



《模式识别》

第一章 课程简介与预备知识

马锦华

<https://cse.sysu.edu.cn/teacher/MaJinhua>

SUN YAT-SEN University



声明：该PPT只供非商业使用，也不可视为任何出版物。由于历史原因，许多图片尚没有标注出处，如果你知道图片的出处，欢迎告诉我们 majh8@mail.sysu.edu.cn.



课程目录（暂定）

❑ 第一章	课程简介与预备知识	6学时
❑ 第二章	特征提取与表示	6学时
❑ 第三章	主成分分析	3学时
❑ 第四章	归一化、判别分析、人脸识别	3学时
❑ 第五章	EM算法与聚类	3学时
❑ 第六章	贝叶斯决策理论	3学时
❑ 第七章	线性分类器与感知机	3学时
❑ 第八章	支持向量机	3学时
❑ 第九章	神经网络、正则项和优化方法	3学时
❑ 第十章	卷积神经网络及经典框架	3学时
❑ 第十一章	循环神经网络	3学时
❑ 第十二章	Transformer	3学时
❑ 第十三章	自监督与半监督学习	3学时
❑ 第十四章	开放世界模式识别	6学时



第二部分：数学背景知识



目标

- 回忆**掌握**相关基本概念和最重要的定理
- 能够熟练**应用**提供的资源列表
- 提高目标
 - 理解相关定理的证明和推导过程
 - 能不看书熟练应用一些重要定理和推导
 - **进一步**：通过查阅资料掌握一些课堂没有讲授的定理，并能应用到学习、研究中遇到的问题中去

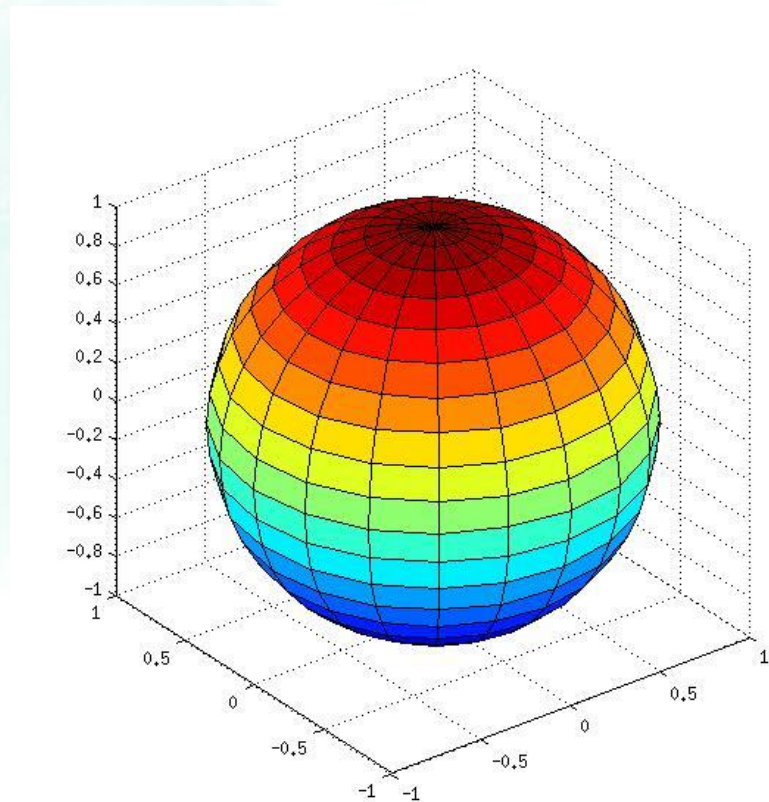


线性代数



向量 (vector)

