Lecture 6

Regression

Information Systems
(Machine Learning)

Andrey Filchenkov

24.10.2017

Lecture plan

- Regression problem
- Nonparametric regression
- Linear regression
- Regularization for regression

- The presentation is prepared with materials of the K.V. Vorontsov's course "Machine Leaning".
- Slides are available online: **goo.gl/fDBgMq**

Lecture plan

- Regression problem
- Nonparametric regression
- Linear regression
- Regularization for regression

Problem formalization

X is an object set, Y is an answer set,

$y: X \to Y$ is an unknown dependency, $Y \in \mathbb{R}$

$X^\ell = \{x_1, \dots, x_\ell\}$ is training sample,

$T^\ell = \{(x_1, y_1), \dots, (x_\ell, y_\ell)\}$ set of instances.

Problem: find $a: X \to Y$.

Problem formalization

X is an object set, Y is an answer set,

$y: X \to Y$ is an unknown dependency, $Y \in \mathbb{R}$

$X^\ell = \{x_1, \dots, x_\ell\}$ is training sample,

$T^\ell = \{(x_1, y_1), \dots, (x_\ell, y_\ell)\}$ set of instances.

Problem: find $a: X \to Y$.

`

$a(x) = f(x, \theta)$ is dependency model, $\theta \in \mathbb{R}^t$.

Ordinary Least Squares

Standard assumptions:

$$y(x_i) = f(x_i, \theta) + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}\left(0, \sigma_i^2\right), \quad i = 1, \dots, \ell.$$

Maximum likelihood:

$$L(\varepsilon_i, \dots, \varepsilon_i | \theta) = \prod_{i=1}^{\ell} \frac{1}{\sigma_i \sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma_i^2} \varepsilon_i^2\right) \to \max_\theta .$$

$$-\ln L(\varepsilon_i, \dots, \varepsilon_i | \theta) =$$

$$= \mathrm{const}(\theta) + \frac{1}{2} \sum_{i=1}^{\ell} \frac{1}{\sigma_i^2} \left(f(x_i, \theta) - y_i\right)^2 \to \min_\theta .$$

Lecture plan

- Regression problem
- Nonparametric regression
- Linear regression
- Regularization for regression

Main idea

Basic idea: let think that $\theta(x) = \theta$ nearby $x \in X$:

$$Q(\theta, T^\ell) = \sum_{i=1}^{\ell} w_i(x)(\theta - y_i)^2 \to \min_{\theta \in \mathbb{R}}.$$

Main idea: use kernel smoothing:

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

where h is window width.

Kernel smoothing

Nadaraya-Watson kernel smoothing:

$$a_h(x, T^\ell) = \frac{\sum_{i=1}^{\ell} y_i w_i(x)}{\sum_{i=1}^{\ell} w_i(x)} = \frac{\sum_{i=1}^{\ell} y_i K\left(\frac{\rho(x_i, x)}{h}\right)}{\sum_{i=1}^{\ell} K\left(\frac{\rho(x_i, x)}{h}\right)}.$$

Basis theorem

Theorem. If

1) sample T^ℓ is simple, distributed with $p(x, y)$;

2) $\int_0^\infty K(r)dr < \infty$, $\lim\limits_{r\to\infty} rK(r) = 0$;

3) $\mathrm{E}(y^2|x) < \infty\ \forall x \in X$;

4) $\lim\limits_{\ell\to\infty} h_\ell = 0$, $\lim\limits_{\ell\to\infty} \ell h_\ell = \infty$,

then $a_h\left(x, T^\ell\right) \to^P \mathrm{E}(y|x)$ in any $x \in X$,

when $E(y|x), p(x), D(y|x)$ are continuing, $p(x) > 0$.

Method discussion

- kernel function has impact on smoothness;
- kernel function has small impact on approximation quality;
- h impacts on approximation quality;
- k can be tuned;
- sensitive to noise.

Lecture plan

- Regression problem
- Nonparametric regression
- Linear regression
- Regularization for regression

Linear regression model

Model of multidimensional linear regression:

$$f(x,\theta) = \sum_{j=1}^{n} \theta_j f_j(x), \quad \theta \in \mathbb{R}^n.$$

Matrix notation:

$$F = \begin{pmatrix} f_1(x_1) & \dots & f_n(x_1) \\ \dots & \dots & \dots \\ f_1(x_\ell) & \dots & f_n(x_\ell) \end{pmatrix}, y = \begin{pmatrix} y_1 \\ \dots \\ y_\ell \end{pmatrix}, \theta = \begin{pmatrix} \theta_1 \\ \dots \\ \theta_n \end{pmatrix}.$$

Quality in matrix notation:

$$Q\left(\theta, T^\ell\right) = \sum_{i=1}^{\ell} (f(x_i,\theta) - y_i) = ||F\theta - y||^2 \to \min_{\theta\in\mathbb{R}}.$$

