



CSC380: Principles of Data Science

Linear Models 1

1

1

Outline

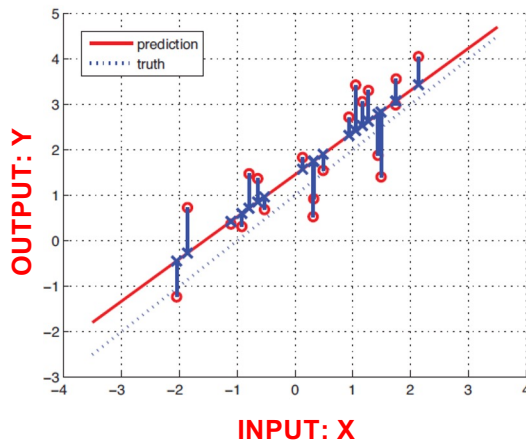
2

- Linear Regression first focus on what is a linear function
- Least Squares Estimation then learn how to train a linear function
- Regularized Least Squares
- Logistic Regression

2

Linear Regression

3



Regression Learn a function that predicts outputs from inputs,

$$y = f(x)$$

Outputs y are real-valued

Linear Regression As the name suggests, uses a *linear function*:

$$y = w^T x + b$$

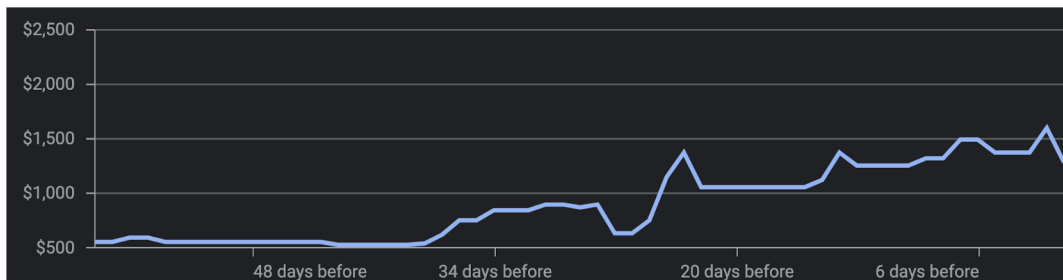
$$w^T x := \sum_{d=1}^D w_d x_d$$

3

Linear Regression

4

When is linear regression useful?



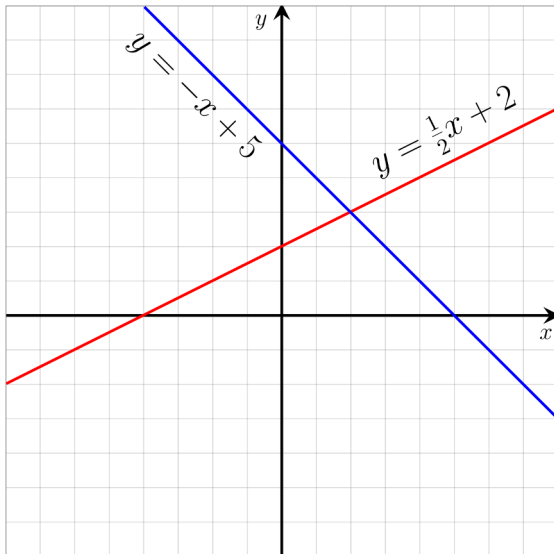
Price of an airline ticket

Used anywhere a linear relationship is assumed between inputs / (real-valued) outputs

4

Line Equation

5



Recall the equation for a line has a *slope* and an *intercept*,

$$y = w \cdot x + b$$

Slope **Intercept**

- Intercept (b) indicates where line crosses y-axis
- Slope controls angle of line
- Positive slope (w) → Line goes up left-to-right
- Negative slope → Line goes down left-to-right

5

Review: inner product

Two vectors:

$$\vec{x} = \langle 2, -3 \rangle \quad \mathbf{x} = \begin{bmatrix} 2 \\ -3 \end{bmatrix}$$

$$\vec{y} = \langle 5, 1 \rangle \quad \mathbf{y} = \begin{bmatrix} 5 \\ 1 \end{bmatrix}$$

Multiply corresponding entries and add:

$$\vec{x} \cdot \vec{y} = \langle 2, -3 \rangle \cdot \langle 5, 1 \rangle = (2)(5) + (-3)(1) = 7$$

$$\mathbf{x}^T \mathbf{y} = \begin{bmatrix} 2 & -3 \end{bmatrix} \begin{bmatrix} 5 \\ 1 \end{bmatrix} = \begin{bmatrix} 7 \end{bmatrix} \quad (\text{or just } 7) \quad (\text{so } \vec{x} \cdot \vec{y} \text{ becomes } \mathbf{x}^T \mathbf{y})$$

6

Moving to higher dimensions...