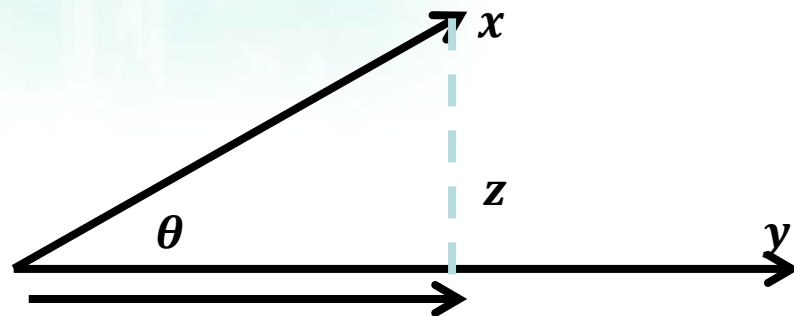
- $\mathbf{x} = (x_1, x_2, \dots, x_d)^T \in \mathbb{R}^d$
- 内积 (dot-product, inner-product, 点积)
 - $\mathbf{x}^T \mathbf{y} = \mathbf{y}^T \mathbf{x} = \sum_{i=1}^d x_i y_i$
- 向量的长度 (vector norm)
 - $\|\mathbf{x}\| = \sqrt{\mathbf{x}^T \mathbf{x}}$, $\|\mathbf{x}\|^2 = \mathbf{x}^T \mathbf{x}$
 - 若 $\|\mathbf{x}\| = 1$, 称 \mathbf{x} 为单位向量
- 正交 (orthogonal)
 - $\mathbf{x}^T \mathbf{y} = 0$
 - \mathbf{x} 和 \mathbf{y} 被称为垂直 (perpendicular): $\mathbf{x} \perp \mathbf{y}$





内积、角度、投影

- x : $\|x\|$ 决定 **长度**, $\frac{x}{\|x\|}$ 决定 **方向**
- 向量之间的夹角(angle):
 - $x^T y = \|x\| \cdot \|y\| \cdot \cos \theta$
 - $\|x^T y\| \leq \|x\| \cdot \|y\|$
- **x 在 y 上的投影(projection)**
 - 方向: $\frac{y}{\|y\|}$ 长度: $\|x\| \cos \theta = \|x\| \frac{x^T y}{\|x\| \cdot \|y\|} = \frac{x^T y}{\|y\|}$
 - 投影 $\text{proj}_y x$: $\frac{x^T y}{\|y\|^2} y$
 - $\text{proj}_y x \perp z$
 - $\text{proj}_y x + z = x$





柯西-施瓦茨不等式

- \mathbb{R}^n

- $(\sum_{k=1}^n a_k b_k)^2 \leq (\sum_{k=1}^n a_k^2)(\sum_{k=1}^n b_k^2)$

- 等号成立当且仅当存在固定实数 c , 使得 $\forall k, a_k = c b_k$

- 平方可积函数空间 L^2

- $\left[\int_a^b f(x)g(x)dx \right]^2 \leq \left[\int_a^b [f(x)]^2 dx \right] \left[\int_a^b [g(x)]^2 dx \right]$

- 等号成立当且仅当存在固定实数 c , 使得 $\forall x \in [a, b], f(x) = c g(x)$ (几乎处处)



矩阵(Matrix)

- $X = \begin{bmatrix} x_{11} & \cdots & x_{1n} \\ \vdots & \ddots & \vdots \\ x_{m1} & \cdots & x_{mn} \end{bmatrix} : m \times n \text{ 的矩阵}$

- $n = m$ 时称为方阵(square matrix)

- 行向量: $m = 1$

- 列向量 (向量) : $n = 1$

- 对角阵(diagonal matrix): 方阵中, 只有对角线非零

- 单位阵(identity matrix): 对角线全部为1的对角阵

- 一般记为 I 或者 I_n



矩阵运算

- 乘法: $X: m \times n, Y: n \times p$
 - X 的列数等于 Y 的行数时矩阵乘法 XY 才有定义
 - 一般来说 $XY \neq YX$
- 矩阵的幂(power)
 - 对方阵有定义: $X^2 = XX, X^3 = XXX, \dots$
- 转置(transpose)
 - $X: m \times n$, 那么 $X^T: n \times m$
 - $X^T X: n \times n, XX^T: m \times m$
- 对称矩阵(symmetric matrix)
 - 是方阵, $X_{ij} = X_{ji}, \forall i, j$



