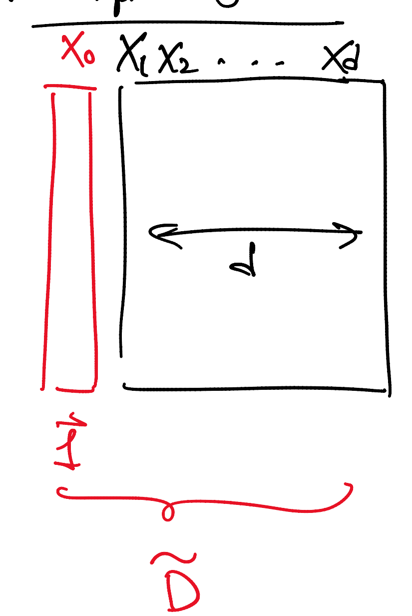
Lecture 10

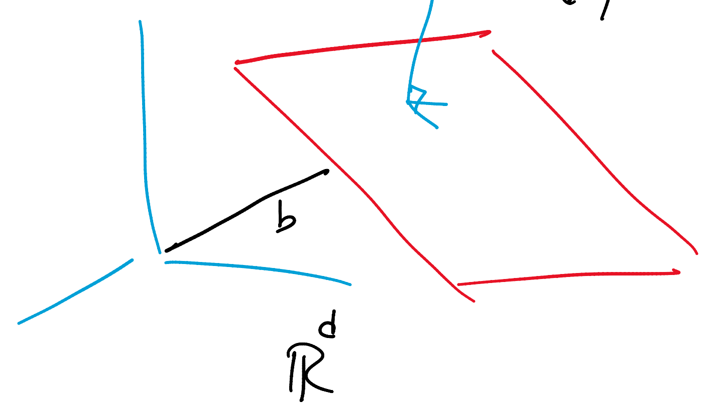
Thursday, September 28, 2023 10:01 AM

Multiple Regression



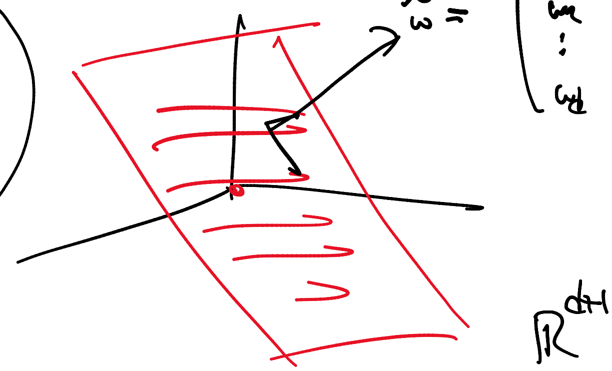
normal weight vector

$$\vec{\omega} = \begin{pmatrix} w_1 \\ w_2 \\ \vdots \\ w_d \end{pmatrix}$$



$\Downarrow$

$$\tilde{\omega} = \begin{pmatrix} b = w_0 \\ w_1 \\ w_2 \\ \vdots \\ w_d \end{pmatrix}$$



$$\hat{y}_i = \tilde{\omega}\,\vec{x}_i \quad \forall i = 1 \ldots n$$

$$= w_0 x_0 + w_1 x_1 + \cdots + w_d x_d$$

$$= b + w_1 x_1 + \cdots + w_d x_d$$

$$\boxed{\hat{Y} = \tilde{D}\,\vec{\tilde{\omega}}}$$

$n \times (d+1)$ $\quad$ $(d+1) \times 1$

point-view

$$\|Y - \hat{Y}\|^2 = SSE$$

$$= \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

$$SSE = \|Y - \hat{Y}\|^2$$

$$= (Y-\hat{Y})^T(Y-\hat{Y})$$

$$= \|Y\|^2 - 2\hat{Y}^T(Y) + \hat{Y}^T\hat{Y}$$

$$= \|Y\|^2 - 2(\tilde{D}\tilde{w})^T Y + (\tilde{D}\tilde{w})^T(\tilde{D}\tilde{w})$$

$$SSE = \|Y\|^2 - 2\tilde{w}^T(\tilde{D}^T Y) + \tilde{w}^T(\tilde{D}^T\tilde{D})\tilde{w}$$

$$\boxed{\frac{\partial SSE}{\partial \tilde{w}} = 0}$$

$$\frac{\partial SSE}{\partial \tilde{w}} = -2\tilde{D}^T Y + 2\tilde{D}^T\tilde{D}\tilde{w} = 0$$

