Leren 2020 sample exercises week1

1 Partial Derivatives

What does it mean to build the partial derivative of a multivariate function?
Direction of steepest descent in one direction

Take the the partial derivative w.r.t. x, y, and z of

$$f(x, y, z) = 2\ln(y - \exp(x^{-1}) - \sin(zx^2))$$

$$\frac{\partial f(x,y,z)}{\partial x} = \frac{2}{(y - \exp(x^{-1}) - sin(zx^2))}(-\exp(x^{-1})(-x^{-2}) - \cos(zx^2)(2zx))$$

$$\frac{\partial f(x,y,z)}{\partial x} = \frac{2(\exp(1/x)/x^2 - 2zx\cos(zx^2))}{(y - \exp(x^{-1}) - sin(zx^2))}$$

$$\frac{\partial f(x,y,z)}{\partial y} = \frac{2}{(y - \exp(x^{-1}) - \sin(zx^2))}$$

$$\frac{\partial f(x,y,z)}{\partial z} = \frac{2}{(y - \exp(x^{-1}) - \sin(zx^2))}(-cos(zx^2)x^2)$$

$$\frac{\partial f(x,y,z)}{\partial z} = \frac{-2x^2cos(zx^2)}{(y - \exp(x^{-1}) - \sin(zx^2))}$$

2 Normal Distribution

Suppose we want a probability density function for the length of all adults alive today. We get a sample of lengths of N adults as measuring everyone is impractical. We make sure the sample is independent and identically distributed (iid), so for example we don't over-represent tall Dutch adults. This gives us our data:

$$X = \{x^t\}_{t=1}^N \tag{1}$$

Suppose we assume the length of adults is normally distributed (the Gaussian distribution, the famous bell curve). Note that this is merely an assumption and may not be the best choice (for example when there happens to be a large group of short adults, a large group of tall adult, but not so many medium length adults). This gives us our density function:

$$\mathcal{N}(\mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] \tag{2}$$

with sufficient statistics:

$$\mu \equiv E(X), \text{ the expected value of X (mean of X)} \tag{3}$$
$$\sigma^2 \equiv Var(X), \text{ the variance of X} \tag{4}$$

Because we haven't measured everyone and just use a sample we can only estimate μ and σ by respectively m and s. You probably remember you can simply compute these as:

$$m = \frac{\sum_{t=1}^N x^t}{N} \tag{5}$$
$$s^2 = \frac{\sum_{t=1}^N (x^t - m)^2}{N} \tag{6}$$

These are given in paragraph 4.3.2 in equation 4.8 in "Introduction to Machine Learning 3rd ed, Ethem Alpaydim". Now show that these are indeed the correct Maximum Likelihood Estimations.

2.1 Maximum Likelihood Estimation (MLE)

With Maximum Likelihood Estimation we want to find the parameter values that give the highest likelihood given the data. Often we can use the following steps for this:

1. Write down the likelihood function that we want to maximize.

$$l(m, s|X) = P(X|m, s) = \prod_{t=1}^{N} \frac{1}{\sqrt{2\pi}s} \exp\left[-\frac{(x^t - m)^2}{2s^2}\right] \tag{7}$$

2. Take the log() of this function when this makes it easier to take the derivatives w.r.t. the parameters (and it does in this case!). This is valid because the log() function is monotonically increasing and therefore the x that maximizes $\log(f(x))$ will also maximize $f(x)$.

$$\begin{aligned}
\mathcal{L}(m, s|X) &= \log(l(m, s|X)) && (8)\\
&= \log\left(\prod_{t=1}^{N} \frac{1}{\sqrt{2\pi}s} \exp\left[-\frac{(x^t - m)^2}{2s^2}\right]\right) && (9)
\end{aligned}$$

using: $\log(ab) = \log(a) + \log(b)$

$$= \sum_{t=1}^{N} \log\left(\frac{1}{\sqrt{2\pi}s} \exp\left[-\frac{(x^t - m)^2}{2s^2}\right]\right) \tag{10}$$

using: $\log(ab) = \log(a) + \log(b)$

$$= \sum_{t=1}^{N} \log\left(\frac{1}{\sqrt{2\pi}s}\right) + \sum_{t=1}^{N} \log\left(\exp\left[-\frac{(x^t - m)^2}{2s^2}\right]\right) \tag{11}$$

using: $\log(\exp(a)) = a$

$$= N \log\left(\frac{1}{\sqrt{2\pi}s}\right) + \sum_{t=1}^{N} \left[-\frac{(x^t - m)^2}{2s^2}\right] \tag{12}$$

using: $\log(ab) = \log(a) + \log(b)$

$$= N \log\left(\frac{1}{\sqrt{2\pi}}\right) + N \log\left(\frac{1}{s}\right) - \frac{1}{2s^2} \sum_{t=1}^{N} (x^t - m)^2 \tag{13}$$

using: $\log(a^b) = b \log(a)$

$$= -N \log(\sqrt{2\pi}) - N \log(s) - \frac{1}{2s^2} \sum_{t=1}^{N} (x^t - m)^2 \tag{14}$$

