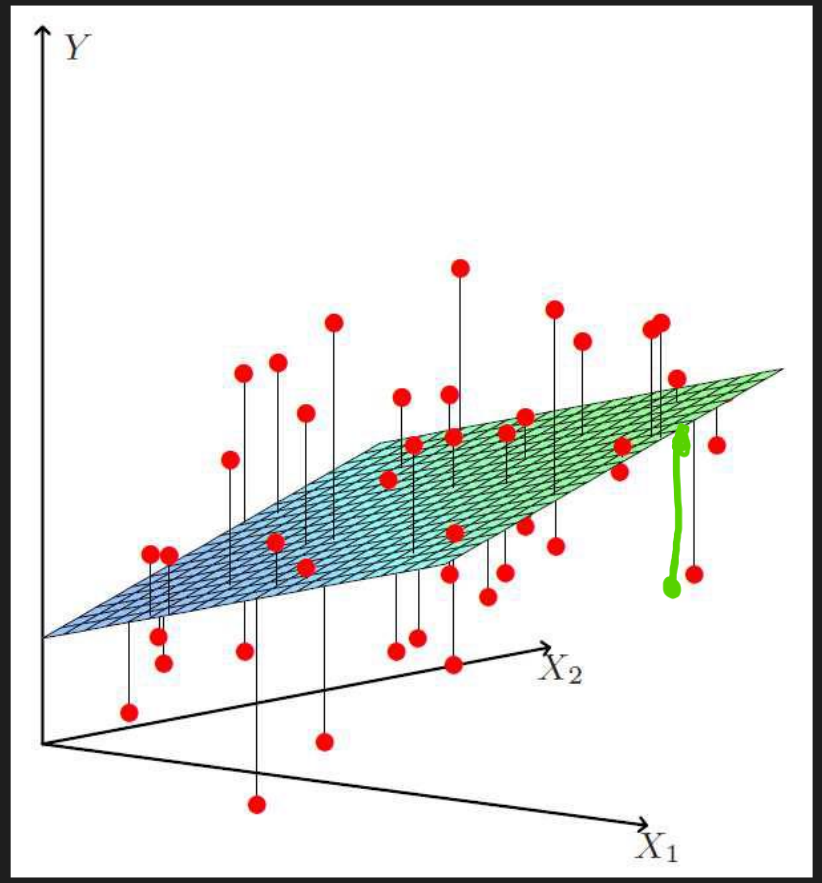
INTRODUCTION TO MACHINE LEARNING COMPSCI 4ML3

LECTURE 3

HASSAN ASHTIANI

ORDINARY LEAST SQUARES (D-DIMENSIONS)

- ASSUME $x \in \mathbb{R}^d$, $y \in \mathbb{R}$
- INSTEAD OF A LINE, WE NEED TO FIT A HYPERPLANE!
- WHY ARE THE LINES VERTICAL?
  - ANY DIFFERENT IF WE MINIMIZE THE DISTANCE TO THE HYPERPLANE?



HTTP://WWW.CS.CORNELL.EDU/COURSES/CS4758/2013SP/MATERIALS/CS4758-LIN-REGRESSION.PDF

MATRIX FORM OLS

$\sum(\Delta_i)^2$

- $\Delta = \begin{pmatrix} \Delta_1 \\ \dots \\ \Delta_n \end{pmatrix} = \begin{pmatrix} x_1^1 & \cdots & x_d^1 \\ \vdots & \ddots & \vdots \\ x_1^n & \cdots & x_d^n \end{pmatrix} \begin{pmatrix} w_1 \\ \dots \\ w_d \end{pmatrix} - \begin{pmatrix} y^1 \\ \dots \\ y^n \end{pmatrix} = \hat{Y}_{n\times 1} - Y_{n\times 1} = \begin{pmatrix} \hat{y}^1 - y^1 \\ \vdots \\ \hat{y}^n - y^n \end{pmatrix}$

$$\underset{W\in\mathbb{R}^{d\times 1}}{\text{MIN}} \sum_{i=1}^{n} (\Delta_i)^2 = \underset{W\in\mathbb{R}^{d\times 1}}{\text{MIN}} \langle \Delta, \Delta \rangle = \underset{W\in\mathbb{R}^{d\times 1}}{\text{MIN}} \|\Delta\|_2^2$$

