Regression and extensions

Victor Kitov

v.v.kitov@yandex.ru

Table of Contents

Linear regression

- Linear model $f(x,\beta) = x^T\beta = \sum_{i=1}^{D}\beta_i x^i$
  - we include constant feature in x
- Assumptions:
  - each x^i has linear impact with weight β_i on y
  - impact of x^i does not depend on other features.
- Benefits:
  - interpretable
  - simple to optimize
  - simple so doesn't overfit less
    - for large D may be optimal model!

Solution

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:

$$L(\beta) = \sum_{n=1}^{N} \left(x_n^T \beta - y_n\right)^2 = \|X\beta - Y\|_2^2 \to \min_{\beta}$$

$$L'(\beta) = 2\sum_{n=1}^{N} x_n \left(x_n^T \beta - y_n\right) = 0$$

In matrix form:

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

so

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

Comments

- Intuition: β_i is proportional to covariance between x_n^i and y_n, normalized by $Var[x^i]$ and $cov[x^i, x^j]$.
- This is the global minimum, because the optimized criteria is convex.
- Solution $\widehat{\beta} = (X^T X)^{-1} X^T Y$ exists when $X^T X$ is non-degenerate.
- Problem occurs when one of the features is a linear combination of the other.
    - because of the property $\forall X : rank(X) = rank(X^T X)$

Comments

- Intuition: β_i is proportional to covariance between x_n^i and y_n, normalized by $Var[x^i]$ and $cov[x^i, x^j]$.
- This is the global minimum, because the optimized criteria is convex.
- Solution $\widehat{\beta} = (X^T X)^{-1} X^T Y$ exists when $X^T X$ is non-degenerate.
- Problem occurs when one of the features is a linear combination of the other.
  - because of the property $\forall X: rank(X) = rank(X^T X)$
  - example: constant unity feature c and one-hot-encoding $e_1, e_2, ...e_K$, because $\sum_k e_k \equiv c$

Linearly dependent features

- Problem may be solved by:
    - feature selection
    - dimensionality reduction
    - imposing additional requirements on the solution (regularization)

Analysis of linear regression

Advantages:

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

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:

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

$$f(x) = \phi(x)^T \beta = \sum_{m=1}^{M} \beta_m \phi_m(x)$$

The model remains to be linear in β, so all advantages of linear regression remain:

- interpretability
- closed form solution
- global optimum

Typical transformations

$\phi_k(x)$	**comments**
$\mathbb{I}\left\{x^i \in [a,b]\right\}$	binarization of feature
$\left(x^i\right)\left(x^j\right)$	interaction of features
$\exp\left\{-\gamma \left\Vert x-z\right\Vert^2\right\}$	closeness to some reference point z
$\ln x^k$	alignment of distribution with heavy tails
$F(x^k)$	convert to uniform distribution with c.d.f. of x^k

Non-linear regression

- Alternatively we can model $\mathcal{X} \to \mathcal{Y}$ with arbitrary non-linear function $\widehat{y} = f(x|\theta)$

$$L(\theta|X, Y) = \sum_{n=1}^{N} (f(x_n|\theta) - y_n)^2$$

$$\widehat{\theta} = \arg\min_{\theta} L(\theta|X, Y)$$

- No analytical solution for $\widehat{\theta}$ will exist in general
    - need numeric optimization methods.

Table of Contents

Regularization

- Overfitting problem: not only *accuracy* matters for the solution but also *model simplicity*!
- Estimate model complexity with regularizer $R(\beta)$:

$$L(\beta)+\lambda R(\beta)=\sum_{n=1}^{N}\left(x_n^T\beta-y_n\right)^2+\lambda R(\beta)\to\min_{\beta}$$

- $\lambda>0$ - hyperparameter (how simple model we want).

$$\begin{array}{ll} R(\beta)=||\beta||_1, & \text{Lasso regression} \\ R(\beta)=||\beta||_2^2 & \text{Ridge regression} \\ R(\beta)=\alpha\left\|\beta\right\|_1+(1-\alpha)\left\|\beta\right\|_2^2 & \text{ElasticNet} \end{array}$$

- λ controls complexity of the model:

Regularization

- Overfitting problem: not only *accuracy* matters for the solution but also *model simplicity*!
- Estimate model complexity with regularizer $R(\beta)$:

$$L(\beta)+\lambda R(\beta)=\sum_{n=1}^{N}\left(x_n^T\beta-y_n\right)^2+\lambda R(\beta)\to\min_{\beta}$$

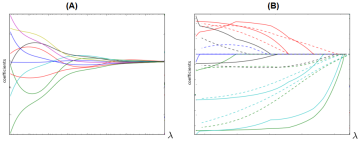
- $\lambda>0$ - hyperparameter (how simple model we want).

$$\begin{array}{ll} R(\beta)=||\beta||_1, & \text{Lasso regression} \\ R(\beta)=||\beta||_2^2 & \text{Ridge regression} \\ R(\beta)=\alpha\left\|\beta\right\|_1+(1-\alpha)\left\|\beta\right\|_2^2 & \text{ElasticNet} \end{array}$$

- λ controls complexity of the model:$\uparrow\lambda\Leftrightarrow$ complexity$\downarrow$.

