

10-301/601: Introduction to Machine Learning

Lecture 7 – Linear Regression

Henry Chai

5/28/24

Front Matter

- Announcements:
 - HW3 released 5/23, due 6/4 at 11:59 PM
- Recommended Readings:
 - Murphy, [Chapters 7.1-7.3](#)

Recall: Regression

- Learning to diagnose heart disease

as a **(supervised)**

regression task

features

targets

data points

x_1 Family History	x_2 Resting Blood Pressure	x_3 Cholesterol	y Heart Disease?
Yes	Low	Normal	\$0
No	Medium	Normal	\$20
No	Low	Abnormal	\$30
Yes	Medium	Normal	\$100
Yes	High	Abnormal	\$5000

Decision Tree Regression

- Learning to diagnose heart disease

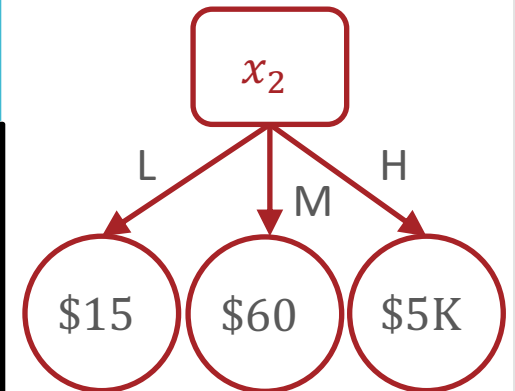
as a **(supervised)** regression task

features

targets

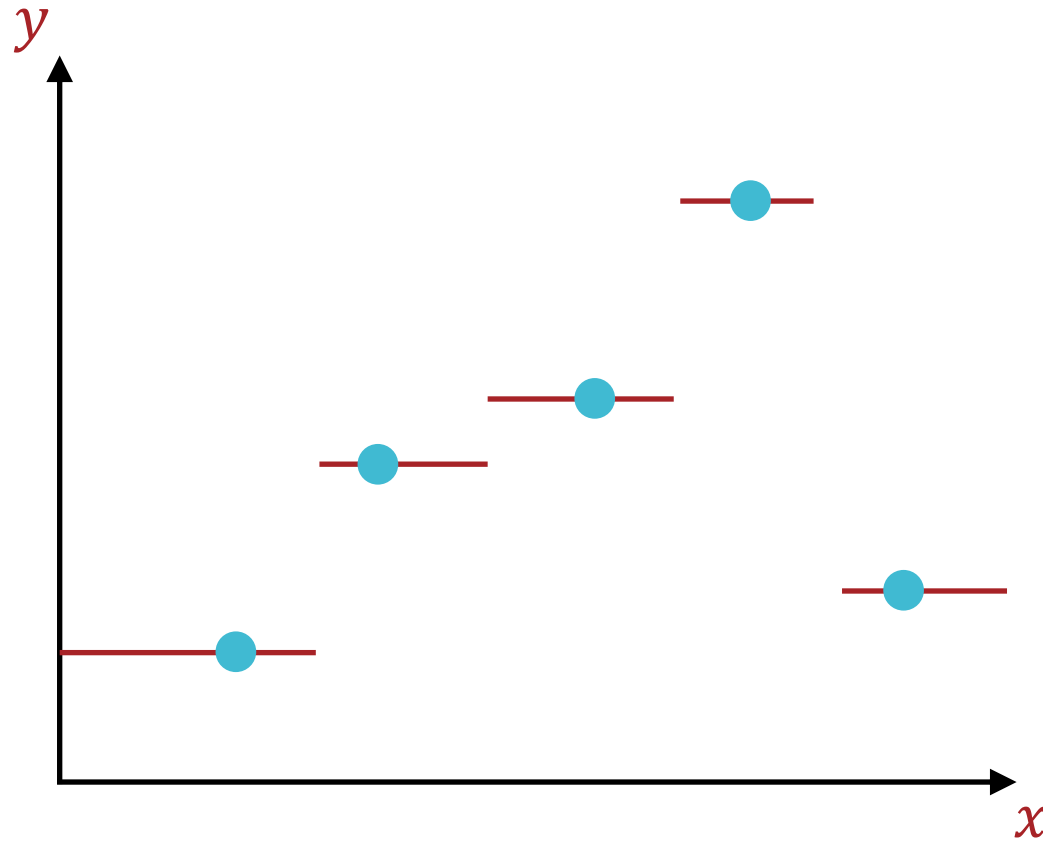
x_1 Family History	x_2 Resting Blood Pressure	x_3 Cholesterol	y Heart Disease?
Yes	Low	Normal	\$0
No	Medium	Normal	\$20
No	Low	Abnormal	\$30
Yes	Medium	Normal	\$100
Yes	High	Abnormal	\$5000

