跟我学人工智能

刘森

2018 年 2 月 15 日

你好，世界hello, world

1 机器学习算法

不同的算法，本身没有好坏之分，有的只是，根据不同的场景选择合适的算法。

线性回归和Logistic回归，虽然听起来都叫作“回归”，但其实两者却是做不一样的事情：一个是做连续数据的预测，一个是做离散数据的预测；一个是真正做回归的，一个是做分类的，它们两个【用途】是完全不一样的。【如何推导出来？】线性回归是用高斯分布的方式推导出来，Logistic回归既然是做分类，就用Bnody分布，两点分布来推导出来。两者大的工具都是【最大似然估计】。在线性回归里面，要讨论一个东西：【最小二乘法的本质是什么】。或者说，为什么有最小二乘法呢？有没有最小三乘法呢？有没有最小四乘法呢？在【线性回归】和【Logistic回归】中强调两个工具：【梯度下降算法】和【极大似然估计】。

1.1 线性回归

高斯分布
极大似然估计MLE
最小二乘法的本质

1.1.1 什么是线性回归

线性回归$y = ax + b$

考虑多个变量情形，例如两个变量，$h_\theta(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2$ ，可以写成如下形式：

$$h_\theta(x) = \sum_{i=0}^{n} \theta_i x_i = \theta^T x$$

其中，θ展开后，呈现如下形式：

$$\begin{pmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \end{pmatrix}$$

其中，x展开后，呈现如下形式：

$$\begin{pmatrix} 1 \\ x_1 \\ x_2 \end{pmatrix}$$

上式中的1就表示x_0，而相应的θ_0表示截距，是比较难以直接解释的。

再把上面的式子拿过来，

$$h_\theta(x) = \sum_{i=0}^{n} \theta_i x_i = \theta^T x$$

在$h_\theta(x)$中，x看起来是【自变量】，但事实上是【样本】，所以x是已知的，而θ是未知的，我们要通过某一种办法来求解出θ，这个就是【线性回归要解决的问题】。

第05课《回归》00:10:30

目前讲的问题是【what】，即什么是线性回归。过一会儿，会讲【how】，用什么样的工具去求，如何去求的问题。

1.1.2 使用极大似然估计解释最小二乘

第05课《回归》00:20:00

使用极大似然估计解释最小二乘

$$y^{(i)} = \theta^T x Y(i) + \epsilon^{(i)}$$

the $\epsilon^{(i)}$ are distributed IID (independently and identically distributed) according to a Gaussian distribution (also called a Normal distribution) with mean *zero* and some variance σ^2.

误差$\epsilon^{(i)}(1 \leq i \leq m)$是独立同分布的，服从均值为0，方差为某定值$\sigma^2$的【高斯分布】。原因：【中心极限定理】，可以查阅一下“中心极限定理的意义”。

似然函数第05课《回归》00:21:21

首先，两边是相等的：

$$y^{(i)} = \theta^T x^{(i)} + \epsilon^{(i)}$$

其中，$x^{(i)}$表示第i个【样本】，$\theta^T x^{(i)}$表示第i个样本的【预测值】，$y^{(i)}$表示第i个样本的【真实值】，而$\epsilon^{(i)}$表示第i个样本的误差。

根据【中心极限定理】，$\epsilon^{(i)}$应该是呈现一个高斯分布的形态。

$$p(\epsilon^{(i)}) = \frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{(\epsilon^{(i)})^2}{2\sigma^2})$$

另外，$\epsilon^{(i)} = y^{(i)} - \theta^T x^{(i)}$，此时将$\epsilon^{(i)}$代入上式：

$$p(y^i|x^i;\theta) = \frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{(y^{(i)} - \theta^T x^{(i)})^2}{2\sigma^2})$$

如此一来，上式当中就没有误差ϵ了，因此只要指定了x和θ，就可以认为是一个y的分布。换句话讲，y其实服从的是【均值是$\theta^T x$，方差是某一个σ的高斯分布（正态分布）】。

那么，用什么可以估计这个θ呢？答：【最大似然估计】。

在上面的公式中，i只是表示第i个样本，假设一共有m个样本，那么，【m个样本的似然估计】就可以表示为：

$$L(\theta) = \prod_{i=1}^{m} p(y^i|x^i;\theta) = \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{(y^{(i)} - \theta^T x^{(i)})^2}{2\sigma^2})$$

如此一来，怎么求θ呢？直接对【似然函数】取对数，然后再想办法。

高斯的对数似然与最小二乘

$$
\begin{aligned}
l(\theta) &= \log L(\theta) \\
&= \log \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{(y^{(i)} - \theta^T x^{(i)})^2}{2\sigma^2}) \\
&= \sum_{i=1}^{m} \log \frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{(y^{(i)} - \theta^T x^{(i)})^2}{2\sigma^2}) \\
&= m \log \frac{1}{\sqrt{2\pi}\sigma} - \frac{1}{\sigma^2} \cdot \frac{1}{2} \sum_{i=1}^{m} (y^{(i)} - \theta^T x^{(i)})^2
\end{aligned} \tag{1}
$$