Comments

- Dependency of β from λ for ridge (A) and LASSO (B):



- LASSO can be used for automatic feature selection.
- λ is usually found using cross-validation on exponential grid, e.g. $[10^{-6}, 10^{-5}, ...10^{5}, 10^{6}]$.
- It's always recommended to use regularization because
  - it gives smooth control over model complexity.
  - removes ambiguity for multiple solutions case.

Ridge regression solution

Ridge regression criterion

$$\sum_{n=1}^{N}\left(x_n^T\beta - y_n\right)^2 + \lambda\beta^T\beta \to \min_{\beta}$$

Stationarity condition can be written as:

$$2\sum_{n=1}^{N}x_n\left(x_n^T\beta - y_n\right) + 2\lambda\beta = 0$$

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

$$\left(X^TX + \lambda I\right)\beta = X^TY$$

so

$$\widehat{\beta} = (X^TX + \lambda I)^{-1}X^TY$$

Comments

- $X^T X + \lambda I$ is always non-degenerate as a sum of:
    - non-negative definite $X^T X$
    - positive definite λI
- Intuition:
    - out of all valid solutions select one giving simplest model
- Other regularizations also restrict the set of solutions.

Different account for different features

- Traditional approach regularizes all features uniformly:

$$\sum_{n=1}^{N}\left(x_n^T\beta - y_n\right)^2 + \lambda R(\beta) \to \min_w$$

- 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 by:

$$\sum_{n=1}^{N}\left(x_n^T\beta - y_n\right)^2 + \lambda_1 R(\{\beta_i | i \in I_1\}) + ... + \lambda_K R(\{\beta_i | i \in I_K\}) \to \min_w$$

- $\lambda_1, \lambda_2, ...\lambda_K$ can be set using cross-validation
- In practice use common regularizer but with different feature scaling.

Linear monotonic regression

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

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

- Examples:
  - in credit scoring we know that salary should be positively correlated with credibility.
  - avaraging of forecasts of different prediction algorithms ($\beta_i = 0$ means, that i-th component does not improve accuracy of forecasting)

Table of Contents

Idea

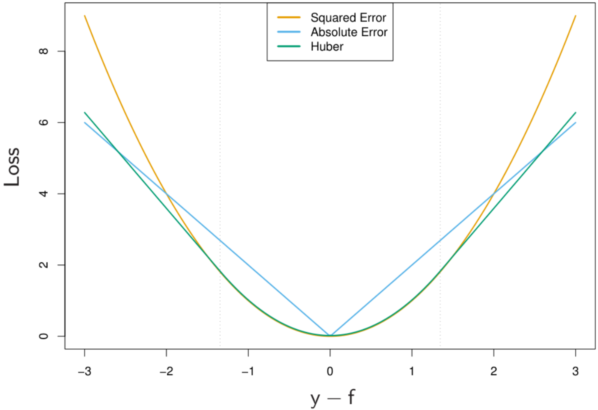
- Generalize quadratic to arbitrary loss:

$$\sum_{n=1}^{N}\left(x^T\beta - y_n\right)^2 \to \min_{\beta} \quad\Longrightarrow\quad \sum_{n=1}^{N}\mathcal{L}(x_n^T\beta - y_n) \to \min_{\beta}$$

LOSS	NAME	PROPERTIES
$\mathcal{L}(\varepsilon) = \varepsilon^2$	quadratic	differentiable
$\mathcal{L}(\varepsilon) = \lvert\varepsilon\rvert$	absolute	robust
$\mathcal{L}(\varepsilon) = \begin{cases} \frac{1}{2}\varepsilon^2, & \lvert\varepsilon\rvert \le \delta \\ \delta\left(\lvert\varepsilon\rvert - \frac{1}{2}\delta\right) & \lvert\varepsilon\rvert > \delta \end{cases}$	Huber	differentiable, robust

- Robust means solution is robust to outliers in the training set.

