CSCI3230 (ESTR3108)
Fundamentals of Artificial Intelligence

Tutorial 1

Yuehao Wang

Email: yhwang21@cse.cuhk.edu.hk
Office: Room 1024, 10/F, SHB

Dept. of Computer Science & Engineering
The Chinese University of Hong Kong

Outline

Part 1. Review linear algebra

Part 2. Least squares

Part 3. Bias-variance decomposition

Part 1. Review linear algebra

Vector

- Notation for vectors in $\mathbb{R}^n$:

$$X = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \quad \text{column vector}$$

$$X' = (x_1, \cdots, x_n) \quad \text{row vector}$$

- Transpose: column vector $\leftrightarrow$ row vector

$$X' = X^T, \quad X'^T = X$$

- In the context of vectors, column vectors and row vectors are the same. But when putting them with matrices, they are totally different.

Vectors

- Scaling: $a \in \mathbb{R}$, $X = (x_1, \ldots, x_n) \in \mathbb{R}$, $aX = (ax_1, \ldots, ax_n)$
- Addition: $X = (x_1, \ldots, x_n)$, $Y = (y_1, \ldots, y_n)$,
  $X + Y = (x_1 + y_1, \ldots, x_n + y_n)$
- Suppose we have m vectors: $X_1 \ldots X_m$, linear combination of these vectors: for any m scalars $a_1 \ldots a_m$,

$$a_1 X_1 + \cdots + a_m X_m$$

- Inner product (dot product): $\forall X \in \mathbb{R}^n, Y \in \mathbb{R}^n$

$X = (\cdots)$

$Y = (\cdots)$

$$X \cdot Y = X^T Y = \sum_{i=1}^{n} x_i y_i$$

- Euclidean norm: $\|X\|_2 = \sqrt{X^T X} = \sqrt{\sum_{i=1}^{n} x_i^2}$

Matrix

- Notation for $m \times n$ matrices:

$$\mathbf{A} = \begin{pmatrix} a_{1,1} & \cdots & a_{1,n} \\ \vdots & \ddots & \vdots \\ a_{m,1} & \cdots & a_{m,n} \end{pmatrix}$$

  $a_{i,j}$ is the element at the i-th row, j-th column of $\mathbf{A}$.

- Denote the i-th column of $\mathbf{A}$ as $\mathbf{A}_i$ (column vectors), and j-th row of $\mathbf{A}$ as $\mathbf{A}^{(j)}$ (row vectors).
- Transpose: columns $\leftrightarrow$ rows

$$\mathbf{A}^T = \begin{pmatrix} a_{1,1} & \cdots & a_{m,1} \\ \vdots & \ddots & \vdots \\ a_{1,n} & \cdots & a_{m,n} \end{pmatrix} = [{A^{(1)}}^T, \ldots, {A^{(m)}}^T]$$

Matrix

- n-dimensional row vector: $1 \times n$ matrix $\quad 1\,\overset{n}{[\,\cdots\,]}$
- n-dimensional column vector: $n \times 1$ matrix $\quad n\,[\,\vdots\,]$
- Scaling: $c \in \mathbb{R}$, $\mathbf{A} = \begin{pmatrix} a_{1,1} & \cdots & a_{1,n} \\ \vdots & \ddots & \vdots \\ a_{m,1} & \cdots & a_{m,n} \end{pmatrix}$, $c\mathbf{A} = \begin{pmatrix} ca_{1,1} & \cdots & ca_{1,n} \\ \vdots & \ddots & \vdots \\ ca_{m,1} & \cdots & ca_{m,n} \end{pmatrix}$
- Addition: $\mathbf{A} = \begin{pmatrix} a_{1,1} & \cdots & a_{1,n} \\ \vdots & \ddots & \vdots \\ a_{m,1} & \cdots & a_{m,n} \end{pmatrix}$, $\mathbf{B} = \begin{pmatrix} b_{1,1} & \cdots & b_{1,n} \\ \vdots & \ddots & \vdots \\ b_{m,1} & \cdots & b_{m,n} \end{pmatrix}$

mxm $\qquad\qquad$ mxn

$$\mathbf{A} + \mathbf{B} = \begin{pmatrix} a_{1,1} + b_{1,1} & \cdots & a_{1,n} + b_{1,n} \\ \vdots & \ddots & \vdots \\ a_{m,1} + b_{m,1} & \cdots & a_{m,n} + b_{m,n} \end{pmatrix}$$