3. Take the partial derivative with respect to each parameter.

$$\frac{\partial \mathcal{L}(m,s|X)}{\partial m} = \frac{\partial - N\log(\sqrt{2\pi}) - N\log(s) - \frac{1}{2s^2}\sum_{t=1}^{N}(x^t - m)^2}{\partial m} \tag{15}$$

$$= \frac{\frac{1}{2s^2}\sum_{t=1}^{N}\partial\left[(x^t - m)^2\right]}{\partial m} \tag{16}$$

using: chain rule

$$= \frac{\frac{1}{2s^2}\sum_{t=1}^{N}\partial\left[(x^t - m)^2\right]}{\partial(x^t - m)}\frac{\partial(x^t - m)}{\partial m} \tag{17}$$

$$= -\frac{1}{s^2}\sum_{t=1}^{N}\left[(x^t - m)\right] \tag{18}$$

$$\frac{\partial \mathcal{L}(m,s|X)}{\partial s} = \frac{\partial - N\log(\sqrt{2\pi}) - N\log(s) - \frac{1}{2s^2}\sum_{t=1}^{N}(x^t - m)^2}{\partial s} \tag{19}$$

$$= \frac{\partial - N\log(s) - \frac{1}{2s^2}\sum_{t=1}^{N}(x^t - m)^2}{\partial s} \tag{20}$$

using: $\frac{\partial log(a)}{\partial a} = \frac{1}{a}$

$$= -\frac{N}{s} + \frac{1}{s^3}\sum_{t=1}^{N}(x^t - m)^2 \tag{21}$$

4. Set the derivatives to 0 and solve for each parameter. When the derivatives can be reduced to second degree (parabola) or lower polynomial functions w.r.t. the parameters this will give us the values that maximize the Likelihood function. Later we will see techniques to deal with more complex functions.

$$-\frac{1}{s^2}\sum_{t=1}^{N}\left[(x^t - m)\right] = 0 \tag{22}$$

assuming: $-\frac{1}{s^2} \neq 0$

$$\sum_{t=1}^{N}[x^t] - Nm = 0 \tag{23}$$

$$m = \frac{\sum_{t=1}^{N} x^t}{N} \tag{24}$$

$$-\frac{N}{s}+\frac{1}{s^3}\sum_{t=1}^{N}(x^t-m)^2 =0 \tag{25}$$

$$-\frac{Ns^2+\sum_{t=1}^{N}(x^t-m)^2}{s^3} =0 \tag{26}$$

assuming: $s^3 \neq 0$

$$s^2 =\frac{\sum_{t=1}^{N}(x^t-m)^2}{N} \tag{27}$$

3 Vector Derivatives and Gradients

In this problem, we will compute basic derivatives of scalars and vectors with respect to vectors. Although these can look simple on the surface, the details and conventions are worth explicitly restating since the results can be somewhat counter-intuitive and are fundamental to modern machine learning (and your homework).

First, given two column vectors $\boldsymbol{y} \in \mathbb{R}^m$ and $\boldsymbol{x} \in \mathbb{R}^n$, we can write them explicitly as:

$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix} = \begin{bmatrix} y_1 & y_2 & \dots & y_m \end{bmatrix}^T \qquad \& \qquad \boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} x_1 & x_2 & \dots & x_n \end{bmatrix}^T$$

We then (according to accepted convention) define the the derivative of the vector $\boldsymbol{y}$ with respect to the vector $\boldsymbol{x}$ to be given by the matrix:

$$\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{x}^T} = \begin{bmatrix} \frac{\partial \boldsymbol{y}}{\partial x_1} & \frac{\partial \boldsymbol{y}}{\partial x_2} & \dots & \frac{\partial \boldsymbol{y}}{\partial x_n} \end{bmatrix} = \begin{bmatrix} \frac{\partial y_1}{\partial x_1} & \frac{\partial y_1}{\partial x_2} & \dots & \frac{\partial y_1}{\partial x_n} \\ \frac{\partial y_2}{\partial x_1} & \frac{\partial y_2}{\partial x_2} & \dots & \frac{\partial y_2}{\partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial y_m}{\partial x_1} & \frac{\partial y_m}{\partial x_2} & \dots & \frac{\partial y_m}{\partial x_n} \end{bmatrix} \tag{28}$$

We note the choice of notation here $\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{x}^T}$ correctly implies that the derivative is with respect to the transpose of $\boldsymbol{x}$. Although many mathematics texts drop this transpose in notation (and instead simply use the notation $\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{x}}$), the vector derivative is always defined as the $m \times n$ matrix given above.