Non-quadratic loss functions[1,2]



[1] What is the value of constant prediction, minimizing sum of squared errors?
[2] What is the value of constant prediction, minimizing sum of absolute errors?

Weighting objects - 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.

Example

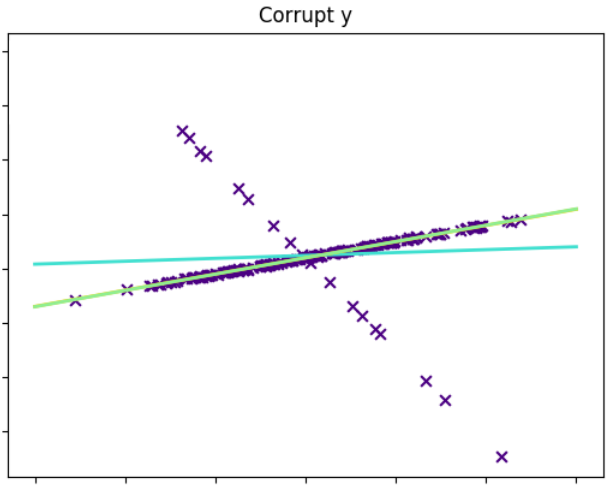


Table of Contents

Minimum squared error estimate

For training sample $(x_1, y_1), \dots (x_N, y_N)$ consider finding constant $\widehat{y} \in \mathbb{R}$:

$$L(\widehat{y}) = \sum_{i=1}^{N} (\widehat{y} - y_i)^2 \to \min_{\widehat{y} \in \mathbb{R}}$$

$$\frac{\partial L}{\partial \widehat{y}} = 2 \sum_{i=1}^{N} (\widehat{y} - y_i) = 0, \text{ so } \widehat{y} = \frac{1}{N} \sum_{i=1}^{N} y_i$$

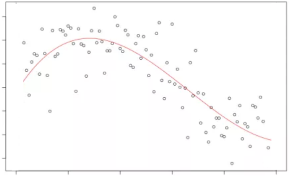
Minimum squared error estimate

For training sample $(x_1, y_1), \dots (x_N, y_N)$ consider finding constant $\widehat{y} \in \mathbb{R}$:

$$L(\widehat{y}) = \sum_{i=1}^{N} (\widehat{y} - y_i)^2 \to \min_{\widehat{y} \in \mathbb{R}}$$

$$\frac{\partial L}{\partial \widehat{y}} = 2 \sum_{i=1}^{N} (\widehat{y} - y_i) = 0, \text{ so } \widehat{y} = \frac{1}{N} \sum_{i=1}^{N} y_i$$

We need to model general curve $y(x)$:



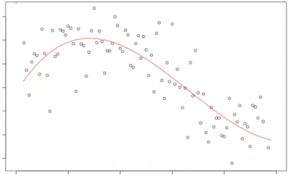
Minimum squared error estimate

For training sample $(x_1, y_1), \dots (x_N, y_N)$ consider finding constant $\widehat{y} \in \mathbb{R}$:

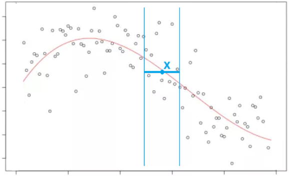
$$L(\widehat{y}) = \sum_{i=1}^{N} (\widehat{y} - y_i)^2 \to \min_{\widehat{y} \in \mathbb{R}}$$

$$\frac{\partial L}{\partial \widehat{y}} = 2 \sum_{i=1}^{N} (\widehat{y} - y_i) = 0, \text{ so } \widehat{y} = \frac{1}{N} \sum_{i=1}^{N} y_i$$

We need to model general curve $y(x)$:



Nadaraya-Watson regression - localized averaging approach.



Nadaraya-Watson regression

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

$$L(\widehat{y}|x) = \sum_{i=1}^{N} w_i(x)(\widehat{y} - y_i)^2 \to \min_{\alpha \in \mathbb{R}}$$

- Weights should $\downarrow$ as $\rho(x, x_i) \uparrow$.

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

- $K(u)$ - some decreasing function, called kernel.
- $h(x)$ - some ≥ 0 function called bandwidth.
    - Intuition: "window width", consider $h(x) = h$, $K(u) = \mathbb{I}[u \leq 1]$.

Parameters

- Typically used $K(u)$[3]:

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

- Typically used $h(x)$:
  - $h(x) = const$
  - $h(x) = \rho(x, x_{i_K})$, where x_{i_K} - K-th nearest neighbour.
    - better for unequal distribution of objects

[3] Compare them in terms of required computation.

