ECE M146 Discussion 2

Introduction to Machine Learning

Instructor: Lara Dolecek

TA: Zehui (Alex) Chen

1. Matrix calculus review

(a) Gradient of differentiable function $f : \mathbb{R}^n \to \mathbb{R}$:

$\nabla f(x) = \left[\frac{\partial}{\partial x_1} f(x), \frac{\partial}{\partial x_2} f(x), \cdots, \frac{\partial}{\partial x_n} f(x)\right]^T.$

- $\nabla_w(w^T b)$

$$\frac{\partial w^T b}{\partial w_i} = \frac{\partial \sum_j w_j b_j}{\partial w_i} = b_i \qquad w_i b_i$$

$$\nabla_w (w^T b) = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix} = b$$

- $\nabla_w(\|w\|^2)$

$$\frac{\partial \|w\|^2}{\partial w_i} = \frac{\partial \sum_j w_j^2}{\partial w_i} = 2w_i \qquad w_i^2$$

$$\nabla_w \|w\|^2 = \begin{bmatrix} 2w_1 \\ 2w_2 \\ \vdots \\ 2w_n \end{bmatrix} = 2w$$

- $\nabla_w(w^T A w)$

$$[w_1, w_2 \ \dots \ w_n] \begin{bmatrix} A_{11} & A_{12} & \cdot \\ A_{21} & & \vdots \\ \vdots & \cdot & \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_n \end{bmatrix}$$

$$\frac{\partial w^T A w}{\partial w_i} = \frac{\partial \sum_j \sum_k w_j A_{jk} w_k}{\partial w_i}$$

$j=k=i$: $\underline{w_i^2 A_{ii}}$ $\quad$ $j\neq i$ $k \neq i$: $w_i \sum_{k \neq i} A_{ik} w_k$ $\quad$ $k=i$ $j \neq i$: $w_i \sum_{j \neq i} w_j A_{ji}$

$$= \frac{\partial A_{ii} w_i^2}{\partial w_i} + \frac{\partial w_i \sum_{k \neq i} A_{ik} w_k}{\partial w_i} + \frac{\partial w_i \sum_{j \neq i} A_{ji} w_j}{\partial w_i}$$

$$= \underline{2 w_i A_{ii}} + A(i, :_{\neq i}) w_{\neq w_i} + A(:_{\neq i}, i)^T w_{\neq i}$$

i-th row of A excluding A_{ii}

i-th column of A excluding A_{ii}

$$= A(i,:) w + A(:,i)^T w$$

- $\nabla_w(w^T X^T X w)$ ✓

$$\nabla_w w^T X^T X w$$

$$= X^T X w + \underbrace{(X^T X)^T}_{X^T X} w$$

$$= 2 X^T X w$$

$$\nabla_w (w^T A w) = \begin{bmatrix} A(1,:) w + A(:,1)^T w \\ A(2,:) w + A(:,2)^T w \\ \vdots \\ \vdots \end{bmatrix}$$

A $\quad$ A^T

$$= \underline{Aw + A^T w}$$

(b) Jacobian/derivative matrix of differentiable function $f : \mathbb{R}^n \to \mathbb{R}^m$:

$f_1, \dots, f_m : R^n \to R$

$$f(x) \begin{bmatrix} f_1(x) \\ f_2(x) \\ \vdots \\ \vdots \end{bmatrix} : R^n \to R^m$$

$$\boldsymbol{J} = \underbrace{\begin{bmatrix} \nabla f_1(x)^T \\ \nabla f_2(x)^T \\ \vdots \\ \nabla f_m(x)^T \end{bmatrix}}_{m \times n}, \boldsymbol{J}_{ij} = \frac{\partial f_i}{\partial x_j}$$

$$\underbrace{\Delta f(x)}_{m \times 1} \doteq \underbrace{J}_{m \times n} \underbrace{\Delta x}_{n \times 1}$$

$$A = \begin{bmatrix} a_1^T \\ a_2^T \\ \vdots \end{bmatrix} \qquad Ax = \begin{bmatrix} a_1^T x \\ a_2^T x \\ \vdots \end{bmatrix} \qquad A = \begin{bmatrix} A_{11} & A_{12} & A_{13} & \cdots \\ A_{21} & \cdots & & \end{bmatrix} \qquad a_1 = \begin{bmatrix} A_{11} \\ A_{12} \\ A_{13} \\ \vdots \end{bmatrix}$$

- Ax

$$J = \begin{bmatrix} \nabla f_1(x)^T \\ \nabla f_2(x)^T \\ \vdots \end{bmatrix} = \begin{bmatrix} a_1^T \\ a_2^T \\ \vdots \\ a_m^T \end{bmatrix} = A$$

- Example: transformation from polar (r, θ) to Cartesian coordinates (x, y): $x = \underset{f_1}{r \cos(\theta)}, y = \underset{f_2}{r \sin(\theta)}$.

$$\begin{bmatrix} \Delta x \\ \Delta y \end{bmatrix} = \underbrace{\begin{bmatrix} \frac{\partial x}{\partial r} & \frac{\partial x}{\partial \theta} \\ \frac{\partial y}{\partial r} & \frac{\partial y}{\partial \theta} \end{bmatrix}}_{J} \begin{bmatrix} \Delta r \\ \Delta \theta \end{bmatrix} \qquad J = \begin{bmatrix} \cos\theta, & -r\sin\theta \\ \sin\theta, & r\cos\theta \end{bmatrix}$$

(c) Hessian matrix for twice differentiable function $f : \mathbb{R}^n \to \mathbb{R}$:
$\nabla^2 f(x)_{ij} = \frac{\partial^2}{\partial x_i \partial x_j} f(x)$.
The Hessian matrix is also the derivative matrix $\boldsymbol{J}$ of the gradient $\nabla f(x)$.