Normal equation sysmet

Minimum condition:

$$\frac{\partial Q(\theta)}{\partial \theta} = 2F^{\top}(F\theta - y) = 0.$$

$F^{+} = (F^{\top}F)^{-1}F^{\top}$ is **pseudo reverse matrix (Moore–Penrose inverse)**

$P_F = FF^{+}$ is **projection matrix**

Solution:

$$\theta^{*} = F^{+}y.$$

Minimum approximation:

$$Q(\theta^{*}) = ||P_F y - y||^2.$$

Singular vector decomposition

Theorem: any matrix F size of $\ell \times n$ can be represented with singular decomposition

$$F = VDU^\top.$$

With

- $V = (v_1, \dots, v_n)$ is size of $\ell \times n$ and orthogonal $V^\top V = I_n$, rows v_j are eigenvectors of matrix $FF^\top$;
- $U = (u_1, \dots, u_n)$ is size of $n \times n$ and ortogonal $U^\top U = I_n$, rows u_j are eigenvectors of matrix $F^\top F$;
- $D = \mathrm{diag}\left(\sqrt{\lambda_1}, \dots, \sqrt{\lambda_n}\right)$ of size $n \times n$, $\sqrt{\lambda_j}$ are **singular numbers**, squares of eigenvalues of matrices $FF^\top$ and $F^\top F$.

OLS with SVD

$$F^{+} = (UDV^{\top}VDU^{\top})UDV^{\top} = UD^{-1}V^{\top} = \sum_{j=1}^{n} \frac{1}{\sqrt{\lambda_j}} u_j v_j^{\top};$$

$$\theta^{*} = F^{+}y = UD^{-1}V^{\top}y = \sum_{j=1}^{n} \frac{1}{\sqrt{\lambda_j}} u_j (v_j^{\top} y);$$

$$F\theta^{*} = P_F y = (VDU^{\top})UD^{-1}V^{\top}y = VV^{\top}y = \sum_{j=1}^{n} v_j (v_j^{\top} y);$$

$$||\theta^{*}||^2 = ||D^{-1}V^{\top}y||^2 = \sum_{j=1}^{n} \frac{1}{\lambda_j} \left(v_j^{\top} y\right)^2.$$

Lecture plan

- Regression problem
- Nonparametric regression
- Linear regression
- Regularization for regression

Ridge regression

Assumption: values of θ have Gaussian distribution with covariance matrix σI_n:

$$Q_\tau(\theta) = ||F\theta - y||^2 + \frac{1}{2\sigma}||\theta||^2 \to \min_\theta,$$

where $\tau = 1/\sigma$ is regularization penalty.

OLS solution:

$$\theta_\tau^* = (F^\top F + \tau I_n)^{-1} F^\tau y.$$

Solution for ridge regression

$$\theta^*_\tau = U(D^2 + \tau I_n)^{-1} D V^\top y = \sum_{j=1}^{n} \frac{\sqrt{\lambda_j}}{\lambda_j + \tau} u_j (v_j^\top y);$$

$$F\theta^*_\tau = (VDU^\top)\theta^*_\tau = V \mathrm{diag}\left(\frac{\lambda_j}{\lambda_j + \tau}\right) V^\top y =$$

$$= \sum_{j=1}^{n} \frac{\lambda_j}{\lambda_j + \tau} v_j (v_j^\top y);$$

$$||\theta^*||^2 = ||D^2 (D^2 + \tau I_n)^{-1} D^{-1} V^\top y||^2 == \sum_{j=1}^{n} \frac{1}{\lambda_j + \tau} \left(v_j^\top y\right)^2.$$

Tibshirani lasso

Assumption: values of vector θ has Laplacian distribution:

$$\begin{cases} Q_\tau(\theta) = ||F\theta - y||^2 \to \min_\theta; \\ \sum_{i=1}^{n} |a_i| \le \kappa. \end{cases}$$

LASSO (least absolute shrinkage and selection operator).

Will lead to feature selection.

