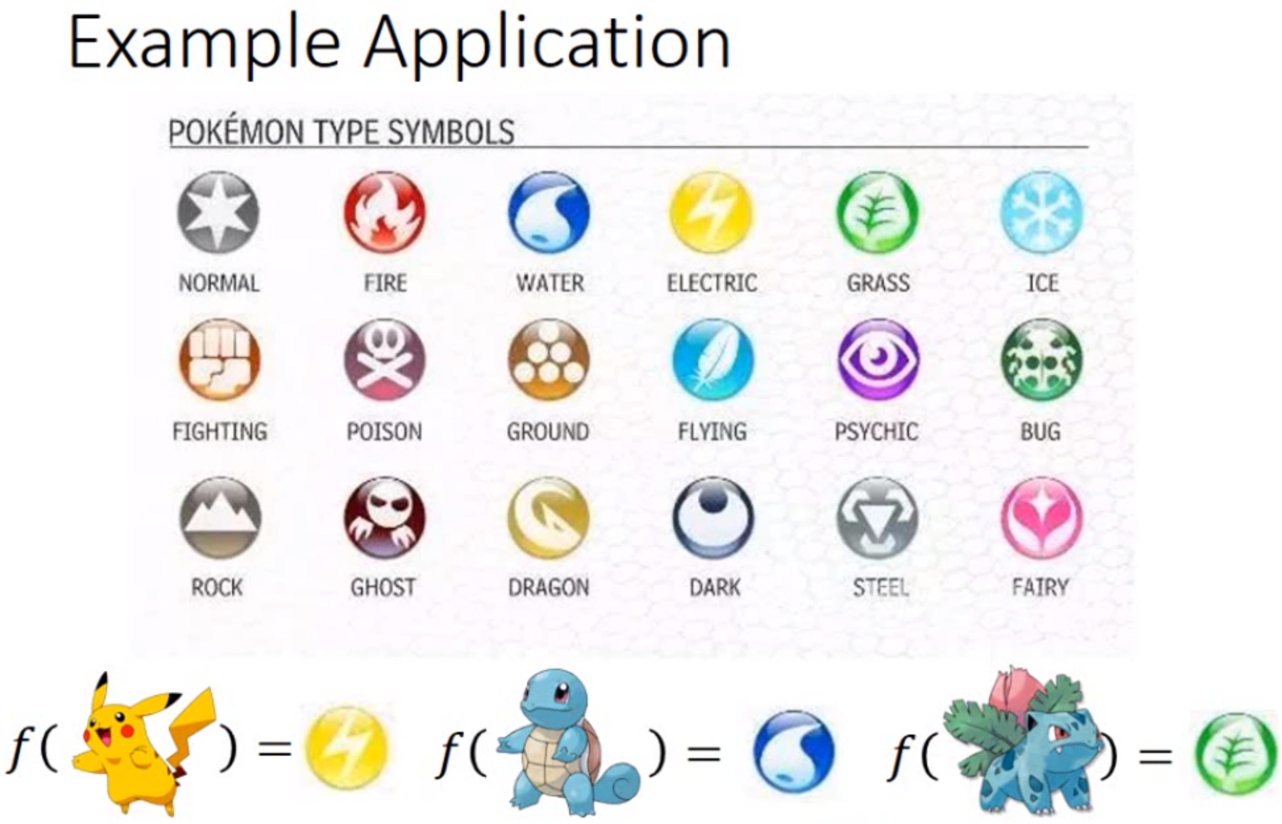
举例

已知有的宝可梦中有18种属性，而我们通过一个函数，将宝可梦精灵放到了这个函数中，而这个函数将会自动显示出该宝可梦所属的类型。



而每个宝可梦精灵都会有以下的属性：

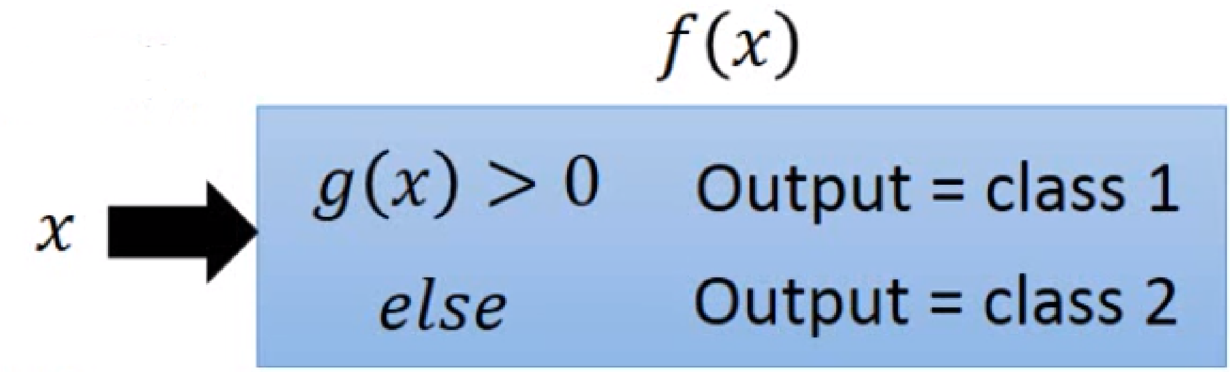
- 总强度（total）
- 生命值（HP）
- 攻击力（Attack）

- 防御力（Defense）
- 特殊攻击力（SP Atk）
- 特殊防御力（SP Def）
- 速度（Speed）

设定某一个预测模型，将宝可梦输入到模型函数中，能够准确预测出这个宝可梦的属性。

理想方法

- 建立函数模型



- 建立损失函数

$$L(f) = \sum_n \boxed{\delta(f(x^n) \neq \hat{y}^n)}$$

这里表示如果预测结果与样本相同则为1

- 通过损失函数找到最佳的函数模型。

高斯分布

正态分布（Normal distribution），也称“常态分布”，又名高斯分布(Gaussian distribution)，

> 若随机变量 X 服从一个位置参数为 μ 、尺度参数为 σ 的概率分布，且其概率密度函数为 [2]
>
> $$f(x) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$
>
> 则这个随机变量就称为**正态随机变量**，正态随机变量服从的分布就称为**正态分布，记作** $X \sim N\left(\mu,\sigma^2\right)$，读作 X 服从 $N\left(\mu,\sigma^2\right)$，或 X 服从正态分布。