行列式值、矩阵的逆

- 方阵的行列式值(determinant)
 - $|X|$, 或写作 $\det(X)$
 - $|X| = |X^T|$
 - $|XY| = |X||Y|$
 - $|\lambda X| = \lambda^n |X|$ ($X: n \times n$)
- 方阵的逆矩阵(inverse matrix)
 - X^{-1} : 满足 $XX^{-1} = X^{-1}X = I_n$
 - X 可逆(invertible) $\iff |X| \neq 0$
 - $(X^{-1})^{-1} = X$, $(\lambda X)^{-1} = \frac{1}{\lambda} X^{-1}$
 - $(XY)^{-1} = Y^{-1}X^{-1}$, $(X^{-1})^T = (X^T)^{-1}$



方阵的特征值、特征向量、迹

- 特征值(eigenvalue)和特征向量(eigenvector)

- $A\mathbf{x} = \lambda\mathbf{x}$ $A: n \times n$

- λ : 特征值 \mathbf{x} : 特征向量

- n 阶方阵有 n 个特征值（复数域）

- 可能存在相等的特征值

- 特征值和对角线的关系

- $\sum_{i=1}^n a_{ii} = \sum_{i=1}^n \lambda_i$

- $\det(A) = \prod_{i=1}^n \lambda_i$

- 方阵的迹(trace)

- $\text{tr}(A) = \sum_{i=1}^n a_{ii} = ??$, $\text{tr}(\mathbf{AB}) = \text{tr}(\mathbf{BA})$



实对称矩阵

- 对称矩阵，每个项都是实数
 - Real symmetric matrix
 - 这门课程中最常用到
- 性质：
 - 所有特征值都是实数，特征向量都是实向量
 - 特征值记为 $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_n$
 - 对应的特征向量记为 $\xi_1, \xi_2, \dots, \xi_n$
 - 特征向量互相垂直: $\xi_i^T \xi_j = 0 \ (i \neq j)$
 - $E = [\xi_1 \ \xi_2 \ \dots \ \xi_n]$ 是 $n \times n$ 的，是满秩(full rank)的， $\text{rank}(E) = n$.



实对称矩阵的分解

- $X: n \times n$ 的实对称矩阵
 - 特征值为 λ_i , 其对应的特征向量为 ξ_i
- $X = \sum_{i=1}^n \lambda_i \xi_i \xi_i^T$
 - 称为谱分解(spectral decomposition)
 - 约定 $\|\xi_i\| = 1$, 则 E 是正交矩阵(orthogonal matrix)
 - $X = E \Lambda E^T$
 - Λ 是一个对角矩阵, $\Lambda_{ii} = \lambda_i$
 - $EE^T = E^T E = I$, $E^{-1} = ?$, $|E| = ?$
- 进一步阅读
 - LU分解, Cholesky分解, QR分解
 - 资源: Horn & Johnson, Matrix Analysis (Second Edition)



正定、半正定

- 对称方阵 A 是正定的(positive definite)当且仅当
 - $\forall \mathbf{x} \neq \mathbf{0} \quad \mathbf{x}^T A \mathbf{x} = \sum_{i,j} x_i x_j A_{ij} > 0$
 - $\forall \mathbf{x} \neq \mathbf{0} \quad \mathbf{x}^T A \mathbf{x} \geq 0$ 则 A 为半正定(positive semi-definite)
 - 分别记为 $A \succ 0$ 或 $A \succeq 0$
- $\mathbf{x}^T A \mathbf{x}$: 称为二次型(quadratic)
 - 这门课程会经常用到, 一般满足 $A \succeq 0$
- 等价关系
 - 1. $A \succ 0$ ($A \succeq 0$)
 - 2. 特征值全部为正数 (非负实数)
- 正定矩阵的任意主子矩阵也是正定矩阵