$\mathbf{A}$ and $\mathbf{B}$ have the same shape: $\mathbb{R}^{m \times n}$

Special matrices

- Diagonal matrix in $\mathbb{R}^{n\times n}$:

$$\Lambda = \begin{pmatrix} \Lambda_1 & 0 & \cdots & 0 \\ 0 & \Lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \Lambda_n \end{pmatrix}$$

$I \cdot 3 = 3$

$I \cdot A = A$

- Identity matrix in $\mathbb{R}^{n\times n}$: $I = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix}$

- Symmetric matrix: $A = A^T$

Matrix-vector multiplication

- Matrix-vector multiplication

$$
\begin{aligned}
\mathbf{A}X &= \begin{pmatrix} a_{1,1} & \cdots & a_{1,n} \\ \vdots & \ddots & \vdots \\ a_{m,1} & \cdots & a_{m,n} \end{pmatrix} \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \\
&= \begin{pmatrix} a_{1,1} \\ \vdots \\ a_{m,1} \end{pmatrix} x_1 + \begin{pmatrix} a_{1,2} \\ \vdots \\ a_{m,2} \end{pmatrix} x_2 + \cdots \begin{pmatrix} a_{1,n} \\ \vdots \\ a_{m,n} \end{pmatrix} x_n \\
&= \mathbf{A}_1 x_1 + \cdots + \mathbf{A}_n x_n
\end{aligned}
$$

- Notice the shape: $\mathbf{A} \in \mathbb{R}^{m \times n}$, $X \in \mathbb{R}^n$, $\mathbf{A}X \in \mathbb{R}^m$.
- Matrix-vector multiplication $=$ linear combination of $\mathbf{A}$'s columns.

Matrix-vector multiplication

Another perspective of matrix-vector multiplication:

- Matrix-vector multiplication

$$\mathbf{A}X = \begin{pmatrix} \mathbf{A}^{(1)} \\ \vdots \\ \mathbf{A}^{(m)} \end{pmatrix} X$$

$$= \begin{pmatrix} \mathbf{A}^{(1)} \cdot X \\ \vdots \\ \mathbf{A}^{(m)} \cdot X \end{pmatrix}$$

scalar

Recall that $\mathbf{A}^{(i)}$ is the i-th row of $\mathbf{A}$.

- Matrix-vector multiplication = dot product of $\mathbf{A}$'s rows and X.

Map two → one value

Matrix-matrix multiplication

- Matrix-matrix multiplication

$$\mathbf{AX} = \begin{pmatrix} a_{1,1} & \cdots & a_{1,n} \\ \vdots & \ddots & \vdots \\ a_{m,1} & \cdots & a_{m,n} \end{pmatrix} \begin{pmatrix} x_{1,1} & \cdots & x_{1,l} \\ \vdots & \ddots & \vdots \\ x_{n,1} & \cdots & x_{n,l} \end{pmatrix}$$

- Let X_i be the i-th column of $\mathbf{X}$, the i-th column of $\mathbf{AX}$ is $\mathbf{A}X_i$:

$$\mathbf{AX} = [\mathbf{A}X_1, \ldots, \mathbf{A}X_l]$$

- Notice the shape: $\mathbf{A} \in \mathbb{R}^{m \times n}$, $\mathbf{X} \in \mathbb{R}^{n \times l}$, $\mathbf{AX} \in \mathbb{R}^{m \times l}$.
- Multiplication for block matrices: $\mathbf{A} = [\mathbf{A}_1, \mathbf{A}_2]$

$$\mathbf{BA} = \mathbf{B}[\mathbf{A}_1, \mathbf{A}_2] = [\mathbf{BA}_1, \mathbf{BA}_2]$$

