Lecture 03: 概率与贝叶斯理论

Kai Yu and Yanmin Qian

Cross Media Language Intelligence Lab (X-LANCE)
Department of Computer Science & Engineering
Shanghai Jiao Tong University

2021

内容纲要

- 面向贝叶斯推理的概率论回顾
- 离散和连续分布
- 统计量
- 贝叶斯公式
- 联合概率和多变量分布
- 高斯混合模型
- 信息和熵
- 随机过程

概率

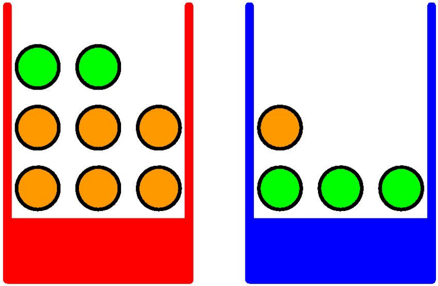


图: 苹果和橘子

概率 是某事件发生或某个声明为真的可能性的度量或估计。

- 不确定性： 来源于非确定性的随机实验的结果
- 可数性 是用于区分离散和连续的关键
- 事件 $\rightarrow$ 随机变量

  离散 数出上海交通大学计算机系不同性别的人数
  连续 学生早上到校的时间

例如：扔一个骰子，可能有有种可能

$$\Omega = \{\mathrm{up}, \mathrm{down}, \mathrm{left}, \mathrm{right}, \mathrm{front}, \mathrm{back}\}$$

映射为一个随机变量 x 的值，$x \in \mathcal{X}$

$$\mathcal{X} = \{1, 2, 3, 4, 5, 6\}$$

- x 的数值比较没有意义
- 表示：**概率质量函数 (Probability Mass Function, PMF)**

$$\sum_{x \in \mathcal{X}} P(x) = 1, \quad P(x) \geqslant 0 \quad \forall x$$

- 离散 PMF 通常是一个查找表

x	1	2	3	4	5	6
P(x)	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

伯努利实验: 考虑抛硬币实验，引入随机变量 $x \in \{0, 1\}$，1 表示成功（“正面”），0 表示失败（“反面”）。

图: 抛硬币

伯努利实验: 考虑抛硬币实验，引入随机变量 $x \in \{0,1\}$，1 表示成功（“正面”），0 表示失败（“反面”）。伯努利分布可以定义为

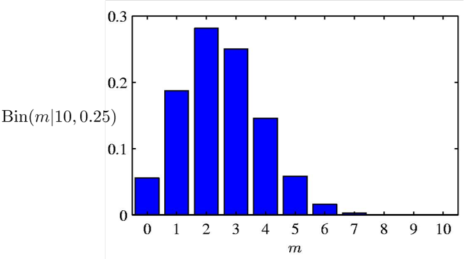
x	1	0
P(x)	μ	$1-\mu$

$$P(x=1|\mu)=\mu$$

$$\mathrm{Bern}(x|\mu)=\mu^x(1-\mu)^{1-x}$$

Q: 抛 N 次硬币的实验结果的分布是什么样的?

离散概率分布
二项分布



抛 N 次硬币:

$$P(m\ heads|N,\mu) = C(N,m)\mu^m(1-\mu)^{N-m}$$

$$C(N,m) = \frac{N!}{m!(N-m)!}$$

Q: 如果抛的不是硬币，而是一个骰子，那么结果的分布是什么样的？

离散概率分布

多项分布

- 1-of-K 编码： $x = (0, 0, 1, 0, 0, 0)^T$

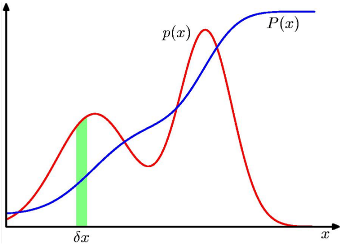
$$P(x|\mu) = \prod_{k=1}^{K} \mu_k^{x_k}$$

$$P(y = k|\mu) = P(x_k = 1) = \mu_k$$

$$\forall k : \mu_k \leq 0, \qquad \sum_{k=1}^{K} \mu_k = 1$$

Q1: 多项分布有多少自由参数？
Q2: 离散变量的取值一定是有限的吗？

连续概率分布



$$P(x \in (a,b)) = \int_a^b p(x)dx$$

$$P(z) = \int_{-\infty}^{z} p(x)dx$$

$$p(x) \geqslant 0, \quad \int_{-\infty}^{\infty} p(x)dx = 1$$

概念对比：

- **概率质量函数** vs **概率密度函数**
- **概率值**（Probability）vs **似然度**（Likelihood）

Q: pdf 的最大值是多少?