We can see this definition yields the desired result that the derivative of a vector with respect to itself is the identity matrix (as is very useful when using

the chain-rule):

$$\frac{\partial \boldsymbol{x}}{\partial \boldsymbol{x}^T} = \begin{bmatrix} \frac{\partial x_1}{\partial x_1} & \frac{\partial x_1}{\partial x_2} & \dots & \frac{\partial x_1}{\partial x_n} \\ \frac{\partial x_2}{\partial x_1} & \frac{\partial x_2}{\partial x_2} & \dots & \frac{\partial x_2}{\partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial x_n}{\partial x_1} & \frac{\partial x_n}{\partial x_2} & \dots & \frac{\partial x_n}{\partial x_n} \end{bmatrix} = \begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{bmatrix} = \boldsymbol{I} \tag{29}$$

We further note, *importantly*, that the gradient of a vector function of $f(\boldsymbol{x}) : \mathbb{R}^n \to \mathbb{R}^n$ with respect to a vector $\boldsymbol{x}$ is conventionally defined as the transpose of the vector derivative, i.e.:

$$\nabla_{\boldsymbol{x}} f(\boldsymbol{x}) = \begin{bmatrix} \frac{\partial f(\boldsymbol{x})}{\partial x_1} & \frac{\partial f(\boldsymbol{x})}{\partial x_2} & \dots & \frac{\partial f(\boldsymbol{x})}{\partial x_n} \end{bmatrix}^T = \left(\frac{\partial f(\boldsymbol{x})}{\partial \boldsymbol{x}^T} \right)^T \tag{30}$$

We see that given this more precise notation, the gradient of a scalar function $s(\boldsymbol{x}) : \mathbb{R}^n \to \mathbb{R}^1$ can then be written as the partial derivative with respect to the column vector $\boldsymbol{x}$ yielding a column vector as desired (i.e. $\nabla_{\boldsymbol{x}} s(\boldsymbol{x}) = \frac{\partial s(\boldsymbol{x})}{\partial \boldsymbol{x}}$)

Another important property is the chain rule for vector derivatives. For two functions defined over the same space f, g: $\mathbb{R}^n \to \mathbb{R}^n$, we can decompose the product as:

$$\frac{\partial f(\boldsymbol{x})^T g(\boldsymbol{x})}{\partial \boldsymbol{x}^T} = g(\boldsymbol{x})^T \frac{\partial f(\boldsymbol{x})}{\partial \boldsymbol{x}^T} + f(\boldsymbol{x})^T \frac{\partial g(\boldsymbol{x})}{\partial \boldsymbol{x}^T} \tag{31}$$

Or equivalently:

$$\nabla_{\boldsymbol{x}} \left[f(\boldsymbol{x})^T g(\boldsymbol{x}) \right] = \nabla_{\boldsymbol{x}} \left[f(\boldsymbol{x}) \right] g(\boldsymbol{x}) + \nabla_{\boldsymbol{x}} \left[g(\boldsymbol{x}) \right] f(\boldsymbol{x}) \tag{32}$$

4 Practice Problems

For use in the following problems, we define a matrix $\boldsymbol{W} \in \mathbb{R}^{m \times n}$ to be given by:

$$\boldsymbol{W} = \begin{bmatrix} w_{1,1} & w_{1,2} & \dots & w_{1,n} \\ w_{2,1} & w_{2,2} & \dots & w_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ w_{m,1} & w_{m,2} & \dots & w_{m,n} \end{bmatrix} \tag{33}$$

and let $\boldsymbol{w}_i = \begin{bmatrix} w_{i,1} & w_{i,2} & \dots & w_{i,n} \end{bmatrix}$ is the i'th row of $\boldsymbol{W}$.

1) Compute $\frac{\partial \boldsymbol{W}\boldsymbol{x}}{\partial \boldsymbol{x}^T}$ and $\nabla_{\boldsymbol{x}} \left[\boldsymbol{W}\boldsymbol{x} \right]$. (Note they are different!)

First, let $\boldsymbol{W}\boldsymbol{x} = \boldsymbol{y}$. Then observe, $\boldsymbol{y} = \begin{bmatrix} \boldsymbol{w}_1^T \boldsymbol{x} & \boldsymbol{w}_2^T \boldsymbol{x} & \dots & \boldsymbol{w}_m^T \boldsymbol{x} \end{bmatrix}^T$.