多元高斯分布

首先在了解多元高斯分布之前，先了解协方差。

协方差

一个男孩子的猥琐程度跟他受女孩子的欢迎程度是否存在一些联系。协方差就是这样一种用来度量两个随机变量关系的统计量，来度量各个维度偏离其均值的程度，协方差可以这样来定义：

$$\mathrm{cov}(X,Y)=\frac{\sum_{i=1}^{n}(X_i-\overline{X})(Y_i-\overline{Y})}{n-1}$$

协方差的结果有什么意义呢？如果结果为正值，则说明两者是正相关的（从协方差可以引出“相关系数”的定义），也就是说一个人越猥琐越受女孩欢迎。如果结果为负值， 就说明两者是负相关，越猥琐女孩子越讨厌。如果为0，则两者之间没有关系，猥琐不猥琐和女孩子喜不喜欢之间没有关联，就是统计上说的“相互独立”。

给出协方差矩阵的定义：

$$C_{n\times n}=(c_{i,j},\qquad c_{i,j}=\mathrm{cov}(\mathrm{Dim}_i,\mathrm{Dim}_j))$$

这个定义还是很容易理解的，我们可以举一个三维的例子，假设数据集有三个维度，则协方差矩阵为：

$$C=\begin{pmatrix}\mathrm{cov}(x,x) & \mathrm{cov}(x,y) & \mathrm{cov}(x,z)\\ \mathrm{cov}(y,x) & \mathrm{cov}(y,y) & \mathrm{cov}(y,z)\\ \mathrm{cov}(z,x) & \mathrm{cov}(z,y) & \mathrm{cov}(z,z)\end{pmatrix}$$