data points



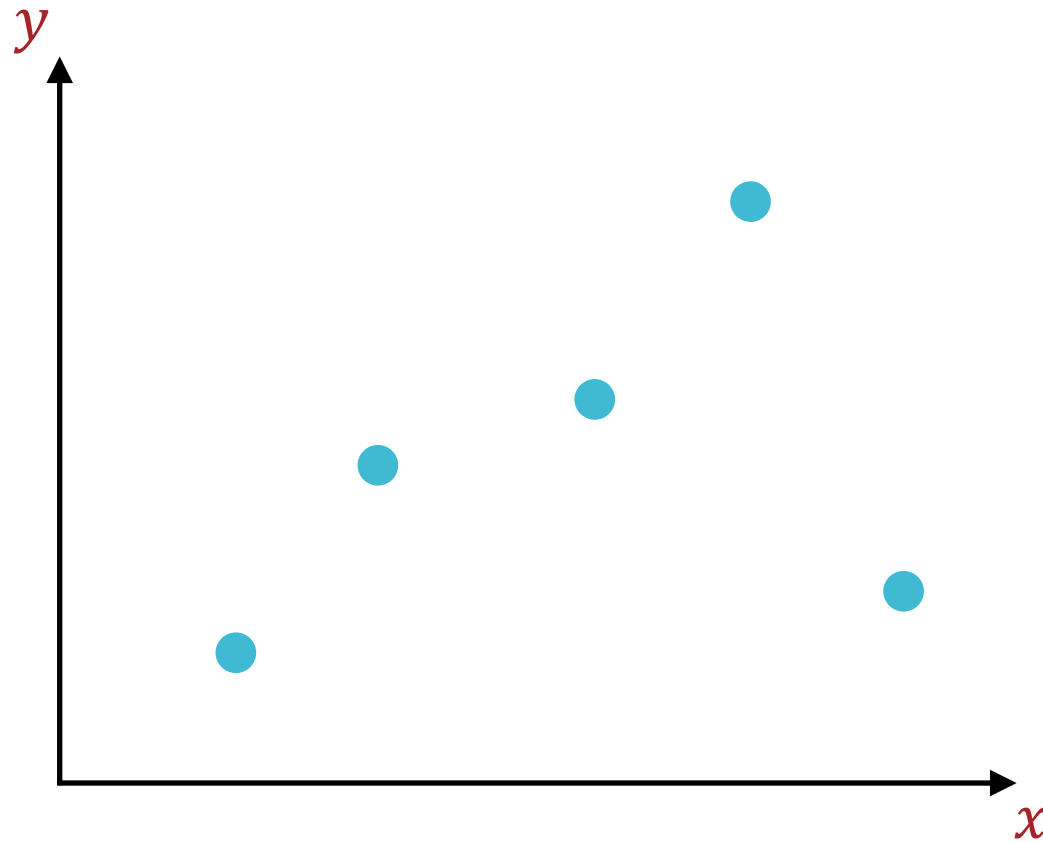
1-NN Regression

- Suppose we have real-valued targets $y \in \mathbb{R}$ and one-dimensional inputs $x \in \mathbb{R}$



2-NN Regression?

- Suppose we have real-valued targets $y \in \mathbb{R}$ and one-dimensional inputs $x \in \mathbb{R}$



Linear Regression

- Suppose we have real-valued targets $y \in \mathbb{R}$ and D -dimensional inputs $\mathbf{x} = [x_1, \dots, x_D]^T \in \mathbb{R}^D$
- **Assume**

$$y = \mathbf{w}^T \mathbf{x} + w_0$$

Linear Regression

- Suppose we have real-valued targets $y \in \mathbb{R}$ and D -dimensional inputs $\mathbf{x} = [1, x_1, \dots, x_D]^T \in \mathbb{R}^{D+1}$
- **Assume**

$$y = \mathbf{w}^T \mathbf{x}$$

General Recipe for Machine Learning

- Define a model and model parameters
- Write down an objective function
- Optimize the objective w.r.t. the model parameters

Recipe for Linear Regression

- Define a model and model parameters
 - Assume $y = \mathbf{w}^T \mathbf{x}$
 - Parameters: $\mathbf{w} = [w_0, w_1, \dots, w_D]$

- Write down an objective function
 - Minimize the mean squared error

$$\ell_{\mathcal{D}}(\mathbf{w}) = \frac{1}{N} \sum_{n=1}^N \ell^{(n)}(\mathbf{w}) = \frac{1}{N} \sum_{n=1}^N (\mathbf{w}^T \mathbf{x}^{(n)} - y^{(n)})^2$$

