COMP5212: Machine Learning

Lecture 1

Minhao Cheng

Term project
Details

- Group of at most 4 students
- Open research projects
- Term project report + term project presentation (online/offline)

Math Basics

Math Basics

Linear Algebra

- Vector, matrix, tensor, inverse
- Norms: measure the size of a vector
  - l_p norm: $\|x\|_p = (\sum_i |x_i|^p)^{1/p}$ $\|x\|_2^2 = x^T x$ also equal to Euclidean distance
  - Frobenius Norm: $\|A\|_F = \sqrt{\sum_{ij} A_{ij}^2}$
  - $x^T y = \|x\|_2 \|y\|_2 cos\theta$
    - Projection
- Trace: $tr(A) = \sum_i A_{ii}$
  - $\|A\|_F = \sqrt{tr(AA^T)}$
  - $a = tr(a), tr(A^T) = tr(A), tr(A \pm B) = tr(A) \pm tr(B), tr(ABC) = tr(CAB) = tr(BCA)$

Math Basics
Linear Algebra

- Linear dependence, span
- Orthogonal, orthonormal,
- Eigendecomposition, quadratic form
  - $f(x) = x^T Ax, s.t \ \|x\|_2 = 1$
- Positive definite: all eigenvalues are positive, positive semidefinite are all positive or zero
  - $\forall x, \ x^T Ax \geq 0$
- Singular Value Decomposition (SVD)
  - $A = UDV^T$, where A is $m \times n$ matrix, U is $m \times m$ matrix, V is $n \times n$ vector

Math Basics

Derivates

- Derivative, chain rule
  - Given a composite function $f(x) = h(g(x))$
  - $\dfrac{df}{dx} = \dfrac{dh}{dg} \cdot \dfrac{dg}{dx}$
- Integral

Math Basics

Matrix Derivates

- Scalar to vector: f is a scalar, $x = [x_1\ x_2 \dots\ x_p]^T$ is a $p \times 1$ vector, then
  - $\dfrac{\partial f}{\partial x} = [\dfrac{\partial f}{\partial x_1}\ \dfrac{\partial f}{\partial x_2}\ \dots\ \dfrac{\partial f}{\partial x_p}]^T$
- Vector to scalar: $f = [f_1\ f_2 \dots\ f_m]^T$ is a $m \times 1$ vector, x is a scalar, then
  - $\dfrac{\partial f}{\partial x} = [\dfrac{\partial f_1}{\partial x}\ \dfrac{\partial f_2}{\partial x}\ \dots\ \dfrac{\partial f_m}{\partial x}]$

Math Basics
Matrix Derivates

- Vector to vector: $f = [f_1\ f_2 \dots\ f_m]^T$ is a $m \times 1$ vector, $x = [x_1\ x_2 \dots\ x_p]^T$ is a $p \times 1$ vector, then
  - $$\frac{\partial f}{\partial x} = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_2}{\partial x_1} & \cdots & \frac{\partial f_m}{\partial x_1} \\ \frac{\partial f_1}{\partial x_2} & \frac{\partial f_2}{\partial x_2} & \cdots & \frac{\partial f_m}{\partial x_2} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_1}{\partial x_p} & \frac{\partial f_2}{\partial x_p} & \cdots & \frac{\partial f_m}{\partial x_p} \end{bmatrix}$$
- Scalar to matrix: f is a scalar, X is a $p \times q$ matrix, then
  - $$\frac{\partial f}{\partial X} = \begin{bmatrix} \frac{\partial f}{\partial X_{11}} & \frac{\partial f}{\partial X_{12}} & \cdots & \frac{\partial f}{\partial X_{1q}} \\ \frac{\partial f}{\partial X_{21}} & \frac{\partial f}{\partial X_{22}} & \cdots & \frac{\partial f}{\partial X_{2q}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f}{\partial X_{p1}} & \frac{\partial f}{\partial X_{p2}} & \cdots & \frac{\partial f}{\partial X_{pq}} \end{bmatrix}$$

Math Basics

Matrix Derivates

- Matrix to scalar: F is a $p \times q$ matrix, x is a scalar, then
- $$\frac{\partial F}{\partial x} = \begin{bmatrix} \frac{\partial F_{11}}{\partial x} & \frac{\partial F_{21}}{\partial x} & \cdots & \frac{\partial F_{m1}}{\partial x} \\ \frac{\partial F_{21}}{\partial x} & \frac{\partial F_{22}}{\partial x} & \cdots & \frac{\partial F_{m2}}{\partial x} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial F_{n1}}{\partial x} & \frac{\partial F_{n2}}{\partial x} & \cdots & \frac{\partial F_{nm}}{\partial x} \end{bmatrix}$$