概率公式

- 加法公式：

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

- 乘法公式：

$$P(X,Y) = P(Y|X)P(X)$$

- 独立性：

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

$$P(X|Y) = P(X)$$

$$P(Y|X) = P(Y)$$

随机变量的和

给定离散变量 $Z = X + Y$, $P(X)$ 和 $P(Y)$, 且 X 和 Y 独立，那么 $P(Z)$ 的分布是什么？

$$\begin{aligned} P(Z=z) &= \sum_x P(X=x, Y=z-x) \\ &= \sum_x P(X=x)P(Y=z-x) \\ &= \sum_y P(X=z-y)P(Y=y) \end{aligned}$$

这就是 $P(X)$ 和 $P(Y)$ 的**卷积**。
卷积就是计算一个函数和另一个函数的镜像的重叠区域。

将 $P(X=a)$ 简记为 $P_x(a)$。令 X 和 Y 是投两个骰子的结果，Z 是这两者的和，那么

$$
\begin{aligned}
P_z(2) &= P_x(1)P_y(1) = \frac{1}{6}\times\frac{1}{6} = \frac{1}{36} \\
P_z(3) &= P_x(1)P_y(2) + P_x(2)P_y(1) = \frac{1}{6}\times\frac{1}{6} + \frac{1}{6}\times\frac{1}{6} = \frac{1}{36} \\
P_z(4) &= P_x(1)P_y(3) + P_x(2)P_y(2) + P_x(3)P_y(1) = \frac{3}{36} \\
&\vdots
\end{aligned}
$$

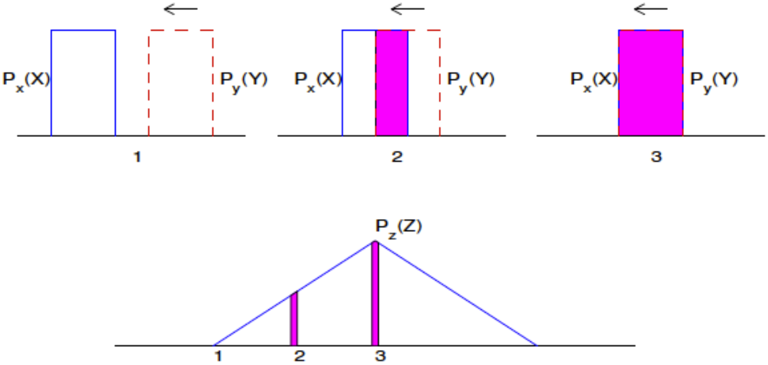
Z	2	3	4	5	6	7	8	9	10	11	12
P(Z)	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

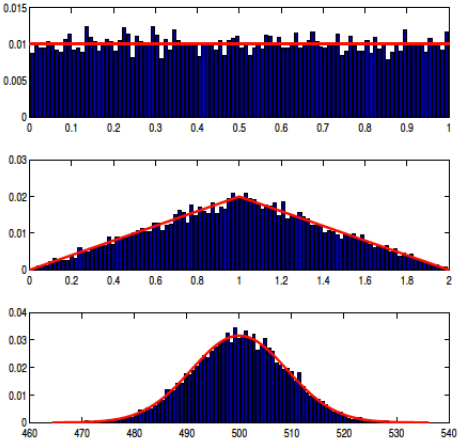
两个均匀分布的随机变量经过**卷积**后的和是一个三角分布。

卷积

可视化

$$P_z(Z) = \sum_Y P_x(Z-Y)P_y(Y) = \sum_X P_x(X)P_y(Z-X)$$





- 随机变量的和
  - u_n 是独立同分布 (i.i.d) 的，$p(u_n)$ 服从均匀分布 $[0,1]$
  - $x = \sum_{n=1}^{N} u_n$
- $p(x)$ 依赖于 N
  - $N = 1$ (上): 均匀分布
  - $N = 2$ (中): 三角分布 [0,2]
  - $N = 1000$ (下): 近似高斯
- 中心极限定理
  随着 $N \to \infty$, $p(x)$ 趋近于高斯分布且与 $p(u)$ 无关