矩阵求导

- 假设一切求导的条件都满足（导数都存在）
- $\frac{\partial \mathbf{a}}{\partial x}$ 是一个向量, $\left(\frac{\partial \mathbf{a}}{\partial x}\right)_i = \frac{\partial a_i}{\partial x}$
- 对于矩阵, $\left(\frac{\partial \mathbf{A}}{\partial x}\right)_{ij} = \frac{\partial A_{ij}}{\partial x}$
- $\left(\frac{\partial x}{\partial \mathbf{a}}\right)_i = \frac{\partial x}{\partial a_i}$ $\left(\frac{\partial x}{\partial \mathbf{A}}\right)_{ij} = \frac{\partial x}{\partial A_{ij}}$ $\left(\frac{\partial \mathbf{a}}{\partial \mathbf{x}}\right)_{ij} = \frac{\partial a_i}{\partial x_j}$
- 如何求导?
 - 能够查表（例如The Matrix Cookbook）并合理应用

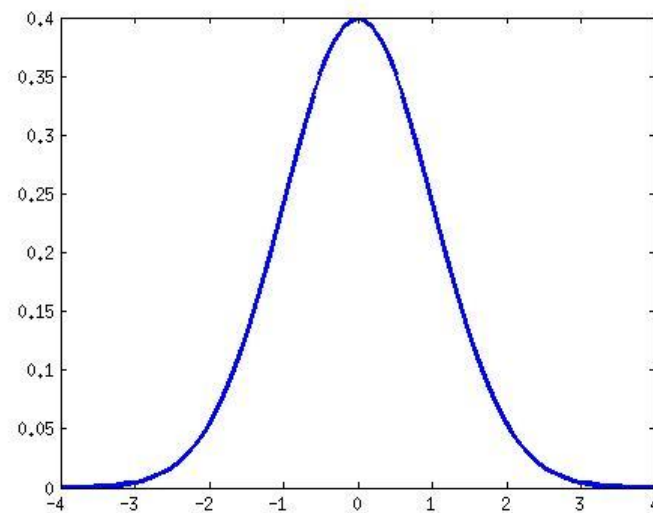
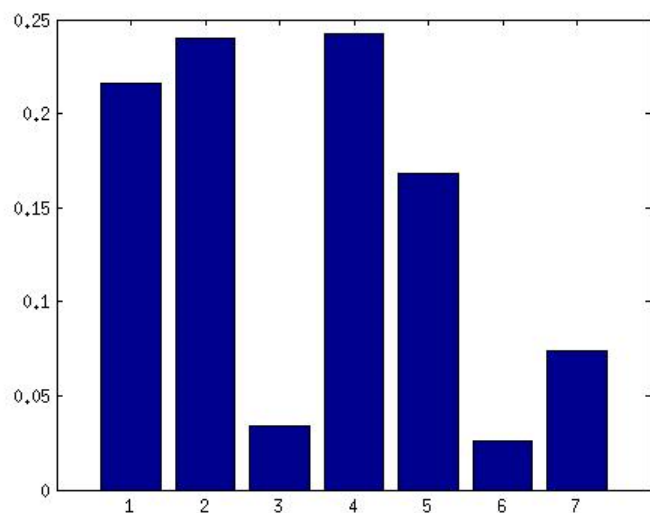


概率与统计



随机变量(Random variable)