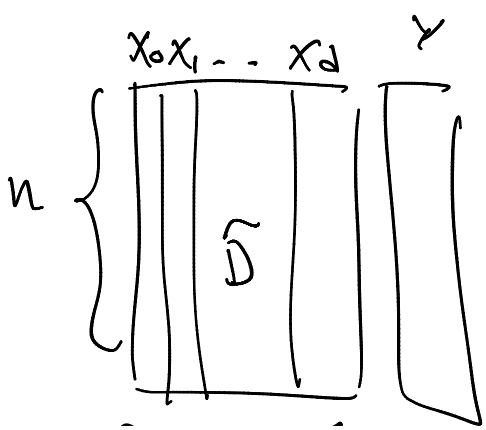
$$\boxed{\tilde{D}^T\tilde{D}\,\tilde{w} = \tilde{D}^T Y}$$

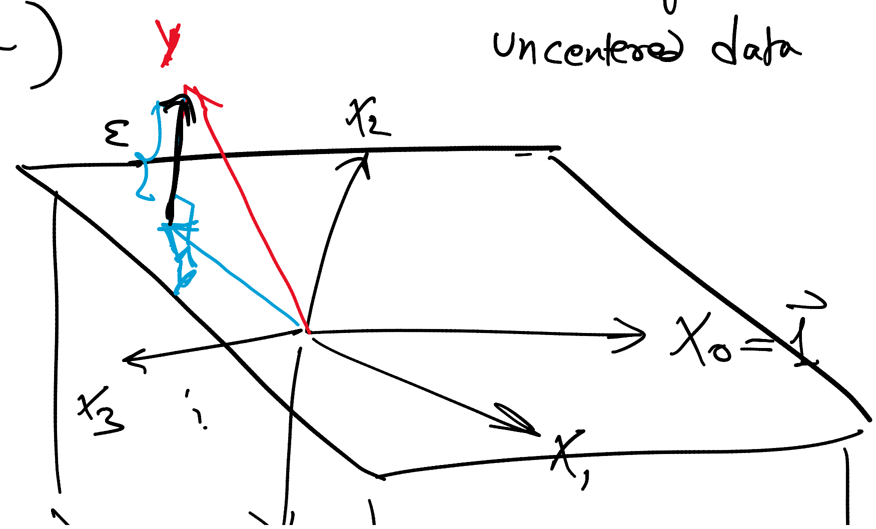
normal equations

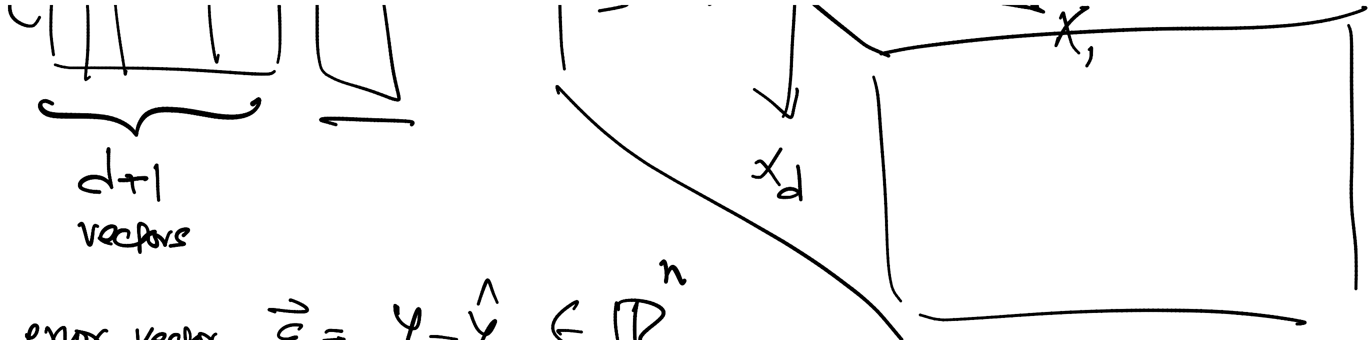
$$\tilde{w} = (\tilde{D}^T\tilde{D})^{-1}\tilde{D}^T Y$$

like Cov! scatter for uncentered data

Column view (Geometric)