Inverse matrix

- For squared matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$, if there is a $\mathbf{B} \in \mathbb{R}^{n\times n}$ such that

$$\mathbf{AB} = I = \mathbf{BA}$$

$A^{-1}A = I$

$\mathbf{A}$ is invertible and $\mathbf{A}^{-1} = \mathbf{B}$

- How to find inverse matrix of $\mathbf{A}$ (if invertible):
  1. Gauss-Jordan elimination: find a row-operating matrix $\mathbf{B}$ which transforms $\mathbf{A}$ to $\mathbf{I}$. (More feasible for human begings)
  2. Use eign-decomposition: $\mathbf{A} = U\Lambda U^{-1}$, $\mathbf{A}^{-1} = U\Lambda^{-1}U^{-1}$
  3. Use the analytic solution:

$$\boxed{\mathbf{A}^{-1} = \frac{1}{|\mathbf{A}|}\mathbf{C}^T}$$

where $\mathbf{C}$ is the adjugate matrix of $\mathbf{A}$.

$$B[A \mid I] = [BA \mid BI] = [I \mid B]$$

Matrix calculus

- Univariate real-valued function $f(x) : \mathbb{R} \to \mathbb{R}$
- Multi-variate real-valued function $f(x_1, \ldots, x_n) : \mathbb{R}^n \to \mathbb{R}$. For convenience, we use $X = (x_1, \ldots, x_n)^T$ to represent the independent variable. So we can write it as $f(X) : \mathbb{R}^n \to \mathbb{R}$.
- We use denominator layout to find the derivative w.r.t. X.

$$\begin{pmatrix} \frac{\partial f(x_1 \cdots x_n)}{\partial x_1} \\ \vdots \\ \partial x_n \end{pmatrix} \frac{\partial f(X)}{\partial X} = \left(\frac{\partial f(X)}{\partial x_1}, \frac{\partial f(X)}{\partial x_2}, \ldots, \frac{\partial f(X)}{\partial x_n} \right)^T$$

Due to denominator layout, the first derivative is a column vector.

Reference: Wikipedia - Matrix calculus

Matrix calculus

$\forall X \in \mathbb{R}^n, Y \in \mathbb{R}^n$:

$f(x) = Y^T X : R^n \to R$

$$\frac{\partial Y^T X}{\partial X} = Y$$

$$\frac{\partial X^T Y}{\partial X} = Y$$

$f(x) = ax$

$\frac{df(x)}{dx} = a$

$Y^T X = \sum_{i=1}^{n} y_i x_i = y_1 x_1 + y_2 x_2 + \dots + y_n x_n$

$\frac{\partial Y^T X}{\partial X} = \left(\frac{\partial Y^T X}{\partial x_i} \dots \frac{\partial Y^T X}{\partial x_n} \right) = Y$

$\downarrow \qquad \downarrow$

$y_i \quad \dots \quad y_n$

Matrix calculus

$\forall X \in \mathbb{R}^n, \mathbf{A} \in \mathbb{R}^{n\times n}$:

$f(x) = X^T \underline{A} X$

$$\frac{\partial X^T \mathbf{A} X}{\partial X} = (\mathbf{A} + \mathbf{A}^T)X$$

$f(x) = \underline{a}x^2 + \underline{b}x + \underline{c}$

$X^T A X$

$f'(x) = 2ax + b$

$\Uparrow$

$\sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} x_i y_j$

$\frac{\partial X^T A X}{\partial X_1} = 2 \sum a_{1j} x_j$

$\boxed{(A + A^T) X}$

Linear algebra materials

1. Introduction to Linear Algebra. Gilbert Strang.
2. Matrix Analysis and Applied Linear Algebra. Carl D. Meyer.
3. Advanced Linear Algebra. Steven Roman.
4. Linear Algebra and Its Applications. Manolis C. Tsakiris.