$X_{n\times d}$ $W_{d\times 1}$ $Y_{n\times 1}$

$$\boldsymbol{\min_{W\in\mathbb{R}^{d\times 1}} \|XW - Y\|_2^2}$$

TAKING THE "DERIVATIVE"

- REAL-VALUED FUNCTION OF A VECTOR

$$f : \mathbb{R}^d \to \mathbb{R}$$

$$f(w) =$$

GRADIENT: $\nabla_w f = \begin{bmatrix} \frac{\partial f}{\partial w_1} \\ \vdots \\ \frac{\partial f}{\partial w_d} \end{bmatrix}$

- VECTOR-VALUED FUNCTION OF A VECTOR

$$g : \mathbb{R}^m \to \mathbb{R}^n$$

JACOBIAN:

$$\nabla_w g = \begin{bmatrix} \frac{\partial g_1}{\partial w_1} & \cdots & \frac{\partial g_1}{\partial w_m} \\ \vdots & & \\ \frac{\partial g_n}{\partial w_1} & \cdots & \frac{\partial g_n}{\partial w_m} \end{bmatrix}$$

$n \times m$

unlike gradients the columns are input variables not the rows.

MATRIX/VECTOR CALCULUS

- $u, v \in R^n$
- $g(u) = u^T v$

$g: \mathbb{R}^n \rightarrow \mathbb{R}$

$u_1 v_1 + u_2 v_2 \cdots$

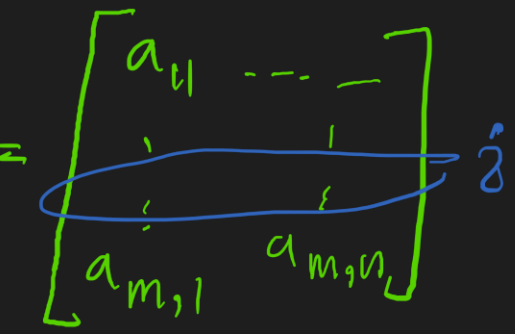
- $\nabla u(g) =$ gradient

$$\begin{bmatrix} \frac{\partial g}{\partial u_1} \\ \vdots \\ \frac{\partial g}{\partial u_n} \end{bmatrix} = \begin{bmatrix} \frac{\partial \sum u_i v_i}{\partial u_1} \\ \vdots \\ \frac{\partial \sum u_i v_i}{\partial u_n} \end{bmatrix} = \begin{bmatrix} v_1 \\ \vdots \\ v_n \end{bmatrix} = V$$

MATRIX/VECTOR CALCULUS

- $A \in R^{m \times n}, u \in R^n$
- $g(u) = Au$

$g : \mathbb{R}^n \rightarrow \mathbb{R}^m$

$A =$ 

- $\nabla u(g) =$ $A_{m \times n}$

$$\frac{\partial g_i}{\partial u_j} = \frac{\partial \left(\sum_{k=1}^{n} a_{i,k} \, u_k \right)}{\partial u_j} = a_{i,j}$$

Jacobian

MATRIX/VECTOR CALCULUS

- $A \in R^{n \times n}, u \in R^n$
- $g(u) = u^T A u$ $\quad g: R^n \rightarrow R$
- $\nabla u(g) =$ $\underbrace{u^T}_{1 \times n} \underbrace{(A + A^T)_{n \times n}}_{1 \times n}$

jacobian

show this as an exercise.

