DDA3020 Homework 1

Due date: Oct 16, 2023

Instructions

- The **deadline** is 23:59, Oct 16, 2023.
- The weight of this assignment in the final grade is 15%.
- **Electronic submission**: Turn in solutions electronically via Blackboard. Be sure to submit your homework as a single file. Please name your solution file as $A1\_studentID\_name$
- Note that **late submissions** will result in discounted scores: 0-24 hours $\rightarrow$ 80%, 24-120 hours $\rightarrow$ 50%, 120 or more hours $\rightarrow$ 0%.
- Answer the questions in English. Otherwise, you'll lose half of the points.
- Collaboration policy: You need to solve all questions independently and collaboration between students is **NOT** allowed.
- If you have any questions concerning this homework, feel free to reach out to TA Dong QIAO (dongqiao@link.cuhk.edu.cn) or Wenrang ZHANG (223040237@link.cuhk.edu.cn). You're also welcome to physically visit them during their office hours with your questions.

1 Written Problems (50 points)

1.1. (Matrix Differential, 20 points) It is a fundamental ability to derive the derivatives of some functions in machine learning. Nine basic cases can be summarized as in Table 1. Nevertheless, only three of them (*i.e.*, cells in gray background) are common in our course. In this problem, you will follow the next steps to derive the derivatives for these cases.

(1) **Layout of Derivatives**

- **Differentiation of a scalar function *w.r.t.* a vector**: If $f : \mathbb{R}^n \mapsto \mathbb{R}$ is a scalar function of n variables, $\boldsymbol{w} \in \mathbb{R}^n$ is a $n \times 1$ vector, then differentiation of $f(\boldsymbol{w})$ *w.r.t.* $\boldsymbol{w}$

Table 1: Vector/Matrix derivatives

	f	$\boldsymbol{f}$	$\boldsymbol{F}$
w	$\frac{f}{w}$	$\frac{d\boldsymbol{f}}{dw}$	$\frac{d\boldsymbol{F}}{dw}$
$\boldsymbol{w}$	$\frac{df}{d\boldsymbol{w}}$	$\frac{d\boldsymbol{f}}{d\boldsymbol{w}}$	$\frac{d\boldsymbol{F}}{d\boldsymbol{w}}$
$\boldsymbol{W}$	$\frac{df}{d\boldsymbol{W}}$	$\frac{d\boldsymbol{f}}{d\boldsymbol{W}}$	$\frac{d\boldsymbol{F}}{d\boldsymbol{W}}$

results in a $n \times 1$ vector;

$$\frac{df}{d\boldsymbol{w}} = \begin{bmatrix} \frac{\partial f}{\partial w_1} \\ \frac{\partial f}{\partial w_2} \\ \vdots \\ \frac{\partial f}{\partial w_n} \end{bmatrix}$$

- **Differentiation of a scalar function *w.r.t.* a matrix**: If $f : \mathbb{R}^{m\times n} \mapsto \mathbb{R}$ is a scalar function of mn variables, $\boldsymbol{w} \in \mathbb{R}^{m\times n}$ is a $m \times n$ matrix, then differentiation of $f(\boldsymbol{w})$ *w.r.t.* $\boldsymbol{w}$ results in a $m \times n$ matrix;

$$\frac{df}{d\boldsymbol{w}} = \begin{bmatrix} \frac{\partial f}{\partial w_{11}} & \cdots & \frac{\partial f}{\partial w_{1n}} \\ \vdots & \cdots & \vdots \\ \frac{\partial f}{\partial w_{m1}} & \cdots & \frac{\partial f}{\partial w_{mn}} \end{bmatrix}$$

- **Differentiation of a vector function *w.r.t.* a vector**: If $\boldsymbol{f} : \mathbb{R}^n \mapsto \mathbb{R}^m$ is a vector function of size $m \times 1$ and $\boldsymbol{w}$ is a $n \times 1$ vectors, then differentiation of $\boldsymbol{f}(\boldsymbol{w})$ *w.r.t.* $\boldsymbol{w}$ results in a $n \times m$ matrix.

$$\frac{d\boldsymbol{f}^T}{d\boldsymbol{w}} = \begin{bmatrix} \frac{\partial f_1}{\partial w_1} & \cdots & \frac{\partial f_h}{\partial w_1} \\ \vdots & \cdots & \vdots \\ \frac{\partial f_1}{\partial w_d} & \cdots & \frac{\partial f_h}{\partial w_d} \end{bmatrix}$$

