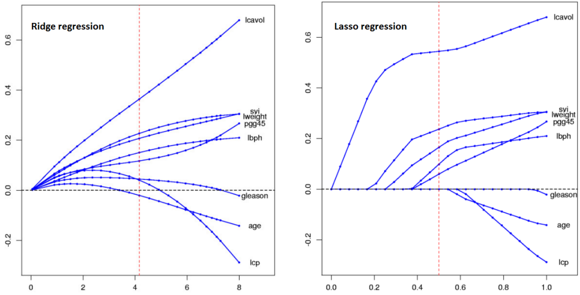


Table of contents

Linear monotonic regression

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

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

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

Weights

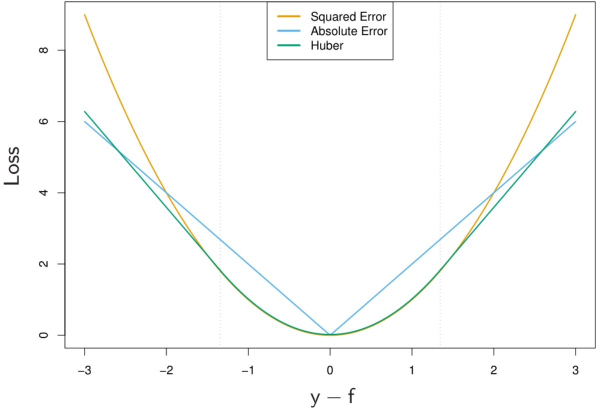
- Weighted account for observations

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

- 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

Table of contents

Non-quadratic loss functions



Averaging in the sum-of-squares sense

Optimizing sum of squared errors

$$\sum_{n=1}^{N} (y_n - \mu)^2 \to \min_{\mu}$$

gives:

$$\mu = \frac{1}{N} \sum_{n=1}^{N} y_i$$

Optimizing sum of squared errors

$$\sum_{n=1}^{N} |y_n - \mu| \to \min_{\mu}$$

gives:

$$\mu = \text{median}_i \, z_i$$

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

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

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

$\mathbb{E}\left\{(\mathbb{E}[y|x]-y)^2\middle|\,x\right\}$ determines the level of irreducible natural noise in the data.

Minimization of expected absolute error

- Let $x, y \sim P(x, y)$. Then

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

$$\mathbb{E}\left\{ |\mu - y| \,|\, x \right\} = \int_{-\infty}^{+\infty} |y - \mu| \, p(y|x) dy =$$

$$\underbrace{\int_{\mu}^{+\infty} (y - \mu) \, p(y|x) dy}_{I(\mu)} + \underbrace{\int_{-\infty}^{\mu} (\mu - y) \, p(y|x) dy}_{J(\mu)}$$

Minimization of expected absolute error

Using the formula for differentiating integrated function $F(\mu)=\int_{\alpha(\mu)}^{\beta(\mu)}f(y,\mu)dy$:

$$F'(\mu)=\int_{\alpha(\mu)}^{\beta(\mu)}f'_{\mu}(y,\mu)dy+\beta'(\mu)f(\beta(\mu),\mu)-\alpha'(\mu)f(\alpha(\mu),\mu)$$

we obtain:

$$\begin{aligned} I'(\mu) &= \int_{\mu}^{+\infty}-p(y|x)dy-(\mu-\mu)p(\mu|x)=-P(y\geq\mu|x) \\ J'(\mu) &= \int_{-\infty}^{\mu}p(y|x)dy+(\mu-\mu)p(\mu|x)=P(y\leq\mu|x) \end{aligned}$$

Stationarity condition becomes:

$$P(y\leq\mu|x)=P(y\geq\mu|x)$$

which means that $\mu=\text{median}\{y|x\}$

Table of contents

Non-linear regression

- $f(x, \alpha)$ may be non-linear function:

$$Q(\alpha, X_{training}) = \sum_{i=1}^{N} (f(x_i, \alpha) - y_i)^2$$

$$\widehat{\alpha} = \arg\min_{\alpha \in \mathbb{R}^D} Q(\alpha, X_{training})$$

- Stationarity condition for α:

$$\frac{\partial Q}{\partial \alpha}(\alpha, X_{training}) = 2 \sum_{i=1}^{N} (f(x_i, \alpha) - y_i) \frac{\partial f}{\partial \alpha}(x_i, \alpha) = 0$$

- Multicollinearity issue, regularization, weighted account for observations apply here as well.