7

- **1d regression**: regression with 1d input:

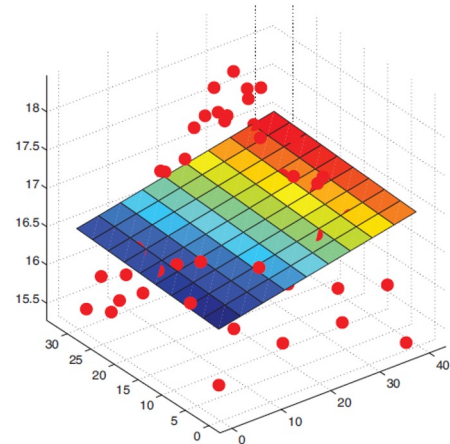
$$y = wx + b$$

- **D-dimensional regression**: input vector is $x \in \mathbb{R}^D$.

Recall the definition of an *inner product*:

$$w^T x = w_1 x_1 + w_2 x_2 + \dots + w_D x_D = \sum_{d=1}^D w_d x_d$$

The model is $y = w^T x + b$



[Image: Murphy, K. (2012)]

7

Moving to higher dimensions...

8

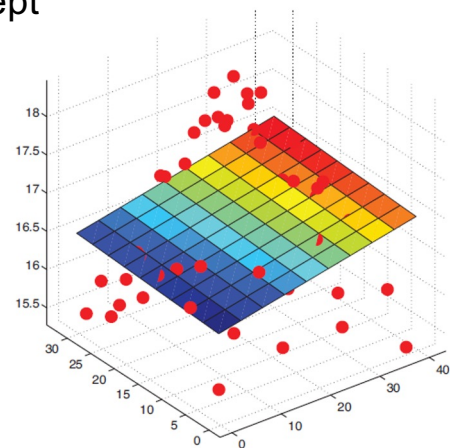
Often we simplify this by including the intercept into the weight vector,

$$\tilde{w} = \begin{pmatrix} w_1 \\ \vdots \\ w_D \\ b \end{pmatrix} \quad \tilde{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_D \\ 1 \end{pmatrix} \quad y = \tilde{w}^T \tilde{x}$$

Since:

$$\begin{aligned} \tilde{w}^T \tilde{x} &= \sum_{d=1}^D w_d x_d + b \cdot 1 \\ &= w^T x + b \end{aligned}$$

from now on, we assume that $w \in \mathbb{R}^D$ and $x \in \mathbb{R}^D$ already has b and 1 in the last coordinate respectively.



8

Learning Linear Regression Models

9

There are several ways to think about fitting regression:

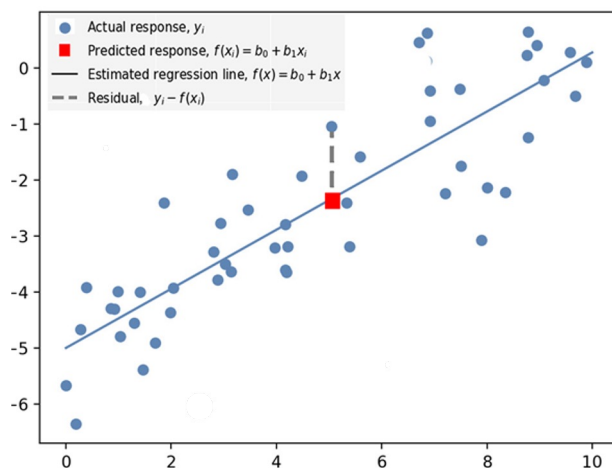
- **Intuitive** Find a plane/line that is close to data
- **Functional** Find a line that minimizes the *least squares* loss
- **Estimation** Find maximum likelihood estimate of parameters

They are all the same thing...

9

Fitting Linear Regression

10



Intuition Find a line that is as *close as possible* to every training data point

The distance from each point to the line is the **residual**

$$y - w^T x$$

Training Output Prediction

Let's find w that will minimize the residual!

<https://www.activestate.com/resources/quick-reads/how-to-run-linear-regressions-in-python-scikit-learn/>

10

Outline

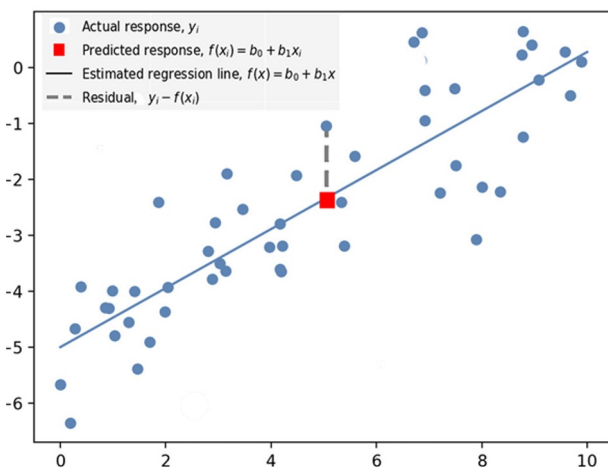
11

- Linear Regression
- Least Squares Estimation
- Regularized Least Squares
- Logistic Regression

11

Least Squares Solution

12



Functional Find a line that minimizes the sum of squared residuals!