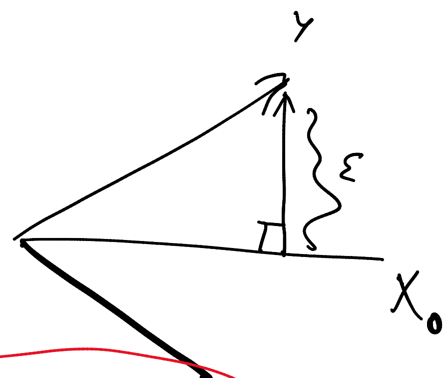


error vector $\vec{\varepsilon} = Y - \hat{Y} \in \mathbb{R}^n$

(d+1) dim subspace of $\mathbb{R}^n$

$\vec{\varepsilon}$ is orthogonal to the subspace

$\vec{\varepsilon}$ is orthogonal to all vectors e.g. X_i



$$\tilde{D}^T \vec{\varepsilon} = 0$$

$$\tilde{D} = \begin{array}{c} \\ n \end{array} \begin{bmatrix} | & | & & | \\ x_0 & x_1 & \cdots & x_d \\ | & | & & | \end{bmatrix}$$

$$\tilde{D}^T \vec{\varepsilon} = \begin{array}{c} \\ d+1 \\ \\ \end{array} \underbrace{\begin{bmatrix} - & x_0^T & - \\ - & x_1^T & - \\ & \vdots & \\ - & x_d^T & - \end{bmatrix}}_{n} \vec{\varepsilon}$$

$$\tilde{D}^T \vec{\varepsilon} = 0$$

$$\begin{bmatrix} x_0^T \vec{\varepsilon} \\ x_1^T \varepsilon \\ \vdots \\ x_d^T \varepsilon \end{bmatrix} = \vec{0}$$

$$\Rightarrow \tilde{D}^T (Y - \hat{Y}) = 0$$

$$\Rightarrow \tilde{D}^T Y = \tilde{D}^T \hat{Y}$$

$$\tilde{D}^T Y = \tilde{D}^T (\tilde{D} \tilde{w})$$

$$\boxed{\tilde{D}^T \tilde{D} \tilde{w} = \tilde{D}^T Y}$$ normal equation

$$\boxed{\tilde{D}^T \tilde{D} \tilde{w} = \tilde{D}^T Y}$$ normal equation

---

$\{X_0\ X_1 \cdots X_d\}$ ← not a good basis

$\Downarrow$

$\{U_0\ U_1 \cdots U_d\}$ ← orthogonal basis

$$\hat{Y} = \underline{\text{Proj}_{U_0}(Y)} \cdot \vec{U_0} + \underline{\text{Proj}_{U_1}(Y)} \cdot \vec{U_1} + \cdots + \underline{\text{Proj}_{U_d}(Y)} \cdot \vec{U_d}$$

via QR

$$\underline{\underline{\tilde{D}^T}} \tilde{D} \tilde{w} = \tilde{D}^T Y$$

$$\tilde{w} = \underline{\underline{(\tilde{D}^T \tilde{D})^{-1}}} (\tilde{D}^T Y)$$

may not exist

Positive semi definite

Ridge regression

$$\boxed{\tilde{w} = \underbrace{\left(\tilde{D}^T\tilde{D}+\alpha I\right)^{-1}}_{\text{always have an inverse}} \tilde{D}^T Y}$$

$\alpha > 0$

full rank matrix

$$\underbrace{\begin{bmatrix} & & & \\ & & & \\ & & & \\ & & & \end{bmatrix}}_{\tilde{D}^T\tilde{D}} + \underbrace{\begin{bmatrix} \alpha & & & 0 \\ & \alpha & & \\ & & \alpha & \\ 0 & & & \alpha \end{bmatrix}}_{\alpha I} = \underbrace{\begin{bmatrix} +\alpha & & & \\ & +\alpha & & \\ & & +\alpha & \\ & & & +\alpha \end{bmatrix}}_{(\tilde{D}^T\tilde{D}+\alpha I)}$$

$\underbrace{\alpha I}_{\text{ridge}}$

$\lambda_i + \alpha$

λ_i

If λ_i is an eigenvalue of $\tilde{D}^T\tilde{D}$ then $\lambda_i+\alpha$ is the corresponding eigenvalue of $\tilde{D}^T\tilde{D}+\alpha I$

$$\lambda_i = 0 \Rightarrow \lambda_i + \alpha > 0$$

Ridge Regression

(Regularization)

$$SSE = \|Y - \hat{Y}\|^2$$

$$\boxed{\min_{\tilde{w}} \; J = \|Y - \tilde{D}\tilde{w}\|^2}$$ original objective

$$\min_{\tilde{\omega}} \; J = \| Y - \tilde{D}\tilde{\omega} \| \qquad \text{original objective}$$

On the training sample

$$\hat{y}_i = \underset{5}{\underline{w_0}} x_0 + \underset{1000}{w_1} x_1 + \cdots + \underset{10}{\underline{w_d}} x_d$$

$$\min_{\tilde{\omega}} \; J = \| Y - \tilde{D}\tilde{\omega} \|^2 + \alpha \| \tilde{\omega} \|^2 \qquad \left(\|\tilde{\omega}\|^2 = \tilde{\omega}^T\tilde{\omega} \right)$$

