Machine Learning and Data Mining I (DS 4400)
Homework 1 Solution

Instructor: Ehsan Elhamifar
Due Date: January 28, 2020, 1:35pm

1) Let $\boldsymbol{a} \in \mathbb{R}^n$ be an n-dimensional vector and let $\boldsymbol{U} \in \mathbb{R}^{n \times n}$ be an orthonormal matrix, i.e., $\boldsymbol{U}^\top \boldsymbol{U} = \boldsymbol{U}\boldsymbol{U}^\top = \boldsymbol{I}_n$. Show the following:

[8 points. 4 pts for each part]

1. $\text{trace}(\boldsymbol{a}\boldsymbol{a}^\top) = \|\boldsymbol{a}\|_2^2$.
2. $\|\boldsymbol{U}\boldsymbol{a}\|_2^2 = \|\boldsymbol{a}\|_2^2$. ~> This means the magnitude of a vector does not change after transformation by orthonormal matrix.

Solution: 1. $\text{trace}(aa^\top) = \text{trace}(a^\top a)$ [we know $\text{trace}(AB) = \text{trace}(BA)$]

$= \text{trace}(\|a\|_2^2) = \|a\|_2^2$. [Since $a^\top a = \|a\|_2^2 \in \mathbb{R}$]

2. $\|Ua\|_2^2 = (Ua)^\top (Ua) = a^\top \underbrace{U^\top U}_{I_n} a = a^\top I_n a$ [Since U is orthonormal: $U^\top U = I_n$]

$= a^\top a = \|a\|_2^2$.

2) Let $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ and $\boldsymbol{B} \in \mathbb{R}^{n \times n}$ be arbitrary but invertible matrices and let α be a scalar. Show the following:

[12 points, 4 pts each part]

1. $(\boldsymbol{A}\boldsymbol{B})^{-1} = \boldsymbol{B}^{-1}\boldsymbol{A}^{-1}$.
2. $(\boldsymbol{A}^\top)^{-1} = (\boldsymbol{A}^{-1})^\top$.
3. $\text{trace}(\alpha \boldsymbol{A}) = \alpha\, \text{trace}(\boldsymbol{A})$.

Solution: 1. The definition of inverse of a matrix Z is Z^{-1} so that $Z^{-1}Z = ZZ^{-1} = I_n$.

For $(B^{-1}A^{-1})$ to be inverse of AB, we investigate $(B^{-1}A^{-1})(AB) \stackrel{?}{=} I_n$

$(B^{-1}A^{-1})(AB) = B^{-1}\underbrace{(A^{-1}A)}_{I_n}B = B^{-1}B = I_n$.

2. We know $AA^{-1} = I_n \xrightarrow[\text{of both sides}]{\text{take transpose}} (AA^{-1})^\top = I_n^\top \longrightarrow (A^{-1})^\top A^\top = I_n$

As a result $(A^{-1})^\top$ must be the inverse of $A^\top$: $(A^\top)^{-1} = (A^{-1})^\top$.

3. $A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \cdots & & & \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix} \longrightarrow \text{trace}(A) = \sum_{i=1}^n a_{ii}$.

$\alpha A = \begin{bmatrix} \alpha a_{11} & \alpha a_{12} & \cdots & \alpha a_{1n} \\ \alpha a_{21} & \alpha a_{22} & \cdots & \alpha a_{2n} \\ \cdots & & & \\ \alpha a_{n1} & \alpha a_{n2} & \cdots & \alpha a_{nn} \end{bmatrix} \longrightarrow \text{trace}(\alpha A) = \sum_{i=1}^n \alpha a_{ii} = \alpha\, \text{trace}(A)$.

3) For vectors $\boldsymbol{x} \in \mathbb{R}^n$, $\boldsymbol{a} \in \mathbb{R}^n$ and matrices $\boldsymbol{X} \in \mathbb{R}^{n\times n}$, $\boldsymbol{A} \in \mathbb{R}^{n\times n}$, show the following.

1. $\frac{\partial \boldsymbol{a}^\top \boldsymbol{A}\boldsymbol{x}}{\partial \boldsymbol{x}} = \boldsymbol{A}^\top \boldsymbol{a}$.
2. $\frac{\partial \,\mathrm{trace}\,(\boldsymbol{A}^\top \boldsymbol{X})}{\partial \boldsymbol{X}} = \boldsymbol{A}$.
3. $\frac{\partial \|\boldsymbol{A}\boldsymbol{x}\|^2}{\partial \boldsymbol{x}} = 2\boldsymbol{A}^\top \boldsymbol{A}\boldsymbol{x}$.

