



LiVE Group
视觉智能与学习研究中心

机器学习 (第3讲)

主讲：张磊

E-mail: leizhang@cqu.edu.cn
Lab Website: <http://www.leizhang.tk>





第三章： 机器学习中的数学基础



第三章：机器学习的数学基础

人工智能、智能信息处理、机器学习、模式识别的数学基础主要包括以下几个部分：

- 高等数学
- 线性代数
- 矩阵论
- 优化理论
- 概率论
- 信息论
- 计算复杂度理论。

学好这门课，这些数学基础是必不可少的。



第三章：机器学习的数学基础

向量

在线性代数中，标量(scalar)是一个实数，而向量(vector)是指n个实数组成的有序数组，称为n维向量。如果没有特别说明，一个n维向量一般表示为一个列向量，即大小为 $n \times 1$ 的列向量。向量的表示形式一般采用粗体、小写。如 \mathbf{a} 。如果写成 a ，则表示一个标量。

$$\mathbf{a} = \begin{bmatrix} \mathbf{a}_1 \\ \mathbf{a}_2 \\ \vdots \\ \mathbf{a}_n \end{bmatrix} \in \mathcal{R}^n \quad (\text{n维欧几里得空间})$$

常见向量

全1向量：指所有元素为1的向量，通常用 $\mathbf{1}_n$ 表示， n 表示向量的维度。

$\mathbf{1}_K = [1, \dots, 1]_{K \times 1}^T$ 表示 K 维的全1向量。



第三章：机器学习的数学基础

向量的模和范数

向量 \mathbf{a} 的模表示为 $\|\mathbf{a}\|$,计算方法为

$$\|\mathbf{a}\| = \sqrt{\sum_{i=1}^n a_i^2}$$

在线性代数中，范数(norm)是一个表示“长度”概念的函数，为向量空间内所有向量赋予非零的正长度或者大小。对于一个 n 维的向量 \mathbf{x} ，其常见的范数有

L_1 范数：

$$|\mathbf{x}|_1 = \sum_{i=1}^n |x_i|.$$

L_2 范数：

$$\|\mathbf{x}\|_2 = \sqrt{\sum_{i=1}^n x_i^2} = \sqrt{\mathbf{x}^T \mathbf{x}}.$$

也叫欧几里得范数，向量的长度；
如果向量 \mathbf{x} 满足 $\|\mathbf{x}\|_2 = 1$ ，
则称向量是归一化的



第三章：机器学习的数学基础

向量的内积

两个具有相同维数 n 的向量 \mathbf{x} 和 \mathbf{y} 的内积记为 $\mathbf{x}^T\mathbf{y}$,是一个**标量**。

$$\mathbf{x}^T\mathbf{y} = \sum_{i=1}^n x_i y_i$$

内积也称作标量积或点积，在泛函里面表示为 $\langle \mathbf{x}, \mathbf{y} \rangle$

向量的夹角

两个 n 维向量的夹角定义

$$\cos \theta = \frac{\mathbf{x}^T\mathbf{y}}{\|\mathbf{x}\|\|\mathbf{y}\|}$$

涵义： 向量内积是两个向量共线性的度量，说明两个向量之间的相似性。
若 \mathbf{x} 和 \mathbf{y} 正交，则 $\mathbf{x}^T\mathbf{y} = 0$ 。由 $\cos \theta \leq 1$ ， $\mathbf{x}^T\mathbf{y} \leq \|\mathbf{x}\|\|\mathbf{y}\|$ (柯西-施瓦茨不等式)



第三章：机器学习的数学基础

向量的外积

两个具有相同维数 n 的向量 \mathbf{x} 和 \mathbf{y} 的外积记为 \mathbf{xy}^T ,是一个矩阵。

$$\begin{aligned}\mathbf{xy}^T &= \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} (y_1 \ y_2 \ \cdots \ y_n) \\ &= \begin{pmatrix} x_1 y_1 & \cdots & x_1 y_n \\ \vdots & \ddots & \vdots \\ x_n y_1 & \cdots & x_n y_n \end{pmatrix}\end{aligned}$$

向量外积也称作矩阵积或二元积



第三章：机器学习的数学基础

矩阵

一个大小为 $m \times n$ 的矩阵(matrix)是一个由 m 行 n 列元素排列成的矩形阵列。矩阵里的元素可以是数字、符号或数学式。在我们的课程里，矩阵一般默认为数字矩阵。