- Optimize the objective w.r.t. the model parameters
 - Solve in *closed form*: take partial derivatives, set to 0 and solve

Minimizing the Squared Error

$$\ell_{\mathcal{D}}(\mathbf{w}) = \frac{1}{N} \sum_{n=1}^N (\mathbf{w}^T \mathbf{x}^{(n)} - y^{(n)})^2$$

$$= \frac{1}{N} \sum_{n=1}^N \left(\sum_{d=0}^D w_d x_d^{(n)} - y^{(n)} \right)^2$$

$$\frac{\partial \ell_{\mathcal{D}}(\mathbf{w})}{\partial w_d} = \frac{1}{N} \sum_{n=1}^N 2 \left(\sum_{d=0}^D w_d x_d^{(n)} - y^{(n)} \right) \frac{\partial}{\partial w_d} \left(\sum_{d=0}^D w_d x_d^{(n)} - y^{(n)} \right)$$

$$= \frac{1}{N} \sum_{n=1}^N 2 \left(\sum_{d=0}^D w_d x_d^{(n)} - y^{(n)} \right) x_d^{(n)}$$

Recipe for Linear Regression

- Define a model and model parameters
 - Assume $y = \mathbf{w}^T \mathbf{x}$
 - Parameters: $\mathbf{w} = [w_0, w_1, \dots, w_D]$

- Write down an objective function
 - Minimize the mean squared error

$$\ell_{\mathcal{D}}(\mathbf{w}) = \frac{1}{N} \sum_{n=1}^N \ell^{(n)}(\mathbf{w}) = \frac{1}{N} \sum_{n=1}^N (\mathbf{w}^T \mathbf{x}^{(n)} - y^{(n)})^2$$

- Optimize the objective w.r.t. the model parameters
 - Solve in *closed form*: take ~~partial derivatives~~ gradient, set to 0 and solve

Linear Regression

- Suppose we have real-valued targets $y \in \mathbb{R}$ and D -dimensional inputs $\mathbf{x} = [1, x_1, \dots, x_D]^T \in \mathbb{R}^{D+1}$

- **Assume**

$$y = \mathbf{w}^T \mathbf{x}$$

- Notation: given training data $\mathcal{D} = \{(\mathbf{x}^{(n)}, y^{(n)})\}_{n=1}^N$

$$\bullet X = \begin{bmatrix} 1 & \mathbf{x}^{(1)T} \\ 1 & \mathbf{x}^{(2)T} \\ \vdots & \vdots \\ 1 & \mathbf{x}^{(N)T} \end{bmatrix} = \begin{bmatrix} 1 & x_1^{(1)} & \cdots & x_D^{(1)} \\ 1 & x_1^{(2)} & \cdots & x_D^{(2)} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_1^{(N)} & \cdots & x_D^{(N)} \end{bmatrix} \in \mathbb{R}^{N \times D+1}$$

is the *design matrix*

- $\mathbf{y} = [y^{(1)}, \dots, y^{(N)}]^T \in \mathbb{R}^N$ is the *target vector*

Minimizing the Squared Error

$$\ell_{\mathcal{D}}(\mathbf{w}) = \frac{1}{N} \sum_{n=1}^N (\mathbf{w}^T \mathbf{x}^{(n)} - y^{(n)})^2 = \frac{1}{N} \sum_{n=1}^N (\mathbf{x}^{(n)T} \mathbf{w} - y^{(n)})^2$$

$$= \frac{1}{N} \|\mathbf{X}\mathbf{w} - \mathbf{y}\|_2^2 \quad \text{where } \|\mathbf{z}\|_2 = \sqrt{\sum_{d=1}^D z_d^2} = \sqrt{\mathbf{z}^T \mathbf{z}}$$

$$= \frac{1}{N} (\mathbf{X}\mathbf{w} - \mathbf{y})^T (\mathbf{X}\mathbf{w} - \mathbf{y})$$

$$= \frac{1}{N} (\mathbf{w}^T \mathbf{X}^T \mathbf{X} \mathbf{w} - 2\mathbf{w}^T \mathbf{X}^T \mathbf{y} + \mathbf{y}^T \mathbf{y})$$

$$\nabla_{\mathbf{w}} \ell_{\mathcal{D}}(\mathbf{w}) = \frac{1}{N} (2\mathbf{X}^T \mathbf{X} \mathbf{w} - 2\mathbf{X}^T \mathbf{y})$$