一个均值为 μ、方差为 σ^2 的高斯 (正态) 分布的表达式为

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

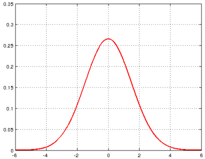


图: 均值为 0，方差为 1.5 的高斯

Q: 事件可以既离散又连续吗？

随机事件的**概率**指重复实验时某一个实验结果出现的相对频率。

- **测度理论**可用于严格定义概率函数
- **前提：** 仔细定义的事件**集合**
  - Ω 是一个任意的非空集合
  - $\mathcal{F}$ 是 Ω 子集的集合
- **定义和性质：** 概率 P 是在 Ω 上定义的一个函数，使得
  1. **非负性：** 对任意 $A \in \mathcal{F}$，$0 \leq P(A) \leq 1$, $P(\emptyset) = 0$
  2. **归一性：** $P(\bigcup_{i=1}^{\infty} A_i) = P(\Omega) = 1$
  3. **可加性：** 如果 $A_i \in \mathcal{F}, i = 1, 2, \cdots$ 且对任意 $i \neq j$ 有 $A_i \cap A_j = \emptyset$，那么

$$P(\bigcup_{i=1}^{N} A_i) = \sum_{i=1}^{N} P(A_i) \qquad (1)$$

- **期望**是概率函数的加权平均值。它是从一个分布中采样后的均值。

$$\mathbb{E}[f(x)] = \sum_x P(x)f(x)$$

$$\mathbb{E}[f(x)] = \int p(x)f(x)\mathrm{d}x$$

- 条件期望（离散）

$$\mathbb{E}_x[f(x)|y] = \sum_x P(x|y)f(x)$$

- 近似期望（离散和连续）

$$\mathbb{E}[f(x)] \simeq \frac{1}{N}\sum_{n=1}^{N} f(x_n)$$

$$\mathrm{var}[f(x)] = \mathbb{E}\left[(f(x) - \mathbb{E}[f(x)])^2\right] = \mathbb{E}\left[f(x)^2\right] - \mathbb{E}[f(x)]^2$$

$$\mathrm{var}[f(x)] \simeq \frac{1}{N}\sum_{n=1}^{N}(f(x_n) - \mathbb{E}[f(x)])^2$$

$$\mathrm{var}[f(x)] \simeq \frac{1}{N-1}\sum_{n=1}^{N}\left(f(x_n) - \frac{1}{N}\sum_{m=1}^{N} f(x_m)\right)^2$$

例

$$p(x) = \begin{cases} 1 - \frac{x}{2}, & 0 \leq x \leq 2 \\ 0, & \text{otherwise} \end{cases}$$

$$\mathbb{E}\left[x\right] = \int_0^2 \left(x - \frac{x^2}{2}\right) dx = \left[\frac{x^2}{2} - \frac{x^3}{6}\right]_0^2 = \frac{2}{3}$$

$$\mathbb{E}\left[x^2\right] = \int_0^2 \left(x^2 - \frac{x^3}{2}\right) dx = \left[\frac{x^3}{3} - \frac{x^4}{8}\right]_0^2 = \frac{2}{3}$$

因此

$$u = \mathbb{E}\left[x\right] = \frac{2}{3}$$

$$\sigma^2 = \mathbb{E}[x^2] - \mu^2 = \frac{2}{9}$$

- 随机变量 X 与 Y 独立： $P(X,Y)=P(X)P(Y)$
- 统计量计算： $\mathbb{E}[X]=\sum_x xP(x)$， $\mathrm{Var}[X]=\mathbb{E}[(x-\mu_X)^2]$
- $Z=XY$， Z 的期望为

$$
\begin{aligned}
\mathbb{E}[XY] &= \sum_x \sum_y xyP(x,y) \\
&= \sum_x \sum_y xyP(x)P(y) \\
&= \sum_x xP(x) \sum_y yP(y) \\
&= E[X]E[Y]
\end{aligned}
$$

