Assignment 1

Min Wang MatrikelN. 3440557 Email: st168632@stud.uni-stuttgart.de

Jiaqi Zhao MatrikelN. 3470190 Email: st169584@stud.uni-stuttgart.de

Delong Kong MatrikelN. 3374801 Email: st164316@stud.uni-stuttgart.de

1. Matrix equations

1.1

$$
\begin{aligned}
&XA + A^T = I \\
&XA = I - A^T \\
&XAA^{-1} = (I - A^T)A^{-1} \\
&X = (I - A^T)A^{-1}
\end{aligned}
$$

1.2

$$
\begin{aligned}
&X^T C = [2A(X + B)]^T \\
&C^T X = 2A(X + B) \\
&(C^T - 2A)X = 2AB \\
&X = 2(C^T - 2A)^{-1}AB
\end{aligned}
$$

1.3

$$
\begin{aligned}
&(Ax - y)^T A = O_n^T \\
&[(Ax - y)^T A]^T = O_n \\
&A^T(Ax - y) = O_n \\
&A^T Ax = A^T y + O_n \\
&x = (A^T A)^{-1}A^T y
\end{aligned}
$$

1.4

$$
\begin{aligned}
&(Ax-y)^T A + x^B = O_n^T \\
&[(Ax-y)^T A + x^B]^T = O_n \\
&A^T(Ax-y) + B^T x = O_n \\
&(A^T A + B^T)x = A^T y \\
&x = (A^T A + B^T)^{-1} A^T y
\end{aligned}
$$

2. Vector derivatives

2.1

$$\mathbf{x} = [x_1, x_2, ..., x_n]^T \epsilon \mathbf{R}^n$$

$$\frac{\partial \mathbf{x}}{\partial \mathbf{x}} = \begin{bmatrix} \frac{\partial x_1}{\partial x_1} & \frac{\partial x_1}{\partial x_2} & ... & \frac{\partial x_1}{\partial x_n} \\ \frac{\partial x_2}{\partial x_1} & \frac{\partial x_2}{\partial x_2} & ... & \frac{\partial x_2}{\partial x_n} \\ ... & ... & ... & ... \\ \frac{\partial x_n}{\partial x_1} & \frac{\partial x_n}{\partial x_2} & ... & \frac{\partial x_n}{\partial x_n} \end{bmatrix} \epsilon \mathbf{R}^{n \times n}$$

2.2

$$let \quad \mathbf{x}^T\mathbf{x} \;= \mathbf{f} \;\; \epsilon \; \mathbf{R}^{1\times 1}$$

$$\frac{\partial \mathbf{x}^T\mathbf{x}}{\partial \mathbf{x}} = \frac{\partial \mathbf{f}}{\partial \mathbf{x}} = \begin{bmatrix} \frac{\partial \mathbf{f}}{\partial x_1} & \frac{\partial \mathbf{f}}{\partial x_2} & ... & \frac{\partial \mathbf{f}}{\partial x_n} \end{bmatrix}^T \epsilon \mathbf{R}^{n\times 1}$$

2.3

At the end of this pdf, which is in written-form.

3. Error Meassures

3.1

$MSE = \dfrac{1}{n}(y-\hat{y})^T(y-\hat{y})$
$MAE = \dfrac{1}{n}\|y-\hat{y}\|$

3.2

Given several examples with the same input feature values, the optimal prediction of regression using MSE will be the mean of target values of these examples.

If use MAE, the optimal prediction will be the median. The difference are the following:

First, MSE is more sensitive to outliers (large errors) than MAE and will punish large errors more, because it squares all the errors.

Second, minimizing MSE is equivalent of maximizing the likelihood of the data, under the assumption that the target comes from a normal distribution, conditioned on the input. If you believe that your target data, conditioned on the input, is normally distributed around a mean value and it's important to penalize outliers extra much, you should use MSE.

If you don't want outliers to play a big role or if you know that your distribution is multimodal and it's desirable to have predictions at one of the modes rather than at the mean of them, you should use MAE.

Reference: https://peltarion.com/knowledge-center/documentation/evaluationview/regression-loss-metrics

3.3

MSE and MAE both calculate in this binary classification problem the probability of the wrong classification.

assumes the numerator layout convention

3. $x \in R^n$, $y \in R^d$, $A \in R^{d \times n}$

$\Rightarrow (Ax - y) \in R^{d \times 1}$

$\Rightarrow \quad f(x) = (Ax - y)^T (Ax - y) + x^T B x$

$f(x) \in R^{1 \times 1} \qquad f: R^n \to R$

$$\frac{\partial f}{\partial x} = \left(\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}, \cdots \frac{\partial f}{\partial x_n} \right) \in R^{1 \times n}$$

using $u = u(x)$, $v = v(x)$

then $\dfrac{\partial u^T v}{\partial x} = u^T \dfrac{\partial v}{\partial x} + v^T \dfrac{\partial u}{\partial x}$

let $u(x) = v(x) = Ax - y$

using A is not a function of x

A is symmetric

then $\dfrac{\partial x^T A x}{\partial x} = 2x^T A$

let $B = A$

$$\Rightarrow \frac{\partial f}{\partial x} = (Ax - y)^T A + (Ax - y)^T A + 2x^T B$$

$$= 2\left[(Ax - y)^T A + x^T B \right]$$

Identities: scalar-by-vector $\frac{\partial y}{\partial \mathbf{x}} = \nabla_{\mathbf{x}} y$