Minimizing the Squared Error

$$\ell_{\mathcal{D}}(\mathbf{w}) = \frac{1}{N} \sum_{n=1}^N (\mathbf{w}^T \mathbf{x}^{(n)} - y^{(n)})^2 = \frac{1}{N} \sum_{n=1}^N (\mathbf{x}^{(n)T} \mathbf{w} - y^{(n)})^2$$

$$= \frac{1}{N} \|X\mathbf{w} - \mathbf{y}\|_2^2 \quad \text{where } \|\mathbf{z}\|_2 = \sqrt{\sum_{d=1}^D z_d^2} = \sqrt{\mathbf{z}^T \mathbf{z}}$$

$$= \frac{1}{N} (X\mathbf{w} - \mathbf{y})^T (X\mathbf{w} - \mathbf{y})$$

$$= \frac{1}{N} (\mathbf{w}^T X^T X \mathbf{w} - 2\mathbf{w}^T X^T \mathbf{y} + \mathbf{y}^T \mathbf{y})$$

$$\nabla_{\mathbf{w}} \ell_{\mathcal{D}}(\hat{\mathbf{w}}) = \frac{1}{N} (2X^T X \hat{\mathbf{w}} - 2X^T \mathbf{y}) = 0$$

$$\rightarrow X^T X \hat{\mathbf{w}} = X^T \mathbf{y}$$

$$\rightarrow \hat{\mathbf{w}} = (X^T X)^{-1} X^T \mathbf{y}$$

Minimizing the Squared Error

$$\ell_{\mathcal{D}}(\mathbf{w}) = \frac{1}{N} \sum_{n=1}^N (\mathbf{w}^T \mathbf{x}^{(n)} - y^{(n)})^2 = \frac{1}{N} \sum_{n=1}^N (\mathbf{x}^{(n)T} \mathbf{w} - y^{(n)})^2$$

$$= \frac{1}{N} \|\mathbf{X}\mathbf{w} - \mathbf{y}\|_2^2 \quad \text{where } \|\mathbf{z}\|_2 = \sqrt{\sum_{d=1}^D z_d^2} = \sqrt{\mathbf{z}^T \mathbf{z}}$$

$$= \frac{1}{N} (\mathbf{X}\mathbf{w} - \mathbf{y})^T (\mathbf{X}\mathbf{w} - \mathbf{y})$$

$$= \frac{1}{N} (\mathbf{w}^T \mathbf{X}^T \mathbf{X} \mathbf{w} - 2\mathbf{w}^T \mathbf{X}^T \mathbf{y} + \mathbf{y}^T \mathbf{y})$$

$$\nabla_{\mathbf{w}} \ell_{\mathcal{D}}(\mathbf{w}) = \frac{1}{N} (2\mathbf{X}^T \mathbf{X} \mathbf{w} - 2\mathbf{X}^T \mathbf{y})$$

$$H_{\mathbf{w}} \ell_{\mathcal{D}}(\mathbf{w}) = \frac{1}{N} 2\mathbf{X}^T \mathbf{X} \quad \text{which is positive semi-definite}$$

$$\hat{\mathbf{w}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$$

1. Is $\mathbf{X}^T \mathbf{X}$ invertible?

2. If so, how computationally expensive is inverting $\mathbf{X}^T \mathbf{X}$?

Closed Form Solution

0 surveys completed



0 surveys underway

Is $X^T X$ always invertible?

Yes

No

Unsure

If $X^T X$ is invertible, how computationally expensive is it to invert?

$O(N^2)$

$O(D^2)$

$O(ND)$

$O(N^3)$

$O(D^3)$

Closed Form Solution

$$\hat{\mathbf{w}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$$

1. Is $\mathbf{X}^T \mathbf{X}$ invertible?
 - When $N \gg D + 1$, $\mathbf{X}^T \mathbf{X}$ is (almost always) full rank and therefore, invertible!
 - If $\mathbf{X}^T \mathbf{X}$ is not invertible (occurs when one of the features is a linear combination of the others), what does that imply about our problem?
2. If so, how computationally expensive is inverting $\mathbf{X}^T \mathbf{X}$?
 - $\mathbf{X}^T \mathbf{X} \in \mathbb{R}^{D+1 \times D+1}$ so inverting $\mathbf{X}^T \mathbf{X}$ takes $O(D^3)$ time...
 - Computing $\mathbf{X}^T \mathbf{X}$ takes $O(ND^2)$ time
 - What alternative optimization method(s) can we use to minimize the mean squared error?

Key Takeaways

- Decision tree and k NN regression
- Closed form solution for linear regression
 - Setting partial derivative/gradients to 0 and solving for critical points
 - Potential issues with the closed form solution: invertibility and computational costs