Q: Z 的期望和方差是多少？ $Z=X+Y$， $Z=nX$

$$
\begin{aligned}
\mathbb{E}[X+Y] &= \sum_x \sum_y (x+y)P(x,y) \\
&= \sum_x \sum_y xP(x,y) + \sum_y \sum_x yP(x,y) \\
&= \sum_x x \sum_y P(x,y) + \sum_y y \sum_x P(x,y) \\
&= \sum_x xP(x) + \sum_y yP(y) = \mathbb{E}[X] + \mathbb{E}[Y]
\end{aligned}
$$

$$
\begin{aligned}
\mathrm{Var}[X+Y] &= \mathbb{E}[(X+Y)^2] - (\mathbb{E}[X+Y])^2 \\
&= \mathbb{E}[X^2 + Y^2 + 2XY] - \left((\mathbb{E}[X])^2 + (\mathbb{E}[Y])^2 + 2\mathbb{E}[Y]\mathbb{E}[X]\right) \\
&= \left(\mathbb{E}[X^2] - (\mathbb{E}[X])^2\right) + \left(\mathbb{E}[Y^2] - (\mathbb{E}[Y])^2\right) \\
&= \mathrm{Var}[X] + \mathrm{Var}[Y]
\end{aligned}
$$

统计量

独立性与期望/方差的关系

$$
\begin{aligned}
\mathbb{E}[nX] &= \sum_x nxP(x) \\
&= n\sum_x xP(x) \\
&= n\mathbb{E}[X]
\end{aligned}
$$

$$
\begin{aligned}
\mathrm{Var}[nX] &= \mathbb{E}[(nX)^2] - (\mathbb{E}[nX])^2 \\
&= n^2\mathbb{E}[X^2] - n^2(\mathbb{E}[X])^2 \\
&= n^2\mathrm{Var}[X]
\end{aligned}
$$

贝叶斯公式给出了两个随机变量之间的关系

$$P(X|Y) = \frac{P(Y|X)P(X)}{P(Y)}$$

- **先验或边缘概率**:
  $P(X)$ 是 X 的**先验或边缘**分布，因为它没有考虑任何 Y 的信息
- **后验或条件概率**:
  $P(X|Y)$ 描述了 X 在 Y 条件下的概率，也就是说给定 Y 发生时 X 的概率

贝叶斯公式中的似然:

- 连续分布 - 2 维图像

$$p(y|x) = \frac{p(x|y)p(y)}{p(x)}$$

- 混合分布 - 语音识别

$$P(Y|x) = \frac{p(x|Y)P(Y)}{p(x)}$$

概率的解释

- **贝叶斯:** 概率度量了置信程度，贝叶斯公式连接了考虑证据前后的置信程度。
- **频率:** 概率度量了实验结果的占比。$P(X|Y)$ 是在得到 Y 的实验结果中出现 X 的占比。

Monty Hall 问题

假如你需要选择一个门：只有一个门后是汽车，其他门后都只有山羊。你选择了一个门，比如第 3 个，知道答案的主持人打开了另一个门，比如第 1 个，发现第 1 个门后是山羊。然后他问：

Q: 你要改选第 2 个门吗？

Monty Hall 问题

- 随机实验：你选择了一个门，主持人打开了一个门
- 实验前（先验）：

$$P(D_2 = car) = P(D_3 = car) = P(D_1 = car) = \frac{1}{3}$$

$$P(D_2 = goat) = P(D_3 = goat) = P(D_1 = goat) = \frac{2}{3}$$

- 实验表示（条件）：
  H：主持人打开的门
  Y：你选择的门

Monty Hall 问题

- 实验后（后验）:

$$
\begin{aligned}
&P(D_2 = car|H = 1, Y = 3)\\
=&\frac{P(H = 1|D_2 = car, Y = 3)P(D_2 = car|Y = 3)}{P(H = 1|Y = 3)}\\
=&\frac{1 \times \frac{1}{3}}{\frac{1}{2}} = \frac{2}{3}
\end{aligned}
$$

$$
\begin{aligned}
&P(H = 1|Y = 3)\\
=&\sum_{D_3 \in car, goat} P(H = 1|D_3, Y = 3)P(D_3|Y = 3)\\
=&\frac{1}{2} \times \frac{1}{3} + \frac{1}{2} \times \frac{2}{3} = \frac{1}{2}
\end{aligned}
$$

联合概率和多变量分布

- **联合概率**是所有随机变量 $x_i \in \mathcal{X}_i, \quad i = 1, \cdots, d$ 同时发生的概率。
- 为了记号方便，经常将联合事件合并为一个向量 $\mathbf{x} \in \mathcal{X}$，其中

$$\mathbf{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_d \end{bmatrix} \quad \mathcal{X} = \begin{bmatrix} \mathcal{X}_1 \\ \vdots \\ \mathcal{X}_d \end{bmatrix}$$

