Machine Learning with Python

MTH786U/P 2020/21

Mathematical preliminaries

Mihail Poplavskyi, Queen Mary University of London (QMUL)

m.poplavskyi@qmul.ac.uk

LINEAR ALGEBRA

Linear algebra

Linear algebra

Matrix: $X = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1d} \\ x_{21} & x_{22} & \dots & x_{2d} \\ \vdots & & \ddots & \vdots \\ x_{s1} & x_{s2} & \dots & x_{sd} \end{pmatrix} \in \mathbb{R}^{s \times d}$

Vectors: $w = \begin{pmatrix} w_1 \\ w_2 \\ \vdots \\ w_d \end{pmatrix} \in \mathbb{R}^{d \times 1}, \ y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_s \end{pmatrix} \in \mathbb{R}^{s \times 1}$

Linear algebra

Matrix: $X = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1d} \\ x_{21} & x_{22} & \cdots & x_{2d} \\ \vdots & & \ddots & \vdots \\ x_{s1} & x_{s2} & \cdots & x_{sd} \end{pmatrix} \in \mathbb{R}^{s\times d}$

Vectors: $w = \begin{pmatrix} w_1 \\ w_2 \\ \vdots \\ w_d \end{pmatrix} \in \mathbb{R}^{d\times 1}, \ y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_s \end{pmatrix} \in \mathbb{R}^{s\times 1}$

Matrix-vector multiplication: $Xw = y \quad \Leftrightarrow \quad \sum_{j=1}^{d} x_{ij} w_j = y_i \quad \forall i \in \{1, \ldots, s\}$

Linear algebra

Matrix: $X = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1d} \\ x_{21} & x_{22} & \cdots & x_{2d} \\ \vdots & & \ddots & \vdots \\ x_{s1} & x_{s2} & \cdots & x_{sd} \end{pmatrix} \in \mathbb{R}^{s \times d}$

Vectors: $w = \begin{pmatrix} w_1 \\ w_2 \\ \vdots \\ w_d \end{pmatrix} \in \mathbb{R}^{d \times 1}, \ y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_s \end{pmatrix} \in \mathbb{R}^{s \times 1}$

Matrix-vector multiplication: $Xw = y \quad \Leftrightarrow \quad \sum_{j=1}^{d} x_{ij} w_j = y_i \quad \forall i \in \{1, \ldots, s\}$

Matrix-matrix multiplication: $XW = Y \quad \Leftrightarrow \quad \sum_{j=1}^{d} x_{ij} w_{jk} = y_{ik}$ for

$X \in \mathbb{R}^{s \times d}, W \in \mathbb{R}^{d \times m}, Y \in \mathbb{R}^{s \times m}$

Linear algebra

Matrix: $X = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1d} \\ x_{21} & x_{22} & \dots & x_{2d} \\ \vdots & & \ddots & \vdots \\ x_{s1} & x_{s2} & \dots & x_{sd} \end{pmatrix} \in \mathbb{R}^{s\times d}$

Vectors: $w = \begin{pmatrix} w_1 \\ w_2 \\ \vdots \\ w_d \end{pmatrix} \in \mathbb{R}^{d\times 1}, \ y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_s \end{pmatrix} \in \mathbb{R}^{s\times 1}$

Matrix-vector multiplication: $Xw = y \quad \Leftrightarrow \quad \sum_{j=1}^{d} x_{ij} w_j = y_i \quad \forall i \in \{1, \dots, s\}$

Matrix-matrix multiplication: $XW = Y \quad \Leftrightarrow \quad \sum_{j=1}^{d} x_{ij} w_{jk} = y_{ik}$ for $X \in \mathbb{R}^{s\times d}, W \in \mathbb{R}^{d\times m}, Y \in \mathbb{R}^{s\times m}$

Inner / Dot product:

$\langle x, y \rangle := \sum_{j=1}^{d} x_j y_j = x^\top y = y^\top x = x \cdot y$ for $x, y \in \mathbb{R}^{d\times 1}$

Linear algebra

Matrix: $X = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1d} \\ x_{21} & x_{22} & \dots & x_{2d} \\ \vdots & & \ddots & \vdots \\ x_{s1} & x_{s2} & \dots & x_{sd} \end{pmatrix} \in \mathbb{R}^{s \times d}$

