Math Recap

Jingwei Tang @ CGL

19.02.2020

jingwei.tang@inf.ethz.ch

- **Bottom-up**: Building up concepts from foundational to more advanced. This strategy has the advantage that the reader at all times is able to rely on their previously learned concepts.

- **Top-down:** Drilling down from practical needs to more basic requirements. This goal-driven approach has the advantage that the readers know at all times why they need to work on a particular concept.

-- <Mathematics for Machine Learning>

- **Bottom-up**: Building up concepts from foundational to more advanced. This strategy has the advantage that the reader at all times is able to rely on their previously learned concepts.

- **Top-down:** Drilling down from practical needs to more basic requirements. This goal-driven approach has the advantage that the readers know at all times why they need to work on a particular concept.

-- <Mathematics for Machine Learning>

1. **Linear Algebra**
2. **(Brief) Vector Calculus**
3. **Probability Theory**

Matrix Multiplications

- Let matrix $A \in \mathbb{R}^{m \times p}$ and $B \in \mathbb{R}^{p \times n}$, matrix-matrix multiplication can be defined as

$$C_{ij} = \sum_{k=1}^{p} A_{ik} B_{kj}$$

  the result matrix $C \in \mathbb{R}^{m \times n}$

- Vectors can be viewed as matrices: $x \in \mathbb{R}^{p \times 1}$

$$y_i = \sum_{k=1}^{p} A_{ik} x_k$$

Matrix Multiplications

- <u>Example</u>: matrix-matrix multiplication:

$$\begin{bmatrix} 1 & 2 & 3 \\ 3 & 2 & 1 \end{bmatrix} \begin{bmatrix} 0 & 2 \\ 1 & -1 \\ 0 & 1 \end{bmatrix} =$$

- <u>Example</u>: matrix-vector multiplication:

$$\begin{bmatrix} 0 & 2 \\ 1 & -1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 2 \\ 3 \end{bmatrix} =$$

Matrix Multiplications

$\forall A \in \mathbb{R}^{m\times n}, B \in \mathbb{R}^{n\times p}, C, D \in \mathbb{R}^{p\times q}$

- Associativity: $(AB)C = A(BC)$
- Distributivity: $(A + B)C = AC + BC$

  $A(C + D) = AC + AD$
- **NO** Commutativity: $AB \neq BA$

Linear Systems

- System of linear equations

$$a_{11}x_1 + \cdots + a_{1n}x_n = b_1$$
$$\dots\dots$$
$$a_{m1}x_1 + \cdots + a_{mn}x_n = b_m$$

can be represented through matrix-vector multiplication:

$$\begin{bmatrix} a_{11} & \dots & a_{1n} \\ \vdots & & \vdots \\ a_{m1} & \dots & a_{mn} \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} b_1 \\ \vdots \\ b_m \end{bmatrix}$$

Linear Systems $Ax = b$

$A \in \mathbb{R}^{m \times n}, x \in \mathbb{R}^n, b \in \mathbb{R}^m$: m equations, n variables

- $m = n$: Square Systems
  - Can have $0, 1, \infty$ solution(s).
- $m < n$: Underdetermined Systems
  - Typically have ∞ solutions.
- $m > n$: Overdetermined Systems
  - Linear Regression: least-squares solution ($\min||Ax - b||^2$)

Linear Systems $Ax = b$

Square system <u>examples</u>:

- $\begin{matrix} 2x + 3y = 5 \\ x + y = 3 \end{matrix} \Rightarrow \begin{matrix} x = 4 \\ y = -1 \end{matrix}$
- $\begin{matrix} 2x + 3y = 5 \\ 4x + 6y = 5 \end{matrix} \Rightarrow$ NO solutions
- $\begin{matrix} 2x + 3y = 5 \\ 4x + 6y = 10 \end{matrix} \Rightarrow x = \frac{5-3y}{2}, \forall y \in \mathbb{R}$

Square system $Ax = b$ has an **unique** solution.

$\Leftrightarrow$

A is **invertible**.

$\Leftrightarrow$

$Ax = 0$ only has trivial solution $x = 0$.

