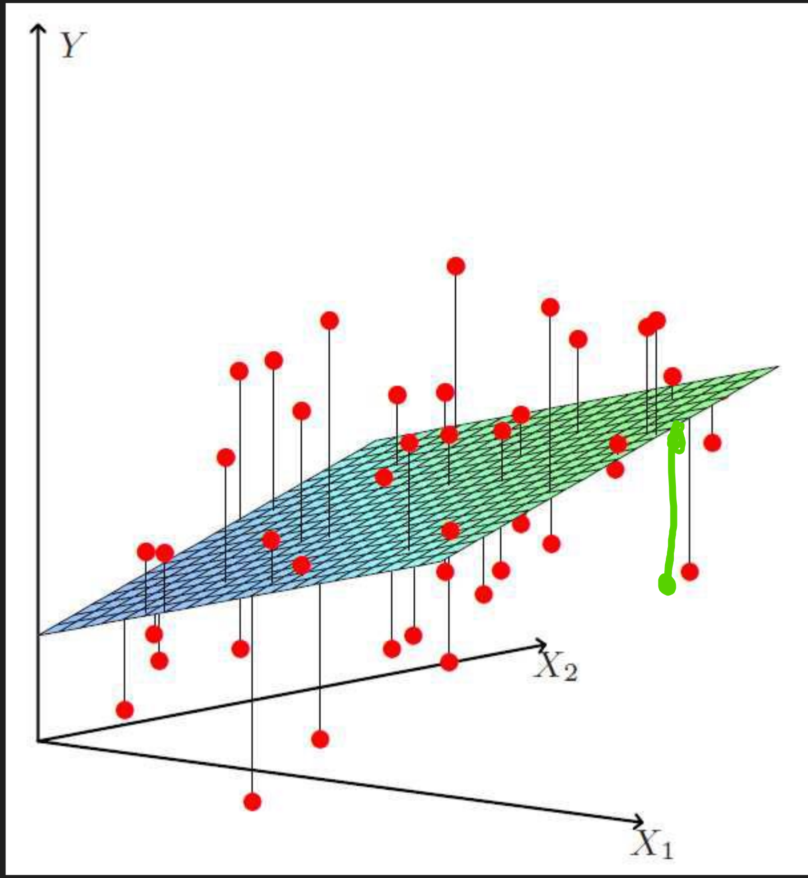
INTRODUCTION TO MACHINE LEARNING COMPSCI 4ML3

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

$\sum(\Delta_i)^2$

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

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

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

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

TAKING THE "DERIVATIVE"

- REAL-VALUED FUNCTION OF A VECTOR

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

$f(w) =$

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

- VECTOR-VALUED FUNCTION OF A VECTOR

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

$\boxed{\text{JACOBIAN:}}$

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

unlike gradients the columns are input variables not the rows.

MATRIX/VECTOR CALCULUS

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

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

$\boxed{u_1 v_1} + u_2 v_2 \cdots$

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

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

jacobian

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

$$A = \begin{bmatrix} a_{11} & \cdots & \\ \vdots & & \vdots \\ a_{m,1} & & a_{m,n} \end{bmatrix} \; i$$

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

MATRIX/VECTOR CALCULUS

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

jacobian

show this as an exercise,

SOLVING OLS

$f(W) = \quad \|XW - Y\|_2^2.$ WHAT IS ∇f?

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

$$\frac{\partial\left[(W^TX^T-Y^T)(XW-Y)\right]}{\partial W} = \frac{\partial W^T X^T X W}{\partial W} + \cancel{\frac{\partial Y^T Y}{\partial W}} \quad 0$$

$$A$$

$$-\frac{\partial W^T X^T Y}{\partial W} - \frac{\partial Y^T X W}{\partial W} = \vec{0}$$

$Y^T X W$

Scaler

$$\Rightarrow \quad w^T(X^TX + X^TX) + 0 - 2Y^TX = 0$$

$$\Rightarrow 2X^TXW = 2X^TY$$

* if X^TX is invertible then

$$\boxed{\bar{W} = (X^TX)^{-1}X^TY}$$

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?

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

EXAMPLE

BIAS/INTERCEPT TERM

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

SOME EXAMPLES

- OLS NOTEBOOK