- **多变量分布**因此写作

$$P(\mathbf{x}) \geqslant 0, \quad \sum_{\mathbf{x} \in \mathcal{X}} P(\mathbf{x}) = 1$$

多变量分布的统计量

以连续分布为例

- **均值**：标量均值的向量扩展

$$\boldsymbol{\mu} = \mathbb{E}[\mathbf{x}] = \int_{\mathbf{x}} \mathbf{x}\, p(\mathbf{x})\, d\boldsymbol{x}$$

- **协方差**：根据维度之间相关性的矩阵扩展

$$\begin{aligned}\boldsymbol{\Sigma} &= \mathbb{E}[(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^\top] = \mathbb{E}[\mathbf{x}\mathbf{x}^\top] - \boldsymbol{\mu}\boldsymbol{\mu}^\top \\ &= \begin{bmatrix} \mathbb{E}[x_1^2]-\mu_1^2 & \cdots & \mathbb{E}[x_1 x_d]-\mu_1\mu_d \\ \vdots & \ddots & \vdots \\ \mathbb{E}[x_d x_1]-\mu_d\mu_1 & \cdots & \mathbb{E}[x_d^2]-\mu_d^2 \end{bmatrix}\end{aligned}$$

协方差矩阵永远对称

相关性

变量 x 和 y 的协方差矩阵可写为

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_{xx} & \sigma_{xy} \\ \sigma_{xy} & \sigma_{yy} \end{bmatrix}$$

其中

$$\sigma_{xy} = \mathbb{E}[(x - \mu_x)(y - \mu_y)] \text{ and } \sigma_x^2 = \sigma_{xx}.$$

协方差系数 ρ 定义为

$$\rho = \frac{\sigma_{xy}}{\sigma_x \sigma_y}, \quad -1 \leq \rho \leq 1$$

当 $\rho = 0$ 时，称两个随机变量为**不相关**.

Q: 不相关 $=$ 独立?

多变量高斯分布

d 维多变量高斯分布的形式为

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

如果协方差矩阵为对角矩阵，那么表达式可以简化为

$$p(\mathbf{x}) = \prod_{i=1}^{d} \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left\{ -\frac{(x_i-\mu_i)^2}{2\sigma_i^2} \right\}$$

多变量高斯分布

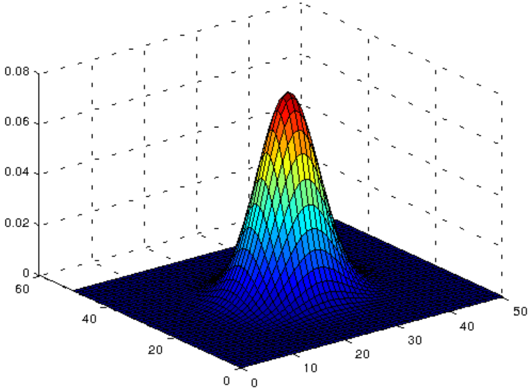
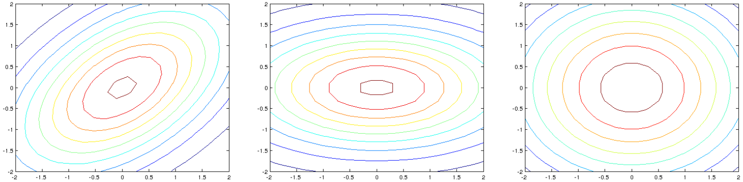


图: 一个 2 维高斯

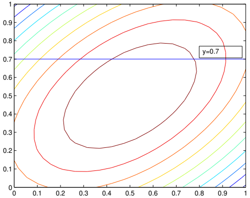
多变量高斯分布

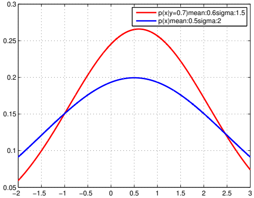


$$\Sigma = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix} \qquad \Sigma = \begin{bmatrix} 3 & 0 \\ 0 & 1 \end{bmatrix} \qquad \Sigma = \begin{bmatrix} 3 & 0 \\ 0 & 3 \end{bmatrix}$$

