Lec02 - Background

基本离散变量分布

1. 伯努利分布
2. 分类分布

联合分布示例

1. 单像素多通道

对于一个单独的像素点，它有三个通道，分别对应R,G,B,每个通道的取值都是$\{0,\dots,255\}$，那么从这个联合分布中采样$(r,g,b)\sim p(R,G,B)$等价于在该像素中随机生成一种颜色。

我们需要多少个参数来唯一确定这种联合分布的情况？256^3-1，因为每一种情况都对应着一个概率，而最后的一种情况则可以由前面的所有情况的概率之和计算得到。

2. 多像素单通道

对于一个28×28像素的图片，我们假设每一个像素都是一个单通道的伯努利变量，即$X_i\in\{0,1\}=\{Black,White\}$

一共有多少种可能的分布情况？$2^{28\times 28}$，对于每一个像素都有两个可能的取值。

需要多少个参数来唯一确定这种联合概率分布的情况？$2^{28\times 28}-1$，因为每一种情况都对应着一个概率，而最后的一种情况可以由前面的所有情况的概率之和计算得到。

通过独立性建模

假设所有随机变量相互独立

此时联合分布就可以表示为边际分布的乘积：

$$p(x_1,\dots,x_n)=p(x_1)p(x_2)\cdots p(x_n)=\prod_{i=1}^n p(x_i)$$

此时对于**多像素单通道**问题，我们仍然有$2^{28\times 28}$种可能性，但是现在我们只需要28×28个参数来唯一确定其联合分布了，因为我只需要确定$p(x_1),\dots,p(x_n)$即可。

但是这种假设**过于绝对**，根据此假设构建出来的模型很难有泛用性。

两个重要规律

1. **链式法则**

   令$S_1,\dots,S_n$为事件，$p(S_i)>0$，则

$$p(S_1 \cap S_2 \cap \cdots \cap S_n) = p(S_1)\, p(S_2|S_1)\, p(S_3|S_2 \cap S_1) \cdots p(S_n|S_1 \cap \cdots \cap S_{n-1})$$

链式法则提供了一种将联合分布写成相对简单的概率分布的乘积的形式，这些相对简单的概率分布是边际分布或条件概率分布。

2. **贝叶斯定理**

令S_1, S_2为事件，$p(S_1), P(S_2) > 0$，则

$$p(S_1|S_2) = \frac{p(S_1 \cap S_2)}{P(S_2)} = \frac{p(S_2|S_1)\, p(S_1)}{p(S_2)}$$

其意义是能够基于先验概率和S_2在S_1条件下发生的似然性为条件写出一个事件相对于另一个事件的条件概率。

通过条件性独立建模

使用链式法则对**多像素单通道**问题进行建模

$$p(x_1, \ldots, x_n) = p(x_1)\, p(x_2|x_1) \cdots p(x_n|x_1, \ldots, x_{n-1})$$

需要多少个参数来唯一确定这种联合分布的情况？ $1 + 2 + \cdots + 2^{n-1} = 2^n - 1$

- $p(x_1)$需要1个参数。
- $p(x_2|x_1 = 0)$需要一个参数，$p(x_2|x_1 = 1)$需要一个参数，因此$p(x_2|x_1)$需要2个参数。
- 以此类推，第n项需要2^{n-1}个参数。

现在假设$X_{i+1} \perp X_1, \ldots, X_{i-1}|X_i$，即在给定第$i$个元素的情况下，第$i+1$个元素仅仅与第$i$个元素相关，而与其他元素相互独立。此时重新审视$(4)$式能够发现：

$$p(x_1, \ldots, x_n) = p(x_1)\, p(x_2|x_1)\, p(x_3|x_2) \cdots p(x_n|x_{n-1})$$

需要多少个参数来唯一确定这种联合分布的情况？ $2n - 1$

- $p(x_1)$需要1个参数。
- $p(x_2|x_1), \ldots, p(x_n|x_{n-1})$需要两个参数。

但是这种假设仍然过于绝对，因此我们采用更先进的办法：贝叶斯网络

贝叶斯网络

- 利用条件性参数化，而非联合参数化。
- 对于每一个随机变量X_i我们都唯一确定X_i关于一个随机变量组$\mathbf{X}_{\mathbf{A}_i}$的条件概率$p(x_i|\mathbf{x}_{\mathbf{A}_i})$，其中随机变量组$\mathbf{X}_{\mathbf{A}_i}$中的所有随机变量称为变量$X_i$的父变量，基于这个假设，我们能够得到以下式子：