Vectors: $w = \begin{pmatrix} w_1 \\ w_2 \\ \vdots \\ w_d \end{pmatrix} \in \mathbb{R}^{d \times 1}, \; y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_s \end{pmatrix} \in \mathbb{R}^{s \times 1}$

Matrix-vector multiplication: $Xw = y \quad \Leftrightarrow \quad \sum_{j=1}^{d} x_{ij} w_j = y_i \quad \forall i \in \{1, \dots, s\}$

Matrix-matrix multiplication: $XW = Y \quad \Leftrightarrow \quad \sum_{j=1}^{d} x_{ij} w_{jk} = y_{ik}$ for $X \in \mathbb{R}^{s \times d}, W \in \mathbb{R}^{d \times m}, Y \in \mathbb{R}^{s \times m}$

Inner / Dot product:

$\langle x, y \rangle := \sum_{j=1}^{d} x_j y_j = x^\top y = y^\top x = x \cdot y$ for $x, y \in \mathbb{R}^{d \times 1}$

Norm: $\|x\| := \sqrt{\langle x, x \rangle}$

Linear algebra

The transpose $X^\top$ of a matrix X is defined as $X^\top = \begin{pmatrix} x_{11} & x_{21} & \cdots & x_{s1} \\ x_{12} & x_{22} & \cdots & x_{s2} \\ \vdots & & \ddots & \vdots \\ x_{1d} & x_{2d} & \cdots & x_{sd} \end{pmatrix} \in \mathbb{R}^{d \times s}$

Linear algebra

The transpose $X^\top$ of a matrix X is defined as $X^\top = \begin{pmatrix} x_{11} & x_{21} & \dots & x_{s1} \\ x_{12} & x_{22} & \dots & x_{s2} \\ \vdots & & \ddots & \vdots \\ x_{1d} & x_{2d} & \dots & x_{sd} \end{pmatrix} \in \mathbb{R}^{d \times s}$

We can find so-called singular vectors u_σ, v_σ and singular values σ that satisfy

$$Xu_\sigma = \sigma v_\sigma$$

$$X^\top v_\sigma = \sigma u_\sigma$$

Linear algebra

The transpose $X^\top$ of a matrix X is defined as $X^\top = \begin{pmatrix} x_{11} & x_{21} & \dots & x_{s1} \\ x_{12} & x_{22} & \dots & x_{s2} \\ \vdots & & \ddots & \vdots \\ x_{1d} & x_{2d} & \dots & x_{sd} \end{pmatrix} \in \mathbb{R}^{d \times s}$

We can find so-called singular vectors u_σ, v_σ and singular values σ that satisfy

$$\begin{aligned} Xu_\sigma &= \sigma v_\sigma \\ X^\top v_\sigma &= \sigma u_\sigma \end{aligned} \quad \Longrightarrow \quad \begin{aligned} X^\top X u_\sigma &= \sigma^2 u_\sigma \\ XX^\top v_\sigma &= \sigma^2 v_\sigma \end{aligned}$$

Linear algebra

The transpose $X^\top$ of a matrix X is defined as $X^\top = \begin{pmatrix} x_{11} & x_{21} & \cdots & x_{s1} \\ x_{12} & x_{22} & \cdots & x_{s2} \\ \vdots & & \ddots & \vdots \\ x_{1d} & x_{2d} & \cdots & x_{sd} \end{pmatrix} \in \mathbb{R}^{d \times s}$

We can find so-called singular vectors u_σ, v_σ and singular values σ that satisfy

$$\begin{aligned} Xu_\sigma &= \sigma v_\sigma \\ X^\top v_\sigma &= \sigma u_\sigma \end{aligned} \quad \Longrightarrow \quad \begin{aligned} X^\top X u_\sigma &= \sigma^2 u_\sigma \\ XX^\top v_\sigma &= \sigma^2 v_\sigma \end{aligned} \quad \Longrightarrow \quad \sigma = \frac{\|Xu_\sigma\|}{\|u_\sigma\|} = \frac{\|X^\top v_\sigma\|}{\|v_\sigma\|}$$