图: 不同 Σ 下的 2 维高斯

多变量高斯的性质





2 维高斯，$\mu = \begin{bmatrix} 0.5 \\ 0.5 \end{bmatrix}$ $\Sigma = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$

一个高斯的条件边缘概率也是一个高斯 (右图)。

p(x|y=0.7)mean:0.6sigma:1.5
p(x)mean:0.5sigma:2

多变量高斯的性质

- 每一个分量的边缘分布 $p(x_i)$ 都是一个高斯。
- 任意一个子集的联合边缘分布 $p(x_i, x_j, ..)$ 都是一个高斯。
- 条件分布 $p(x_i|x_j)$ 是一个高斯。
- 如果 $\mathbf{x}$ 服从高斯分布，$\mathbf{y} = A\mathbf{x} + \mathbf{b}$，那么 $\mathbf{y}$ 是一个均值为 $A\boldsymbol{\mu}_x + b$，方差为 $A\Sigma_x A^\top$ 的高斯。

多变量高斯的期望和方差

$$\mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\mathbf{x}-\boldsymbol{\mu})\right\}$$

$$\begin{aligned}\mathbb{E}[\mathbf{x}] &= \frac{1}{(2\pi)^{D/2}} \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} \int \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\mathbf{x}-\boldsymbol{\mu})\right\} \mathbf{x} \mathrm{d}\mathbf{x} \\ &= \frac{1}{(2\pi)^{D/2}} \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} \int \exp\left\{-\frac{1}{2}\mathbf{z}^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} \mathbf{z}\right\} (\mathbf{z}+\boldsymbol{\mu}) \mathrm{d}\mathbf{z}\end{aligned}$$

$\mathbf{z}$ 有反对称性，因此有 $\mathbb{E}[\mathbf{x}] = \mu$

$$\mathbb{E}\left[\mathbf{x}\mathbf{x}^{\mathrm{T}}\right] = \boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}} + \boldsymbol{\Sigma}$$

$$\mathrm{cov}[\mathbf{x}] = \mathbb{E}\left[(\mathbf{x}-\mathbb{E}[\mathbf{x}])(\mathbf{x}-\mathbb{E}[\mathbf{x}])^{\mathrm{T}}\right] = \boldsymbol{\Sigma}$$

近似任意分布

- 有参数形式的 $p(x)$ 存在缺点
- 可以用分布作为基，近似任意分布
- **Product of Experts:**

$$p(x) = \frac{1}{Z}\prod_{m=1}^{M} p_m(x)$$

- **Mixture of Experts:**

$$p(x) = \sum_{m=1}^{M} c_m p_m(x)$$

Q: 如果 $p_m(x)$ 都是高斯，那么 PoE 和 MoE 都是高斯吗？

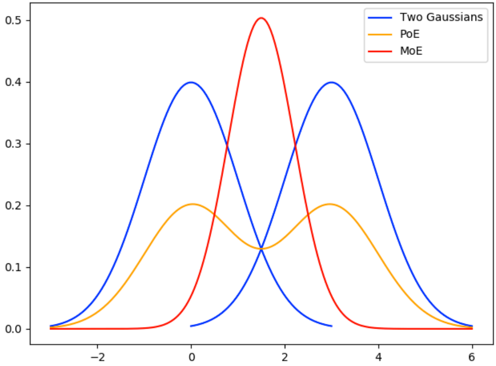
近似任意分布

- **Product of Experts:**

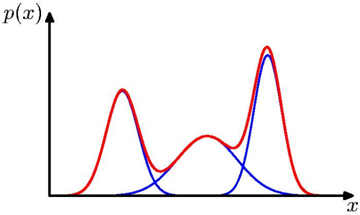
$$p(x) = \frac{1}{Z}\prod_{m=1}^{M} p_m(x)$$

- **Mixture of Experts:**

$$p(x) = \sum_{m=1}^{M} c_m p_m(x)$$



高斯混合模型



将 M 个高斯模型合并为一个复杂模型

$$
\begin{aligned}
p(\mathbf{x}) &= \sum_{m=1}^{M} P(m)p(\mathbf{x}|m) \\
&= \sum_{m=1}^{M} c_m \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m)
\end{aligned}
$$

- $c_m = P(m)$：**混合系数**，$\forall c_m > 0$，$\sum_{m=1}^{M} c_m = 1$
- $\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m) = p(\mathbf{x}|m)$：**混合成分**
- **参数集为**：**权重** $c_1, \ldots, c_M$，**均值** $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_M$，**方差** $\boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_M$