Condition	Expression	Numerator layout, i.e. by x^T; result is row vector	Denominator layout, i.e. by x; result is column vector

u = u(x), v = v(x)	$\frac{\partial(\mathbf{u}\cdot\mathbf{v})}{\partial\mathbf{x}} = \frac{\partial\mathbf{u}^\top\mathbf{v}}{\partial\mathbf{x}} =$	$\mathbf{u}^\top\frac{\partial\mathbf{v}}{\partial\mathbf{x}} + \mathbf{v}^\top\frac{\partial\mathbf{u}}{\partial\mathbf{x}}$ $\frac{\partial\mathbf{u}}{\partial\mathbf{x}}, \frac{\partial\mathbf{v}}{\partial\mathbf{x}}$ in numerator layout	$\frac{\partial\mathbf{u}}{\partial\mathbf{x}}\mathbf{v} + \frac{\partial\mathbf{v}}{\partial\mathbf{x}}\mathbf{u}$ $\frac{\partial\mathbf{u}}{\partial\mathbf{x}}, \frac{\partial\mathbf{v}}{\partial\mathbf{x}}$ in denominator layout

A is not a function of **x** **A** is symmetric	$\frac{\partial \mathbf{x}^\top \mathbf{A}\mathbf{x}}{\partial \mathbf{x}} =$	$2\mathbf{x}^\top \mathbf{A}$	$2\mathbf{A}\mathbf{x}$

$$x = (x_1, x_2, x_3 \cdots x_n)^T$$

$$\frac{\partial f}{\partial x} = \left(\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}, \cdots \frac{\partial f}{\partial x_n}\right) \in R^{1\times n}$$

$$H(f) = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f}{\partial x_2 x_1} & \frac{\partial^2 f}{\partial x_2 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_2 \partial x_n} \\ \vdots & \vdots & & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1} & \frac{\partial^2 f}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_n^2} \end{bmatrix} \in R^{n\times n}$$

In this case should use denominator-layout notation, since the dimension of H(f)

$$\frac{\partial f}{\partial x} = (Ax-y)^T A + (Ax-y)^T A + 2x^T B$$

$$= 2\left[(Ax-y)^T A + x^T B\right]$$

$$= 2\left[x^T A^T A - y^T A + x^T B\right]$$

use formula

Condition	Expression	Numerator layout, i.e. by y and x^T	Denominator layout, i.e. by y^T and x
a is not a function of **x**	$\frac{\partial \mathbf{a}}{\partial \mathbf{x}} =$	$\mathbf{0}$	
	$\frac{\partial \mathbf{x}}{\partial \mathbf{x}} =$	$\mathbf{I}$	
A is not a function of **x**	$\frac{\partial \mathbf{A}\mathbf{x}}{\partial \mathbf{x}} =$	$\mathbf{A}$	$\mathbf{A}^\top$
A is not a function of **x**	$\frac{\partial \mathbf{x}^\top \mathbf{A}}{\partial \mathbf{x}} =$	$\mathbf{A}^\top$	$\mathbf{A}$

$$H(f) = 2\left[\frac{\partial x^T(A^TA)}{\partial x} - \frac{\partial y^TA}{\partial x} + x^TB\right]$$

$$= 2\left[A^TA - 0 + B\right]$$

$$= 2\left[A^TA + B\right] \in R^{n\times n}$$

Here it's an assumption

$\star$ if A^TA is PD.

$\Rightarrow H(f)$ is positive definite

a is critical point of f

$\text{grad} f(a) = 0 \implies a$ is local minimum let $\frac{\partial f}{\partial x} = 0$

$$\Rightarrow 2\left[(Ax-y)^T A + x^T B\right] = 0_n^T$$

$$\left[(Ax-y)^T A\right]^T + (x^T B)^T = 0_n$$

$$A^T(Ax-y) + B^T x = 0_n$$

$$(A^T A + B^T)x = A^T y$$

Cholesky decomposition $A^T A$ is positive semidefinite $\neq$ PD !

if $A^T A$ is PD

$\because$ B is PD $\therefore$ $B^T = B$

$\therefore$ $A^T A + B^T = A^T A + B$

is positive definite

$\Rightarrow$ invertible

let

$$\frac{\partial f}{\partial x} = 2\left[(Ax-y)^T A + x^T B\right] = 0$$

$$\Rightarrow (Ax-y)^T A + x^T B = 0 \quad \Rightarrow (Ax-y)^T = -x^T B$$

$$\Rightarrow (A^T A + B^T) x = A^T y \quad \Rightarrow x = (A^T A + B^T)^{-1} A^T y$$

$$x^T = (A^T y)^T \left[(A^T A + B^T)^{-1}\right]^T$$

$$= y^T A \left[(A^T A + B^T)^T\right]^{-1}$$

$$= y^T A \left[A^T A + B\right]^{-1}$$

f(x)= (Ax−y)⊤(Ax− y) + x⊤Bx

$$f(x) = (Ax-y)^T Ax - (Ax-y)^T y + x^T Bx$$

$$= \left[(Ax-y)^T A - x^T B\right] x - (Ax-y)^T y$$

$$= \quad - (Ax-y)^T y$$

$$= -(x^T A^T - y^T) y$$

$$= y^T y - x^T A^T y$$

$$= y^T y - y^T A \left[A^T A + B \right]^{-1} A^T y$$

We need discussion and help.

② if (A^TA+B^T) not invertible

$\Leftrightarrow |A^TA+B^T| = 0$

$A^TA+B^T \in R^{n\times n}$

More generally, regardless of whether m=n or not and regardless of the rank of A, all solutions (if any exist) are given using the Moore-Penrose pseudoinverse of A