Linear algebra

We can find so-called singular vectors u_σ, v_σ and singular values σ that satisfy

$$\begin{aligned} Xu_\sigma &= \sigma v_\sigma \\ X^\top v_\sigma &= \sigma u_\sigma \end{aligned} \quad \Longrightarrow \quad \begin{aligned} X^\top X u_\sigma &= \sigma^2 u_\sigma \\ XX^\top v_\sigma &= \sigma^2 v_\sigma \end{aligned} \quad \Longrightarrow \quad \sigma = \frac{\|Xu_\sigma\|}{\|u_\sigma\|} = \frac{\|X^\top v_\sigma\|}{\|v_\sigma\|}$$

It can be shown that for every matrix $X \in \mathbb{R}^{s\times d}$ there exist $\{\sigma_i\}_{i=1}^{\min(s,d)}$ with $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_{\min(s,d)}$ and vectors $\{u_{\sigma_i}\}_{i=1}^{\min(s,d)}$ and $\{v_{\sigma_i}\}_{i=1}^{\min(s,d)}$ such that

$$Xw = \sum_{j=1}^{\min(s,d)} \sigma_j \langle w, u_j \rangle v_j \quad \text{and} \quad X^\top y = \sum_{j=1}^{\min(s,d)} \sigma_j \langle y, v_j \rangle u_j$$

for all $w \in \mathbb{R}^d$ and $y \in \mathbb{R}^s$

CALCULUS

Calculus

Calculus

Assume we have a function $f(x_1, x_2, \ldots, x_d)$

and let $x = (x_1, x_2, \ldots, x_d)$

Calculus

Assume we have a function $f(x_1, x_2, \ldots, x_d)$ and let $x = (x_1, x_2, \ldots, x_d)$. Then the gradient of f is the vector of partial derivatives:

$$\nabla f(x)^\top = \begin{pmatrix} \frac{\partial}{\partial x_1} f(x) \\ \frac{\partial}{\partial x_2} f(x) \\ \vdots \\ \frac{\partial}{\partial x_d} f(x) \end{pmatrix}$$

Calculus

Assume we have a function $f(x_1, x_2, \ldots, x_d)$ and let $x = (x_1, x_2, \ldots, x_d)$. Then the gradient of f is the vector of partial derivatives:

$$\nabla f(x)^\top = \begin{pmatrix} \frac{\partial}{\partial x_1} f(x) \\ \frac{\partial}{\partial x_2} f(x) \\ \vdots \\ \frac{\partial}{\partial x_d} f(x) \end{pmatrix}$$

Example: $f(x_1, x_2) = (x_1 x_2 - y)^2$

Calculus

Assume we have a function $f(x_1, x_2, \ldots, x_d)$

and let $x = (x_1, x_2, \ldots, x_d)$. Then the gradient

of f is the vector of partial derivatives:

$$\nabla f(x)^\top = \begin{pmatrix} \frac{\partial}{\partial x_1} f(x) \\ \frac{\partial}{\partial x_2} f(x) \\ \vdots \\ \frac{\partial}{\partial x_d} f(x) \end{pmatrix}$$

Example: $f(x_1, x_2) = (x_1 x_2 - y)^2$

Then

$$\frac{\partial}{\partial x_1} f(x) = 2(x_1 x_2 - y) x_2$$

$$\frac{\partial}{\partial x_2} f(x) = 2x_1 (x_1 x_2 - y)$$

Calculus

Assume we have a function $f(x_1, x_2, \ldots, x_d)$

and let $x = (x_1, x_2, \ldots, x_d)$. Then the gradient

of f is the vector of partial derivatives:

$$\nabla f(x)^\top = \begin{pmatrix} \frac{\partial}{\partial x_1} f(x) \\ \frac{\partial}{\partial x_2} f(x) \\ \vdots \\ \frac{\partial}{\partial x_d} f(x) \end{pmatrix}$$

Example: $f(x_1, x_2) = (x_1 x_2 - y)^2$

Then

$$\frac{\partial}{\partial x_1} f(x) = 2(x_1 x_2 - y) x_2$$

$$\frac{\partial}{\partial x_2} f(x) = 2x_1(x_1 x_2 - y)$$

$$\Rightarrow \quad \nabla f(x)^\top = 2 \begin{pmatrix} x_1 x_2^2 - y x_2 \\ x_1^2 x_2 - x_1 y \end{pmatrix}$$