LASSO regression

The resulting optimization problem

$$Q_\tau(\theta) = ||F\theta - y||^2 + \tau\|\theta\|_1 \to \min_{\theta},$$

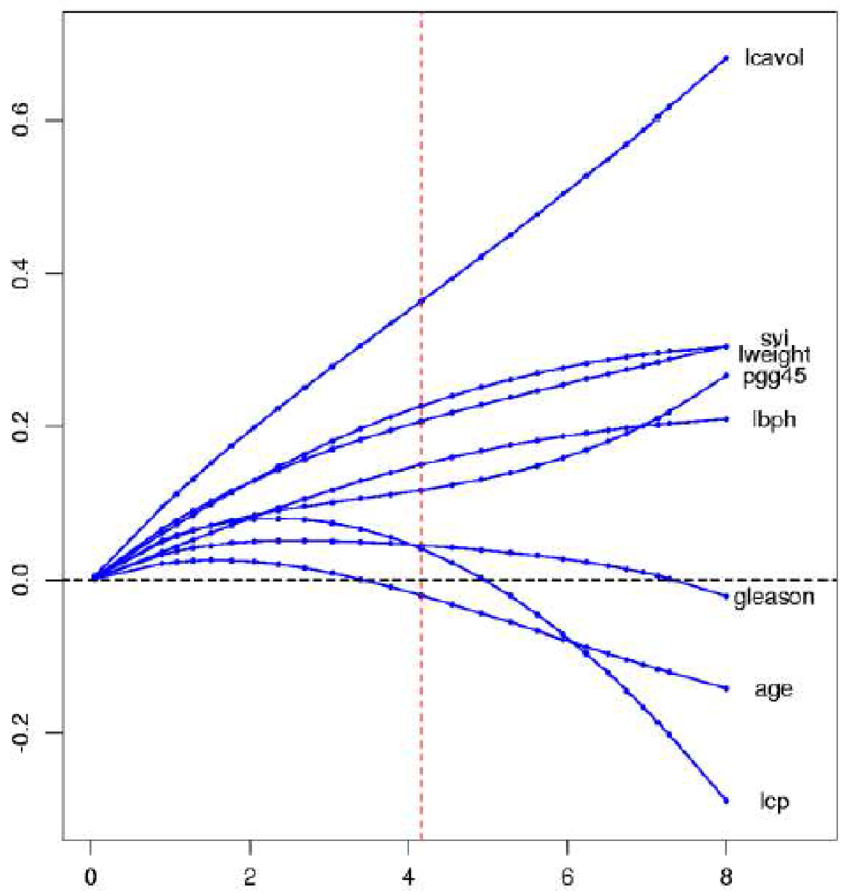
where $\|\theta\|_1$ is l_1-norm:$\|\theta\|_1 = \sum|\theta_i|$.

No nice analytical solution exist.

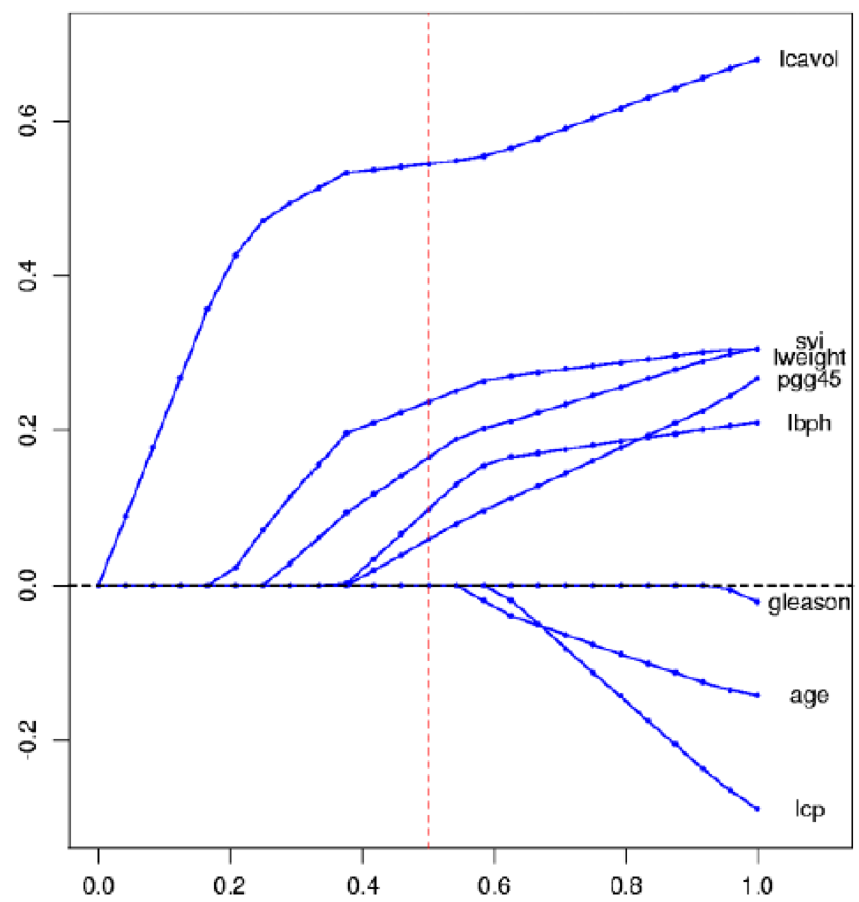
However, a nice computational solution exist.

Comparison

Ridge regression



Lasso



Regularizer discussion

- l_1-norm and l_2-norm regularizers are the most popular
- **ElasicNet**, which is sum of the previous two is also popular
- Many other may be used with respect to initial assumptions
- Some techniques are de-facto regularization or can be interpreted as regularization