$$p(x_1, \ldots, x_n) = \prod_i p(x_i|\mathbf{x}_{\mathbf{A}_i})$$

- 一个贝叶斯网络实际上可以由一幅**有向无环图**(directed acyclic graph, DAG)来表示。图$G = (V, E)$具有以下性质：
  - 每个事件(元素)X_i都由图中的一个节点$i \in V$来表示。
  - 每个节点的条件概率分布都由$p(x_i|\mathbf{x}_{Pa(i)})$表示。
- 定义其联合分布：

$$p(x_1,\dots,x_n)=\prod_{i\in V}p(x_i|\mathbf{x}_{Pa(i)})$$

- **复杂度与$|Pa(i)|$有关，而与$|V|$无关。**

朴素贝叶斯理论用于单标签预测

检测垃圾邮件(用标签Y表示，$Y=1$表示垃圾邮件，$Y=0$表示正常邮件)

- 令索引$1\dots n$表示每一个英文单词（或中文汉字也可以）。
- 如果某单词i在邮件内容中出现，则$X_i=1$，如果没有出现，则$X_i=0$
- 训练集中的邮件遵循某种分布$p(Y,X_1,\dots,X_n)$

假设每个单词X_i在给定标签Y的情况下与其他单词条件独立(朴素贝叶斯)，则有：

$$p(y,x_1,\dots,x_n)=p(y)\prod_{i=1}^{n}p(x_i|y)$$

接下来利用贝叶斯法则进行预测：

$$p(Y=1|x_1,\dots,x_n)=\frac{p(Y=1)\prod_{i=1}^{n}p(x_i|Y=1)}{\sum_{y=\{0,1\}}p(Y=y)\prod_{i=1}^{n}p(x_i|Y=y)}$$