Calculus

We can extend this to functions $f : \mathbb{R}^d \to \mathbb{R}^s$ with multiple outputs via the Jacobian matrix $J_f : \mathbb{R}^d \to \mathbb{R}^{s\times d}$ defined as

$$J_f(x) := \begin{pmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_d} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_s}{\partial x_1} & \cdots & \frac{\partial f_s}{\partial x_d} \end{pmatrix}$$

Calculus

We can extend this to functions $f : \mathbb{R}^d \to \mathbb{R}^s$ with multiple outputs via the Jacobian matrix $J_f : \mathbb{R}^d \to \mathbb{R}^{s \times d}$ defined as

$$J_f(x) := \begin{pmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_d} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_s}{\partial x_1} & \cdots & \frac{\partial f_s}{\partial x_d} \end{pmatrix}$$

And we can even define a second-order derivative matrix (known as Hessian) via

$$H_f(x) := J_{\nabla f}(x)$$

PROBABILITY & STATISTICS

Probability & statistics

Assume we have a random variable X with a finite no. of outcomes $x_1, x_2, \ldots, x_s$ and probabilities $\rho_1 = P(X = x_1), \rho_2 = P(X = x_2), \ldots, \rho_s = P(X = x_s)$. The expectation of X is defined as

$$\mathbb{E}_x[x_i] := \sum_{i=1}^{s} x_i \, \rho_i$$

Probability & statistics

Assume we have a random variable X with a finite no. of outcomes $x_1, x_2, \ldots, x_s$ and probabilities $\rho_1 = P(X = x_1), \rho_2 = P(X = x_2), \ldots, \rho_s = P(X = x_s)$. The expectation of X is defined as

$$\mathbb{E}_x[x_i] := \sum_{i=1}^{s} x_i \, \rho_i$$

Example: $s = 3$, $x_1 = 1$, $x_2 = 11/10$, $x_3 = 1/2$ and $\rho_1 = 1/2$, $\rho_2 = 1/3$, $\rho_3 = 1/6$

Probability & statistics

Assume we have a random variable X with a finite no. of outcomes $x_1, x_2, \ldots, x_s$ and probabilities $\rho_1 = P(X = x_1), \rho_2 = P(X = x_2), \ldots, \rho_s = P(X = x_s)$. The expectation of X is defined as

$$\mathbb{E}_x[x_i] := \sum_{i=1}^{s} x_i \, \rho_i$$

Example: $s = 3$, $x_1 = 1$, $x_2 = 11/10$, $x_3 = 1/2$ and $\rho_1 = 1/2$, $\rho_2 = 1/3$, $\rho_3 = 1/6$

$$\Longrightarrow \qquad \mathbb{E}_x[x_i] = \sum_{i=1}^{3} x_i \, \rho_i = \frac{1}{2} + \frac{11}{30} + \frac{1}{12} = \frac{19}{20} = 0.95$$

(weighted average)

Probability & statistics

Assume we have an absolutely continuous random variable X with probability density function ρ. The expectation of X is defined as

$$\mathbb{E}_x[x] := \int_{\mathbb{R}} x\,\rho(x)\,dx$$

Probability & statistics

Assume we have an absolutely continuous random variable X with probability density function ρ. The expectation of X is defined as

$$\mathbb{E}_x[x] := \int_{\mathbb{R}} x\,\rho(x)\,dx$$

Example: uniform random variable X in $[a, b]$ with $\rho(x) = \begin{cases} \frac{1}{b-a} & x \in [a, b] \\ 0 & \text{otherwise} \end{cases}$

Probability & statistics

Assume we have an absolutely continuous random variable X with probability density function ρ. The expectation of X is defined as

$$\mathbb{E}_x[x] := \int_{\mathbb{R}} x\,\rho(x)\,dx$$

Example: uniform random variable X in $[a, b]$ with $\rho(x) = \begin{cases} \frac{1}{b-a} & x \in [a, b] \\ 0 & \text{otherwise} \end{cases}$

$$\implies \qquad \mathbb{E}_x[x] = \int_{\mathbb{R}} x\,\rho(x)\,dx = \frac{1}{b-a}\int_a^b x\,dx$$