现在，其实是通过【最大似然估计】加上【高斯分布】来得到了【最小二乘法】目标函数。换句话说，这就是解释的“为什么会有最小二乘法”这个概念。

$$
J(\theta) = \frac{1}{2} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)})^2 \tag{2}
$$

1.1.3 θ的解析式的求解过程

θ的解析式的求解过程第05课《回归》00:29:52

将M个N维样本组成矩阵X：（1）X的每一行对应一个样本，共M个样本（measurements）；（2）X的每一列对应样本的一个维度，共N维（regressors）（3）还有额外的一维常数项$x_0^{(i)}$，全为1。

目标函数：

$$
J(\theta) = \frac{1}{2} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)})^2 = \frac{1}{2} (X\theta - y)^T (X\theta - y) \tag{3}
$$

梯度：

$$
\begin{aligned}
\nabla_\theta J(\theta) &= \nabla_\theta (\frac{1}{2} (X\theta - y)^T (X\theta - y)) \\
&= \nabla_\theta (\frac{1}{2} (\theta^T X^T - y^T)(X\theta - y)) \\
&= \nabla_\theta (\frac{1}{2} (\theta^T X^T X\theta - \theta^T X^T y - y^T X\theta + y^T y)) \\
&= \frac{1}{2} (2X^T X\theta - X^T y - (y^T X)^T) \\
&= X^T X\theta - X^T y
\end{aligned} \tag{4}
$$

在上式中，求驻点，令$X^TX\theta - X^Ty = 0$。
所以θ取值如下：

$$\theta = (X^TX)^{-1} \cdot X^Ty \tag{5}$$

1.1.4 最小二乘法意义下的参数最优解

参数的解析解

$$\theta = (X^TX)^{-1}X^Ty \tag{6}$$

若X^TX不可逆或防止过拟合，增加λ扰动

$$\theta = (X^TX + \lambda I)^{-1}X^Ty \tag{7}$$

第05课《回归》00:38:36
“简单”方法记忆结论

$$\begin{aligned} & X\theta = y \\ & \Rightarrow X^TX\theta = X^Ty \\ & \Rightarrow \theta = (X^TX)^{-1}X^Ty \end{aligned} \tag{8}$$

1.1.5 梯度下降算法Gradient Descent

第05课《回归》01:14:49
事实上，通过解析解的方式$\theta = (X^TX)^{-1}X^Ty$来求解θ是没有问题的，真的是对的；但是，我们往往在机器学习中，习惯在这个地方引出“梯度下降算法”，并且也习惯使用“梯度下降算法”来求解θ。我们这里还是求目标函数$J(\theta) = \frac{1}{2}\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})^2$的梯度，延着负梯度方向不停的下降，降到某一个值，降不下去了，我们就可以知道，得到了一个局部的极小值，这个局部的极小值点，或许就是我们想要的θ。
初始化θ（随机初始化）
沿着负梯度方向迭代，更新后的θ使$J(\theta)$更小

$$\theta_j = \theta_j - \alpha \cdot \frac{\partial J(\theta)}{\partial \theta} \tag{9}$$

其中，α表示学习率、步长。这个步长α事实上是有办法可以指定比较优的，后续再讲怎么选α会更优。

1.1.6 梯度方向

梯度方向第05课《回归》01:15:37

$$\begin{aligned}\frac{\partial}{\partial\theta_j}J(\theta) &= \frac{\partial}{\partial\theta_j}\frac{1}{2}(h_\theta(x)-y)^2 \\ &= 2\cdot\frac{1}{2}(h_\theta(x)-y)\cdot\frac{\partial}{\partial\theta_j}(h_\theta(x)-y) \\ &= (h_\theta(x)-y)\cdot\frac{\partial}{\partial\theta_j}(\sum_{i=0}^{n}\theta_i x_i - y) \\ &= (h_\theta(x)-y)x_j\end{aligned} \tag{10}$$

这里是求偏导，一共有n个特征，当前求解的是第j个特征的偏导。得到这个梯度之后，我们就可以延着负梯度的方向下降下去就可以了。

这里有个需要注意的地方，我们求出来的是梯度，而我们要用的是**负梯度**，而**负梯度**应该是$-(h_\theta(x)-y)x_j$，因此可以调换一下$h_\theta(x)$和y的位置，写成这种形式$(y-h_\theta(x))x_j$。

1.1.7 批量梯度下降算法

Repeat until convergence(会聚; 集收敛)

$$\theta_j := \theta_j + \alpha\sum_{i=1}^{m}(y^{(i)}-h_\theta(x^{(i)}))x_j^{(i)} \tag{11}$$

gradient descent. Note that, while gradient descent can be susceptible to local minima in general, the optimization problem we have posed here for linear regression has only one global, and no other local, optima; thus gradient always converges (assuming the learning rate α is not too large) to the global minimum. Indeed, J is a convex(凸形; 凸的) quadratic(二次的;) function.