一个 n 维向量，可以看作是一个 $n \times 1$ 的矩阵。

一个 $m \times n$ 的矩阵 \mathbf{M} ，表达为 $\mathbf{M} \in \mathfrak{R}^{m \times n}$ (欧几里得空间)

矩阵的数学符号通常有粗体、大写表示。



第三章：机器学习的数学基础

矩阵的基本运算

如果 A 和 B 都为 $m \times n$ 的矩阵，则 A 和 B 的加减为

$$(A + B)_{ij} = A_{ij} + B_{ij},$$

$$(A - B)_{ij} = A_{ij} - B_{ij}.$$

A 和 B 的点乘 $A \odot B \in \mathbb{R}^{m \times n}$ 为

$$(A \odot B)_{ij} = A_{ij}B_{ij}.$$

一个标量 c 与矩阵 A 乘积为

$$(cA)_{ij} = cA_{ij}.$$

若 A 是 $m \times p$ 矩阵和 B 是 $p \times n$ 矩阵，则乘积 AB 是一个 $m \times n$ 的矩阵

$$(AB)_{ij} = \sum_{k=1}^p A_{ik}B_{kj}$$



第三章：机器学习的数学基础

矩阵的基本运算

$m \times n$ 矩阵 A 的**转置** (Transposition) 是一个 $n \times m$ 的矩阵, 记为 A^T , A^T 第 i 行第 j 列的元素是原矩阵 A 第 j 行第 i 列的元素,

$$(A^T)_{ij} = A_{ji}.$$

如果 A 是对称矩阵, 则 $A=A^T$

矩阵的**向量化**是将矩阵表示为一个列向量。这里, vec 是向量化算子。设 $A = [a_{ij}]_{m \times n}$, 则

$$\text{vec}(A) = [a_{11}, a_{21}, \dots, a_{m1}, a_{12}, a_{22}, \dots, a_{m2}, \dots, a_{1n}, a_{2n}, \dots, a_{mn}]^T.$$



第三章：机器学习的数学基础

矩阵的基本运算

矩阵的迹

矩阵 \mathbf{A} 的迹，数学表示为 $\text{Tr}(\mathbf{A})$ 或 $\text{Trace}(\mathbf{A})$,即对角线元素的和。

$$\text{Tr}(\mathbf{A}\mathbf{A}^\top) = \text{Tr}(\mathbf{A}^\top\mathbf{A}) = \|\mathbf{A}\|_F^2 = \sum_{i,j} A_{i,j}^2$$

矩阵的逆 \mathbf{A}^{-1}

矩阵的转置 \mathbf{A}^\top



第三章：机器学习的数学基础

矩阵的基本运算

矩阵与向量的乘积

$$\begin{bmatrix} m_{11} & \cdots & m_{1d} \\ \vdots & \ddots & \vdots \\ m_{n1} & \cdots & m_{nd} \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_d \end{bmatrix} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$$

可以写另一种表达方式

$$\mathbf{M}\mathbf{x} = \mathbf{y}$$

其中，

$$y_i = \sum_{j=1}^d m_{ij}x_j \text{ (线性组合)}$$



第三章：机器学习的数学基础

常见矩阵

对称矩阵指其转置等于自己的矩阵，即满足 $A = A^T$ 。

对角矩阵（Diagonal Matrix）是一个主对角线之外的元素皆为 0 的矩阵。对角线上的元素可以为 0 或其他值。一个 $n \times n$ 的对角矩阵矩阵 A 满足：

$$A_{ij} = 0 \text{ if } i \neq j \quad \forall i, j \in \{1, \dots, n\}$$

对角矩阵 A 也可以记为 $\text{diag}(\mathbf{a})$ ， \mathbf{a} 为一个 n 维向量，并满足

$$A_{ij} = a_j.$$



第三章：机器学习的数学基础

常见矩阵

$n \times n$ 的对角矩阵 $A = \text{diag}(\mathbf{a})$ 和 n 维向量 \mathbf{b} 的乘积为一个 n 维向量

$$A\mathbf{b} = \text{diag}(\mathbf{a})\mathbf{b} = \mathbf{a} \odot \mathbf{b},$$

其中 \odot 表示点乘，即 $(\mathbf{a} \odot \mathbf{b})_i = a_i b_i$ 。

单位矩阵是一种特殊的的对角矩阵，其主对角线元素为 1，其余元素为 0。
 n 阶单位矩阵 I_n ，是一个 $n \times n$ 的方形矩阵。可以记为 $I_n = \text{diag}(1, 1, \dots, 1)$ 。
一个矩阵和单位矩阵的乘积等于其本身。