Q: D 维变量 M 个高斯混合成分的分布有多少个参数？

高斯混合模型 (Gaussian Mixture Model, GMM) 可以用于近似任意分布。

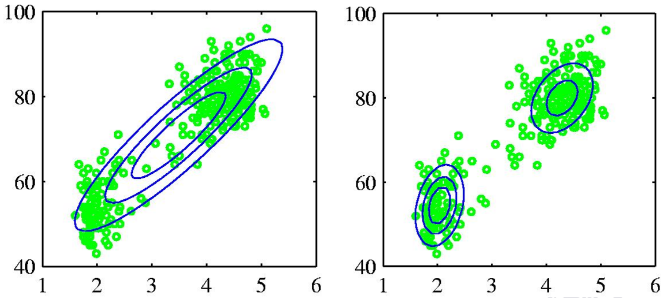
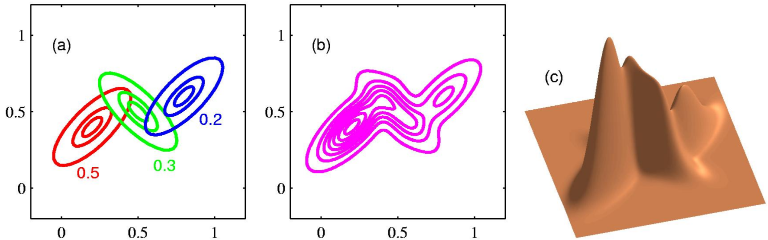


图: 单高斯和双高斯建模

高斯混合模型

从高斯混合模型中采样



从高斯混合模型中采样可分为 2 步：

1. 根据 $c_1, \ldots, c_M$，从 M 个混合成分中采样出第 m 个高斯。
2. 从 $\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m)$ 中采样出一个样本。

- **信息:** “传达的消息”，它包含了不确定性。
- **不确定性:** 由事件发生的概率度量，且与之成比例。一个事件越不确定，就需要越多的信息来消除它的不确定性。
- 离散变量 x 中的信息是:

$$I(x) = -\log_2 P(x)$$

- **熵**: 整个信源的平均信息就是对整体不确定性的度量，单位是比特 (bits)。

$$H = \mathbb{E}[-\log_2 P(x)] = -\sum_{x \in X} P(x) \log_2 P(x)$$

$$H = \mathbb{E}[I(x)] = -\int_{-\infty}^{\infty} p(x) \log_e(p(x))dx$$

- 熵在编码理论、统计物理、机器学习中都是一个重要的量。
- 熵是随机变量的信息的期望。
- 熵是分布 $p(x)$ 的函数。

Q: x 是一个有 8 种可能状态的离散变量，需要用多少 bits 来传输 x 的状态?

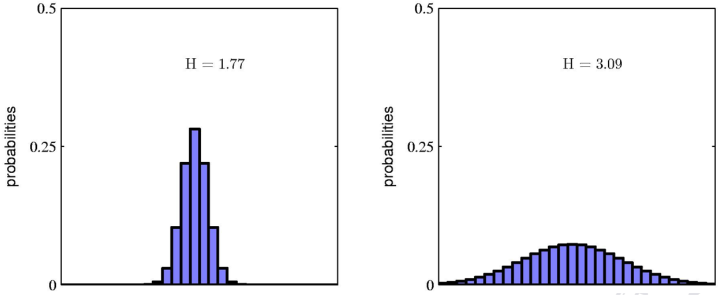
x	a	b	c	d	e	f	g	h
$p(x)$	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{1}{64}$	$\frac{1}{64}$	$\frac{1}{64}$	$\frac{1}{64}$
code	0	10	110	1110	111100	111101	111110	111111

$$\begin{aligned} \mathrm{H}[x] &= -\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{4}\log_2\frac{1}{4} - \frac{1}{8}\log_2\frac{1}{8} - \frac{1}{16}\log_2\frac{1}{16} - \frac{4}{64}\log_2\frac{1}{64} \\ &= 2 \text{ bits} \end{aligned}$$

$$\begin{aligned} \text{average code length} &= \frac{1}{2}\times 1 + \frac{1}{4}\times 2 + \frac{1}{8}\times 3 + \frac{1}{16}\times 4 + 4\times\frac{1}{64}\times 6 \\ &= 2 \text{ bits} \end{aligned}$$