We also call this matrix the gradient matrix of $\boldsymbol{f}$ *w.r.t.* $\boldsymbol{w}$. The tranpose of this matrix is exactly the Jacobian matrix $\frac{d\boldsymbol{f}}{d\boldsymbol{w}^T}$(*i.e.*, $\frac{d\boldsymbol{f}}{d\boldsymbol{w}}$ in convention) as

$$\frac{d\boldsymbol{f}}{d\boldsymbol{w}^T} = \begin{bmatrix} \frac{\partial f_1}{\partial w_1} & \cdots & \frac{\partial f_1}{\partial w_d} \\ \vdots & \cdots & \vdots \\ \frac{\partial f_h}{\partial w_1} & \cdots & \frac{\partial f_h}{\partial w_d} \end{bmatrix}$$

(2) **(10 points) Derivation by definition**: Use definition of multivariate calculus to derive the derivatives

Example 1: For $f(\boldsymbol{w}) = \boldsymbol{a}^T\boldsymbol{w}$ with $\boldsymbol{w} \in \mathbb{R}^n$, we have $\frac{df}{d\boldsymbol{w}} = \boldsymbol{a}$ since

$$\frac{df}{d\boldsymbol{w}} = \frac{\partial \sum_{i=1}^n a_i x_i}{\partial x_i} = \frac{\partial a_i x_i}{\partial x_i} = a_i$$

Derive the following derivatives by definition.

- (5 pts) $f(\boldsymbol{w}) = \boldsymbol{w}^T\boldsymbol{A}\boldsymbol{w}$ with $\boldsymbol{w} \in \mathbb{R}^n$ and $\boldsymbol{A} \in \mathbb{R}^{n\times n}$;
- (5 pts) $\boldsymbol{f}(\boldsymbol{w}) = \boldsymbol{A}\boldsymbol{w}$ with $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ and $\boldsymbol{w} \in \mathbb{R}^n$.

(3) **(Bonus task: 5 points) Derivation by differentiation**: derive the derivatives based on the relation between differentiation and gradient as

$$f : \mathbb{R}^n \mapsto \mathbb{R} : \qquad df = \sum_{i=1}^{n} \frac{\partial f}{\partial w_i} dw_i = (\frac{df}{d\boldsymbol{w}})^T d\boldsymbol{w} = tr((\frac{df}{d\boldsymbol{w}})^T d\boldsymbol{w})$$

$$f : \mathbb{R}^{m\times n} \mapsto \mathbb{R} : \qquad df = \sum_{i,j} \frac{\partial f}{\partial W_{ij}} dW_{ij} = tr((\frac{df}{d\boldsymbol{W}})^T d\boldsymbol{W})$$

Example 2: For $f(\boldsymbol{W}) = \boldsymbol{a}^T\boldsymbol{W}\boldsymbol{b}$ with $\boldsymbol{W} \in \mathbb{R}^{m\times n}$, $\boldsymbol{a} \in \mathbb{R}^m$, and $\boldsymbol{b} \in \mathbb{R}^n$, we have $\frac{df}{d\boldsymbol{W}} = \boldsymbol{a}\boldsymbol{b}^T$ since

$$\begin{aligned} df &= d(\boldsymbol{a}^T\boldsymbol{W}\boldsymbol{b}) = d(\boldsymbol{a}^T)\boldsymbol{W}\boldsymbol{b} + \boldsymbol{a}^T d(\boldsymbol{W})\boldsymbol{b} + \boldsymbol{a}^T\boldsymbol{W}d(\boldsymbol{b}) = \boldsymbol{a}^T d(\boldsymbol{W})\boldsymbol{b} \\ &= tr(\boldsymbol{a}^T d(\boldsymbol{W})\boldsymbol{b}) = tr(\boldsymbol{b}\boldsymbol{a}^T d\boldsymbol{W}) = tr((\boldsymbol{a}\boldsymbol{b}^T)^T d\boldsymbol{W}) \end{aligned}$$

Hint: Some useful formulas are here.

$$d(X+Y) = dX + dY \qquad d(XY) = (dX)Y + Xd(Y) \qquad dX^T = (dX)^T$$
$$tr(X^T) = tr(X) \qquad tr(XY) = tr(YX) \qquad d(tr(X)) = tr(dX)$$

Derive the following derivatives by differentiation.