第三章：机器学习的数学基础

导数

对于定义域和值域都是实数域的函数 $y=f(x)$,若 $f(x)$ 在点 x_0 的某个邻域 Δx 内, 极限 $f'(x_0) = \lim_{\Delta x \rightarrow 0} \frac{f(x_0 + \Delta x) - f(x_0)}{\Delta x}$ 存在

则称函数 $y=f(x)$ 在点 x_0 处可导, 其导数为 $f'(x_0)$

函数 $f(x)$ 的导数 $f'(x)$ 也可记作 $\nabla_x f(x)$, $\frac{\partial f(x)}{\partial x}$ 或 $\frac{\partial}{\partial x} f(x)$ 。



第三章：机器学习的数学基础

向量和矩阵的导数

- 对于一个 p 维的向量 $\mathbf{x} \in \mathbb{R}^p$, 函数 $y = f(\mathbf{x}) = f(x_1, x_2, \dots, x_p)$ 是一个标量, 则 y 关于 \mathbf{x} 的导数为

$$\frac{\partial f(\mathbf{x})}{\partial \mathbf{x}} = \begin{bmatrix} \frac{\partial f(\mathbf{x})}{\partial x_1} \\ \vdots \\ \frac{\partial f(\mathbf{x})}{\partial x_p} \end{bmatrix} \in \mathbb{R}^p$$



相当于梯度, 是由 p 个偏导数组成的矢量

- 如果 $\mathbf{y} = f(\mathbf{x}) = f(x_1, x_2, \dots, x_p) \in \mathbb{R}^q$ 是一个向量,

$$\frac{\partial f(\mathbf{x})}{\partial \mathbf{x}} = \begin{bmatrix} \frac{\partial f_1(\mathbf{x})}{\partial x_1} & \dots & \frac{\partial f_q(\mathbf{x})}{\partial x_1} \\ \vdots & \vdots & \vdots \\ \frac{\partial f_1(\mathbf{x})}{\partial x_p} & \dots & \frac{\partial f_q(\mathbf{x})}{\partial x_p} \end{bmatrix} \in \mathbb{R}^{p \times q}$$



第三章：机器学习的数学基础

导数法则

加（减）法则

$y = f(\mathbf{x}), z = g(\mathbf{x})$ 则

$$\frac{\partial(\mathbf{y} + \mathbf{z})}{\partial \mathbf{x}} = \frac{\partial \mathbf{y}}{\partial \mathbf{x}} + \frac{\partial \mathbf{z}}{\partial \mathbf{x}}$$

乘法法则

(1) 若 $\mathbf{x} \in \mathbb{R}^p$, $\mathbf{y} = f(\mathbf{x}) \in \mathbb{R}^q$, $\mathbf{z} = g(\mathbf{x}) \in \mathbb{R}^q$, 则

$$\frac{\partial \mathbf{y}^\top \mathbf{z}}{\partial \mathbf{x}} = \frac{\partial \mathbf{y}}{\partial \mathbf{x}} \mathbf{z} + \frac{\partial \mathbf{z}}{\partial \mathbf{x}} \mathbf{y}$$



第三章：机器学习的数学基础

导数法则

(2) 若 $\mathbf{x} \in \mathbb{R}^p$, $\mathbf{y} = f(\mathbf{x}) \in \mathbb{R}^s$, $\mathbf{z} = g(\mathbf{x}) \in \mathbb{R}^t$, $\mathbf{A} \in \mathbb{R}^{s \times t}$ 和 \mathbf{x} 无关, 则

$$\frac{\partial \mathbf{y}^\top \mathbf{A} \mathbf{z}}{\partial \mathbf{x}} = \frac{\partial \mathbf{y}}{\partial \mathbf{x}} \mathbf{A} \mathbf{z} + \frac{\partial \mathbf{z}}{\partial \mathbf{x}} \mathbf{A}^\top \mathbf{y}$$

(3) 若 $\mathbf{x} \in \mathbb{R}^p$, $y = f(\mathbf{x}) \in \mathbb{R}$, $\mathbf{z} = g(\mathbf{x}) \in \mathbb{R}^p$, 则

$$\frac{\partial y \mathbf{z}}{\partial \mathbf{x}} = y \frac{\partial \mathbf{z}}{\partial \mathbf{x}} + \frac{\partial y}{\partial \mathbf{x}} \mathbf{z}^\top$$