Math Basics

Matrix Derivates

- In the vector view:
  - Scalar to vector: $df = \sum_{i=1}^{n} \frac{\partial f}{\partial x_i} dx_i = \frac{\partial f}{\partial x}^T dx$ where $\frac{\partial f}{\partial x}$ and dx are $n \times 1$ vector
  - Similarly, scalar to matrix: $df = \sum_{i=1}^{m} \sum_{j=1}^{n} \frac{\partial f}{\partial X_{ij}} dX_{ij} = tr(\frac{\partial f}{\partial X}^T dX)$
  - For the derivate, we also have $d(X \pm Y) = dX \pm dY$, $d(XY) = (dX)Y + XdY$, $d(X^T) = (dX)^T$, $dtr(X) = tr(dX)$, $dX^{-1} = -X^{-1}dXX^{-1}$
  - For the trace operation, we also have $a = tr(a)$, , $tr(A \pm B) = tr(A) \pm tr(B)$, $tr(AB) = tr(BA)$, $tr(A^T(B \odot C)) = tr((A \odot B)^T C)$

Math Basics

Matrix Derivates

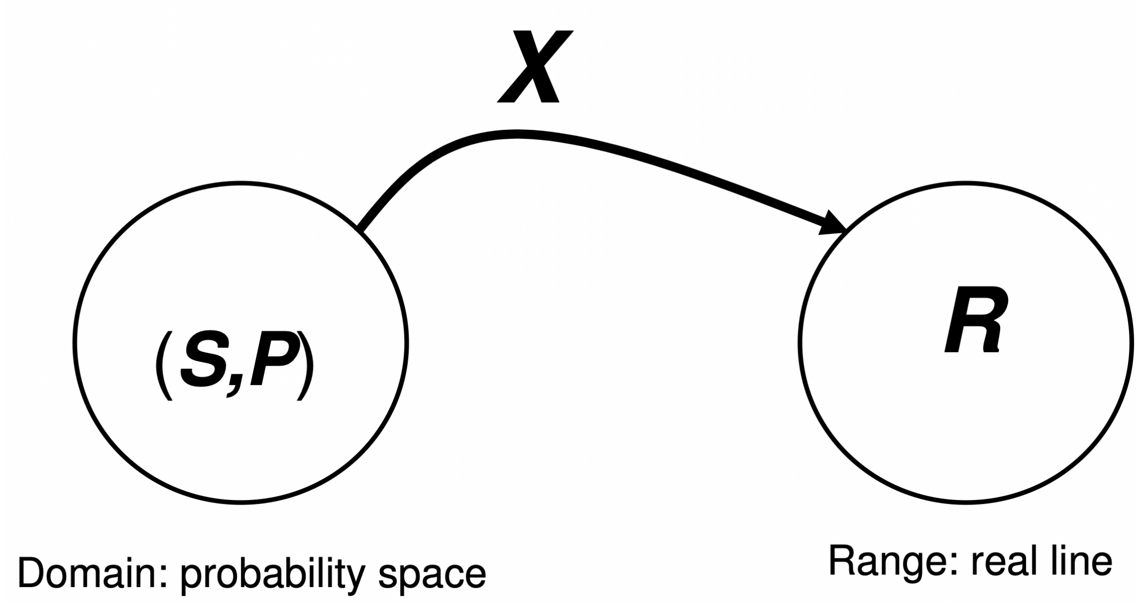
- Chain rule: f is a function of Y, let Y=AXB, to get $\frac{\partial f}{\partial X}$
  - $df = tr(\frac{\partial f}{\partial Y}^T dY) = tr(\frac{\partial f}{\partial Y}^T AdXB) = tr(B\frac{\partial f}{\partial Y}^T AdX) = tr((A^T\frac{\partial f}{\partial Y}B^T)^T dX)$
    - Since $dY = d(A)XB + AdXB + AXdB = AdXB$ as $dA = 0, dB = 0$
  - So we get $\frac{\partial f}{\partial X} = A^T\frac{\partial f}{\partial Y}B^T$

Math Basics

Matrix Derivates

- Ex 1: $f = a^T X b$, solve $\dfrac{\partial f}{\partial X}$, where a is $m \times 1$ vector, X is $m \times n$ matrix, b is $n \times 1$ vector
- Ex 2: $f = a^T exp(Xb)$, solve $\dfrac{\partial f}{\partial X}$, where a is $m \times 1$ vector, X is $m \times n$ matrix, b is $n \times 1$ vector
- Ex 3: $f = \|Xw - y\|^2$, solve $\dfrac{\partial f}{\partial w}$, where y is $m \times 1$ vector, X is $m \times n$ matrix, w is $n \times 1$ vector
-

