Linear and Kernel Regression

Linear Models. Regularization. Connection to integral equations.

Machine Learning, Skoltech

Alexey Artemov[1]

[1] Skoltech

Practical overview

- Linear models for regression (a recap)
- This practical: connection with integral equations
- Regularization: connection with PCA
- Trying it out in codes

Multivariate linear regression

- Multiple features (regressors) $\mathbf{x}_i = (x_{1i}, \ldots x_{di})$ available for each y_i
- The model:

$$
\begin{aligned}
y_1 &= w_1 x_{11} + \ldots w_d x_{d1} + \varepsilon_1, \\
y_2 &= w_1 x_{12} + \ldots w_d x_{d2} + \varepsilon_2, \\
&\ldots \\
y_\ell &= w_1 x_{1\ell} + \ldots w_d x_{d\ell} + \varepsilon_\ell,
\end{aligned}
$$

is often written in matrix-vector form as

$$
\begin{bmatrix} y_1 \\ \vdots \\ y_\ell \end{bmatrix} = \begin{bmatrix} x_{11} & x_{12} & \ldots & x_{d1} \\ \vdots & \vdots & \ddots & \vdots \\ x_{1\ell} & x_{2\ell} & \ldots & x_{d\ell} \end{bmatrix} \begin{bmatrix} w_1 \\ \vdots \\ w_d \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_\ell \end{bmatrix} \quad \longleftrightarrow \quad \boldsymbol{y} = \boldsymbol{X}\mathbf{w} + \boldsymbol{\varepsilon}
$$

Multivariate linear regression: the solution

- The problem: minimize MSE

$$Q(h, X^l) = \sum_{i=1}^{\ell} \Big(y_i - \sum_{k=1}^{d} w_k x_{ki}\Big)^2 \equiv \|\boldsymbol{y} - \boldsymbol{X}\mathbf{w}\|^2 \to \min_{\mathbf{w} \in \mathbb{R}^d}$$

- Solve analytically via computing the gradient

$$\nabla_{\mathbf{w}} \|\boldsymbol{y} - \boldsymbol{X}\mathbf{w}\|^2 = 2(\boldsymbol{y} - \boldsymbol{X}\mathbf{w})\boldsymbol{X} = 0$$

- The solution

$$\mathbf{w}^* = (\boldsymbol{X}^\intercal \boldsymbol{X})^{-1} \boldsymbol{X}^\intercal \boldsymbol{y}$$

Connection with integral equations

- The setting (stemming from physics):

$$\boldsymbol{A}v(y) = u(x),$$

where commonly $\mathbf{A}$ is an integral operator:

$$\boldsymbol{A}v(y) = \int K(x,y)v(y)dy = u(x),$$

with scalar or vector-valued x, y.

- Commonly in practice measurements for $u(x)$ taken at points $x_1, x_2, \ldots, x_\ell$: $u_i = u(x_i)$.
- The problem: estimate $v(y)$ from $u_i, i = 1, \ldots, \ell$. Rewriting the integral equation,

$$\sum_{j=1}^{n} K(x_i, y_j) v_j \Delta y_j = u_i, \qquad i = 1, \ldots, \ell.$$

Connection with integral equations

- Letting $\mathbf{u} = (u_1, \ldots, u_\ell), \mathbf{v} = (v_1, \ldots, v_n), \boldsymbol{K} = (K(x_i, v_j)\Delta y_j)_{ij}$, we get
$$\boldsymbol{K}\mathbf{v} = \mathbf{u}.$$
- In general $n \neq \ell$, this direct inversion of $\boldsymbol{K}$ is not possible.
- Commonly to create a square matrix from non-square $\boldsymbol{K}$, multiply the equation by $\boldsymbol{K}^\top$:
$$\boldsymbol{K}^\top\boldsymbol{K}\mathbf{v} = \boldsymbol{K}^\top\mathbf{u}.$$
Still making $\boldsymbol{K}^\top\boldsymbol{K}$ square does not guarantee invertibility (or even a good condition number)!

Ridge regression

- **Ridge regression:** adding a “ridge” to the diagonal of $\boldsymbol{K}^\intercal \boldsymbol{K}$ would improve conditioning:

$$(\boldsymbol{K}^\intercal \boldsymbol{K} + \alpha \boldsymbol{I})\mathbf{v} = \boldsymbol{K}^\intercal \mathbf{u}.$$

- You all know that solution:

$$\mathbf{v} = (\boldsymbol{K}^\intercal \boldsymbol{K} + \alpha \boldsymbol{I})^{-1} \boldsymbol{K}^\intercal \mathbf{u}.$$

- Analogously, Kernel regression can be modified from

$$u(x) = \mathbf{r}^\intercal \boldsymbol{R}_0^{-1} \mathbf{u} \qquad (\boldsymbol{R}_0)_{ij} = R(x_i, x_j), \quad (\mathbf{r})_i = R(x, x_i)$$

  to

$$u(x) = \mathbf{r}^\intercal (\boldsymbol{R}_0 + \alpha \boldsymbol{I})^{-1} \mathbf{u}.$$

Regularizers?