Invertibility and Determinant

- Matrix $A \in \mathbb{R}^{n\times n}$ is called **invertible** if there exists $B \in \mathbb{R}^{n\times n}$ s.t. $AB = I = BA$, B is then called the inverse of A, $B = A^{-1}$.
- $A \in \mathbb{R}^{n\times n}$ is **invertible** or **nonsingular** if and only if it is square and full rank. Equivalently, having $\det(A) \neq 0$.
- $\det(A) : \mathbb{R}^{n\times n} \to \mathbb{R}$

  e.g. $\det \begin{bmatrix} a & b \\ c & d \end{bmatrix} = \begin{vmatrix} a & b \\ c & d \end{vmatrix} = ad - cb$

$$\begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix} = a_{11} \begin{vmatrix} a_{22} & a_{23} \\ a_{32} & a_{33} \end{vmatrix} - a_{12} \begin{vmatrix} a_{21} & a_{23} \\ a_{31} & a_{33} \end{vmatrix} + a_{13} \begin{vmatrix} a_{21} & a_{22} \\ a_{31} & a_{32} \end{vmatrix}$$

Eigenvalues and Eigenvectors

- Let $A \in \mathbb{R}^{n \times n}$ be a square matrix. $\lambda \in \mathbb{R}$ is an **eigenvalue** of A and $x \in \mathbb{R}^n \setminus \{0\}$ is the corresponding **eigenvector** if

$$Ax = \lambda x$$

$\Leftrightarrow (A - \lambda I_n)x = 0$ has solutions other than $x = 0$.

$\Leftrightarrow det(A - \lambda I_n) = 0$. (Polynomial of degree n)

A is **invertible**.
Equivalently, $\det(A) \neq 0$

$\Leftrightarrow$

$Ax = 0$ only has trivial solution $x = 0$.

Eigenvalues and Eigenvectors

<u>Example</u>: find the eigenvalues and eigenvectors of

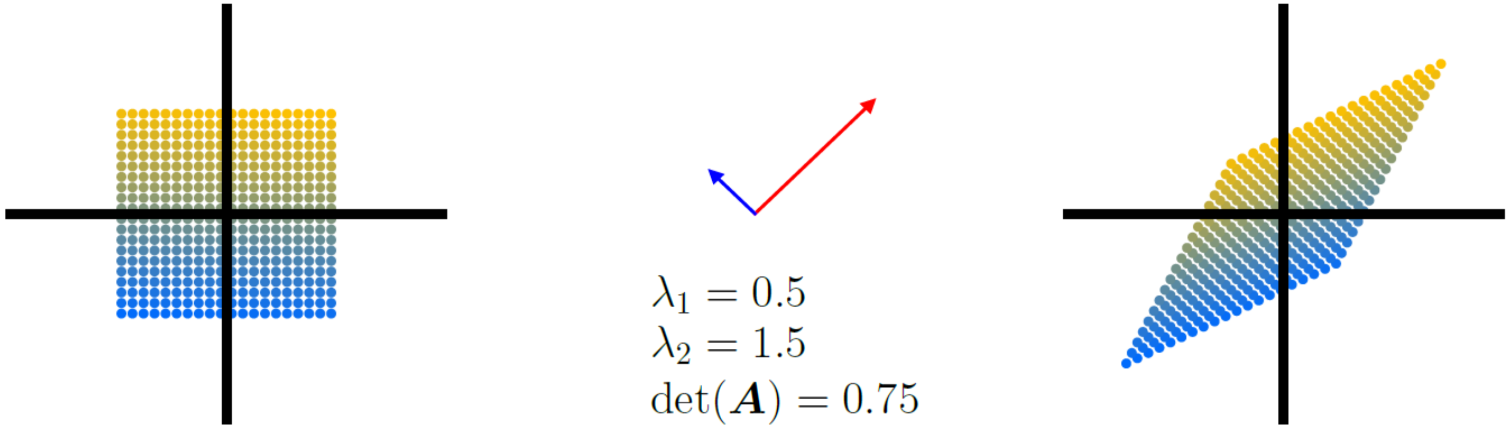
$$A = \begin{bmatrix} 1 & \frac{1}{2} \\ \frac{1}{2} & 1 \end{bmatrix}$$

Eigenvalues and Eigenvectors

Solution:

$$\lambda_1 = \frac{1}{2}, \qquad E_{\lambda_1} = \operatorname{span}\left\{\begin{bmatrix} 1 \\ -1 \end{bmatrix}\right\},$$

$$\lambda_2 = \frac{3}{2}, \qquad E_{\lambda_2} = \operatorname{span}\left\{\begin{bmatrix} 1 \\ 1 \end{bmatrix}\right\}.$$



