ML Erasmus+ course on Machine Learning. Oktober 2, 2023

Material for exercise 1

$$y = Ax$$

$$y \in \mathbb{R}^m \qquad x \in \mathbb{R}^n$$

$$A \in \mathbb{R}^{m \times n}$$

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ \vdots & & & \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}$$

$$\frac{\partial y}{\partial x}$$

write out element by element

$$y_i = \sum_{k=1}^{n} a_{ik} x_k$$

$$\frac{\partial y_i}{\partial x_j} = a_{ij} \quad \text{for all} \quad \begin{matrix} i = 1, 2, \dots m \\ j = 1, 2, \dots n \end{matrix}$$

Jacobian

$$\frac{\partial y}{\partial x} = \begin{bmatrix} \frac{\partial y_1}{\partial x_1} & \frac{\partial y_1}{\partial x_2} & \cdots & \frac{\partial y_1}{\partial x_n} \\ \vdots & & & \\ \frac{\partial y_m}{\partial x_1} & \cdots & & \frac{\partial y_m}{\partial x_n} \end{bmatrix}$$

$$\Rightarrow \frac{\partial y}{\partial x} = A$$

Define $y = Ax$

A does not depend on x

x may depend on z

$$\frac{\partial y}{\partial z} = A \frac{\partial x}{\partial z}$$

$$y_i = \sum_{k=1}^{n} a_{ik} x_k$$

$$\frac{\partial y_i}{\partial z_j} = \sum_{k=1}^{n} a_{ik} \frac{\partial x_k}{\partial z_j}$$

element ij of $A \cdot \frac{\partial x}{\partial z}$

$$\frac{\partial y}{\partial z} = \frac{\partial y}{\partial x}\frac{\partial x}{\partial z} = A\frac{\partial x}{\partial z}$$

small digression:

$$y = \underbrace{Ax}\ [\ \vec{y} = X\vec{\beta}\ ]$$

Design matrix

X does not depend on β

Define a scalar α $[C(\beta)]$ is also a number

$$\alpha = y^T A x$$

A is independent of y and x

$y \in \mathbb{R}^m \quad x \in \mathbb{R}^n \quad A \in \mathbb{R}^{m \times n}$

$$\frac{\partial \alpha}{\partial y} = x^T A^T \quad , \quad \frac{\partial \alpha}{\partial x} = y^T A$$

Proof: define $w^T = y^T A$

$$\frac{\partial \alpha}{\partial x} = w^T = y^T A$$

$\alpha = \alpha^T$ (α is a scalar)

$$\alpha = y^T A x = \alpha^T = \underbrace{x^T A^T}_{w^T} y$$

$$\frac{\partial \alpha}{\partial y} = x^T A^T$$

$$\alpha = x^T A x$$

$$A \in \mathbb{R}^{n \times n}$$

$$\frac{\partial \alpha}{\partial x} = x^T (A + A^T)$$

proof : write out components

$$\alpha = \sum_{j=1}^{n} \sum_{i=1}^{n} a_{ij} x_i x_j$$

if A is symmetric, $A^T = A$

$$\frac{\partial \alpha}{\partial x} = 2 x^T A$$

MSE :

$$C(\beta) = \frac{1}{n} \sum_{i=1}^{n} (y_i - \tilde{y}_i)^2$$

$$\tilde{y}_i = \sum_{j=1}^{p} x_{ij} \beta_j$$

$$\tilde{y} = X\beta \qquad \beta^T = [\beta_1 \; \dots \; \beta_p]$$

$$X = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ \vdots & & & \\ x_{n1} & \dots & & x_{np} \end{bmatrix}$$

$$X \in \mathbb{R}^{n \times p}$$

n = # data samples

$$C(\beta) = \frac{1}{n}\underbrace{(y - X\beta)^T}_{w^T}\underbrace{(y - X\beta)}_{w}$$

$$= \frac{1}{n} w^T w = \alpha$$

$$\frac{\partial C}{\partial \beta} = \frac{\partial \alpha}{\partial \beta}$$

$$\frac{\partial C}{\partial \beta} = \frac{2}{n} w^T \frac{\partial w}{\partial \beta}$$

$$\frac{\partial w}{\partial \beta} = -X$$

$$= -\frac{2}{n}(y - X\beta)^T X$$

$$\frac{\partial C}{\partial \beta^T} = 0 = -\frac{2}{n} X^T (y - X\beta)$$

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

$\hat{\beta}$: optimal β

$$\hat{\beta} = \underset{\beta \in \mathbb{R}^p}{\arg\min}\ C(\beta)$$

$$\frac{\partial^2 C(\beta)}{\partial \beta \, \partial \beta^T} = \frac{2}{n} X^T X$$