SOLVING OLS

$f(W) =$  $\|XW - Y\|_2^2$. What is ∇f?

$$\frac{\partial \|XW-Y\|_2^2}{\partial W} = \frac{\partial\left[(XW-Y)^T(XW-Y)\right]}{\partial W}$$

$$\frac{\partial\left[(W^TX^T-Y^T)(XW-Y)\right]}{\partial W} = \frac{\partial W^TX^TXW}{\partial W} + \frac{\partial Y^TY}{\partial W}$$

$$-\frac{\partial W^TX^TY}{\partial W} - \frac{\partial Y^TXW}{\partial W} = \vec{0}$$

A

Y^TXW scaler

$$\Rightarrow \quad w^T\left(x^T x + x^T x\right) + 0 - 2\,y^T x = 0$$

$$\Rightarrow \quad 2\,x^T x\,w = 2\,x^T y$$

* if $x^T x$ is invertible then

$$\boxed{\bar{w} = (x^T x)^{-1}\, x^T y}$$

SOLVING OLS

$$W^{LS} = (X^T X)^{-1} X^T Y$$

- DEGENERATE CASE WHEN $X^T X$ IS NOT INVERTIBLE?

BIAS/INTERCEPT TERM

- We are missing the bias term (w_0)

$$\min_{w_0, w_1, \ldots, w_d \in \mathbb{R}} \sum_{i=1}^{n} \left(w_1 x_1^i + \cdots + w_d x_d^i + w_0 - y^i \right)^2$$

- Matrix form with the bias term?

$$\min_{W \in \mathbb{R}^{d \times 1}, w_0 \in \mathbb{R}} \left\| \mathrm{X}\mathrm{W} + \begin{pmatrix} w_0 \\ w_0 \\ \cdots \\ w_0 \end{pmatrix} - \mathrm{Y} \right\|_2^2$$

EXAMPLE

$$\hat{y} = w_0 + w_1 x_1 + w_2 x_2$$

$$X_{n \times 2} = \begin{pmatrix} x_1^1 & x_2^1 \\ x_1^2 & x_2^2 \\ x_1^3 & x_2^3 \end{pmatrix} \longrightarrow X'_{n \times 3} = \begin{pmatrix} x_1^1 & x_2^1 & 1 \\ x_1^2 & x_2^2 & 1 \\ x_1^3 & x_2^3 & 1 \end{pmatrix}$$

$$W = \begin{pmatrix} w_1 \\ w_2 \end{pmatrix} \longrightarrow W' = \begin{pmatrix} w_1 \\ w_2 \\ w_0 \end{pmatrix}$$

$$X' W' = \begin{pmatrix} w'^T x'^1 \\ w'^T x'^2 \\ \vdots \end{pmatrix} = \begin{pmatrix} w_0 + \sum w_i x_i^1 \\ \vdots \\ w_0 + \sum w_i x_i^n \end{pmatrix}$$

BIAS/INTERCEPT TERM

- ADD A NEW AUXILIARY DIMENSION TO THE DATA
- $\mathrm{X}'_{n\times(d+1)} = \begin{pmatrix} x_1^1 & \cdots & x_d^1 & 1 \\ \vdots & \ddots & \vdots & 1 \\ x_1^n & \cdots & x_d^n & 1 \end{pmatrix}, \mathrm{W}'_{(d+1)\times 1} = \begin{pmatrix} w_1 \\ \cdots \\ w_d \\ w_0 \end{pmatrix}$
- SOLVE OLS: $\min_{\mathrm{W}'\in\mathbb{R}^{(\mathrm{D}+1)\times 1}} \|\mathrm{X}'\mathrm{W}' - \mathrm{Y}\|_2^2$
- w_0 WILL BE THE BIAS TERM!

test point $x = (x_1, x_2)$

$d = 2$, you know w'

$\hat{y} = ?$ $x' = (x_1, x_2, 1)$

$\hat{y} = w'^T x'$

SOME EXAMPLES

- OLS NOTEBOOK