Eigendecomposition

$$A \begin{bmatrix} | & & | \\ x_1 & \cdots & x_n \\ | & & | \end{bmatrix} = \begin{bmatrix} | & & | \\ x_1 & \cdots & x_n \\ | & & | \end{bmatrix} \begin{bmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_n \end{bmatrix}$$

$$AP = PD$$

$$\Leftrightarrow$$

$A = PDP^{-1}$ (if and only if eigenvectors of A form a basis of $\mathbb{R}^n$)

Eigendecomposition

Example: find the Eigendecomposition of

$$A = \begin{bmatrix} 1 & \frac{1}{2} \\ \frac{1}{2} & 1 \end{bmatrix}$$

$$\lambda_1 = \frac{1}{2}, \qquad E_{\lambda_1} = \mathrm{span}\left\{\begin{bmatrix} 1 \\ -1 \end{bmatrix}\right\},$$

$$\lambda_2 = \frac{3}{2}, \qquad E_{\lambda_2} = \mathrm{span}\left\{\begin{bmatrix} 1 \\ 1 \end{bmatrix}\right\}.$$

Eigendecomposition

$$A = PDP^{-1}$$

$$D = \begin{bmatrix} \frac{1}{2} & 0 \\ 0 & \frac{3}{2} \end{bmatrix}, \quad P = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix}, \quad P^{-1} = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix}$$

Optional, to form a set of (nice) normalized basis

Matrix Decompositions

- **Eigendecomposition**: $A \in \mathbb{R}^{n\times n}$, eigenvectors of A form a basis of $\mathbb{R}^n$

$$A = PDP^{-1}$$

- QR/**QU** Decomposition (from Gram-Schmidt process): $A \in \mathbb{R}^{n\times n}$

$$A = QU$$

- **LU** Decomposition (from Gaussian Elimination): $A \in \mathbb{R}^{m\times n}$

$$A = LU$$

- Singular Value Decomposition (**SVD**):

$A \in \mathbb{R}^{m\times n}, U \in \mathbb{R}^{m\times m}, V \in \mathbb{R}^{n\times n}. \Sigma \in \mathbb{R}^{m\times n}$ is a diagonal matrix.

$$A = U\Sigma V^T$$

- **Cholesky** Decomposition: $A \in \mathbb{R}^{n\times n}$, symmetric and positive definite

$$A = LL^T$$

Positive Definiteness of Matrices

- Symmetric: $A \in \mathbb{R}^{n \times n}$ is symmetric if $A_{ij} = A_{ji}, \forall i, j \in [1, n]$.

  e.g. $\begin{bmatrix} 1 & 2 \\ 2 & 0 \end{bmatrix}$, $\begin{bmatrix} 1 & 2 & 3 \\ 2 & -1 & 5 \\ 3 & 5 & 0 \end{bmatrix}$

- Symmetric Positive Definite: A Symmetric matrix $A \in \mathbb{R}^{n \times n}$ is called **symmetric, positive definite** if

$$\forall x \in \mathbb{R}^n \setminus \{0\}: \quad x^T A x > 0$$

  If only $\geq$ holds, A is called **symmetric, positive semidefinite**.

Positive Definiteness of Matrices

<u>Example</u>: find out whether the following matrix is symmetric positive definite