Part 2. Least squares

Problem settings

- Recall that our data matrix $\mathbf{X}$ and observed labels Y:

$$\mathbf{X} = \begin{pmatrix} \mathbf{1} & x_1^{(1)} & \dots & x_n^{(1)} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{1} & x_1^{(m)} & \dots & x_n^{(m)} \end{pmatrix} \quad Y = \begin{pmatrix} y^{(1)} \\ \vdots \\ y^{(m)} \end{pmatrix}$$

- We use linear function to fit a linear mapping from $\mathbf{X}$ to Y:

$$\hat{Y} = \mathbf{X}\Theta$$

where $\Theta = \begin{pmatrix} \theta_{\mathbf{0}} \\ \theta_1 \\ \vdots \\ \theta_n \end{pmatrix}$, such that $\hat{Y}$ is very close to Y.

Ordinary least squares

- Recall that we use $\|\hat{Y} - Y\|_2^2$ to measure the distance between real labels and estimated labels.
- Note that:

$$\|\hat{Y} - Y\|_2^2 = \sum_{i=0}^{m} (\hat{y}^{(i)} - y^{(i)})^2$$

  which is exactly the residual sum of squares (RSS).
- Ordinary least squares (OLS) estimator:

$$\hat{\Theta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T Y$$

  is the optimal solution that minimizes $\|\hat{Y} - Y\|_2$.

Derivation

Analytic solution:

$$
\begin{aligned}
J(\Theta) = \|\hat{f}_\Theta(\mathbf{X}) - Y\|_2^2 &= (\mathbf{X}\Theta - Y)^T(\mathbf{X}\Theta - Y) \\
&= \Theta^T\mathbf{X}^T\mathbf{X}\Theta - Y^T\mathbf{X}\Theta - \Theta^T\mathbf{X}^T Y - Y^T Y
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial J(\Theta)}{\partial \Theta} &= 2\mathbf{X}^T\mathbf{X}\Theta - \mathbf{X}^T Y - \mathbf{X}^T Y \\
&= 2\mathbf{X}^T(\mathbf{X}\Theta - Y) = 0 \ \cdots\cdots\cdots\ (1) \\
\frac{\partial^2 J(\Theta)}{\partial \Theta^2} &= 2\mathbf{X}^T\mathbf{X} \succeq 0 \ \text{ is for true.}
\end{aligned}
$$

By (1), $\mathbf{X}^T\mathbf{X}\Theta = \mathbf{X}^T Y \implies \boxed{\hat{\Theta} = \Theta^* = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T Y}$

Example

We have 3 samples with bi-variate features:

$$X^{(1)} = \begin{pmatrix} 0 \\ 1 \end{pmatrix} \quad X^{(2)} = \begin{pmatrix} 1 \\ 0 \end{pmatrix} \quad X^{(3)} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

with corresponding observed labels: $y^{(1)} = 1, y^{(2)} = 2, y^{(3)} = 3$. Suppose we use linear regression model to predict the relationship between the featuers and labels. Please use OLS to find the $\hat{\Theta}$.

Example (cont.)

Example (cont.)

Example (cont.)

Part 3. Bias-variance decomposition

Problem settings

- Features: X. True relationship: $f(X)$.
- Observed labels: $y = f(X) + \varepsilon$, where ε is a noise, with $\mathrm{E}(\varepsilon) = 0$.
- $\hat{f}(X)$ is the estimate of $f(X)$ by some estimators, e.g., OLS.
- Bias-variance decomposition:

$$\mathrm{E}\left[(y - \hat{f})^2\right] = \mathrm{Bias}[\hat{f}]^2 + \sigma^2 + \mathrm{Var}\left[\hat{f}\right]$$

which says the expectation of errors between observed labels and estimated labels is sum of squared bias, squared irreducible errors, and the variance of estimated relationship.