$\alpha \equiv$ regularization constant

$\alpha \geq 0$

$$\frac{\partial J}{\partial \tilde{\omega}} = -2\tilde{D}^T Y + 2\tilde{D}^T\tilde{D}\tilde{\omega} + 2\alpha\tilde{\omega} = 0$$

$$\Rightarrow \quad \tilde{D}^T\tilde{D}\tilde{\omega} + \alpha\tilde{\omega} = \tilde{D}^T Y$$

$$\left(\tilde{D}^T\tilde{D} + \alpha I\right)\tilde{\omega} = \tilde{D}^T Y$$

$$\tilde{\omega} = \underbrace{\left(\tilde{D}^T\tilde{D} + \alpha I\right)^{-1}}_{\text{ridge matrix}} \tilde{D}^T Y$$

Exam I

In person / in class

In person / in class
paper - pencil
Open notes

<u>Chapter 1</u>: dot product
norms
projections
prob concepts

<u>Chapter 2</u>

Mean
Variance
Covariance matrix
eigenvalues / eigenvectors | power iteration
Col / point view
geometric interpretation of μ, Σ, σ, ρ (correlation)
total variance & relation to σ_i^2 & change of basis & eigenvalues

Normal Distribution
$\rightarrow$ Σ, Σ^{-1}, basis vectors, PCA

<u>Chapter 7</u>: PCA (projected variance / error)
7.1 & 7.2
max variance $\equiv$ min MSE
eigen decomposition
selecting best subspace

$\boxed{\Sigma u = \lambda u}$

(no kernels, SVD)

Chapter 20 : LDA (projected means & variances)

Sec 20.1

optimization objective $\frac{(\mu_1 - \mu_2)^2}{s_1^2 + s_2^2}$

$\Rightarrow \boxed{Bw = \lambda Sw} \rightarrow$ eq. 20.10

$\vec{w}$: what it means.

$\Downarrow$

$\vec{w} = S^{-1}(\mu_1 - \mu_2)$

Chapter 6 : high dimensional data

hyper - cubes, spheres, planes

Q! bondary, corners, diagonals, normal distribution

$d = 2$

$d = 3$

finally $d \to \infty$?

$\lim d \to \infty$

Chapter 23 : linear regression

23.1, 23.2, 23.3, 23.4

bivariate
multiple

QR ← decomposition

$\vec{w}$: backsolve

adding ridge

SGD : gradient

$\vec{w}_0 \triangleq$ initial vector

SGD: gradient

↳ mean

↳ $\vec{\omega} \leftarrow$ via SGD

PCA LDA Linear regression

$\hat{\omega}_0 \triangleq$ initial vector

∇_ω, η

$\hat{\omega}_t = \hat{\omega}_{t-1} - \eta \cdot \nabla_{\omega_{t-1}}$

Logistic Regression

→ binary logistic regression

$$
\begin{array}{c|c|c}
 & X_1 \, X_2 \cdots X_d & Y \\
n & \vec{x}_i^T \quad D & y_i
\end{array}
$$

$\underbrace{\quad\quad\quad}_{d}$

$y_i \in \{P, N\}$

binary 'class'

$\underbrace{\boxed{1 \quad 0}}_{\text{encoding}}$

$+1 \quad -1$

$$\boxed{\begin{gathered} D = \{(\vec{x}_i^T, y_i)\}_{i=1}^n \\ y_i \in \{0, 1\} \end{gathered}}$$

Q: Should we be using numeric predictions?

A: Not really

$y_i = 1$, Prediction $\hat{y}_i = 1000$

$$(y_i - \hat{y}_i)^2 = (999)^2$$

---

2 classes

$P(1 \mid \vec{x}_i)$ vs $P(0 \mid \vec{x}_i)$

given point/instance $\vec{x}_i$

Task: predict the probability of each class given $\vec{x}_i$

$$\boxed{P(1 \mid \vec{x}_i) + P(0 \mid \vec{x}_i) = 1}$$

$$P(0 \mid \vec{x}_i) = 1 - \underbrace{P(1 \mid \vec{x}_i)}$$

given $\vec{x}_i$, predict $P(1 \mid \vec{x}_i) \equiv \pi(x_i)$

prob of positive class

unknown!

$$\underbrace{P(1 \mid \vec{x}_i) = \pi(x_i)}_{\text{prob } [0,1]} \cancel{\equiv} \underbrace{w_0 + w_1 x_{i1} + w_2 x_{i2} + \dots + w_d x_{id}}_{\hat{w}^T \vec{x}_i}$$