第三章：机器学习的数学基础

导数法则

链式法则

(1) 若 $\mathbf{x} \in \mathbb{R}^p$, $\mathbf{y} = \mathbf{g}(\mathbf{x}) \in \mathbb{R}^s$, $\mathbf{z} = \mathbf{f}(\mathbf{y}) \in \mathbb{R}^t$, 则

$$\frac{\partial \mathbf{z}}{\partial \mathbf{x}} = \frac{\partial \mathbf{y}}{\partial \mathbf{x}} \cdot \frac{\partial \mathbf{z}}{\partial \mathbf{y}}$$

(2) 若 $X \in \mathbb{R}^{p \times q}$ 为矩阵, $Y = \mathbf{g}(X) \in \mathbb{R}^{s \times t}$, $\mathbf{z} = \mathbf{f}(Y) \in \mathbb{R}$,

$$\frac{\partial \mathbf{z}}{\partial X_{ij}} = \text{tr} \left(\left(\frac{\partial \mathbf{z}}{\partial Y} \right)^\top \frac{\partial Y}{\partial X_{ij}} \right)$$

(3) 若 $X \in \mathbb{R}^{p \times q}$ 为矩阵, $\mathbf{y} = \mathbf{g}(X) \in \mathbb{R}^s$, $\mathbf{z} = \mathbf{f}(\mathbf{y}) \in \mathbb{R}$, 则

$$\frac{\partial \mathbf{z}}{\partial X_{ij}} = \left(\frac{\partial \mathbf{z}}{\partial \mathbf{y}} \right)^\top \frac{\partial \mathbf{y}}{\partial X_{ij}}$$



第三章：机器学习的数学基础

导数法则

如果矩阵 $\mathbf{M} \in \mathbb{R}^{n \times d}$ 的每一个元素都是关于标量参数 θ 的函数，那么矩阵 \mathbf{M} 对参数 θ 的导数为

$$\frac{\partial \mathbf{M}}{\partial \theta} = \begin{pmatrix} \frac{\partial m_{11}}{\partial \theta} & \dots & \frac{\partial m_{1d}}{\partial \theta} \\ \vdots & \ddots & \vdots \\ \frac{\partial m_{n1}}{\partial \theta} & \dots & \frac{\partial m_{nd}}{\partial \theta} \end{pmatrix}$$

矩阵对标量的导数！



第三章：机器学习的数学基础

导数法则

按位运算的向量函数及其导数：

$$\text{定义 } \mathbf{x} = [x_1, \dots, x_d]^T \in \mathbb{R}^d, \mathbf{z} = [z_1, \dots, z_d]^T \in \mathbb{R}^d$$
$$\mathbf{z} = f(\mathbf{x})$$

其中 $f(\mathbf{x})$ 是按位运算的，即 $z_i = f(x_i)$

那么 $f(\mathbf{x})$ 对 \mathbf{x} 的导数为

$$\frac{\partial f(\mathbf{x})}{\partial \mathbf{x}} = \begin{pmatrix} \frac{\partial f(x_1)}{\partial x_1} & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \frac{\partial f(x_d)}{\partial x_d} \end{pmatrix}$$



第三章：机器学习的数学基础

随机变量

随机变量通常用于刻画误差或噪声的随机值，如预测一个人的体重。

考查抛硬币的例子：

设 X 是一个随机变量，对于一个正常的硬币， X 的取值为 $x=1$ （正面）或 $x=0$ （反面）

所谓“随机”，是指抛硬币事件发生之前，我们无法知道 X 的取值。

随机变量的类型：**离散**随机变量、**连续**随机变量

注：离散随机变量的相关概念可以扩展到连续随机变量。



第三章：机器学习的数学基础

概率和概率分布

为了刻画一个随机变量的可能取值，我们只能通过随机变量取值的**概率**（即可能性） $P(X=x)$ 。（注：大写表示随机变量，小写表示取值）

假设硬币本身没有问题，抛的过程是公平的，则出现正反面的可能性是一致的，即

$$P(X=1)=P(X=0)=0.5$$

也就是说，抛硬币1000次，正反面朝上的次数各占500次。

- 1) 概率是介于0到1之间的值；
- 2) 所有可能结果的概率之和为1；
- 3) 所有可能取值的集合(所有的 x)以及它们的概率 $P(X=x)$ 称为**分布**