- (2 pts) $f(\boldsymbol{w}) = \boldsymbol{w}^T\boldsymbol{A}\boldsymbol{w}$ with $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ and $\boldsymbol{w} \in \mathbb{R}^n$;
- (3 pts) $f(\boldsymbol{W}) = tr(\boldsymbol{W}^T\boldsymbol{A}\boldsymbol{W})$ with $\boldsymbol{A} \in \mathbb{R}^{m\times m}$ and $\boldsymbol{W} \in \mathbb{R}^{m\times n}$.

(4) **(10 points) Derivation by chain rule**: In the realm of machine learning, the input is commonly a feature vector while the loss is almost always a scalar objective function. Therefore, chain rule is useful for gradient update.

(a) (10 pts) Vector chain ending with a scalar: For $\boldsymbol{x} \to \boldsymbol{y}_1 \to \boldsymbol{y}_2 \to \cdots \to \boldsymbol{y}_n \to z$, the chain rule is

$$\frac{dz}{d\boldsymbol{x}^T} = \frac{dz}{d\boldsymbol{y}_n^T}\frac{\partial \boldsymbol{y}_n}{\partial \boldsymbol{y}_{n-1}^T}\cdots\frac{\partial \boldsymbol{y}_2}{\partial \boldsymbol{y}_1^T}\frac{\partial \boldsymbol{y}_1}{\boldsymbol{x}^T}$$

Assume $\ell = \|\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}\|_2^2$, let $\boldsymbol{z} = \boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}$, then $\ell = \|\boldsymbol{z}\|_2^2 = \boldsymbol{z}^T\boldsymbol{z}$. Consequently, here is a chain like $\boldsymbol{w} \to \boldsymbol{z} \to \ell$. Follow the chain rule, the derivative $\frac{\partial \ell}{\partial \boldsymbol{w}^T} = \frac{\partial \ell}{\partial \boldsymbol{z}^T}\frac{\partial \boldsymbol{z}}{\partial \boldsymbol{w}^T}$ which implies that

$$\frac{\partial \ell}{\partial \boldsymbol{w}} = (\frac{\partial \boldsymbol{z}}{\partial \boldsymbol{w}^T})^T\frac{\partial \ell}{\partial \boldsymbol{z}}$$

Determine the closed-form solution of $\frac{\partial \ell}{\partial \boldsymbol{w}}$.

(b) (Bonus Task, 5 points) Matrix chain ending with a scalar: For $\boldsymbol{X} \to \boldsymbol{Y} \to \ell$, the chain rule is

$$\frac{\partial \ell}{\partial X_{ij}} = \sum_{kl} \frac{\partial \ell}{\partial Y_{kl}} \frac{\partial Y_{kl}}{\partial X_{ij}}$$

Assume $\ell = f(\boldsymbol{Y})$ where $\boldsymbol{Y} = \boldsymbol{A}\boldsymbol{X} + \boldsymbol{B}$, derive the derivative of $\frac{\partial \ell}{\partial \boldsymbol{X}}$.

Hint: Firstly, it has

$$\frac{\partial Y_{kl}}{\partial X_{ij}} = \frac{\partial \sum_s A_{ks} X_{sl}}{\partial X_{ij}} = \frac{\partial A_{ki} X_{il}}{\partial X_{ij}} = A_{ki}\delta_{lj}$$

where $\delta_{lj} = 1$ for $l = j$; otherwise, $\delta_{lj} = 0$.

Secondly, we have

$$\frac{\partial \ell}{\partial X_{ij}} = \sum_{kl} \frac{\partial \ell}{\partial Y_{kl}} A_{ki}\delta_{lj} = \sum_{k} \frac{\partial \ell}{\partial Y_{kj}} A_{ki} = \boldsymbol{A}^T_{:,i} (\frac{\partial \ell}{\partial \boldsymbol{Y}})_{:,j}$$

Lastly, you can derive the form of $\frac{\partial \ell}{\partial \boldsymbol{X}}$.

1.2. (Convexity of functions, 10 points) Prove that: (1) $f(x) = \text{ReLU}(x) = (x)^+ = \max(0, x)$ is convex for $x \in \mathbb{R}$; (2) $f(x) = |x|$ is convex for $x \in \mathbb{R}$; (3) $f(\boldsymbol{x}) = \|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\|_2^2$ is convex where $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ and $\boldsymbol{x} \in \mathbb{R}^n$.