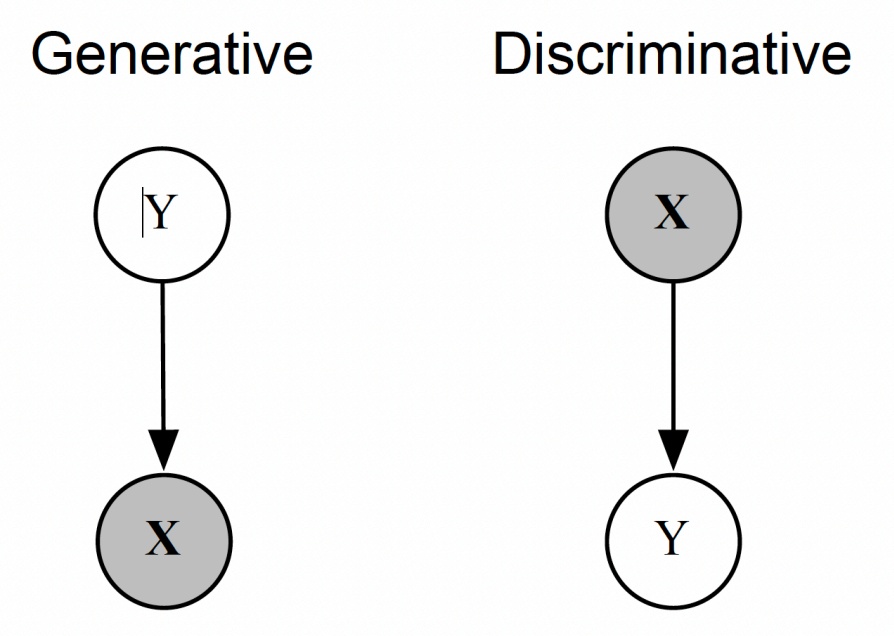
判别模型和生成模型

利用链式法则：

$$p(Y,\mathbf{X})=p(\mathbf{X}|Y)p(Y)=p(Y|\mathbf{X})p(\mathbf{X})$$

其中Y是输出，$\mathbf{X}$是输入。

这两种情况分别对应了生成模型和判别模型。



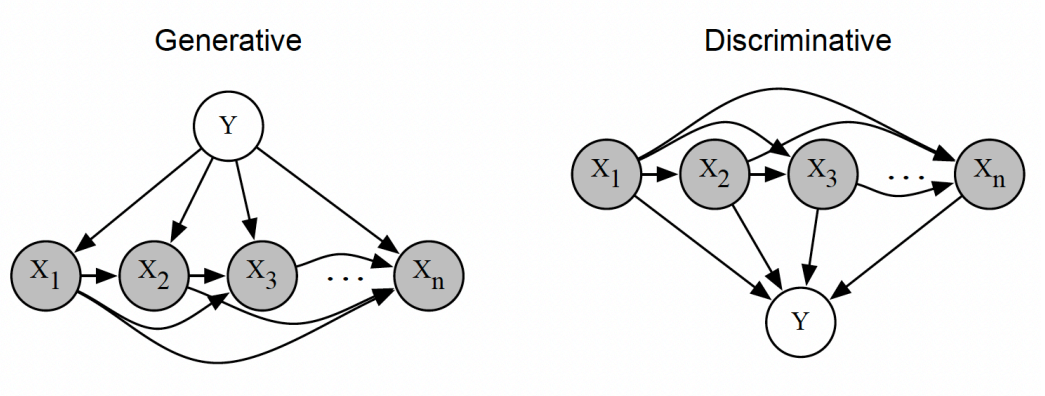
- 实际上我们所需要得到的仅仅是$p(Y|\mathbf{X})$，因为我们无法得知输入的概率分布是什么。
- 在生成模型中，我们需要学习$p(Y)$和$p(\mathbf{X}|Y)$，然后再利用贝叶斯定理计算$p(Y|\mathbf{X})$。

- 在判别模型中，我们只需要学习$p(Y|\mathbf{X})$。
- **判别模型的局限性**：我们只能在**给定输入**$\mathbf{X}$的情况下，才能得到输出Y。我们无法学习输入与输入之间的关系。

接下来我们将上述式子进行进一步链式法则展开：

$$p(Y,\mathbf{X})=p(Y)p(X_1|Y)p(X_2|Y,X_1)\cdots p(X_n|Y,X_1,\ldots,X_{n-1})$$

$$p(Y,\mathbf{X})=p(X_1)p(X_2|X_1)p(X_3|X_1,X_2)\cdots p(Y|X_1,\ldots,X_{n-1},X_n)$$



- 在生成模型中，$p(Y)$可以简单表示，但是我们需要解决怎样表示$p(X_i|X_{Pa(i)},Y)$。
- 在判别模型中，不需要考虑$p(X)$，但是我们需要解决怎样表示$p(Y|\mathbf{X})$。

逻辑回归

什么是回归?

对于一个随机伯努利向量$\mathbf{x}$，它有2^n种可能的取值，最终会形成一个超级大的表格，机器无法装下这么大的表格，因此我们需要假设存在一种简单的函数$f(\mathbf{x},\alpha)$，我能够将其应用于$\mathbf{x}$变量的不同取值，并将其映射到一个条件概率$p(Y|\mathbf{x})$的值，即

$$p(Y=1|\mathbf{x};\alpha)=f(\mathbf{x},\alpha)$$

这就是回归，$f(\mathbf{x},\alpha)$被称为回归函数，其满足以下性质：

- $f(\mathbf{x},\alpha)\in[0,1]$
- 对于输入$x_1,\ldots,x_n$能找到简单且足够合理的依赖关系。
- 函数的形式由一个含有$n+1$个元素的向量α唯一确定。

逻辑回归函数

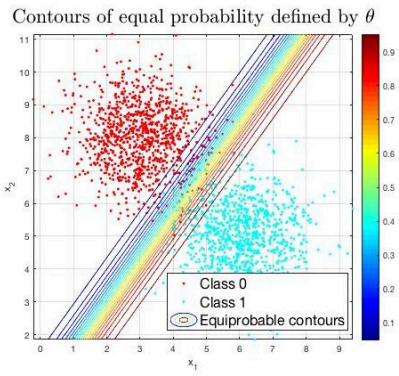
令$z(\alpha,\mathbf{x})=\alpha_0+\sum_{i=1}^{n}\alpha_i x_i$，即线性依赖关系，则