Solution

$$L(\widehat{y}|x) = \sum_{i=1}^{N} w_i(x)(\widehat{y} - y_i)^2 \to \min_{\alpha \in \mathbb{R}}$$

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

Solution

$$L(\widehat{y}|x) = \sum_{i=1}^{N} w_i(x)(\widehat{y} - y_i)^2 \to \min_{\alpha \in \mathbb{R}}$$

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

- From stationarity condition $\frac{\partial L}{\partial \widehat{y}} = 0$ obtain optimal $\widehat{y}(x)$:

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

Comments

- Under general regularity conditions $\widehat{y}(x) \xrightarrow{P} E[y|x]$
- The specific form of the kernel function does not affect the accuracy much.
  - but may affect efficiency[4]
- Compared to K-NN: may use all objects, bandwidth controls smoothness.
  - under what selection of $K(u)$ and $h(x)$ it reduces to basic K-NN?

[4] how?

Comments

Instead of optimizing local constant $\widehat{y}$

$$L(\widehat{y}|x) = \sum_{i=1}^{N} w_i(x)(\widehat{y} - y_i)^2 \to \min_{\alpha \in \mathbb{R}}$$

we could have optimized **local linear regression**

$$L(\widehat{\beta}|x) = \sum_{i=1}^{N} w_i(x)(x^T \beta - y_i)^2 \to \min_{\alpha \in \mathbb{R}}$$

This better handles approximation on the edges of domain.

Table of Contents

Support vector regression

Idea: don't care about small deviations, catch only the large ones + regularization.

$$\begin{cases} \frac{1}{2}\left\|w\right\|^2 \to \min_w \\ \langle w, x_n\rangle + w_0 - y_n \le \varepsilon \quad n=\overline{1,N} \\ y_n - \langle w, x_n\rangle - w_0 \le \varepsilon \quad n=\overline{1,N} \end{cases}$$

Since fitting any dataset with error$\in[-\varepsilon,\varepsilon]$ may be infeasible use penalization of excessive deviations:

$$\begin{cases} \frac{1}{2}\left\|w\right\|^2 + C\sum_{n=1}^{N}(\xi_n+\xi_n^*) \to \min_{w,\xi_n,\xi_n^*} \\ \langle w, x_n\rangle + w_0 - y_n \le \varepsilon + \xi_n, \quad \xi_n \ge 0 \qquad n=\overline{1,N} \\ y_n - \langle w, x_n\rangle - w_0 \le \varepsilon + \xi_n^*, \quad \xi_n^* \ge 0 \qquad n=\overline{1,N} \end{cases}$$

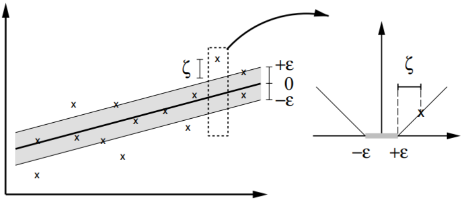
C controls how much errors should matter more than model simplicity.

Support vector regression

Equivalent unconstrained formulation:

$$\frac{1}{2}\|w\|^2 + C\sum_{n=1}^{N}\mathcal{L}(\langle w, x_n\rangle + w_0 - y_n) \to \min_w$$

with ε insensitive loss $\mathcal{L}(u) = \begin{cases} 0, & \text{if } |u| \le \varepsilon \\ |u| - \varepsilon & \text{otherwise} \end{cases}$



Solution will depend only on objects with $|\text{error}| \ge \varepsilon$, called *support vectors*.

Orthogonal matching pursuit (details)

- Approximates the problem ($\|w\|_0 = \#$[non-zero weights]):

$$\begin{cases} \|Xw - Y\|_2^2 \to \min_w \\ \|w\|_0 \leq K \end{cases}$$

or alternatively

$$\begin{cases} \|w\|_0 \to \min \\ \|Xw - Y\|_2^2 \leq \varepsilon \end{cases}$$

- Algorithm: iteratively:
    1. add feature having maximum correlation with residuals
    2. fit multivariate regression: selected features vs. residuals
    3. update residuals by full account of features

Summary

- Linear regression gives interpretable analytic solution.
- Non-linear dependencies can be modeled by adding non-linear features.
- When features are linearly dependent, it fails.
- Regularized versions are always preferable:
  - work in case of linearly dependent features
  - are more robust in close to linear dependency case
  - λ gives a convenient way to control model complexity
- Robust regression is robust to outliers.
  - we may also use robust loss-functions instead of MSE.