For a given y_i, we see $y_i = \boldsymbol{w}_i^T \boldsymbol{x} = \sum_{j=1}^{n} w_{i,j} x_j$, and thus it's partial derivative with respect to a given x_j is given by:

$$\frac{\partial y_i}{\partial x_j} = w_{i,j} \tag{34}$$

Therefore,

$$\frac{\partial \boldsymbol{W}\boldsymbol{x}}{\partial \boldsymbol{x}^T} = \begin{bmatrix} \frac{\partial y_1}{\partial x_1} & \frac{\partial y_1}{\partial x_2} & \cdots & \frac{\partial y_1}{\partial x_n} \\ \frac{\partial y_2}{\partial x_1} & \frac{\partial y_2}{\partial x_2} & \cdots & \frac{\partial y_2}{\partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial y_m}{\partial x_1} & \frac{\partial y_m}{\partial x_2} & \cdots & \frac{\partial y_m}{\partial x_n} \end{bmatrix} = \begin{bmatrix} w_{1,1} & w_{1,2} & \cdots & w_{1,n} \\ w_{2,1} & w_{2,2} & \cdots & w_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ w_{m,1} & w_{m,2} & \cdots & w_{m,n} \end{bmatrix} = \boldsymbol{W} \tag{35}$$

And thus applying equation 30, we get:

$$\nabla_{\boldsymbol{x}} \left[\boldsymbol{W}\boldsymbol{x}\right] = \left(\frac{\partial \boldsymbol{W}\boldsymbol{x}}{\partial \boldsymbol{x}^T}\right)^T = \boldsymbol{W}^T \tag{36}$$

2) For $\boldsymbol{x}, \boldsymbol{z} \in \mathbb{R}^n$ compute $\frac{\partial \boldsymbol{x}^T \boldsymbol{z}}{\partial \boldsymbol{x}^T}$ and $\frac{\partial \boldsymbol{z}^T \boldsymbol{x}}{\partial \boldsymbol{x}^T}$ and their corresponding gradients.

$$\begin{aligned} &\frac{\partial \boldsymbol{x}^T \boldsymbol{z}}{\partial \boldsymbol{x}^T} = \begin{bmatrix} \frac{\partial \boldsymbol{x}^T \boldsymbol{z}}{\partial x_1} & \frac{\partial \boldsymbol{x}^T \boldsymbol{z}}{\partial x_2} & \cdots & \frac{\partial \boldsymbol{x}^T \boldsymbol{z}}{\partial x_n} \end{bmatrix} = \begin{bmatrix} z_1 & z_2 & \cdots & z_n \end{bmatrix} = \boldsymbol{z}^T \\ &\therefore \ \nabla_{\boldsymbol{x}} \left[\boldsymbol{x}^T \boldsymbol{z}\right] = \boldsymbol{z} \end{aligned} \tag{37}$$

Similarly,

$$\begin{aligned} &\frac{\partial \boldsymbol{z}^T \boldsymbol{x}}{\partial \boldsymbol{x}^T} = \begin{bmatrix} \frac{\partial \boldsymbol{z}^T \boldsymbol{x}}{\partial x_1} & \frac{\partial \boldsymbol{z}^T \boldsymbol{x}}{\partial x_2} & \cdots & \frac{\partial \boldsymbol{z}^T \boldsymbol{x}}{\partial x_n} \end{bmatrix} = \begin{bmatrix} z_1 & z_2 & \cdots & z_n \end{bmatrix} = \boldsymbol{z}^T \\ &\therefore \ \nabla_{\boldsymbol{x}} \left[\boldsymbol{z}^T \boldsymbol{x}\right] = \boldsymbol{z} \end{aligned} \tag{38}$$

3) For $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ use the product rule to compute $\frac{\partial \boldsymbol{x}^T \boldsymbol{A}\boldsymbol{x}}{\partial \boldsymbol{x}^T}$ and $\nabla_{\boldsymbol{x}} \left[\boldsymbol{x}^T \boldsymbol{A}\boldsymbol{x}\right]$

Using the product rule for vector derivatives with $f(\boldsymbol{x}) = \boldsymbol{x}$ and $g(\boldsymbol{x}) = \boldsymbol{A}\boldsymbol{x}$ we get:

$$\begin{aligned} \frac{\partial \boldsymbol{x}^T \boldsymbol{A}\boldsymbol{x}}{\partial \boldsymbol{x}^T} &= g(\boldsymbol{x})^T \frac{\partial f(\boldsymbol{x})}{\partial \boldsymbol{x}^T} + f(\boldsymbol{x})^T \frac{\partial g(\boldsymbol{x})}{\partial \boldsymbol{x}^T} \\ &= \boldsymbol{x}^T \boldsymbol{A}^T \boldsymbol{I} + \boldsymbol{x}^T \boldsymbol{A} \\ &= \boldsymbol{x}^T \left(\boldsymbol{A}^T + \boldsymbol{A}\right) \end{aligned} \tag{39}$$

5 References

1. See The Matrix Cookbook for many useful identities but note they use a difference in convention for vector derivatives.

2. See ["On the concept of matrix derivative" by Jan R. Magnus] for more details on a generalized form of the matrix derivative and it's associated product and chain rules.