- X : 可以是离散(discrete)、连续(continuous)、或者混合(hybrid)的





概率密度函数

- (古典) 离散(discrete):
 - 可数的(countable)不相容的若干事件 c_1, c_2, \dots
 - $p(X = x_i) = c_i$ -- probability mass function
 - $c_i \geq 0, \sum_i c_i = 1$
 - 伯努利分布、二项分布、几何分布、泊松分布等
- 连续(continuous): 为简化, 只考虑 $X \in (-\infty, \infty)$
 - $p(x)$: 概率密度函数probability density function (pdf)
 - $p(x) \geq 0, \int_{-\infty}^{+\infty} p(x) dx = 1$
 - 均匀分布、正态分布、指数分布、 Γ 分布等
- 随机变量可以看成是一个(可测)函数, 而不是一个数学分析意义上的变量



分布函数(连续)

- Cumulative distribution function (cdf)

- $F(x) = P(X \leq x) = \int_{-\infty}^x p(x) dx$

- $F(-\infty) = 0 \leq F(x) \leq F(+\infty) = 1$

- 非减性(non-decreasing)

如果 $x \leq y$, 那么 $F(x) \leq F(y)$

- $P(X = x) = ?$

- $P(x_1 < x < x_2) = F(x_2) - F(x_1) = \int_{x_1}^{x_2} p(x) dx$

- PDF和CDF的关系

- $p(x) = F'(x)$



联合、条件分布、变换

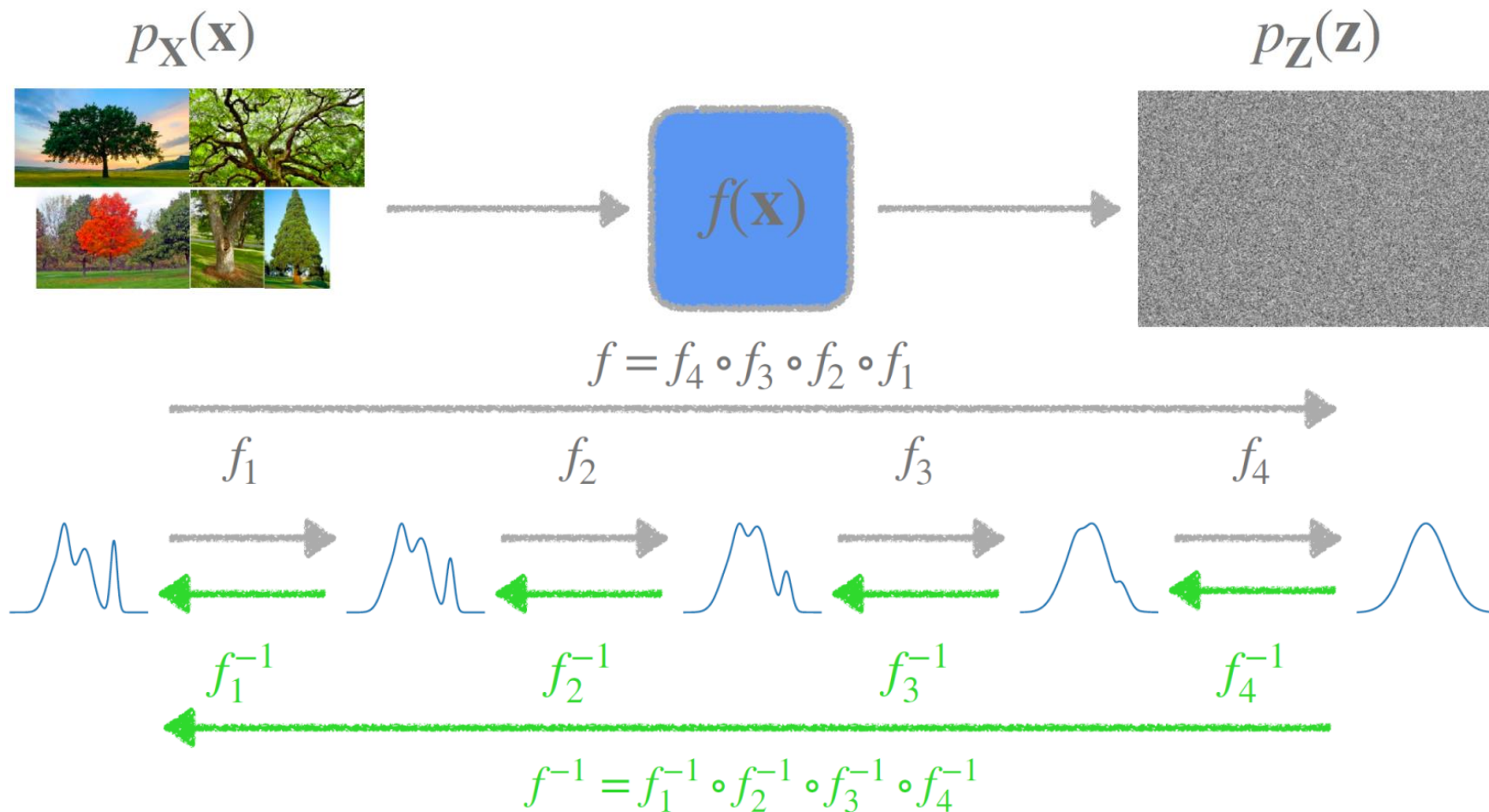
- 联合(joint distribution): $P(X = \mathbf{x})$
 - $p(\mathbf{x}) \geq 0$ $\int p(\mathbf{x})d\mathbf{x} = 1$
- 条件(conditional distribution): $P(X = x|Y = y)$
- $p(x, y) = p(y)p(x|y)$
- $p(x) = \int_y p(x, y)dy$ --marginal (边缘) 分布
- 假设 $x = g(y)$, 那么