Figure: Large parameter space

Figure: Regularized models

Ad-hoc regularization: motivation

- Consider the multivariate linear regression problem with $\boldsymbol{X} \in \mathbb{R}^{d \times d}$

$$\|\boldsymbol{y} - \boldsymbol{X}\mathbf{w}\|^2 \to \min_{\mathbf{w} \in \mathbb{R}^d}$$

Ad-hoc regularization: motivation

- Consider the multivariate linear regression problem with $\boldsymbol{X} \in \mathbb{R}^{d \times d}$

$$\|\boldsymbol{y} - \boldsymbol{X}\mathbf{w}\|^2 \to \min_{\mathbf{w} \in \mathbb{R}^d}$$

- Analytic solution involves computing the product $\boldsymbol{R} = (\boldsymbol{X}^\intercal \boldsymbol{X})^{-1} \boldsymbol{X}^\intercal$

Ad-hoc regularization: motivation

- Consider the multivariate linear regression problem with $\boldsymbol{X} \in \mathbb{R}^{d \times d}$

$$\|\boldsymbol{y} - \boldsymbol{X}\mathbf{w}\|^2 \to \min_{\mathbf{w} \in \mathbb{R}^d}$$

- Analytic solution involves computing the product $\boldsymbol{R} = (\boldsymbol{X}^\intercal \boldsymbol{X})^{-1} \boldsymbol{X}^\intercal$
- If $\boldsymbol{X} = \text{diag}(\lambda_1, \ldots, \lambda_d)$ with $\lambda_1 > \lambda_2 > \ldots > \lambda_d \to 0$ (meaning we're in eigenbasis of $\boldsymbol{X}$) then

$$\begin{aligned} \boldsymbol{R} &= (\boldsymbol{X}^\intercal \boldsymbol{X})^{-1} \boldsymbol{X}^\intercal = \\ &= \big(\text{diag}(\lambda_1, \ldots, \lambda_d)\text{diag}(\lambda_1, \ldots, \lambda_d)\big)^{-1} \text{diag}(\lambda_1, \ldots, \lambda_d) = \\ &= \text{diag}\left(\frac{1}{\lambda_1}, \ldots, \frac{1}{\lambda_d}\right), \quad \text{leading to huge diagonal values in } \boldsymbol{R} \end{aligned}$$

Eigenbasis of X? Connection with PCA

› Let $\boldsymbol{X} = \boldsymbol{USV}^\intercal$ be the SVD of $\boldsymbol{X}$ where $\boldsymbol{V}^\intercal\boldsymbol{V} = \boldsymbol{I}_d, \boldsymbol{UU}^\intercal = \boldsymbol{U}^\intercal\boldsymbol{U} = \boldsymbol{I}_d$, and $\boldsymbol{S}$ is a diagonal $d \times d$ matrix. Then:

$$
\begin{aligned}
\boldsymbol{R} &= (\boldsymbol{X}^\intercal\boldsymbol{X})^{-1}\boldsymbol{X}^\intercal = \\
&= \left((\boldsymbol{USV}^\intercal)^\intercal \cdot \boldsymbol{USV}^\intercal\right)^{-1}(\boldsymbol{USV}^\intercal)^\intercal = && (\boldsymbol{AB})^\intercal = \boldsymbol{B}^\intercal\boldsymbol{A}^\intercal \\
&= \left(\boldsymbol{VSU}^\intercal \cdot \boldsymbol{USV}^\intercal\right)^{-1}\boldsymbol{VSU}^\intercal = && \boldsymbol{U}^\intercal\boldsymbol{U} = \boldsymbol{I}_d \\
&= \left(\boldsymbol{VS}^2\boldsymbol{V}^\intercal\right)^{-1} \cdot \boldsymbol{VSU}^\intercal = && (\boldsymbol{AB})^{-1} = \boldsymbol{B}^{-1}\boldsymbol{A}^{-1} \\
&= (\boldsymbol{V}^\intercal)^{-1}\boldsymbol{S}^{-2}\boldsymbol{V}^\intercal \cdot \boldsymbol{VSU}^\intercal = && \text{orthogonality } \boldsymbol{V}^\intercal = \boldsymbol{V}^{-1} \\
&= \boldsymbol{VS}^{-2}\boldsymbol{SU}^\intercal = \\
&= \boldsymbol{V}\widetilde{\boldsymbol{S}}_{\text{linear}}\boldsymbol{U}^\intercal && \widetilde{\boldsymbol{S}}_{\text{linear}} = \boldsymbol{S}^{-1}.
\end{aligned}
$$