Derivation

$$
\begin{aligned}
\mathrm{E}\left[(y-\hat{f})^2\right] &= \mathrm{E}\left[(f+\varepsilon-\hat{f})^2\right] \\
&= \mathrm{E}\left[(f+\varepsilon-\hat{f}+\mathrm{E}[\hat{f}]-\mathrm{E}[\hat{f}])^2\right] \\
&= \mathrm{E}\left[(f-\mathrm{E}[\hat{f}])^2\right]+\mathrm{E}[\varepsilon^2]+\mathrm{E}\left[(\mathrm{E}[\hat{f}]-\hat{f})^2\right]+2\,\mathrm{E}\left[(f-\mathrm{E}[\hat{f}])\varepsilon\right] \\
&\quad +2\,\mathrm{E}\left[\varepsilon(\mathrm{E}[\hat{f}]-\hat{f})\right]+2\,\mathrm{E}\left[(\mathrm{E}[\hat{f}]-\hat{f})(f-\mathrm{E}[\hat{f}])\right] \\
&= (f-\mathrm{E}[\hat{f}])^2+\mathrm{E}[\varepsilon^2]+\mathrm{E}\left[(\mathrm{E}[\hat{f}]-\hat{f})^2\right]+2(f-\mathrm{E}[\hat{f}])\,\mathrm{E}[\varepsilon] \\
\\
&\quad +2\,\mathrm{E}[\varepsilon]\,\mathrm{E}\left[\,\mathrm{E}[\hat{f}]-\hat{f}\,\right]+2\,\mathrm{E}\left[\,\mathrm{E}[\hat{f}]-\hat{f}\,\right](f-\mathrm{E}[\hat{f}])
\end{aligned}
$$

Derivation

$$
\begin{aligned}
\mathrm{E}\left[(y-\hat{f})^2\right] &= \mathrm{E}\left[(f+\varepsilon-\hat{f})^2\right] \\
&= \mathrm{E}\left[(f+\varepsilon-\hat{f}+\mathrm{E}[\hat{f}]-\mathrm{E}[\hat{f}])^2\right] \\
&= \mathrm{E}\left[(f-\mathrm{E}[\hat{f}])^2\right]+\mathrm{E}[\varepsilon^2]+\mathrm{E}\left[(\mathrm{E}[\hat{f}]-\hat{f})^2\right]+2\,\mathrm{E}\left[(f-\mathrm{E}[\hat{f}])\varepsilon\right] \\
&\quad +2\,\mathrm{E}\left[\varepsilon(\mathrm{E}[\hat{f}]-\hat{f})\right]+2\,\mathrm{E}\left[(\mathrm{E}[\hat{f}]-\hat{f})(f-\mathrm{E}[\hat{f}])\right] \\
&= (f-\mathrm{E}[\hat{f}])^2+\mathrm{E}[\varepsilon^2]+\mathrm{E}\left[(\mathrm{E}[\hat{f}]-\hat{f})^2\right]+2(f-\mathrm{E}[\hat{f}])\,\mathrm{E}[\varepsilon] \\
&\quad +2\,\mathrm{E}[\varepsilon]\,\mathrm{E}\left[\mathrm{E}[\hat{f}]-\hat{f}\right]+2\,\mathrm{E}\left[\mathrm{E}[\hat{f}]-\hat{f}\right](f-\mathrm{E}[\hat{f}]) \\
&= (f-\mathrm{E}[\hat{f}])^2+\mathrm{E}[\varepsilon^2]+\mathrm{E}\left[(\mathrm{E}[\hat{f}]-\hat{f})^2\right] \\
&= (f-\mathrm{E}[\hat{f}])^2+\mathrm{Var}[\varepsilon]+\mathrm{Var}\left[\hat{f}\right] \\
&= \mathrm{Bias}[\hat{f}]^2+\mathrm{Var}[\varepsilon]+\mathrm{Var}\left[\hat{f}\right] \\
&= \mathrm{Bias}[\hat{f}]^2+\sigma^2+\mathrm{Var}\left[\hat{f}\right]
\end{aligned}
$$