可见，协方差矩阵是一个对称的矩阵，而且对角线是各个维度的方差。

多元高斯分布

让我们首先来介绍多元高斯分布的数学形式：

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

多元高斯分布和一元高斯分布是十分相似的，我们用加粗的$\boldsymbol{x}$来表示变量（一个向量），D表示维度（元的数目），加粗的$\boldsymbol{\mu}$表示平均向量，

大写的Σ表示协方差矩阵（Covariance Matrix，是一个方阵），$|\Sigma|$表示Σ的行列式值，$(\boldsymbol{x}-\boldsymbol{\mu})^T$表示矩阵$(\boldsymbol{x}-\boldsymbol{\mu})$的转置。

求解多元高斯分布：最大似然估计

和求解一元高斯分布类似，我们将问题描述为：给定观测值$\{\boldsymbol{x}_i\}$，求$\boldsymbol{\mu}$和Σ，使得似然函数最大：

$$\hat{\boldsymbol{\mu}}, \hat{\Sigma} = \arg\ \max_{\boldsymbol{\mu}, \Sigma}\ p(\{\boldsymbol{x}_i\} | \boldsymbol{\mu}, \Sigma)$$

同样，假设观测值两两相互独立，根据独立概率公式，我们有：

$$\hat{\boldsymbol{\mu}}, \hat{\Sigma} = \arg\ \max_{\boldsymbol{\mu}, \Sigma}\ \prod_{i=1}^{N} p(\boldsymbol{x}_i | \boldsymbol{\mu}, \Sigma)$$

同样（1）取对数，（2）将多元高斯分布的形式带入，我们有：

$$\begin{aligned}
\hat{\boldsymbol{\mu}}, \hat{\Sigma} &= \arg\ \max_{\boldsymbol{\mu}, \Sigma}\ \ln \prod_{i=1}^{N} p(\boldsymbol{x}_i | \boldsymbol{\mu}, \Sigma) \\
&= \arg\ \max_{\boldsymbol{\mu}, \Sigma}\ \sum_{i=1}^{N} \ln p(\boldsymbol{x}_i | \boldsymbol{\mu}, \Sigma) \\
&= \arg\ \max_{\boldsymbol{\mu}, \Sigma}\ \sum_{i=1}^{N} (-\frac{1}{2}(\boldsymbol{x}_i - \boldsymbol{\mu})^T \Sigma^{-1} (\boldsymbol{x}_i - \boldsymbol{\mu}) - \frac{1}{2} \ln |\Sigma| - \frac{D}{2} \ln(2\pi)) \\
&= \arg\ \min_{\boldsymbol{\mu}, \Sigma}\ \sum_{i=1}^{N} (\frac{1}{2}(\boldsymbol{x}_i - \boldsymbol{\mu})^T \Sigma^{-1} (\boldsymbol{x}_i - \boldsymbol{\mu}) + \frac{1}{2} \ln |\Sigma|)
\end{aligned}$$

我们给目标函数做个记号，令

$$J(\boldsymbol{\mu},\Sigma)=\sum_{i=1}^{N}(\frac{1}{2}(\boldsymbol{x}_i-\boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{x}_i-\boldsymbol{\mu})+\frac{1}{2}\ln|\Sigma|)$$

我们仍然分别对$\boldsymbol{\mu}$和Σ求偏导来计算$\hat{\boldsymbol{\mu}}$和$\hat{\Sigma}$。（这里需要矩阵求导的知识，可以参考Matrix Calculus Manual）