Given: $\{(x^{(i)}, y^{(i)})\}_{i=1}^m$

Compute:

$$w^* = \arg \min_w \sum_{i=1}^m (y^{(i)} - w^T x^{(i)})^2$$

Least squares regression

<https://www.activestate.com/resources/quick-reads/how-to-run-linear-regressions-in-python-scikit-learn/>

12

Least Squares

13

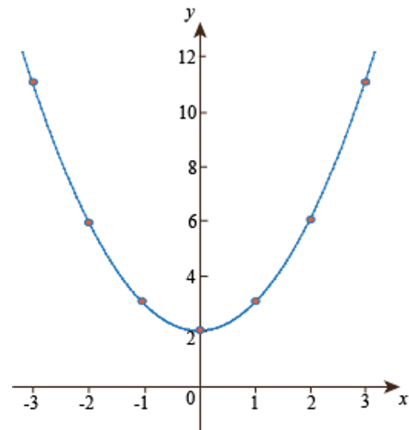
$$\min_w \sum_{i=1}^N (y^{(i)} - w^T x^{(i)})^2$$

This is just a quadratic function...

- *Convex*, unique minimum
- Minimum given by zero-derivative
- Can find a closed-form solution

Let's see for scalar case with no bias,

$$y = wx$$



13

Least Squares : Simple Case

14

$$\frac{d}{dw} \sum_{i=1}^N (y^{(i)} - wx^{(i)})^2 =$$

Derivative (+ chain rule)

$$= \sum_{i=1}^N 2(y^{(i)} - wx^{(i)})(-x^{(i)}) = 0 \Rightarrow$$

**Distributive Property
(and multiply -1 both sides)**

$$0 = \sum_{i=1}^N y^{(i)} x^{(i)} - w \sum_{j=1}^N (x^{(j)})^2$$

Algebra

$$w = \frac{\sum_i y^{(i)} x^{(i)}}{\sum_j (x^{(j)})^2}$$

14

Least Squares: Higher Dimensions

15

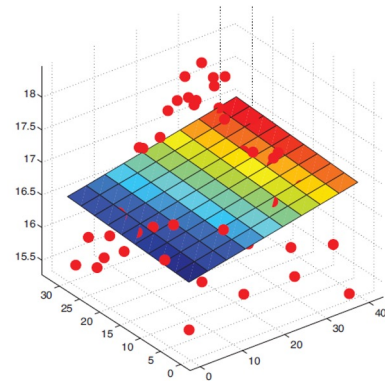
Things are a bit more complicated in higher dimensions and involve more linear algebra,

$$\mathbf{X} = \begin{pmatrix} x_1^{(1)} & \dots & x_D^{(1)} & 1 \\ x_1^{(2)} & \dots & x_D^{(2)} & 1 \\ \vdots & \vdots & \vdots & \vdots \\ x_1^{(m)} & \dots & x_D^{(m)} & 1 \end{pmatrix}$$

Design Matrix
(each row is a data point)

$$\mathbf{y} = \begin{pmatrix} y^{(1)} \\ \vdots \\ y^{(N)} \end{pmatrix}$$

Vector of labels



Can write regression over *all training data* more compactly...

$$\mathbf{y} \approx \mathbf{X}\mathbf{w} \quad \leftarrow \text{mx1 Vector}$$

$$= \begin{pmatrix} (x^{(1)})^T \mathbf{w} \\ \vdots \\ (x^{(m)})^T \mathbf{w} \end{pmatrix}$$

15

Least Squares: Higher Dimensions

16

Least squares can also be written more compactly, $\|\mathbf{x}\| := \sqrt{\mathbf{x} \cdot \mathbf{x}}$

$$\min_w \sum_{i=1}^N (y^{(i)} - w^T x^{(i)})^2 = \|\mathbf{y} - \mathbf{X}\mathbf{w}\|^2$$

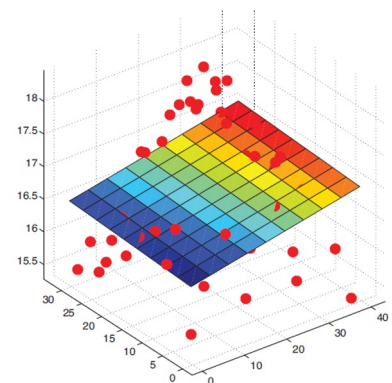
Some slightly more advanced linear algebra gives us a solution,

$$\mathbf{w} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y} \quad \text{compare with the 1d version: } w = \frac{\sum_i y^{(i)} x^{(i)}}{\sum_j (x^{(j)})^2}$$

Ordinary Least Squares (OLS) solution

Derivation a bit advanced for this class, but enough to know

- it has a closed-form and why
- we can evaluate it
- generally know where it comes from.



16