1.绪论

寻找数据中的模式，应用十分广泛，比如行星运行规律为经典力学带来的的跳板作用，原子光谱规律对于量子力学发展的作用。

模式识别关注的重点：发现数据中隐含的规律并利用这些规律进行数据分类等任务。

1.6 信息论

目前已知微分熵的形式：$\lim_{\Delta->0}\{-\sum_i p(x_i)\Delta\ln p(x_i)\} = -\int p(x\ln p(x))dx = \mathbf{H}[x]$,当我们想知道使微分熵取最大值时，p(x)的分布形式时，首先进行以下限制：

$$\int_{-\infty}^{\infty} p(x)dx = 1$$

$$\int_{-\infty}^{\infty} xp(x)dx = \mu$$

$$\int_{-\infty}^{\infty} (x-\mu)^2 p(x)dx = \sigma^2$$

利用拉格朗日乘数法，我们可以得到：

$$-\int_{-\infty}^{\infty} p(x)\ln p(x)dx + \lambda_1(\int_{-\infty}^{\infty} xp(x)dx - \mu) + \lambda_2(\int_{-\infty}^{\infty}(x-\mu)^2 p(x)dx - \sigma^2) + \lambda_3(\int_{-\infty}^{\infty} p(x)dx - 1)$$

然后求导可以得到：

$$p(x) = e^{-1+\lambda_1+\lambda_2 x+\lambda_3(x-\mu)^2}$$

最终可以得到结果为：

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

所以可以得到结论：在给定一阶矩和二阶矩限制的情况下，使得微分熵最大的分布是高斯分布。我们没有限制概率非负这个条件，但最后求出的形式并没有违反这个规则，所以没有必要加这个限制。
高斯分布的微分熵：$H[x] = \frac{1}{2}(1+\ln(2\pi\sigma^2))$,所以方差越大，也就是分布越平均，微分熵就越大。而且也可以看到，与离散形式不同，微分熵可能是负值。
此外，对于联合概率分布$p(x,y)$,已知x的情况下，额外需要的确定y所需的附加信息量为$-\ln p(y|x)$,所以

$$H[y|x] = -\iint p(y,x)\ln p(y|x)dydx$$

就是给定x的情况下，y的条件熵。
所以描述x和y的信息总量等于描述x的信息量加上给定x的情况下描述y需要的额外信息量。

1.6.1 相对熵和互信息

对于某个未知的分布p(x),我们想用一个近似分布q(x)来进行建模，由于我们使用的不是真正的p(x),所以在传递信息时，需要添加额外的信息量：

$$\mathbf{KL}(p||q) = -\int p(x)\ln q(x)dx - (-\int p(x)\ln p(x)dx) = -\int p(x)\ln\frac{q(x)}{p(x)}dx$$

这就是KL散度。
性质：

$$\mathbf{KL}(p||q) \geq 0$$

证明：

$$\begin{aligned}\mathbf{KL}(p||q) &= -\int p(x)\ln\frac{q(x)}{p(x)}dx \\ &\geq -\int p(x)\frac{q(x)}{p(x)}dx = \int q(x)dx = 0\end{aligned}$$

也可以使用Jenson不等式进行证明。
由于$-\ln x$是凸函数，所以由于对于任意的凸函数$f(x)$,$E(f(x)) \geq f(E(x))$,所以$-\int p(x)\ln\frac{q(x)}{p(x)}dx \geq -\ln\int p(x)\frac{q(x)}{p(x)}dx = 0$
对于想要拟合的真实分布$p(x)$,我们使用一个带有参数θ的分布$q(x|\theta)$,想让二者尽可能相似，因此最小化二者间的KL散度。
最小化KL散度等价于最大化似然函数
对于联合分布$p(x,y)$,想知道x，y之间是否接近相互独立，如果独立，则有$p(x,y)=p(x)p(y)$.
因此我们可以通过衡量$\mathbf{KL}(p(x,y)||p(x)p(y)) = -\int p(x,y)\ln\frac{p(x)p(y)}{p(x,y)}dx$即可.
这一项被称为x,y的互信息$\mathbf{I}[\mathbf{x},\mathbf{y}] = \mathbf{H}[\mathbf{x}] - \mathbf{H}[\mathbf{x}|\mathbf{y}] = \mathbf{H}[\mathbf{y}] - \mathbf{H}[\mathbf{y}|\mathbf{x}]$,表示一个新观测y对于x不确定性的减小。

$$\begin{aligned}\mathbf{I}[\mathbf{x},\mathbf{y}] &= -\iint p(x,y)\ln p(y)dxdy + \iint p(x,y)\ln p(y|x)dxdy \\ &= -\int p(y)\ln p(y)dy + \iint p(x,y)\ln p(y|x)dxdy \\ &= \mathbf{H}(\mathbf{y}) - \mathbf{H}[\mathbf{y}|\mathbf{x}]\end{aligned}$$