[18 points , 6 pts each part]

Solution: 1. $a^TAx = \underbrace{(A^Ta)}_{\triangleq b}{}^T x = b^Tx$ (by defining $b \triangleq A^Ta$)

$$\frac{\partial b^Tx}{\partial x} = \frac{\partial \sum_{i=1}^n b_i x_i}{\partial x} = \begin{pmatrix} \partial(b_1x_1+\cdots+b_nx_n)/\partial x_1 \\ \vdots \\ \partial(b_1x_1+\cdots+b_nx_n)/\partial x_n \end{pmatrix} = \begin{pmatrix} b_1 \\ \vdots \\ b_n \end{pmatrix} = b$$

$$\rightarrow \frac{\partial (a^TAx)}{\partial x} = \frac{\partial ((A^Ta)^Tx)}{\partial x} = \frac{\partial (b^Tx)}{\partial x} = b = A^Ta. \checkmark$$

2. $f(X) \triangleq \mathrm{trace}(A^TX) = \mathrm{tr}\left[\begin{pmatrix} a_{11} & \cdots & a_{n1} \\ \vdots & & \\ a_{1n} & \cdots & a_{nn} \end{pmatrix} \begin{pmatrix} x_{11} & \cdots & x_{1n} \\ \vdots & & \vdots \\ x_{n1} & \cdots & x_{nn} \end{pmatrix} \right] = \sum_{i=1}^n \sum_{j=1}^n a_{ij}x_{ij}$

$$\frac{\partial f}{\partial x} = \begin{pmatrix} \frac{\partial f}{\partial x_{11}} & \cdots & \frac{\partial f}{\partial x_{1n}} \\ \vdots & & \vdots \\ \frac{\partial f}{\partial x_{n1}} & \cdots & \frac{\partial f}{\partial x_{nn}} \end{pmatrix} = \begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & & \vdots \\ a_{n1} & \cdots & a_{nn} \end{pmatrix} = A. \checkmark$$

3. $f(x) \triangleq \|Ax\|_2^2 = (Ax)^T(Ax) = x^TA^TAx = x^TBx$ (here, for simplicity, we difined $B \triangleq A^TA$).

$$= [x_1\ x_2\ \cdots\ x_n] \begin{bmatrix} b_{11} & b_{12} & \cdots & b_{1n} \\ b_{21} & b_{22} & \cdots & b_{2n} \\ \vdots & & & \\ b_{n1} & b_{n2} & \cdots & b_{nn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = [x_1\ x_2\ \cdots\ x_n] \begin{bmatrix} b_{11}x_1+b_{12}x_2+\cdots+b_{1n}x_n \\ b_{21}x_1+b_{22}x_2+\cdots+b_{2n}x_n \\ \vdots \\ b_{n1}x_1+b_{n2}x_2+\cdots+b_{nn}x_n \end{bmatrix}$$

$$= x_1(b_{11}x_1+\cdots+b_{1n}x_n) + x_2(b_{21}x_1+\cdots+b_{2n}x_n) + \cdots + x_n(b_{n1}x_1+\cdots+b_{nn}x_n)$$

$$= b_{11}x_1^2 + \cdots + b_{1n}x_1x_n + b_{21}x_1x_2 + \cdots + b_{2n}x_2x_n + \cdots + b_{n1}x_nx_1 + \cdots + b_{nn}x_n^2 = \sum_{i=1}^n \sum_{j=1}^n b_{ij}x_ix_j$$

$$\frac{\partial f}{\partial x} = \begin{pmatrix} \partial f/\partial x_1 \\ \vdots \\ \partial f/\partial x_n \end{pmatrix} = \begin{pmatrix} 2b_{11}x_1 + (b_{12}+b_{21})x_2 + \cdots + (b_{1n}+b_{n1})x_n \\ \vdots \\ (b_{n1}+b_{1n})x_1 + (b_{n2}+b_{2n})x_2 + \cdots + 2b_{nn}x_n \end{pmatrix} = (B+B^T)x \overset{\text{using } B \triangleq A^TA}{=} \left[A^TA + (A^TA)^T\right]x$$

$$= \left[A^TA + A^TA\right]x = 2A^TAx \checkmark$$

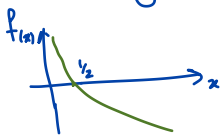
4) Determine whether each of the following functions is convex or not.