Probability & statistics

Assume we have an absolutely continuous random variable X with probability density function ρ. The expectation of X is defined as

$$\mathbb{E}_x[x] := \int_{\mathbb{R}} x\,\rho(x)\,dx$$

Example: uniform random variable X in $[a, b]$ with $\rho(x) = \begin{cases} \frac{1}{b-a} & x \in [a, b] \\ 0 & \text{otherwise} \end{cases}$

$$\implies \quad \mathbb{E}_x[x] = \int_{\mathbb{R}} x\,\rho(x)\,dx = \frac{1}{b-a}\int_a^b x\,dx = \frac{b^2 - a^2}{2(b-a)}$$

Probability & statistics

Assume we have an absolutely continuous random variable X with probability density function ρ. The expectation of X is defined as

$$\mathbb{E}_x[x] := \int_{\mathbb{R}} x\,\rho(x)\,dx$$

Example: uniform random variable X in $[a, b]$ with $\rho(x) = \begin{cases} \frac{1}{b-a} & x \in [a, b] \\ 0 & \text{otherwise} \end{cases}$

$$\implies \quad \mathbb{E}_x[x] = \int_{\mathbb{R}} x\,\rho(x)\,dx = \frac{1}{b-a}\int_a^b x\,dx = \frac{b^2 - a^2}{2(b-a)} = \frac{(b-a)(b+a)}{2(b-a)}$$

Probability & statistics

Assume we have an absolutely continuous random variable X with probability density function ρ. The expectation of X is defined as

$$\mathbb{E}_x[x] := \int_{\mathbb{R}} x\,\rho(x)\,dx$$

Example: uniform random variable X in $[a, b]$ with $\rho(x) = \begin{cases} \frac{1}{b-a} & x \in [a, b] \\ 0 & \text{otherwise} \end{cases}$

$$\implies \quad \mathbb{E}_x[x] = \int_{\mathbb{R}} x\,\rho(x)\,dx = \frac{1}{b-a}\int_a^b x\,dx = \frac{b^2 - a^2}{2(b-a)} = \frac{(b-a)(b+a)}{2(b-a)} = \frac{a+b}{2}$$

Probability & statistics

The variance of a random variable X is defined as

$$\mathrm{Var}_x[x] := \mathbb{E}_x\left[\left(x - \mathbb{E}_x[x]\right)^2\right]$$

$$= \mathbb{E}_x[x^2] - \mathbb{E}_x[x]^2$$

Probability & statistics

The variance of a random variable X is defined as

$$\mathrm{Var}_x[x] := \mathbb{E}_x\left[\left(x - \mathbb{E}_x[x]\right)^2\right]$$

$$= \mathbb{E}_x[x^2] - \mathbb{E}_x[x]^2$$

Its square-root

$$\sigma_x := \sqrt{\mathrm{Var}_x[x]}$$

is known as standard deviation

Probability & statistics

Two random variables X and Y are independent if their joint PDF factors, i.e.

$$\rho(x, y) = \rho_X(x)\,\rho_Y(y)$$

Probability & statistics

Two random variables X and Y are independent if their joint PDF factors, i.e.

$$\rho(x, y) = \rho_X(x)\,\rho_Y(y)$$

An arbitrary no. of n random variables $\{X_i\}_{i=1}^n$ is independent if

$$\rho(x_1, \ldots, x_n) = \prod_{i=1}^{n} \rho_{X_i}(x_i)$$

Probability & statistics

Two random variables X and Y are independent if their joint PDF factors, i.e.

$$\rho(x, y) = \rho_X(x)\,\rho_Y(y)$$

An arbitrary no. of n random variables $\{X_i\}_{i=1}^{n}$ is independent if

$$\rho(x_1, \ldots, x_n) = \prod_{i=1}^{n} \rho_{X_i}(x_i)$$

The collection of random variables is independent and identically distributed (i.i.d.) if in addition we have

$$\rho_{X_1} = \rho_{X_2} = \cdots = \rho_{X_n}$$