$$\begin{aligned}
\frac{\partial J}{\partial \boldsymbol{\mu}}&=\frac{\partial}{\partial \boldsymbol{\mu}}\sum_{i=1}^{N}(\frac{1}{2}(\boldsymbol{x}_i-\boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{x}_i-\boldsymbol{\mu})+\frac{1}{2}|\Sigma|)\\
&=\frac{\partial}{\partial \boldsymbol{\mu}}\frac{1}{2}\sum_{i=1}^{N}(\boldsymbol{x}_i^T\Sigma^{-1}\boldsymbol{x}_i-\boldsymbol{x}_i^T\Sigma^{-1}\boldsymbol{\mu}-\boldsymbol{\mu}^T\Sigma^{-1}\boldsymbol{x}_i+\boldsymbol{\mu}^T\Sigma^{-1}\boldsymbol{\mu})\\
&=\frac{\partial}{\partial \mu}\sum_{i=1}^{N}(\frac{1}{2}\boldsymbol{\mu}^T\Sigma^{-1}\boldsymbol{\mu}-\boldsymbol{\mu}\Sigma^{-1}\boldsymbol{x}_i)\\
&=\Sigma^{-1}\sum_{i=1}^{N}(\boldsymbol{\mu}-\boldsymbol{x}_i)\\
&=\mathbf{0}\\
\frac{\partial J}{\partial \Sigma}&=\frac{\partial}{\partial \Sigma}\sum_{i=1}^{N}(\frac{1}{2}(\boldsymbol{x}_i-\hat{\boldsymbol{\mu}})^T\Sigma^{-1}(\boldsymbol{x}_i-\hat{\boldsymbol{\mu}})+\frac{1}{2}\ln|\Sigma|)\\
&=\frac{1}{2}\sum_{i=1}^{N}(-\Sigma^{-1}(\boldsymbol{x}_i-\hat{\boldsymbol{\mu}})(\boldsymbol{x}_i-\hat{\boldsymbol{\mu}})^T\Sigma^{-1}+\Sigma^{-1})\\
&=\frac{1}{2}\Sigma^{-1}(-(\sum_{i=1}^{N}(\boldsymbol{x}_i-\hat{\boldsymbol{\mu}})(\boldsymbol{x}_i-\hat{\boldsymbol{\mu}})^T)\Sigma^{-1}+N\boldsymbol{I})\\
&=\mathbf{0}
\end{aligned}$$

求得，

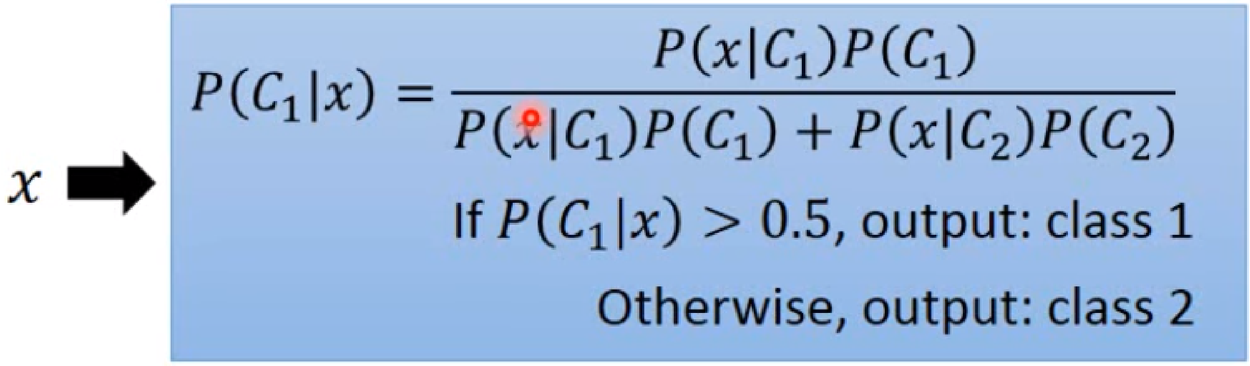
$$\hat{\boldsymbol{\mu}}=\frac{1}{N}\sum_{i=1}^{N}\boldsymbol{x}_i$$

$$\hat{\Sigma}=\frac{1}{N}\sum_{i=1}^{N}(\boldsymbol{x}_i-\hat{\boldsymbol{\mu}})(\boldsymbol{x}_i-\hat{\boldsymbol{\mu}})^T$$