1. $f(x) = (x-a)^2$, for any real number a.
2. $f(x) = -\log(2x)$, with the domain $x \in (0, +\infty)$.
3. $f(x) = e^{g(x)}$, where $g(x)$ is convex.

[15 points ; part 1 : 4 pts , part 2 : 4 pts , part 3 : 7 pts]

Solution: 1. $f'(x) = 2(x-a) \rightarrow f''(x) = 2 \geq 0$, Since $f''(x) \geq 0 \ \forall x$, $f(x)$ is convex.

2. $f(x) = -\log 2x \rightarrow f'(x) = -\frac{2}{2x} = -\frac{1}{x} \rightarrow f''(x) = \frac{1}{x^2} \geq 0$. Since $f''(x) \geq 0 \ \forall x$, $f(x)$ is convex.



3. $f(x) = e^{g(x)} \rightarrow f'(x) = g'(x)\, e^{g(x)}$

$\rightarrow f''(x) = \underbrace{g''(x)}_{\geq 0}\, e^{g(x)} + \underbrace{(g'(x))^2}_{\geq 0}\, \underbrace{e^{g(x)}}_{> 0}$ (From convexity of g, we know $g''(x) \geq 0 \ \forall x$)

$\rightarrow f''(x) \geq 0 \ , \forall x \rightarrow f(x)$ is convex.

5) Consider the function $f(x_1, x_2) = (x_1 - 2)^2 + (x_2 - 3)^2$. Write down the expression of the gradient of f.

[31 points , part 2 : 15 pts , rest each 4 pts]

1. Plot the function.
2. Write a python code that gets as input i) an initial point (x_1^0, x_2^0), ii) the maximum number of iterations of the gradient descent, iii) learning rate ρ; and runs the gradient descent algorithm starting from the initial point until convergence or until the maximum number of iterations is achieved. The output of the code must be the sequence of points (starting from initial point and ending with the last point) obtained by the gradient descent.
3. Use the learning rate $\rho = 0.01$ and the initial point $(x_1, x_2) = (1, 2)$. Plot the sequence of obtained points. After how many iterations does GD converge?
4. Use the learning rate $\rho = 0.5$ and the initial point $(x_1, x_2) = (1, 2)$. Plot the sequence of obtained points. After how many iterations does GD converge?
5. Use the learning rate $\rho = 10$ and the initial point $(x_1, x_2) = (1, 2)$. Plot the sequence of obtained points. After how many iterations does GD converge?

$f(x_1, x_2) = \underbrace{(x_1-2)^2}_{\geq 0} + \underbrace{(x_2-3)^2}_{\geq 0} \geq 0$, when $x_1 = 2, x_2 = 3$, $f(2,3) = 0 \rightarrow (x_1 = 2, x_2 = 3)$ is global minimizer.

$\frac{\partial f}{\partial x} = \begin{pmatrix} \partial f / \partial x_1 \\ \partial f / \partial x_2 \end{pmatrix} = \begin{pmatrix} 2(x_1 - 2) \\ 2(x_2 - 3) \end{pmatrix}$

Running GD at iteration t : $\begin{pmatrix} x_1^{(t)} \\ x_2^{(t)} \end{pmatrix} = \begin{pmatrix} x_1^{(t-1)} \\ x_2^{(t-1)} \end{pmatrix} - \rho \begin{pmatrix} 2(x_1^{(t-1)} - 2) \\ 2(x_2^{(t-1)} - 3) \end{pmatrix}$.

6) Consider the training dataset [16 points, part 1: 12 pts, part 2: 4 pts].

$$\{(0.10, 0.15), (0.50, 0.40), (0.90, 0.85), (1.50, 1.62), (-0.20, -0.17), (-0.5, -0.42)\},$$

where in $(\cdot, \cdot)$, the first entry is the input variable, x, and the second entry is the output variable (response), y. Consider the regression model $y = \theta x$.

1. Write a python code that inputs the above data and outputs the optimal regression value θ^*, using the closed-form solution.
2. Plot the data in 2D and plot the estimated line $y = \theta^* x$.

Homework Submission Instructions: Please submit the analytical part of the homework in class (put a paper copy of the HW on the instructor's desk before the class starts). Handwritten or printed documents generated by Latex/Word are both acceptable. Please submit all your plots and your Python code (.py file) via email, by the DEADLINE. To submit, please send an email to the instructor and cc the TAs.

– The title of your email must be "DS4400: HW01:Your-Last-Name".

– Please attach a single zip file to your email that contains all python codes and plots and a readme file on how to run your files.

– Please name your zip file as "HW01:Your-Last-Name".