$[0,1]$

$[-\infty, \infty]$

$\Downarrow \; \theta \leftarrow$ some function (link)

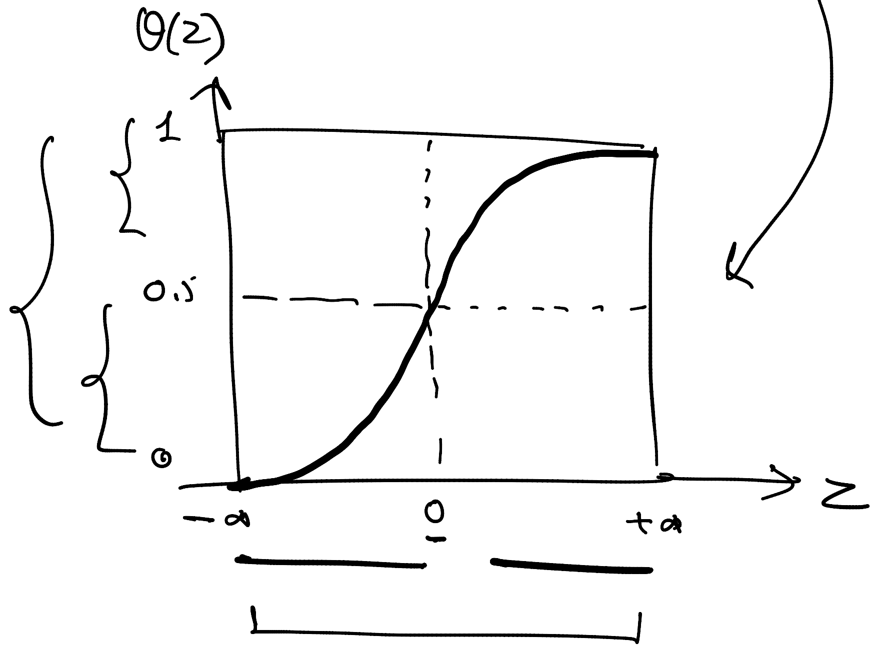
$\theta: [-\infty, \infty] \Rightarrow [0,1]$

$[0, 1]$

$\theta \equiv$ logistic / sigmoid

$$\theta(z) = \frac{1}{1+e^{-z}}$$

scalar



$$\pi(\vec{x}_i) = \theta\left(\boxed{\tilde{w}\,\vec{x}_i}\right) \quad \forall \; i = 1 \ldots n$$

n-equations

$\vec{x}_i$ is one of the input points

parameters: $\tilde{w}$ unknown

non-linear model

Q: how to learn $\tilde{w}$?

A: maximum likelihood estimation (MLE)

$$\max_{\tilde{w}} \; L(Y \mid \tilde{w}) \equiv \text{likelihood of the data}$$

Choose $\tilde{\omega}$ that maximizes the probability of the data that is observed

$$L(Y|\tilde{\omega}) \equiv P(Y|\tilde{\omega})$$

Prob of observed data given $\tilde{\omega}$

$Y \in \mathbb{R}^n$ n-dim joint probability

$$\equiv P((y_1, y_2 \dots y_n)|\tilde{\omega}) \; ?$$

lets make some assumptions!

Assume all points are independent

$$\equiv P(y_1|\tilde{\omega}) \times P(y_2|\tilde{\omega}) \times \dots \times P(y_n|\tilde{\omega})$$

$$\boxed{L(\tilde{\omega}) = \prod_{i=1}^{n} P(y_i|\tilde{\omega})}$$

$$P(y_i = 1|\tilde{\omega}) = \theta(\tilde{\omega}^T \tilde{x}_i)$$

Case :

$$P(y_i|\tilde{\omega}) = \begin{cases} \theta(\tilde{\omega}^T x_i) & \text{if } y_i = 1 \\ 1 - \theta(\tilde{\omega}^T x_i) & \text{if } y_i = 0 \end{cases}$$

$\vec{x}_i$

$$\boxed{f(y_i \mid \tilde{w}) = \theta(\tilde{w}^T x_i)^{y_i} \cdot \left(1 - \theta(\tilde{w}^T x_i)\right)^{1-y_i}}$$

$y_i = 0$

$y_i = 1$

$$\max_{\tilde{w}} \quad \boxed{L(\tilde{w}) = \prod_{i=1}^{n} \theta(\tilde{w}^T x_i)^{y_i} \cdot \left(1 - \theta(\tilde{w}^T x_i)\right)^{1-y_i}}$$

$$\boxed{\frac{\partial L(\tilde{w})}{\partial \tilde{w}} = 0}$$

Solve!