最后求得 μ和 $\sum$ 作为正态分布的位置参数和尺度参数

模型确立的三个步骤

- 设立模型



- 选择最优函数

最后求得μ和$\sum$作为正态分布的位置参数和尺度参数,即求最大似然估计。

- 确立最优函数

通过求得的值来分析比较获得最优的函数模型。

伯努利分布与高斯分布的区别

伯努利分布，只有0，1两种结果，所以是在二分类问题中会用到，也就是之前提到的分类问题；

而高斯分布，是我们在求解最小化损失函数的时候，当时用最小二乘法表示，是因为假设误差函数满足高斯分布的时候的最大似然函数中部分的 loglog 形式

如果假设所有的结果，都是独立的。如：

$P(x_2|C_2)$、$P(x_2|C_2)$、$P(x_2|C_2)$、$P(x_2|C_2)$……

都是独立的，那么我们称并可以使用简单贝叶斯分类器。

后验概率

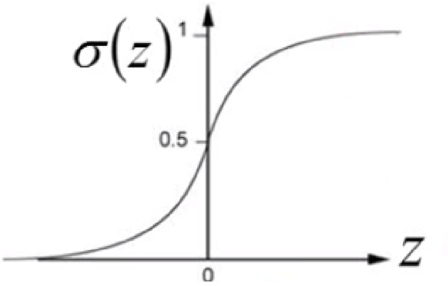
sigmoid函数的由来

$$P(C_1|x) = \frac{P(x|C_1)P(C_1)}{P(x|C_1)P(C_1) + P(x|C_2)P(C_2)}$$

$$= \frac{1}{1 + \frac{P(x|C_2)P(C_2)}{P(x|C_1)P(C_1)}} = \frac{1}{1 + exp(-z)} = \sigma(z)$$

Sigmoid function

$$z = ln\frac{P(x|C_1)P(C_1)}{P(x|C_2)P(C_2)}$$



先验概率

事件发生前的预判概率。可以是基于历史数据的统计，可以由背景常识得出，也可以是人的主观观点给出。一般都是单独事件概率，如P(x),P(y)。

后验概率

事件发生后求的反向条件概率；或者说，基于先验概率求得的反向条件概率。概率形式与条件概率相同。

条件概率

一个事件发生后另一个事件发生的概率。一般的形式为P(x|y)表示y发生的条件下x发生的概率。

我们设A为加了醋的概率,B为吃了之后是酸的概率.C为肉变质的概率

$$P(A|B) = \frac{P(A,B)}{P(B)} = \frac{P(B|A) * P(A)}{P(B|A) * P(A) + P(B|C) * P(C)}$$

P(A|B)就是后验概率,其中 P(B)的展开是运用了全概率公式

关于贝叶斯公式的解释:

"如果我们把事件A看做'结果',把诸事件B1,B2...看做导致这个结果的可能的'原因',则可以形象地把全概率公式看做成为'由原因推结果';而贝叶斯公式则恰好相反,其作用于'由结果推原因':现在有一个'结果'A以发生,在众多可能的'原因'中,到底是哪一个导致了这结果"

$$\mathrm{P}\left(\text{原因 1}\middle|\text{结果}\right)=\frac{\mathrm{P}(\text{原因 1 导致结果})}{\mathrm{P}(\text{结果})}$$

$$=\frac{\mathrm{P}\left(\text{结果}\middle|\text{原因 1}\right)*\mathrm{P}(\text{原因 1})}{\mathrm{P}\left(\text{结果}\middle|\text{原因 1}\right)*\mathrm{P}\left(\text{原因 1}\right)+\mathrm{P}\left(\text{结果}\middle|\text{原因 2}\right)*\mathrm{P}\left(\text{原因 2}\right)+\mathrm{P}\left(\text{结果}\middle|\text{原因 3}\right)*\mathrm{P}\left(\text{原因 3}\right)+\cdots}$$