第三章：机器学习的数学基础

条件概率

一个随机事件通常会影晌另一个随机事件的结果。

对于抛硬币问题，我们思考“你来抛，我来报”的游戏。

假设硬币本身没有问题，抛的过程也是公平的，

- 1) 你抛硬币的过程定义为随机事件1，用随机变量 X 表示；
- 2) 我猜正反面的过程定义随机事件2，用随机变量 Y 表示；

那么条件概率 $P(Y=y|X=x)$ 表示 X 取特定值时， Y 也取某特定值的概率。

假设我总是报真实结果，那么 $P(Y=1|X=1)=1$ ， $P(Y=0|X=0)=1$ 。

那么 $P(Y=0|X=1)=0$ ， $P(Y=1|X=0)=0$ 。



第三章：机器学习的数学基础

条件概率

假设我只有60%的机会实报“正面朝上”的结果，那么
 $P(Y=1|X=1)=0.6$ ， $P(Y=0|X=0)=1$ 。

那么 $P(Y=0|X=1)=0.4$ ， $P(Y=1|X=0)=0$ 。

可以看到

$$P(Y=1|X=1)+P(Y=0|X=1)=1,$$

$$P(Y=1|X=0)+P(Y=0|X=0)=1$$



第三章：机器学习的数学基础

联合概率

1) 对于两个独立事件， $P(Y=y, X=x)=P(Y=y)*P(X=x)$

2) 若两个事件相互依赖， $P(Y=y, X=x)= P(Y=y | X=x)*P(X=x)=?$

Case 1: 如果我总是报真实结果，你抛出的硬币正面朝上且我也报正面的概率： ???

$$P(Y=1, X=1)=P(Y=1 | X=1)*P(X=1)=1*0.5=0.5$$

Case 2: 如果我只有60%的可能报真实结果，你抛出的硬币正面朝上且我也报正面的概率： ???

$$P(Y=1, X=1)=P(Y=1 | X=1)*P(X=1)=0.6*0.5=0.3$$



第三章：机器学习的数学基础

联合概率

对应抛硬币的例子， X 和 Y 有四种可能的组合，那么有下列等式：

$$\sum_{x,y} P(X = x, Y = y) = 1$$

即，概率加和为1。



第三章：机器学习的数学基础

边缘概率（概率的边缘化）

对于抛硬币事件，如果只考查“我报正面的次数所占的比例”，而无关抛硬币的真实结果，就相当于计算 $P(Y=1)$ ，如何计算？

答：可以从联合概率中对 X 进行边缘化得到，也就是对联合概率在 X 的所有取值下的概率之和：

$$P(Y = y) = \sum_{\mathbf{x}} P(Y = y, X = \mathbf{x}) = \sum_{\mathbf{x}} P(Y = y|X = \mathbf{x})P(X = \mathbf{x})$$

因此， $P(Y = 1) = \sum_{\mathbf{x}} P(Y = 1, X = \mathbf{x}) = P(Y = 1, X = 1) + P(Y = 1, X = 0) = P(Y = 1|X = 1)P(X = 1) + P(Y = 1|X = 0)P(X = 0) = 0.3$



第三章：机器学习的数学基础

例：假设我们有一个公平的硬币和两个骰子（其中1号骰子正常，2号骰子存在瑕疵）。定义抛硬币事件为 X ，掷骰子事件为 Y 。

首先，抛硬币，若正面朝上($X=1$)，掷1号骰子；若反面朝上($X=0$)，掷2号骰子。上述条件概率表格如下：

	1	2	3	4	5	6	
1号骰子	$1/6$	$1/6$	$1/6$	$1/6$	$1/6$	$1/6$	$P(Y=y X=1)$
2号骰子	$1/12$	$1/6$	$1/4$	$1/4$	$1/6$	$1/12$	$P(Y=y X=0)$

