Regression

V. Kitov

Yandex School of Data Analysis

Imperial College London
Department of Physics

January 2015

Linear regression

- $g(x, \alpha) = \sum_{j=1}^{n} \alpha_j x_j$
- Denote $X \in \mathbb{R}^{nxd}, \{X\}_{ij}$ is j-th feature of i-th observation, $Y \in \mathbb{R}^n$, $\{Y\}_i$ is i-th output observation.
- In matrix notation

$$Q(\alpha) = ||X\alpha - Y||^2 \to \min_{\alpha}$$

$$\frac{\partial Q}{\partial \alpha}(\alpha) = 2X^T(X\alpha - Y) = 0$$

$$\widehat{\alpha} = (X^T X)^{-1} XY$$

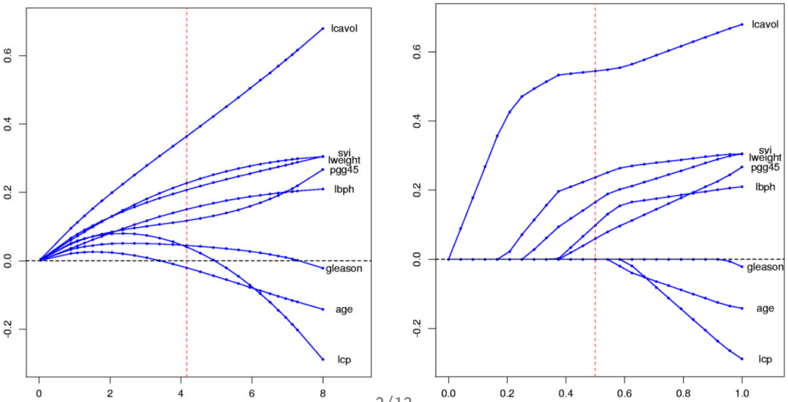
- Caveats:
  - correlation matrix $\Sigma = X^T X$ may be degenerate
    - occurs when some features are linearly dependent
    - solved by feature selection, feature extraction or regularization.

Regularization

- Lasso, ridge, elastic-net

$$Q(\alpha) = ||X\alpha - Y||^2 + \tau||\alpha||_p$$

- Coefficients behave differently:



Linear monotone regression

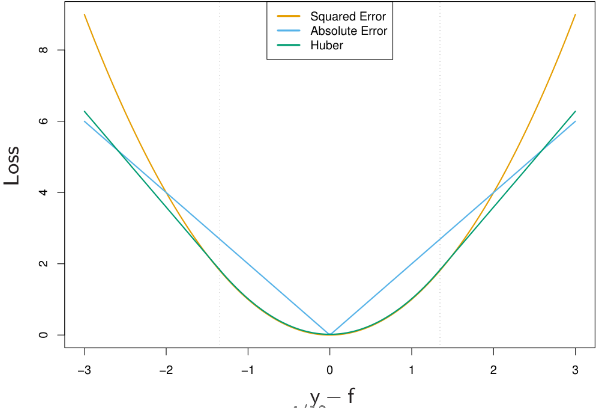
- From expert knowledge the effects of features may be assumed positive:

$$\begin{cases} Q(\alpha) = ||X\alpha - Y||^2 \to \min_\alpha \\ \alpha_j \geq 0, \quad j = 1, 2, ...n \end{cases}$$

- Example: algorithms composition
- Active constraint means feature exclusion.

Comments

- Weighted estimation
- Robust estimation
- Non-quadratic loss functions.



Non-linear regression

- Regression: reconstruct a continuous output using arbitrary inputs.
- Find $\alpha \in \mathbb{R}^d$, when $g(x, \alpha)$ matches continuous output y most accurately on training sample.

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

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

- Found from system of d equations:

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

- Multicollinearity, regularization, weighted estimation, robustness apply here as well.

Kernel regression

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

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

Weights are location dependent:

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

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

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

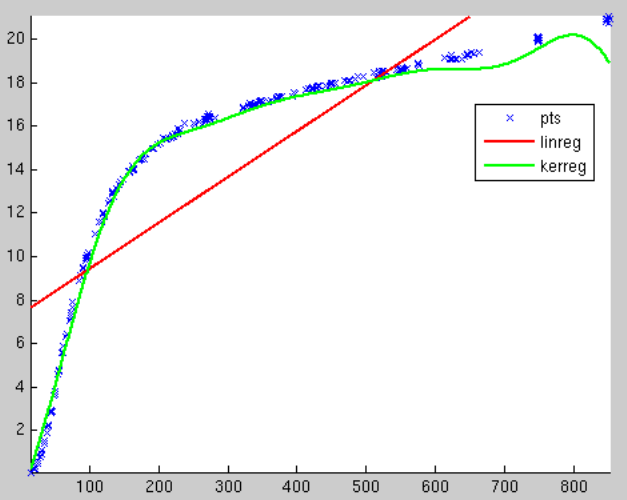
Comments

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

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

- Kernel function selection does not affect much
  - mainly performance if domain is $\mathbb{R}$.
- h controls bias/variance tradeoff
  - can be fixed or variable (for non-uniform samples concentrations)

Example



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

Robust non-parametric regression

- Robust to outliers algorithm
- Outliers are observations for which $\varepsilon_i = y_i - g(x_i, \alpha)$ is large
- Idea: kernel as product of kernels: $K(x, x_i) = D(\varepsilon_i)K(x, x_i)$
- Selection of $D(\varepsilon)$:
  - $D(\varepsilon_i) = \mathbb{I}[\varepsilon_i \leq t]$, where t may be taken as p-quantile value of ε series.
  - $D(\varepsilon_i) = K_P\left(\frac{\varepsilon_i}{6\text{med}\varepsilon_i}\right)$

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

Algorithm

- apply ordinary non-parametric regression to get initial estimates of y_i
  - repeat until convergence of ε_i:
    - estimate $\varepsilon_i = y_i - \alpha(x)$
    - apply robust non-parametric regression and estimate $\alpha(x)$

Non-parametric linear approximation

- Local (in neighbourhood of x) approximation
  $g(u) = \alpha(u - x) + \beta$
- Solve

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

- From $\frac{\partial Q}{\partial \alpha} = 0$ and $\frac{\partial Q}{\partial \beta} = 0$ obtain the prediction (using $w_i = w_i(x)$, $d_i = x_i - x$)

$$\widehat{y}(x) = \frac{\sum_i w_i d_i^2 \sum w_i y_i - \sum_i w_i d_i \sum_i w_i d_i y_i}{\sum_i w_i \sum_i w_i d_i^2 - (\sum_i w_i d_i)^2}$$

Benefits combared to non-parametric constant approximation:

- better predicts local extremums
- better predicts functions at edges