$$p_Y(y) = p_X(x) \left| \frac{dx}{dy} \right| = p_X(g(y)) |g'(y)|$$

- 如果 \mathbf{x} 和 \mathbf{y} 是向量?
- 对 g 的具体要求?
- 应用: normalizing flows



Normalizing Flows





多维分布的期望

- 假设有函数 $f(\mathbf{x})$ ，在 \mathbf{x} 服从分布 $p(\mathbf{x})$ 时：
- f 的期望(expectation)，记为 $E[f(X)]$
 - $E[f(X)] = \sum_{\mathbf{x}} f(\mathbf{x}) \cdot p(X = \mathbf{x})$ ，或
 - $E[f(X)] = \int f(\mathbf{x})p(\mathbf{x})d\mathbf{x}$
 - **条件期望** $E(f(\mathbf{x})|Y = \mathbf{y}) = \sum_{\mathbf{x}} f(\mathbf{x}) \cdot p(\mathbf{x}|\mathbf{y})$
- X 的方差(Variance, 一维)或协方差(covariance, 多维)
 - $\text{Var}(X) = E[(X - E(X))^2]$
 - **$\text{Var}(X) = E[X^2] - (E(X))^2$** 向量形式时怎么样？
- 当 $p(\mathbf{x})$ 和 $f(\mathbf{x})$ 固定时
 - 期望、方差是一个确定的数（或向量、矩阵）
 - $g(\mathbf{y}) = E(X|Y = \mathbf{y})$ 是什么？



估计均值和协方差矩阵

- 训练样本: $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n$
- 均值的估计estimation:

$$\bar{\mathbf{x}} = \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i$$

- Covariance的估计

$$\text{Cov}(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^n (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T$$

无偏估计unbiased estimation

$$\text{Cov}(\mathbf{x}) = \frac{1}{n-1} \sum_{i=1}^n (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T$$



两个随机变量的独立、相关

- 一般说来 $p(x, y) \neq p(x)p(y)$
- 如果 $\forall x, y, p(x, y) = p(x)p(y)$, 则 X 和 Y 互相独立 (independent)
- 协方差 $\text{Cov}(X, Y) = E[(X - E(X))(Y - E(Y))]$
- Pearson 相关系数 (Pearson's correlation coefficient):
 $-1 \leq \rho_{XY} \leq 1$

$$\rho_{XY} = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}},$$

- $\rho_{XY} = 0$, 称为不相关 (not correlated)
- $\rho_{XY} = \pm 1$, 称为完全相关, 如存在线性关系
- 独立保证一定不相关, 但是, 不相关不一定能保证独立