Eigenbasis of X? Connection with PCA

- Doing the same calculations with regularized solution gives:

$$\begin{aligned} \boldsymbol{R}_{\text{ridge}} &= (\boldsymbol{X}^{\intercal}\boldsymbol{X} + \alpha \boldsymbol{I}_d)^{-1}\boldsymbol{X}^{\intercal} = \\ &= \boldsymbol{V}(\boldsymbol{S}^2 + \alpha \boldsymbol{I}_d)^{-1}\boldsymbol{S}\boldsymbol{U}^{\intercal} = \\ &= \boldsymbol{V}\widetilde{\boldsymbol{S}}_{\text{ridge}}\boldsymbol{U}^{\intercal} \end{aligned}$$

same math here

$$\widetilde{\boldsymbol{S}}_{\text{ridge}} = (\boldsymbol{S}^2 + \alpha \boldsymbol{I}_d)^{-1}\boldsymbol{S}.$$

- Predictions with this kind of operators are given by:

$$\widehat{\mathbf{y}} = \boldsymbol{X}\boldsymbol{R}\mathbf{y} = \boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^{\intercal} \cdot \boldsymbol{V}\widetilde{\boldsymbol{S}}_{\text{method}}\boldsymbol{U}^{\intercal} \cdot \mathbf{y} = \boldsymbol{U}\widetilde{\boldsymbol{S}}\boldsymbol{U}^{\intercal}\mathbf{y}$$

- In terms of directions $\boldsymbol{U} = (\mathbf{u}_1, \ldots, \mathbf{u}_d)$ in data,

$$\widehat{\mathbf{y}} = \sum_{j=1}^{d} \mathbf{u}_j \widetilde{S}_{jj} \mathbf{u}_j \mathbf{y}$$

Eigenbasis of X? Connection with PCA

- Values $\widetilde{S}_{jj}$ define behavior of predictions:
  - low (near-zero) values of $\widetilde{S}_{jj}$ suppress any influence of $\mathbf{u}_j$ on the prediction;
  - higher values serve as amplifiers, etc.
- For the least squares,

$$\widetilde{S}_{jj} \equiv 1$$

- For the ridge regression,

$$\widetilde{S}_{jj} \equiv [(\boldsymbol{S}^2 + \alpha \boldsymbol{I}_d)^{-1} \boldsymbol{S}]_{jj} = \frac{\lambda_j^2}{\lambda_j^2 + \alpha}$$

  where λ_j are the singular values of $\boldsymbol{X}$.
- If λ_j is small compared to α, then direction $\mathbf{u}_j$ is suppressed.

Why ridge regression works

- Analytic solution: compute the regularized operator

$$\boldsymbol{R} = (\boldsymbol{X}^\intercal \boldsymbol{X} + \alpha \boldsymbol{I})^{-1} \boldsymbol{X}^\intercal$$

Why ridge regression works

- Analytic solution: compute the regularized operator

$$\boldsymbol{R} = (\boldsymbol{X}^\intercal \boldsymbol{X} + \alpha \boldsymbol{I})^{-1} \boldsymbol{X}^\intercal$$

- If $\boldsymbol{X} = \text{diag}(\lambda_1, \ldots, \lambda_d)$ with $\lambda_1 > \lambda_2 > \ldots > \lambda_d \to 0$ (meaning we're in eigenbasis of $\boldsymbol{X}$) then

$$\begin{aligned}\boldsymbol{R} &= (\boldsymbol{X}^\intercal \boldsymbol{X} + \alpha \boldsymbol{I})^{-1} \boldsymbol{X}^\intercal = \\ &= \big(\text{diag}(\lambda_1, \ldots, \lambda_d)\text{diag}(\lambda_1, \ldots, \lambda_d) + \text{diag}(\alpha, \ldots, \alpha)\big)^{-1}\text{diag}(\lambda_1, \ldots, \lambda_d) = \\ &= \text{diag}\left(\frac{\lambda_1}{\lambda_1^2 + \alpha}, \ldots, \frac{\lambda_d}{\lambda_d^2 + \alpha}\right),\end{aligned}$$

  smoothing diagonal values in $\boldsymbol{R}$