Lecture 3: Linear Regression
CS480/680 Intro to Machine Learning

2023-1-17

Pascal Poupart
David R. Cheriton School of Computer Science

UNIVERSITY OF
WATERLOO

Linear model for regression

- Simple form of regression
- Picture:

Problem

- Data: $\{(\boldsymbol{x_1}, y_1), (\boldsymbol{x_2}, y_2), \ldots, (\boldsymbol{x_N}, y_N)\}$
  - $\boldsymbol{x} = < x_1, x_2, \ldots, x_M >$: input vector
  - y: target (continuous value)
- Problem: find hypothesis h that maps $\boldsymbol{x}$ to y
  - Assume that h is linear:

$$h(\boldsymbol{x}, \boldsymbol{w}) = w_0 + w_1 x_1 + \cdots + w_M x_M = \boldsymbol{w^T} \begin{pmatrix} 1 \\ \boldsymbol{x} \end{pmatrix}$$

- Objective: minimize some loss function
  - Euclidean loss: $L_2(\boldsymbol{w}) = \frac{1}{2} \sum_{n=1}^{N} (h(\boldsymbol{x_n}, \boldsymbol{w}) - y_n)^2$

Optimization

- Find best $\boldsymbol{w}$ that minimizes Euclidean loss

$$\boldsymbol{w}^* = argmin_{\boldsymbol{w}} \frac{1}{2} \sum_{n=1}^{N} \left(y_n - \boldsymbol{w}^{\boldsymbol{T}} \begin{pmatrix} 1 \\ \boldsymbol{x_n} \end{pmatrix} \right)^2$$

- Convex optimization problem

  $\Longrightarrow$ unique optimum (global)

Solution

- Let $\bar{\boldsymbol{x}} = \begin{pmatrix} 1 \\ \boldsymbol{x} \end{pmatrix}$ then $\min_{\boldsymbol{w}} \frac{1}{2} \sum_{n=1}^{N} \left(y_n - \boldsymbol{w}^T \bar{\boldsymbol{x}}_n \right)^2$
- Find $\boldsymbol{w}^*$ by setting the derivative to 0

$$\frac{\partial L_2}{\partial_{w_j}} = \sum_{n=1}^{N} \left(y_n - \boldsymbol{w}^T \bar{\boldsymbol{x}}_n \right) \bar{x}_{nj} = 0 \quad \forall j$$

$$\Longrightarrow \sum_{n=1}^{N} \left(y_n - \boldsymbol{w}^T \bar{\boldsymbol{x}}_n \right) \bar{\boldsymbol{x}}_n = 0$$

- This is a linear system in $\boldsymbol{w}$, therefore we rewrite it as $\boldsymbol{A}\boldsymbol{w} = \boldsymbol{b}$

  where $\boldsymbol{A} = \sum_{n=1}^{N} \bar{\boldsymbol{x}}_n \bar{\boldsymbol{x}}_n^T$ and $\boldsymbol{b} = \sum_{n=1}^{N} y_n \bar{\boldsymbol{x}}_n$

Solution

- If training instances span $\Re^{M+1}$ then $\boldsymbol{A}$ is invertible:

$$\boldsymbol{w} = \boldsymbol{A}^{-1}\boldsymbol{b}$$

- In practice it is faster to solve the linear system $\boldsymbol{A}\boldsymbol{w} = \boldsymbol{b}$ directly instead of inverting $\boldsymbol{A}$
  - Gaussian elimination
  - Conjugate gradient
  - Iterative methods

Picture

Regularization

- Least square solution may not be stable
  - i.e., slight perturbation of the input may cause a dramatic change in the output
  - Form of **overfitting**

Example 1

- Training data: $\bar{\boldsymbol{x}}_1 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$ $\bar{\boldsymbol{x}}_2 = \begin{pmatrix} 1 \\ \epsilon \end{pmatrix}$

  $y_1 = 1$ $y_2 = 1$

- $\boldsymbol{A} =$

- $\boldsymbol{A^{-1}} =$ $\boldsymbol{b} =$

- $\boldsymbol{w} =$

UNIVERSITY OF WATERLOO

Example 2

- Training data: $\bar{\boldsymbol{x}}_1 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$ $\bar{\boldsymbol{x}}_2 = \begin{pmatrix} 1 \\ \epsilon \end{pmatrix}$

  $y_1 = 1 + \epsilon$ $y_2 = 1$

- $\boldsymbol{A} =$

- $\boldsymbol{A^{-1}} =$ $\boldsymbol{b} =$

- $\boldsymbol{w} =$

UNIVERSITY OF WATERLOO

Picture

Regularization

- Idea: favor smaller values
- Tikhonov regularization: add $||\boldsymbol{w}||_2^2$ as a penalty term
- Ridge regression:

$$\boldsymbol{w}^* = argmin_{\boldsymbol{w}} \frac{1}{2} \sum_{n=1}^{N} \left(y_n - \boldsymbol{w}^T \bar{\boldsymbol{x}}_n\right)^2 + \frac{\lambda}{2} ||\boldsymbol{w}||_2^2$$

where λ is a weight to adjust the importance of the penalty

Regularization

- Solution: $(\lambda \boldsymbol{I} + \boldsymbol{A})\boldsymbol{w} = \boldsymbol{b}$

- Notes
  - Without regularization: eigenvalues of linear system may be arbitrarily close to 0 and the inverse may have arbitrarily large eigenvalues.
  - With Tikhonov regularization, eigenvalues of linear system are $\geq \lambda$ and therefore bounded away from 0. Similarly, eigenvalues of inverse are bounded above by $1/\lambda$.

Regularized Examples

Example 1

Example 2