高斯分布

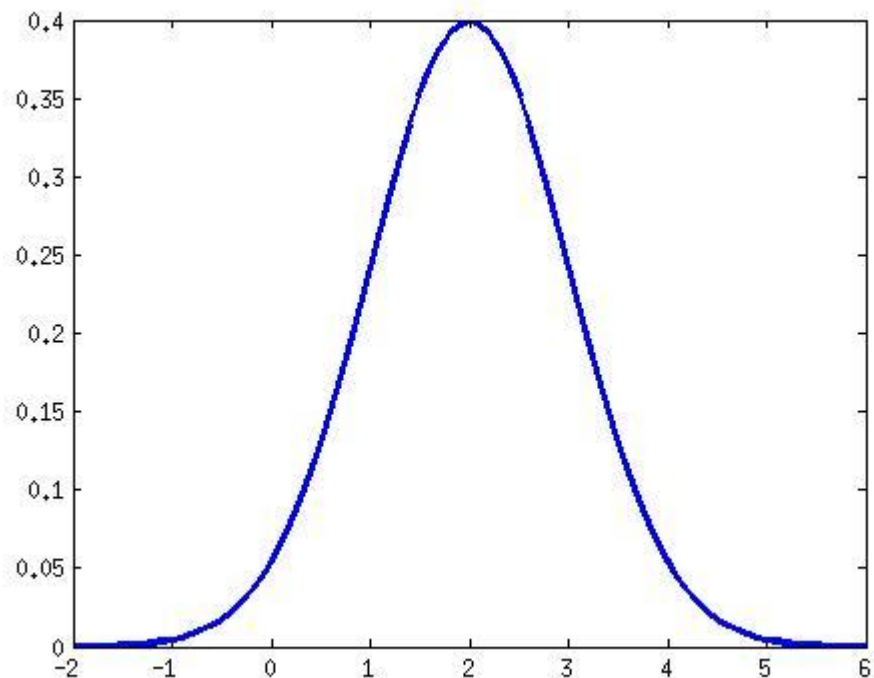
- 又叫正态分布, normal distribution, Gaussian distribution

- 单变量或一维高斯分布 $N(\mu, \sigma^2)$

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

或者 $\frac{1}{\sqrt{2\pi}\sigma} \exp \left(-\frac{(x-\mu)^2}{2\sigma^2} \right)$ 更眼熟?

- μ : 期望, 或称均值
- σ^2 : 方差
 - σ : 标准差 (standard deviation)



- ✓ 图例中 $\mu = 2, \sigma = 1$,
- ✓ Markov不等式: 若 $X \geq 0$ (非负随机变量), 则 $P(X \geq a) \leq \frac{E(X)}{a}$
- ✓ Chebyshev不等式: 对任何分布, $P((X - \mu)^2 \geq k^2) \leq \frac{\sigma^2}{k^2}$ 或 $P(|X - \mu| > k) \leq \frac{\sigma^2}{k^2} (k > 0)$
- ✓ 如果 $k = 3$, 这个界是多少? 正态分布上的实际值是多少?