$$A = \begin{bmatrix} 9 & 6 \\ 6 & 5 \end{bmatrix}$$

1. **Linear Algebra**

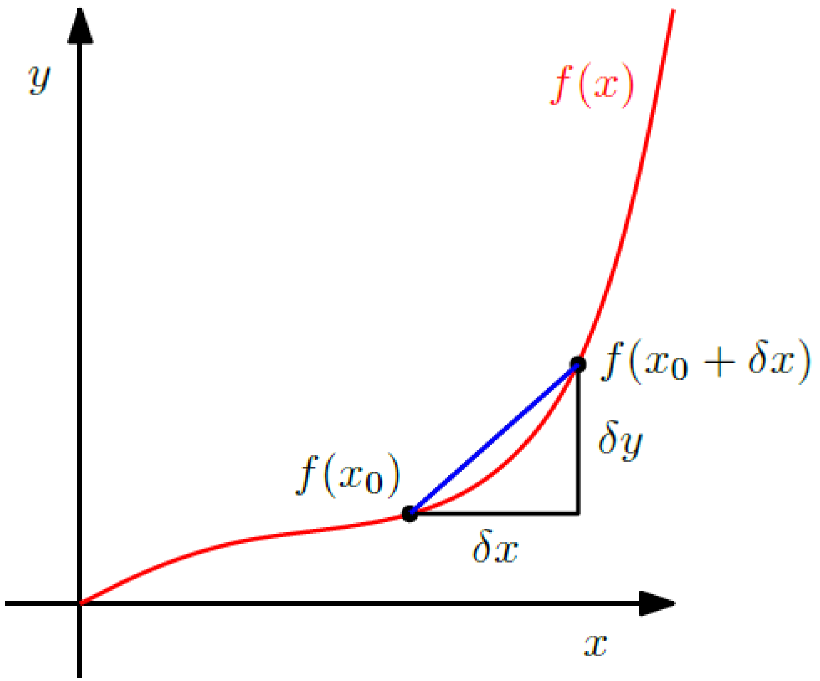
2. **(Brief) Vector Calculus**

3. **Probability Theory**

Notion of Derivatives

<u>Derivative</u>: Let $f: \mathbb{R} \to \mathbb{R},\ x \to f(x)$, the **derivative** is defined as:

$$\frac{df}{dx} := \lim_{\delta x \to 0} \frac{f(x + \delta x) - f(x)}{\delta x}$$



Notion of Derivatives

Partial Derivative: Let $f: \mathbb{R}^n \to \mathbb{R}, \boldsymbol{x} \to f(\boldsymbol{x}), \boldsymbol{x} \in \mathbb{R}^n$ of n variables, the **partial derivative** is defined as:

$$\frac{\partial f}{\partial x_i} := \lim_{\delta x \to 0} \frac{f(x_1, \ldots, x_i + \delta x, \ldots x_n) - f(x_1, \ldots, x_i, \ldots x_n)}{\delta x}$$

Gradient: Collect partial derivatives of all variables and form a row vector

$$\nabla f = \mathrm{grad} f = \frac{df}{d\boldsymbol{x}} = \begin{bmatrix} \frac{\partial f}{\partial x_1} & \frac{\partial f}{\partial x_2} & \ldots & \frac{\partial f}{\partial x_n} \end{bmatrix} \in \mathbb{R}^{1 \times n}$$

Notion of Derivatives

Jacobian: Let $f: \mathbb{R}^n \to \mathbb{R}^m, \boldsymbol{x} \to \boldsymbol{f}(\boldsymbol{x}), \boldsymbol{x} \in \mathbb{R}^n, \boldsymbol{f}(\boldsymbol{x}) \in \mathbb{R}^m$, stacking all the gradient of components of $\boldsymbol{f}(\boldsymbol{x})$ into a matrix:

$$J = \frac{d\boldsymbol{f}(\boldsymbol{x})}{d\boldsymbol{x}} = \begin{bmatrix} \frac{df_1(\boldsymbol{x})}{d\boldsymbol{x}} \\ \vdots \\ \frac{df_m(\boldsymbol{x})}{d\boldsymbol{x}} \end{bmatrix} = \begin{bmatrix} \frac{df_1(\boldsymbol{x})}{dx_1} & \cdots & \frac{df_1(\boldsymbol{x})}{dx_n} \\ \vdots & & \vdots \\ \frac{df_m(\boldsymbol{x})}{dx_1} & \cdots & \frac{df_m(\boldsymbol{x})}{dx_n} \end{bmatrix} \in \mathbb{R}^{m \times n}$$

Notion of Derivatives

- Derivative: $f: \mathbb{R} \to \mathbb{R}, \quad \frac{df}{dx} \in \mathbb{R}$

- Gradient: $f: \mathbb{R}^n \to \mathbb{R}, \quad \nabla f \in \mathbb{R}^{1 \times n}$

- Jacobian: $f: \mathbb{R}^n \to \mathbb{R}^m, \quad J \in \mathbb{R}^{m \times n}$