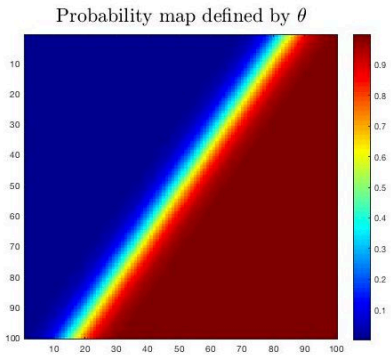
$$p(Y=1|\mathbf{x};\alpha)=\sigma(z(\alpha,\mathbf{x}))$$

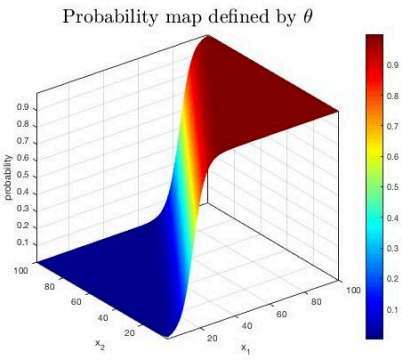
$$\sigma(z)=\frac{1}{1+e^{-z}}$$

被称为**逻辑回归函数**，它具有以下几点性质：

- 决策边界$p(Y=1|\mathbf{x};\alpha)>0.5$在输入空间$\mathbf{x}$中是线性的。
- 相同的概率值都是直线。
- 概率变化率有非常具体的形式。







连续变量

一维情况

如果X是一个连续随机变量，我们可以用概率密度函数(PDF) $p_X : \mathcal{R} \to \mathcal{R}^+$来表示，考虑两个最基本的参数化概率密度函数：

- **高斯分布**：

$$X \sim \mathcal{N}(\mu, \sigma)$$

$$p_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

- **均匀分布**：

$$X \sim \mathcal{U}(a, b)$$

$$p_X(x) = \frac{1}{b-a} \quad [a \le x \le b]$$

高维情况

如果$\mathbf{X}$是一个连续的随机变量向量，我们可以用联合概率密度函数表示：

- **高斯分布**：

$$\mathbf{X} \sim \mathcal{N}(\mu, \mathbf{\Sigma})$$

$$p_{\mathbf{X}}(x) = \frac{1}{\sqrt{(2\pi)^n |\mathbf{\Sigma}|}} \exp(-\frac{1}{2}(x-\mu)^T \mathbf{\Sigma}^{-1}(x-\mu))$$

在高维情况下，贝叶斯法则和链式法则仍然有效。

$$p_{X,Y,Z}(x,y,z) = p_X(x)\, p_{Y|X}(y|x)\, p_{Z|\{X,Y\}}(z|x,y)$$

举例：

- 高斯分布与其他分布混合
  1. 网络$Z \to X$由方程$p_{Z,X}(z,x) = p_Z(z)p_{X|Z}(x|z)$表示，其中
     - $Z \sim Bernoulli(p)$
     - $X|(Z=0) \sim \mathcal{N}(\mu_0, \sigma_0)$，$X|(Z=1) \sim \mathcal{N}(\mu_1, \sigma_1)$

     因此该网络的参数为$p, \mu_0, \mu_1, \sigma_0, \sigma_1$

2. 网络$Z \to X$由方程$p_{Z,X}(z,x) = p_Z(z)p_{X|Z}(x|z)$表示，其中
    - $Z \sim \mathcal{U}(a,b)$
    - $X|(Z=z) \sim \mathcal{N}(z,\sigma)$

    因此该网络的参数为a,b,σ

- **变分自编码器(VAE)**：网络$Z \to X$由方程$p_{Z,X}(z,x) = p_Z(z)p_{X|Z}(x|z)$表示，其中
    - $Z \sim \mathcal{N}(0,1)$
    - $X|(Z=z) \sim \mathcal{N}(\mu_\theta(z), e^{\sigma_\phi(z)})$，其中$\mu_\theta$和$\sigma_\phi$都是通过神经网络学习的对象，$\theta$和$\phi$是相应的权重。
    - **注意**：即使μ_θ和σ_ϕ都十分复杂，整体的函数仍然遵循高斯分布。