- 1) 计算出现正面朝上且是3点的概率；
- 2) 计算掷出3点的概率。



第三章：机器学习的数学基础

贝叶斯规则（贝叶斯公式）

与前面提到的概率不同，贝叶斯公式解决的是一个**逆向问题**。考查的是根据观测结果，预测真实发生的结果（机器学习）。

例：在抛硬币问题，如果“我报的是正面结果即 $Y=1$ ”，那么贝叶斯逆问题是要回答“**该硬币实际的朝向是正面还是反面**”？

方法：计算 $P(X=1|Y=1)$ 以及 $P(X=0|Y=1)$ 的概率，然后对比两个概率值的大小，有

若 $P(X=1|Y=1) > P(X=0|Y=1)$ ，则该硬币实际朝向是正面；

若 $P(X=1|Y=1) < P(X=0|Y=1)$ ，则该硬币实际朝向是反面；



第三章：机器学习的数学基础

贝叶斯规则（贝叶斯公式）

回顾联合概率表达式：

$$P(Y=y, X=x)=$$

$$P(Y=y|X=x)*P(X=x)=P(X=x|Y=y)*P(Y=y)$$

$$\text{所以, } P(X = x|Y = y) = \frac{P(Y = y|X = x)P(X=x)}{P(Y=y)}$$

经验分析： $P(X=1|Y=1)=1$ （如果我报正面，实际一定是正面）

理论结果： $P(X=1|Y=1)=P(Y=1|X=1)P(X=1)/P(Y=1)=0.6*0.5/0.3=1$

问： $P(X=0|Y=0)=?$



第三章：机器学习的数学基础

期望值

当处理随机变量时，使用一个或多个值代表一个分布的特征非常有用，比如期望值。

期望值表示随机变量 X 的函数的期望取值，对于离散随机变量，

$$E_{P(x)}\{f(X)\} = \sum_x f(x)P(x)$$

如果我们对 x 的期望值感兴趣，那么 $f(X) = X$ ，有

$$E_{P(x)}\{X\} = \sum_x xP(x)$$

问题： X 的期望值一定是随机变量可能取值中的某一个值吗？(骰子)



第三章：机器学习的数学基础

期望值

性质1：随机变量 X 的函数的期望值通常不是函数在 X 的期望值的取值，即 $E_{P(x)}\{f(X)\}$ 不一定等于 $f(E_{P(x)}\{X\})$ 。

性质2：当 $f(X) = aX$ 时， $E_{P(x)}\{f(X)\} = f(E_{P(x)}\{X\})$ ；

性质3：函数和的期望值等于每个函数期望值的和。

$$E_{P(x)}\{f(X) + g(X)\} = E_{P(x)}\{f(X)\} + E_{P(x)}\{g(X)\}$$



第三章：机器学习的数学基础

期望值

两种最常见的期望：均值和方差。

1) 均值定义为 $E_{P(x)}\{X\}$;

2) 方差用于度量随机变量的变化程度，定义为实际值与均值之差的平方的期望值，即

$$\text{var}\{X\} = E_{P(x)} \left\{ (X - E_{P(x)}\{X\})^2 \right\}$$

通常，方差越大，则随机变量的取值离均值越远；



第三章：机器学习的数学基础

连续型随机变量-概率密度函数

- 与离散随机变量不同，我们无法写出连续随机变量所有可能的结果，从而也无法给出随机变量取特定值的概率。
- 对于连续随机变量，需要定义概率分布的连续模拟，即 $p(x)$ ，也叫概率密度函数。
- 计算连续随机变量 X 落入某一特定区间的概率，可通过计算 $p(x)$ 关于 x 在这个区间上的定积分：

$$P(x_1 \leq X \leq x_2) = \int_{x_1}^{x_2} p(x) dx$$



第三章：机器学习的数学基础

连续型随机变量-概率密度函数

- 性质1: 如果 $x_1 \leq X \leq x_2$, $\int_{x_1}^{x_2} p(x)dx = 1$
- 性质2: $p(x) \geq 0$ (概率密度函数不能为负)
- 性质3: 概率密度函数不是概率, 没有上界, $p(x)$ 可以大于1

联合概率密度

$$P(x_1 \leq X \leq x_2, y_1 \leq Y \leq y_2) = \int_{x=x_1}^{x_2} \int_{y=y_1}^{y_2} p(x, y) dx dy$$

条件概率密度

$$P(x_1 \leq X \leq x_2 | Y = y) = \int_{x=x_1}^{x_2} p(x | Y = y) dx$$

边缘概率密度

$$p(y) = \int_{x=x_1}^{x_2} p(y, x) dx$$



第三章：机器学习的数学基础

几种常见的连续概率密度函数 $p(y)$

□ 均匀密度函数

$$p(y) = \begin{cases} r, a \leq y \leq b \\ 0, \text{其他} \end{cases}$$

问题：r等于多少？(利用性质1)

□ 高斯密度函数

高斯概率密度函数(高斯分布或正态分布)最常用。

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