Hessian matrix

1) $X^T X$ has only positive ($\lambda_i > 0$) eigenvalues → convex optimization

2) Plays important in gradient descent methods

3) $X^T X \cdot 1/n$ covariance of X?

$$f(x) \approx \tilde{y}(x) = X\beta$$

$$\tilde{y}_i = \tilde{y}(x_i) = \sum_{j=0}^{p-1} \beta_j x_i^j$$

$$= \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \cdots + \beta_{p-1} x_i^{p-1}$$

$$\tilde{y}^T = [\tilde{y}_0 \; \tilde{y}_1 \; \cdots \; \tilde{y}_{n-1}]$$

$$\tilde{y}_0 = \beta_0 + \beta_1 x_0 + \beta_2 x_0^2 + \cdots + \beta_{p-1} x_0^{p-1}$$

$$\tilde{y}_1 = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \cdots + \cdots,$$

$$\vdots$$

$$\tilde{y}_{n-1} = \beta_0 + \beta_1 x_{n-1} + \beta_2 x_{n-1}^2 + \cdots,$$

$$\Rightarrow \hat{y} = X\beta$$

$$X \in \mathbb{R}^{n \times p}$$

$$X = \begin{bmatrix} 1 & x_0 & x_0^2 & \cdots & x_0^{p-1} \\ 1 & x_1 & x_1^2 & \cdots & x_1^{p-1} \\ \vdots & & & & \vdots \\ 1 & & & & x_{n-1}^{p-1} \end{bmatrix}$$

$$= \begin{bmatrix} x_{00} & x_{01} & \cdots & x_{0\,p-1} \\ x_{10} & x_{11} & \cdots & x_{1\,p-1} \\ \vdots & & & \\ x_{n-1\,0} & \cdots & & x_{n-1\,p-1} \end{bmatrix} =$$

x_1

$$= \left[\underline{x}_0 \; \underline{x}_1 \; \cdots \; \underline{x}_{p-1} \right]$$

matrix with p-column vectors

we have p-features/predictors

we have n-data entries

$$p << n.$$

$$\hat{\beta} = \underbrace{(X^T X)^{-1}}\, X^T y$$

$$\underbrace{X^T X}_{\mathbb{R}^{p \times p}} + \lambda \underbrace{I}_{\mathbb{R}^{p \times p}} \qquad I = \begin{bmatrix} 1 & & 0 \\ & \ddots & \\ 0 & & 1 \end{bmatrix}$$

$$\hat{\beta} = (X^T X + \lambda I)^{-1} X^T y$$

Ridge regression

$$C(\beta) = \frac{1}{n} \sum_{i=0}^{n-1} (y_i - \hat{y}_i)^2 + \lambda \sum_{j=0}^{p-1} \beta_j^2$$

SVD :

$$X = U \Sigma V^T$$

$$X \in \mathbb{R}^{n \times p}$$

$$U \in \mathbb{R}^{n \times n} \quad U^T U = U U^T = \mathbb{1}$$

$$V \in \mathbb{R}^{p \times p} \quad V^T V = V V^T = \mathbb{1}$$

$$\Sigma \in \mathbb{R}^{n \times p}$$

$$\Sigma = \begin{bmatrix} \sigma_0 & & & \\ & \sigma_1 & & \\ & & \ddots & \\ & & & \sigma_{p-1} \\ 0 & & & \end{bmatrix} \quad \begin{matrix} \sigma_0 > \sigma_1 > \\ \cdots > \sigma_{p-1} \\ > 0 \end{matrix}$$