Chain Rule

- Real-valued functions: $f, g: \mathbb{R} \to \mathbb{R}, x \to f(x), x \to g(x)$

$$\frac{dg(f(x))}{dx} = \frac{dg(f(x))}{df(x)} \frac{df(x)}{dx}$$

- Multi-variable functions: $\boldsymbol{f}: \mathbb{R}^n \to \mathbb{R}^m, \boldsymbol{x} \to \boldsymbol{f}(\boldsymbol{x}), g: \mathbb{R}^m \to \mathbb{R}, \boldsymbol{x} \to g(\boldsymbol{x})$

$$\frac{dg(f(\boldsymbol{x}))}{d\boldsymbol{x}} = \frac{dg(\boldsymbol{f}(\boldsymbol{x}))}{d\boldsymbol{f}(\boldsymbol{x})} \frac{d\boldsymbol{f}(\boldsymbol{x})}{d\boldsymbol{x}}$$

$1 \times n$ $\quad$ $1 \times m$ $\quad$ $m \times n$

More Complicated Derivatives

- Derivative of Matrices
- Higher-order Derivatives

[See references]

Matrix Cookbook

1 Basics

$$
\begin{aligned}
(\mathbf{AB})^{-1} &= \mathbf{B}^{-1}\mathbf{A}^{-1} && (1)\\
(\mathbf{ABC}...)^{-1} &= ...\mathbf{C}^{-1}\mathbf{B}^{-1}\mathbf{A}^{-1} && (2)\\
(\mathbf{A}^T)^{-1} &= (\mathbf{A}^{-1})^T && (3)\\
(\mathbf{A}+\mathbf{B})^T &= \mathbf{A}^T+\mathbf{B}^T && (4)\\
(\mathbf{AB})^T &= \mathbf{B}^T\mathbf{A}^T && (5)\\
(\mathbf{ABC}...)^T &= ...\mathbf{C}^T\mathbf{B}^T\mathbf{A}^T && (6)\\
(\mathbf{A}^H)^{-1} &= (\mathbf{A}^{-1})^H && (7)\\
(\mathbf{A}+\mathbf{B})^H &= \mathbf{A}^H+\mathbf{B}^H && (8)\\
(\mathbf{AB})^H &= \mathbf{B}^H\mathbf{A}^H && (9)\\
(\mathbf{ABC}...)^H &= ...\mathbf{C}^H\mathbf{B}^H\mathbf{A}^H && (10)
\end{aligned}
$$

2.4 Derivatives of Matrices, Vectors and Scalar Forms

2.4.1 First Order

$$
\begin{aligned}
\frac{\partial \mathbf{x}^T\mathbf{a}}{\partial \mathbf{x}} &= \frac{\partial \mathbf{a}^T\mathbf{x}}{\partial \mathbf{x}} = \mathbf{a} && (69)\\
\frac{\partial \mathbf{a}^T\mathbf{X}\mathbf{b}}{\partial \mathbf{X}} &= \mathbf{a}\mathbf{b}^T && (70)\\
\frac{\partial \mathbf{a}^T\mathbf{X}^T\mathbf{b}}{\partial \mathbf{X}} &= \mathbf{b}\mathbf{a}^T && (71)\\
\frac{\partial \mathbf{a}^T\mathbf{X}\mathbf{a}}{\partial \mathbf{X}} &= \frac{\partial \mathbf{a}^T\mathbf{X}^T\mathbf{a}}{\partial \mathbf{X}} = \mathbf{a}\mathbf{a}^T && (72)\\
\frac{\partial \mathbf{X}}{\partial X_{ij}} &= \mathbf{J}^{ij} && (73)\\
\frac{\partial (\mathbf{XA})_{ij}}{\partial X_{mn}} &= \delta_{im}(\mathbf{A})_{nj} = (\mathbf{J}^{mn}\mathbf{A})_{ij} && (74)\\
\frac{\partial (\mathbf{X}^T\mathbf{A})_{ij}}{\partial X_{mn}} &= \delta_{in}(\mathbf{A})_{mj} = (\mathbf{J}^{nm}\mathbf{A})_{ij} && (75)
\end{aligned}
$$

1. Linear Algebra
2. (Brief) Vector Calculus
3. **Probability Theory**

Probability Space $(\Omega, \mathcal{A}, P)$

- Sample Space Ω: set of all possible outcomes of an experiment

- Event Space $\mathcal{A}$: space of potential results of the experiment (a collection of all subsets of Ω in discrete setting)

- Probability P: with each event $A \in \mathcal{A}$, we associate a number $P(A)$ that measures the 'degree of belief' that the event will occur.