- Affine function $f(x) = a^T x + b$.

$$\nabla f(x) = a$$

$$\nabla_x^2 f(x) = 0_{n \times n}$$

$$x \in R^n$$

- Least squares cost: $\|Ax - b\|^2$.

$$\nabla f(x) = \underline{2A^TAx - 2A^Tb}$$

$$\nabla_x^2 f(x) = 2A^TA$$

- Example: $4x_1^2 + 4x_1x_2 + x_2^2 + 10x_1 + 9x_2$

$$\nabla f(x) = \begin{bmatrix} \frac{\partial f(x)}{\partial x_1} \\ \frac{\partial f(x)}{\partial x_2} \end{bmatrix} = \begin{bmatrix} 8x_1 + 4x_2 + 10 \\ 4x_1 + 2x_2 + 9 \end{bmatrix}$$

$$\nabla^2 f(x) = \begin{bmatrix} \frac{\partial^2 f(x)}{\partial x_1 \partial x_1} & \frac{\partial^2 f(x)}{\partial x_1 \partial x_2} \\ \frac{\partial^2 f(x)}{\partial x_2 \partial x_1} & \frac{\partial^2 f(x)}{\partial x_2 \partial x_2} \end{bmatrix} = \begin{bmatrix} 8 & 4 \\ 4 & 2 \end{bmatrix}$$

2. We now try to provide a probabilistic interpretation of the linear regression problem. Consider a model where each of the N samples is independently drawn according to a normal distribution

pdf

$$\boxed{P(y_n|x_n, w) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_n - \boxed{w^T x_n})^2}{2\boxed{\sigma^2}}\right).} \sim N(w^T x_n, \sigma^2)$$

In this model, each y_n is drawn from a normal distribution with mean $w^T x_n$ and variance σ^2. The σ are **known**. Write the log likelihood of this model as a function of w. Show that finding the maximum likelihood estimate of w leads to the same answer as solving a linear regression problem.

LS Problem : $\boxed{\underset{w}{\arg\min} \sum_{i=1}^{N} (y_i - w^T x_i)^2}$

Maximum Likelihood Estimation for w. give observation $\{x_1\ y_1\}\ \{x_2\ y_2\}$ - - - -

$$\underset{w}{\arg\max}\ P(y_1, \dots y_N | x_1, \dots x_N ; w)$$ MLE of w

$$= \underset{w}{\arg\max} \prod_{i=1}^{N} P(y_i | x_i ; y)$$

$$= \underset{w}{\arg\max} \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - w^T x_i)^2}{2\sigma^2}\right)$$

$$= \underset{w}{\arg\max} \sum_{i=1}^{N} -\frac{(y_i - w^T x_i)^2}{\cancel{2\sigma^2}}\ \cancel{+\ \text{constant}}$$

$$= \boxed{\underset{w}{\arg\min} \sum_{i=1}^{N} (y_i - w^T x_i)^2}$$

3. We now try to provide a probabilistic interpretation of the weighted linear regression. Consider a model where each of the N samples is independently drawn according to a normal distribution

$$P(y_n|x_n, w) = \frac{1}{\sqrt{2\pi\boxed{\sigma_n^2}}} \exp\left(-\frac{(y_n - \boxed{w^T x_n})^2}{2\boxed{\sigma_n^2}}\right). \sim N(w^T x_n, \sigma_n^2)$$

In this model, each y_n is drawn from a normal distribution with mean $w^T x_n$ and variance σ_n^2. The σ_n^2 are **known**. Write the log likelihood of this model as a function of w. Show that finding the maximum likelihood estimate of w leads to the same answer as solving a weighted linear regression. How do σ_n^2 relate to α_n?

Weighted LS Problem

$$\boxed{\underset{w}{\arg\min} \sum_{i=1}^{N} \alpha_i (y_i - w^T x_i)^2}$$

$$\underset{w}{\arg\max}\ P(y_1, \dots y_N | x_1, \dots x_N; w)$$

$$= \underset{w}{\arg\max} \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left(-\frac{(y_i - w^T x_i)^2}{2\sigma_i^2}\right)$$

$$= \underset{w}{\arg\max} \sum_{i=1}^{N} - \frac{(y_i - w^T x_i)^2}{2\sigma_i^2} + \cancel{\text{constant}}$$

$$= \boxed{\underset{w}{\arg\min} \sum_{i=1}^{N} \frac{1}{2\sigma_i^2} (y_i - w^T x_i)^2}$$

$$\alpha_i = \frac{1}{2\sigma_i^2}$$