信息和熵

熵



- 伯努利分布 $\mathcal{B}(1)$

$$\mathrm{Bern}(x|\mu) = \mu^x (1-\mu)^{(1-x)}$$

$$\mathrm{H}[\mathrm{Bern}(x|\mu)] = -(1-\mu)\ln(1-\mu) - \mu \ln \mu$$

- 高斯分布 $\mathcal{N}(\mu, \sigma^2)$

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

$$\mathrm{H}[\mathcal{N}(x|\mu,\sigma^2)] = \frac{1}{2}\ln 2\pi e\sigma^2$$

- 条件熵是条件分布 $p(y|x)$ 的熵的期望。

$$\begin{aligned} \mathrm{H}[y|x] &= \sum_{x'} P(x)\mathrm{H}(y|x=x') \\ &= -\sum_{x} P(x) \sum_{y} P(y|x) \ln P(y|x) \end{aligned}$$

$$\mathrm{H}[y|x] = -\iint p(y,x)\ \ln p(y|x)\ \mathrm{d}y\ \mathrm{d}x$$

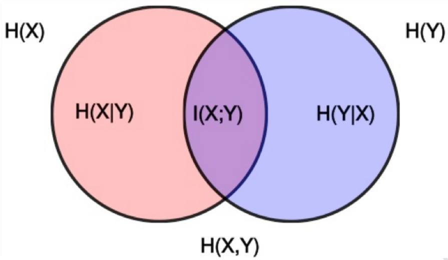
- 联合熵是条件熵和边缘熵之和

$$\mathrm{H}[x,y] = \mathrm{H}[y|x] + \mathrm{H}[x]$$

互信息是对称的，它是边缘熵和条件熵的差。

$$\mathrm{I}[x,y] \equiv -\iint p(x,y)\ln\left(\frac{p(x)p(y)}{p(x,y)}\right)\mathrm{d}x\ \mathrm{d}y$$

$$\mathrm{I}[x,y] = \mathrm{H}[x] - \mathrm{H}[x|y] = \mathrm{H}[y] - \mathrm{H}[y|x]$$



信息和熵
KL 距离

- KL 距离广泛用于描述两个分布的差异

$$\begin{aligned}\mathrm{KL}(p\|q) &= -\int p(x)\ln q(x)\mathrm{d}x - \left(-\int p(x)\ln p(x)\mathrm{d}x\right)\\ &= -\int p(x)\ln\left\{\frac{q(x)}{p(x)}\right\}\mathrm{d}x\end{aligned}$$

$$\mathrm{KL}(p\|q) \le 0 \qquad \mathrm{KL}(p\|q) \neq \mathrm{KL}(q\|p)$$

- 互信息和 KL 距离的关系

$$\begin{aligned}\mathrm{I}[x,y] &\equiv \mathrm{KL}(p(x,y)\|p(x)p(y))\\ &= -\iint p(x,y)\ln\left(\frac{p(x)p(y)}{p(x,y)}\right)\mathrm{d}x\,\mathrm{d}y\end{aligned}$$

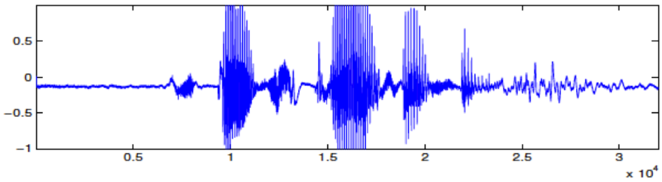
- 交叉熵是两个分布 $q(x)$ 和 $p(x)$ 之间信息的期望

$$\mathrm{H_c}(P,Q) = -\sum_x P(x)\log_2 Q(x)$$

$$\mathrm{H_c}(p,q) = -\int_x p(x)\ln q(x)dx$$

- 交叉熵是非对称的
- 交叉熵广泛用于深度学习的训练准则：$-\log_2 Q_{NN}(x=\mathrm{label})$

随机过程



- x_t 有两层随机性
- 确定性部分：时间（或空间）索引
- 随机过程是由随机变量在任意索引下的联合概率决定的
- 平稳性：联合概率是时不变的

$$P_X(x_{t_1}, \ldots, x_{t_k}) = P_X(x_{t_1+\tau}, \ldots, x_{t_k+\tau})$$