Nadaraya-Watson kernel regression

$f(x,\alpha)=\alpha,\ \alpha\in\mathbb{R}.$

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

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 $\widehat{\alpha}(x)$:

$$f(x,\alpha)=\widehat{\alpha}(x)=\frac{\sum_i y_i w_i(x)}{\sum_i w_i(x)}=\frac{\sum_i y_i K\left(\frac{\rho(x,x_i)}{h}\right)}{\sum_i K\left(\frac{\rho(x,x_i)}{h}\right)}$$

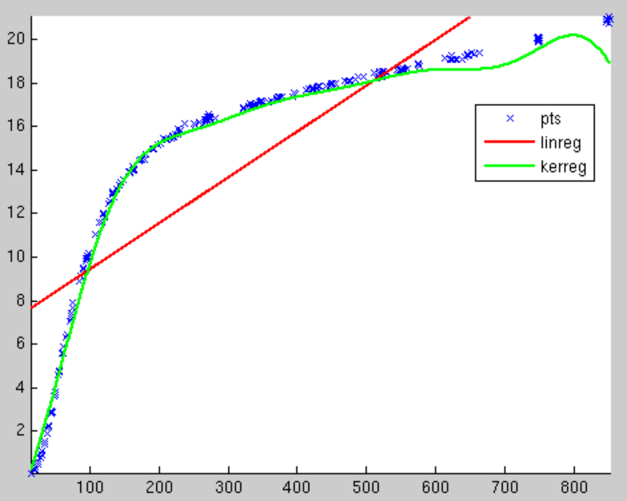
Comments

Under certain regularity conditions $g(x,\alpha) \xrightarrow{P} E[y|x]$
Usually the following kenel functions are used:

$$
\begin{aligned}
K_G(r) &= e^{-\frac{1}{2}r^2} - \text{Gaussian kernel} \\
K_P(r) &= (1-r^2)^2\mathbb{I}[|r|<1] - \text{quartic kernel}
\end{aligned}
$$

- The specific form of the kernel function does not affect accuracy much
- Solution with Gaussian kernel depends on all objects, and with a quadratic kernel - only on objects $\{i : \rho(x, x_i) < h\}$.
- h controls the adaptability of the model to local changes in data
    - can obtain undertrained/overtrained model
    - h can be constant or depend on x (if concentration of objects changes significantly)

Example


pts
linreg
kerreg
20
18
16
14
12
10
8
6
4
2
100
200
300
400
500
600
700
800

Robust kernel regression

- Robustness means that algorithm does not change output significantly in the presence of outliers.
- For outliers $\varepsilon_i = |y_i - f(x_i, \alpha)|$ is big.
- Idea - add weights to objects which encourage regular observations: $K(x, x_i) = D(\varepsilon_i)K(x, x_i)$
- Possible selection of $D(\varepsilon)$:
    - $D(\varepsilon_i) = \mathbb{I}[\varepsilon_i \le t]$, where t may be selected as 95% quantile for series $\varepsilon_1, \varepsilon_2, ...\varepsilon_N$.
    - $D(\varepsilon_i) = K_P\left(\frac{\varepsilon_i}{6\text{med}\varepsilon_i}\right)$

$$f(x, \alpha) = \widehat{\alpha}(x) = \frac{\sum_i y_i w_i(x)}{\sum_i w_i(x)} = \frac{\sum_i y_i D(\varepsilon_i) K\left(\frac{\rho(x, x_i)}{h}\right)}{\sum_i D(\varepsilon_i) K\left(\frac{\rho(x, x_i)}{h}\right)}$$

Algorithm

- apply normal kernel regression for initial forecasts y_i
  - repeat until convergence of ε_i:
    - re-estimate $\varepsilon_i = y_i - \widehat{\alpha}(x_i)$, $i = 1, 2, ...N$.
    - recalculate $\widehat{\alpha}(x_i)$ with $\varepsilon_1, ...\varepsilon_N$
- this idea can be used for all ML methods.

Kernel linear regression

- Local (in neighbourhood of x) approximation $f(u) = (u - x)^T \beta + \beta_0$
- Solve

$$Q(\beta, \beta_0 | X_{training}) = \sum_{i=1}^{N} w(x)((x_i - x)^T \beta + \beta_0 - y_i)^2 \to \min_{\beta, \beta_0 \in \mathbb{R}}$$

- From stationarity conditions $\frac{\partial Q}{\partial \beta} = 0$ and $\frac{\partial Q}{\partial \beta_0} = 0$ obtain the values of the parameters β and β_0.

Advantages of kernel linear regression

- Compared to constant kernel regression, kernel linear regression better predicts:
  - local local minima and maxima
  - linear change at the edges of the training set