Probability Space $(\Omega, \mathcal{A}, P)$

- Sample Space Ω: set of all possible outcomes of an experiment

- Event Space $\mathcal{A}$: space of potential results of the experiment (a collection of all subsets of Ω in discrete setting)

- Probability P: with each event $A \in \mathcal{A}$, we associate a number $P(A)$ that measures the 'degree of belief' that the event will occur.

- Random Variable X: A function/mapping $X: \Omega \rightarrow \mathcal{T}$. We are interested in the probabilities on elements of $\mathcal{T}$.

Probability Space $(\Omega, \mathcal{A}, P)$

Example: tossing coins

- Experiment: tossing coins for two consecutive times.
- Sample Space: $\Omega = \{hh, tt, ht, th\}$ (h for head and t for tails)
- Random Variable: X maps the event to number of heads. $\mathcal{T} = \{0,1,2\}$.

$$X(hh) = 2, X(tt) = 0, X(ht) = X(th) = 1$$

- Probabilities (on $\mathcal{T}$):

$$P(X = 0) = 0.25, P(X = 1) = 0.5, P(x = 2) = 0.25$$

PDF and CDF

- Probability Density Function (PDF):

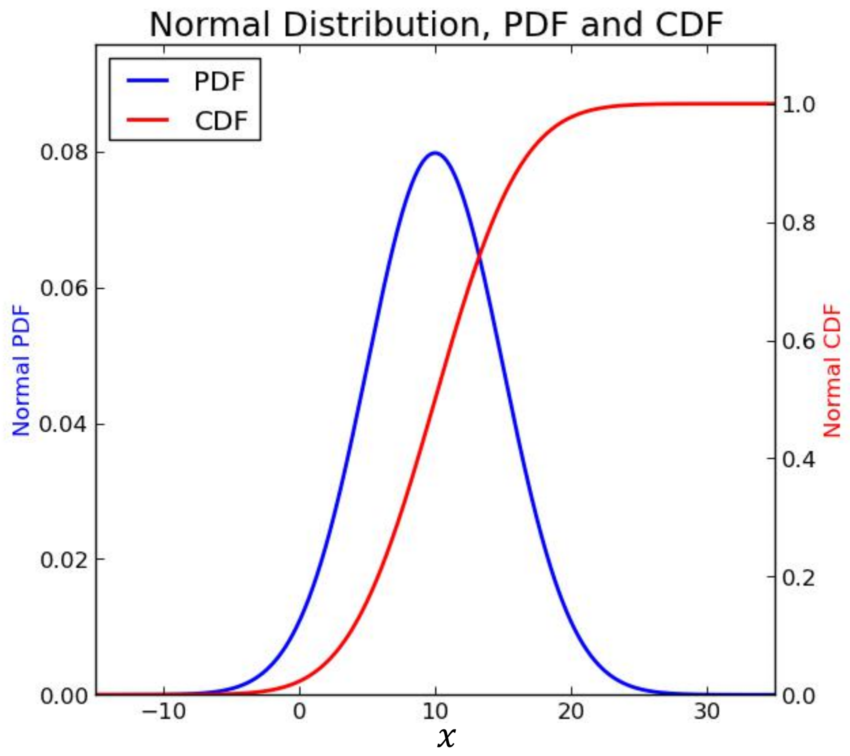
$$f: \mathbb{R} \rightarrow \mathbb{R} \text{ s.t. } \forall \boldsymbol{x} \in \mathbb{R}, f(x) \geq 0 \text{ and}$$
$$\int_{\mathbb{R}} f(x)dx = 1$$

We can associate a random variable X with PDF:

$$P(a \leq X \leq b) = \int_a^b f(x)dx$$

- Cumulative Distribution Function(CDF):

$$F_X(x) = P(X \leq x)$$
$$F_X(x) = \int_{-\infty}^{x} f(z)dz\,, \qquad f(x) = \frac{dF_X(x)}{dx}$$



