CS489/698 Machine Learning
Lecture 3: Jan 11, 2017

Linear Regression

[RN] Sec. 18.6.1, [HTF] Sec. 2.3.1,
[D] Sec. 7.6, [B] Sec. 3.1, [M] Sec. 1.4.5

Linear model for regression

- Simple form of regression
- Picture:

Problem

- Data: $\{(\boldsymbol{x_1}, t_1), (\boldsymbol{x_2}, t_2), \dots, (\boldsymbol{x_N}, t_N)\}$
  - $\boldsymbol{x} = < x_1, x_2, \dots, x_D >$: input vector
  - t: target (continuous value)
- Problem: find hypothesis h that maps $\boldsymbol{x}$ to t
  - Assume that h is linear:

$$y(\boldsymbol{x}, \boldsymbol{w}) = w_0 + w_1 x_1 + \cdots + w_D x_D = \boldsymbol{w}^{\boldsymbol{T}} \binom{1}{\boldsymbol{x}}$$

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

Optimization

- Find best w that minimizes Euclidean loss

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

- Convex optimization problem

$\Longrightarrow$ unique optimum (global)

Solution

- Let $\overline{\boldsymbol{x}} = \binom{1}{\boldsymbol{x}}$ then $\min_{\boldsymbol{w}} \frac{1}{2} \sum_{n=1}^{N} \left(t_n - \boldsymbol{w}^{\boldsymbol{T}} \overline{\boldsymbol{x}}_{\boldsymbol{n}}\right)^2$
- Find $\boldsymbol{w}^*$ by setting the derivative to 0

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

$$\Longrightarrow \sum_{n=1}^{N} \left(t_n - \boldsymbol{w}^{\boldsymbol{T}} \overline{\boldsymbol{x}}_{\boldsymbol{n}}\right) \overline{\boldsymbol{x}}_{\boldsymbol{n}} = 0$$

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

where $\boldsymbol{A} = \sum_{n=1}^{N} \overline{\boldsymbol{x}}_{\boldsymbol{n}} \overline{\boldsymbol{x}}_{\boldsymbol{n}}^{\boldsymbol{T}}$ and $\boldsymbol{b} = \sum_{n=1}^{N} t_n \overline{\boldsymbol{x}}_{\boldsymbol{n}}$

Solution

- If training instances span $\Re^{D+1}$ 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: $\bar{\boldsymbol{x}}_{\mathbf{1}} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$ $\quad \bar{\boldsymbol{x}}_{\mathbf{2}} = \begin{pmatrix} 1 \\ \epsilon \end{pmatrix}$

  $t_1 = 1 \qquad t_2 = 1$

- $\boldsymbol{A} =$

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

- $\boldsymbol{w} =$

Example 2

- Training data: $\bar{\boldsymbol{x}}_{\mathbf{1}} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$ $\quad \bar{\boldsymbol{x}}_{\mathbf{2}} = \begin{pmatrix} 1 \\ \epsilon \end{pmatrix}$

  $t_1 = 1 + \epsilon \quad t_2 = 1$

- $\boldsymbol{A} =$

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

- $\boldsymbol{w} =$

Picture

Regularization

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

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

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