多维高斯分布

- 一维：

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

- 多维

$$p(\mathbf{x}) = (2\pi)^{-\frac{D}{2}}|\Sigma|^{-\frac{1}{2}}\exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}$$

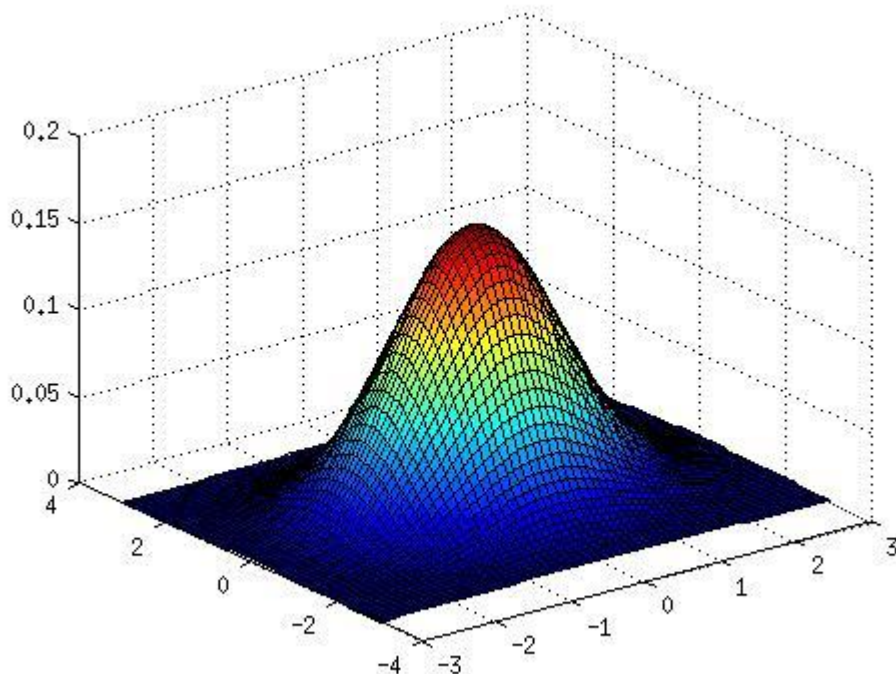
□ D : 维数 记为 $\mathbf{x} \sim N(\boldsymbol{\mu}, \Sigma)$

□ Σ : 协方差矩阵

□ $\boldsymbol{\mu}$: 均值



多维高斯PDF示意图



- ✓ 图例中 $\mu = (0,0)$, $\Sigma = I_2$
- ✓ 更多相关知识，将在PCA中讲授



高斯分布中的相关性和独立

- 一般来说，两变量
 - 独立保证一定不相关
 - 不相关不一定保证独立
- 但是，对于多维高斯分布
 - 不相关意味着协方差矩阵中非对角线项是0
$$\begin{matrix} c_{ii} & \dots & 0 \\ \vdots & \vdots & \vdots \\ 0 & \dots & c_{jj} \end{matrix}$$
 - 在正态分布中，不相关就等价于独立



多维与一维高斯的关系

- 多维高斯 $X = \begin{pmatrix} X_a \\ X_b \end{pmatrix}$
 - 条件分布: $\mathbf{x}_a | \mathbf{x}_b$ 还是高斯分布
 - 边缘分布(margin distribution):
 $p(\mathbf{x}_a) = \int p(\mathbf{x}_a, \mathbf{x}_b) d\mathbf{x}_b$ 也是高斯分布
- 两个高斯分布的加权和也是高斯分布
 - $aX + bY$
- 为什么大家用高斯分布?



优化基础知识

- 本课程主要利用已有优化软件，不讲授优化算法或者理论
 - 最优化自身是一门复杂的课程
 - 吴建鑫《模式识别》第2.3、2.4节（2.3.1节之后）
- 知识要点：
 - 凸集、凸函数
 - 吴建鑫《模式识别》第2.3.2节
 - 拉格朗日乘子法（等式约束）
 - 吴建鑫《模式识别》第2.3.3节
 - 算法复杂性
 - 吴建鑫《模式识别》第2.4节



进一步的阅读

- 如果对本节内容感兴趣，可以参考：
 - 吴建鑫 《模式识别》第13章：正态分布
 - PRML的相关章节（第二章和附录）
 - 其他数学课程：实分析、矩阵分析、概率论、最优化等