Math Basics
Probability



- Random variable: a **function** mapping a probability space (S, P) into a real line $\mathbb{R}$
  - Discrete variable, Probability mass function (PMF)
    - PMF maps a state of a random variable to the probability of the random variable taking on that state
    - $P(\mathrm{x} = x_i) = \dfrac{1}{k}$
  - Continuous variable, Probability density function (PDF)

Math Basics

Probability

- Marginal Probability
  - For discrete random variable x and y, and we know $P(\mathrm{x}, \mathrm{y})$, we can find $\forall x \in \mathrm{x}, P(\mathrm{x} = x) = \sum_y P(\mathrm{x} = x, \mathrm{y} = y)$
  - For continuous …, $p(x) = \int p(x, y) dy$
- Conditional Probability
  - $P(\mathrm{y} = y \mid \mathrm{x} = x) = \dfrac{P(\mathrm{y} = y, \mathrm{x} = x)}{P(\mathrm{x} = x)}$

Math Basics

Probability

- Chain rule
    - $P(x^{(1)}, \ldots, x^{(n)}) = P(x^{(1)}) \prod_{i=2}^{n} P(x^{(i)} \mid x^{(1)}, \ldots, x^{(i-1)})$
- Independence, conditional independence
    - $\forall x \in \mathrm{x}, y \in \mathrm{y}, p(\mathrm{x} = x, \mathrm{y} = y) = p(\mathrm{x} = x)p(\mathrm{y} = y)$
    - $\forall x \in \mathrm{x}, y \in \mathrm{y}, z \in \mathrm{z}, p(\mathrm{x} = x, \mathrm{y} = y, \mathrm{z} = z) = p(\mathrm{x} = x \mid \mathrm{z} = z)p(\mathrm{y} = y \mid \mathrm{z} = z)$
- Exception, Variance, Covariance

Math Basics

Probability

- Exception
  - Discrete: $\mathbb{E}_{\mathrm{x}\sim P}[f(x)] = \sum_x P(x)f(x)$, Continuous: $\mathbb{E}_{\mathrm{x}\sim p}[f(x)] = \int p(x)f(x)dx$
- Variance
  - $Var(f(x)) = \mathbb{E}[(f(x) - \mathbb{E}[f(x)])^2]$
- Covariance
  - $Cov(f(x), g(y)) = \mathbb{E}[(f(x) - \mathbb{E}[f(x)])(g(y) - \mathbb{E}[g(y)])]$

Math Basics

Probability

- Common probability distribution
  - Bernoulli distribution:
    - $P(\mathrm{x}=1)=\phi$, $P(\mathrm{x}=0)=1-\phi$, $P(\mathrm{x}=x)=\phi^x(1-\phi)^{1-x}$, $\mathbb{E}_{\mathrm{x}}[\mathrm{x}]=\phi$, $Var_{\mathrm{x}}[\mathrm{x}]=\phi(1-\phi)$
  - Multinoulli distribution
  - Gaussian distribution
    - $\mathcal{N}(x;\mu,\sigma^2)=\sqrt{\frac{1}{2\pi\sigma^2}}exp(-\frac{1}{2\sigma^2}(x-\mu)^2)$
    - Multivariate normal distribution: $\mathcal{N}(x;\mu,\Sigma)=\sqrt{\frac{1}{(2\pi)^n det(\Sigma)}}exp(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu))$
  - Exponential distribution
    - $p(x;\lambda)=\lambda exp(-\lambda x)$
  - Dirac distribution
    - Dirac delta function: It is zero valued everywhere except 0, yet integrates to 1

Math Basics

Probability

- Mixtures of distribution
  - $$P(x) = \sum_{i} \underbrace{P(c = i)}_{\text{Prior}} P(\mathrm{x} \mid c = i)$$
  - Gaussian Mixture: $p(\mathrm{x} \mid c = i)$ are Gaussians with a separately parameterized mean and covariance
- Bayes rule
  - $$p(\mathrm{x} \mid \mathrm{y}) = \frac{P(\mathrm{x})P(\mathrm{y} \mid \mathrm{x})}{P(\mathrm{y})}$$

Math Basics

Some useful function

- Logistic sigmoid
  - $\sigma(x) = \dfrac{1}{1 + \text{exp}(-x)}$
  - Useful property:
    - $\sigma(x) = \dfrac{\text{exp}(x)}{\text{exp}(x) + \text{exp}(0)}$
    - $\dfrac{d}{dx}\sigma(x) = \sigma(x)(1 - \sigma(x))$
    - $1 - \sigma(x) = \sigma(-x)$
- ReLU
  - $x^+ = \max(0,x)$