EE M146 Discussion 2
Introduction to Machine Learning
Instructor: Lara Dolecek
TA: Zehui (Alex) Chen, Ruiyi (John) Wu

1. Matrix calculus review

   (a) Gradient of differentiable function $f : \mathbb{R}^n \to \mathbb{R}$:
   $\nabla f(x) = \left[\frac{\partial}{\partial x_1} f(x), \frac{\partial}{\partial x_2} f(x), \cdots, \frac{\partial}{\partial x_n} f(x)\right]^T$.

   - $\nabla_w(w^T b)$

     $$\frac{\partial w^T b}{\partial w_i} = \frac{\partial \sum_j w_j b_j}{\partial w_i} = b_i \qquad \nabla_w(w^T b) = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix} = b$$

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

     $$\frac{\partial \|w\|^2}{\partial w_i} = \frac{\partial\, w_1^2 + w_i^2 + \dots w_n^2}{\partial w_i} = 2w_i \qquad \nabla_w \|w\|^2 = 2w$$

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

     $1\times n \quad n\times n \quad n\times 1$

     $$[w_1 \dots w_n]\begin{bmatrix} A_{11} & \dots & \\ & & \\ & & \end{bmatrix}\begin{bmatrix} w_1 \\ \vdots \\ w_n \end{bmatrix}$$

     $j=i \qquad k=i \qquad f\ 0\dots$

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

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

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

     $A \qquad A^T$

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

     $A = X^T X$

     $$= X^T X w + (X^T X)^T w = 2X^T X w$$

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

     $f: \ f_1, \dots, f_m : \mathbb{R}^n \to \mathbb{R}$

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

   - Ax

     $m\times n \quad n\times 1$

     $$A = \begin{bmatrix} a_1^T \\ \vdots \\ a_m^T \end{bmatrix} \qquad J_{Ax} \begin{bmatrix} \nabla a_1^T x^T \\ \vdots \\ \nabla a_m^T x^T \end{bmatrix} = \begin{bmatrix} a_1^T \\ \vdots \\ a_m^T \end{bmatrix} = A$$

     $$\partial Ax = J \partial x$$

     $m\times 1 \quad m\times n \quad n\times 1$

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

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

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

$$\nabla f(x) = a \qquad \nabla_x^2 f(x) = 0$$

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

$$\nabla f(x) = 2A^TAx - 2A^Tb \qquad \nabla^2 f(x) = 2A^TA$$

$$f(x_1, x_2) \quad (Ax-b)^T(Ax-b) = x^TA^TAx - 2A^Tbx + b^Tb$$

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

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

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

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

w ?

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 ~~Loss~~ Problem : $\arg\min_w \sum_{i=1}^{N} (y_i - w^T x_i)^2$

Maximum Likelihood ~~Estimate~~ Estimate of w, give our observation $(x_1, \dots, x_N)\ (y_1 \dots y_N)$

$$\arg\max_w \underline{P(y_1, \dots, y_N \mid x_1, \dots, x_N ; w)}$$

$$= \prod_{i=1}^{N} P(y_i \mid x_i ; w)$$

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

$$\Downarrow$$

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

$$\Downarrow$$

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

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

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

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

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

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

$$\Downarrow$$

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

$$\Downarrow$$

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

$$\Downarrow$$

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

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