1.3. (Gradient Descent, 10 points) Suppose we have training data $\{(\boldsymbol{x}_1, \boldsymbol{y}_1), (\boldsymbol{x}_2, \boldsymbol{y}_2), \ldots, (\boldsymbol{x}_N, \boldsymbol{y}_N)\}$, where $\boldsymbol{x}_i \in \mathbb{R}^d$ and $\boldsymbol{y}_i \in \mathbb{R}^k$, $i = 1, 2, \ldots, N$. (1) Find the closed-form solution of the following problem.

$$\min_{\boldsymbol{W}, \boldsymbol{b}} \sum_{i=1}^{N} \alpha_i \|\boldsymbol{y}_i - \boldsymbol{W}\boldsymbol{x}_i - \boldsymbol{b}\|_2^2,$$

where the diagonal of diagonal matrix $\text{diag}(A) = (\alpha_1, \alpha_2, \ldots, \alpha_N)^T$ are weights for different sample; (2) Show how to use gradient descent to solve the problem.

Hint: You can use either definition or differentiation method to derive the derivatives. If you use differentiation method, please note that

$$\sum_{i=1}^{N} \alpha_i \|\boldsymbol{y}_i - \boldsymbol{W}\boldsymbol{x}_i - \boldsymbol{b}\|_2^2 = tr[(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{W})^T \boldsymbol{A} (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{W})]$$

where $\boldsymbol{Y} = (\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_N)^T \in \mathbb{R}^{N \times k}$, $\boldsymbol{X} = [(\boldsymbol{x}_1^T, 1)^T, (\boldsymbol{x}_2^T, 1)^T, \ldots, (\boldsymbol{x}_N^T, 1)^T]^T \in \mathbb{R}^{N \times (d+1)}$, $\boldsymbol{W} = (\boldsymbol{W}, \boldsymbol{b})^T \in \mathbb{R}^{(d+1) \times k}$, and $\boldsymbol{A} = \text{diag}(\alpha_1, \alpha_2, \ldots, \alpha_N)$.

1.4. (Maximum Likelihood Estimation, 10 points) Suppose $x_1, x_2, \ldots, x_N$ are drawn from $\mathcal{N}(\mu, \sigma^2)$. Show that the maximum likelihood estimation (MLE) of σ^2 is $\hat{\sigma}^2_{MLE} = \frac{1}{N} \sum_{n=1}^{N} (x_n - \mu_{MLE})^2$.

2 Programming (50 points)

2.1. (Polynomial regression, 20 points) In this exercise, we will try to fit a non-linear function g with polynomial regression on the feasible space $\mathbf{X} = [0, 11]$:

$$\text{Unknown} \quad g(x) = ?$$

$$\text{Construct} \quad f(x) = \sum_{i=0}^{n} \alpha_i x^i \quad \Longleftrightarrow \quad f(x) = w^T x', \quad x' = \begin{bmatrix} 1 \\ x \\ x^2 \\ \vdots \\ x^n \end{bmatrix}, \quad s.t. \quad \forall x \in \mathbf{X}, \quad f(x) \approx g(x)$$

Where n is the polynomial degree of freedom and is manually chosen.
Follow the instructions given in the jupyter notebook. At the end of the exercise, you will retrieve an estimation of the desired function and make some comment on this method.

2.2. (Linear regression, 30 points) The CSV or XLS file contains a dataset for regression. There are 7750 samples with 25 features (described in the doc file). This data is for the purpose of bias correction of next-day maximum and minimum air temperatures forecast of the LDAPS model operated by the Korea Meteorological Administration over Seoul, South Korea.

This data consists of summer data from 2013 to 2017. The input data is largely composed of the LDAPS model's next-day forecast data, in-situ maximum and minimum temperatures of present-day, and geographic auxiliary variables. There are two outputs (i.e. next-day maximum and minimum air temperatures) in this data. Hindcast validation was conducted for the period from 2015 to 2017.

You need to delete the first two attributes (station and date), and use attributes 3-23 to predict attributes 24 and 25. Randomly split the data into two parts, one contains 80% of the samples and the other contains 20% of the samples. Use the first part as training data and train a linear regression model and make prediction on the second part. Report the training error and testing error in terms of RMSE.

Repeat the splitting, training, and testing for 10 times. Use a loop and print the RMSEs in each trial.

Note that you need to write the codes of learning the parameters by yourself. Do not use the classification or regression packages of Sklearn. You can use their tools to shuffle the data randomly for splitting.