Joint Distribution

- Let X, Y be two random variables over the same probability space. Joint distribution is defined as

$$F_{X,Y}(x, y) = P(X \leq x, Y \leq y)$$

joint density:

$$f(x, y) = \frac{\partial^2 F_{X,Y}(x, y)}{\partial x \partial y}$$

- Marginalization:

$$f_X(x) = \int_{-\infty}^{\infty} f(x, y) dy, \qquad F_X(x) = F_{X,Y}(x, \infty)$$

Independence

- Two events A and B are independent if

$$P(A \cap B) = P(A)P(B)$$

- Two Random Variables X and Y are independent if their joint distribution function factorizes, i.e.

$$F_{X,Y}(x, y) = F_X(x)F_Y(y)$$

Conditional Probability

Probability of A given B has occurred:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Can also be described through PDF and CDF

- Laws regarding conditional probability:
  - Law of total probability: $P(B) = \sum_{i=1}^{n} P(B|A_i)P(A_i)$
  - Bayes Rule: $P(A|B) = P(B|A)\frac{P(A)}{P(B)}$
  - Chain Rule: $P(A_1, \ldots A_n) = P(A_1)P(A_2|A_1)P(A_3|A_1, A_2)\cdots P(A_n|A_1, \ldots, A_{n-1})$

[See references]

Expectation

- The **expected value** of a function $g: \mathbb{R} \to \mathbb{R}$ of a univariate continuous random variable $X \sim p(x)$ is given by

$$\mathbb{E}_X[g(x)] = \int_{\mathcal{X}} g(x)p(x)dx$$

- The **mean** of a random variable X is defined as

$$\mathbb{E}_X[x] = \int_{\mathcal{X}} xp(x)dx$$

- Linearity:

$$\mathbb{E}_X[af(x) + bg(x)] = a\mathbb{E}_X[f(x)] + b\mathbb{E}_X[g(x)]$$

(Co)variance

- Covariance between two univariate random variables $X, Y \in \mathbb{R}$:
  $Cov_{X,Y}[x, y] = \mathbb{E}_{X,Y}[(x - \mathbb{E}_X[x])(y - \mathbb{E}_Y[y])] = \mathbb{E}[xy] - \mathbb{E}[x]\mathbb{E}[y]$

- Variance is the covariance with itself:
$$\mathbb{V}[x] = Cov_{X,X}[x, x] = \mathbb{E}_X[(x - \mathbb{E}_X[x])^2] = \mathbb{E}_X[x^2] - \mathbb{E}_X[x]^2$$

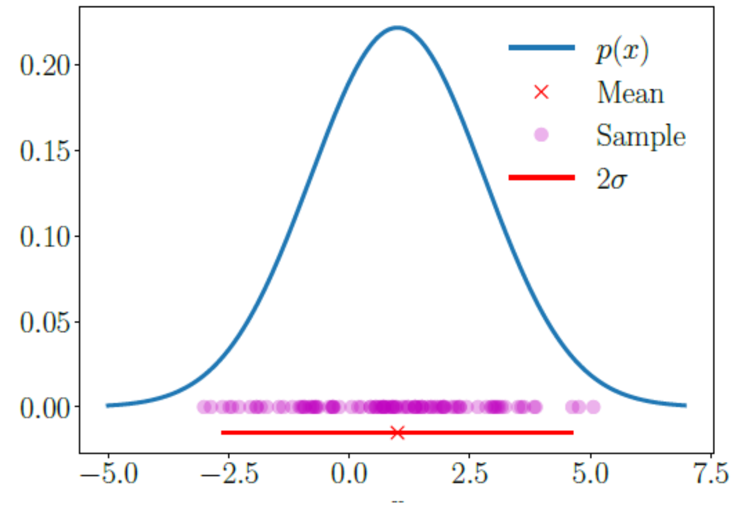
- NOT Linear:
$$\mathbb{V}[x + y] = \mathbb{V}[x] + \mathbb{V}[y] + Cov[x, y] + Cov[y, x]$$

Guassian (Normal) Distribution

- The Gaussian distribution has a density

$$p(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

Denoted as $X \sim N(x|\mu,\sigma^2)$



PDF, CDF, Joint distribution, Expectation, Covariance, Gaussian Distribution can all be **extended** with some efforts to **higher dimensions.**

[see references]

References

- Mathematics for Machine Learning
- Matrix Cookbook

End of Presentation

Start of Q&A Session