1.1.8 随机梯度下降算法

for $i=1$ to m

$$\theta_j := \theta_j + \alpha(y^{(i)} - h_\theta(x^{(i)}))x_j^{(i)} \tag{12}$$

This algorithm is called **stochastic gradient descent** (also **incremental gradient descent**). Whereas batch gradient descent has to scan through the entire training set before taking a single step – a costly operation if m is large – stochastic gradient descent can start making progress right away, and continues to make progress with each example it looks at. Often, stochastic gradient descent gets θ "close" to the minimum much faster than batch gradient descent. (Note however that it may never "converge" to the minimum, and the parameters θ will keep oscillating(振荡;（使）摆动) around the minimum of $J(\theta)$; but in practice most of the values near the minimum will be reasonably good approximations to the true minimum.) For these reasons, particularly when the training set is large, stochastic gradient descent is often preferred over batch gradient descent.

在没有明确**批量梯度下降**和**随机梯度下降**哪个更优的情况下，优先选择“随机梯度下降”。

在求解θ的时候，不一定非要是最好的，它只能说是**堪用**的，能够work的，可用的就行了。有的时候，不要追求完美，完美往往是达不到的，只要能够过得去，还可以，就行了。

1.1.9 折中：mini-batch

第05课《回归》01:22:37

如果不是每拿到一个样本即更改梯度，而是若干个样本的平均梯度作为更新方向，则是mini-batch梯度下降算法。

机器学习(Machine learning)，第一，要会理论，它是一个能够走多远的基础；第二，是要会写代码，它是保证当前的工作能够持续下去。“渔”和“鱼”都得要。

机器学习中，很多时间都花在了**如何建模型、如何调参、如何选特征、如何优化模型**这些事情上，写代码并没有那么的困难。

1.1.10 权值的设置

高斯核函数第05课《回归》01:45:26

我们希望用“线性回归”求得模型的“残差”服从高斯分布（正态分布）；如果不服从高斯分布，我们就去重新选“特征”。如果不服从高斯分布，说明有些“特征”没有被考虑进来。

1.2 逻辑回归：分类问题的首选算法

第05课《回归》02:00:30

1.2.1 Logistic/sigmoid函数

$$g(z) = \frac{1}{1+e^{-z}} \tag{13}$$

$$h_\theta(x) = g(\theta^T x) = \frac{1}{1+e^{-\theta^T x}} \tag{14}$$

$$\begin{aligned} g'(x) &= (\frac{1}{1+e^{-x}})' = \frac{e^{-x}}{(1+e^{-x})^2} \\ &= \frac{1}{1+e^{-x}} \cdot \frac{e^{-x}}{1+e^{-x}} \\ &= \frac{1}{1+e^{-x}} \cdot (1 - \frac{1}{1+e^{-x}}) \\ &= g(x) \cdot (1 - g(x)) \end{aligned} \tag{15}$$

1.2.2 Logistic回归参数估计

假定

$$\begin{aligned} P(y=1|x;\theta) &= h_\theta(x) \\ P(y=0|x;\theta) &= 1 - h_\theta(x) \end{aligned} \tag{16}$$

上面的两个式子，可以用一个式子来表示：

$$p(y|x;\theta) = (h_\theta(x))^y (1-h_\theta(x))^{1-y} \tag{17}$$

那么似然函数$L(\theta)$这样求：

$$\begin{aligned} L(\theta) &= p(\vec{y}|X;\theta) \\ &= \prod_{i=1}^{m} p(y^{(i)}|x^{(i)};\theta) \\ &= \prod_{i=1}^{m} (h_\theta(x^{(i)}))^{y^{(i)}} (1-h_\theta(x^{(i)}))^{1-y^{(i)}} \end{aligned} \tag{18}$$

1.2.3 对数似然函数

第05课《回归》02:01:23

$$\begin{aligned} l(\theta) &= \log L(\theta) \\ &= \sum_{i=1}^{m} y^{(i)} \log h_\theta(x^{(i)}) + (1-y^{(i)}) \log(1-h_\theta(x^{(i)})) \end{aligned} \tag{19}$$

$$\begin{aligned} \frac{\partial}{\partial\theta_j} l(\theta) &= (y\frac{1}{g(\theta^T x)} - (1-y)\frac{1}{1-g(\theta^T x)})\frac{\partial}{\partial\theta_j} g(\theta^T x) \\ &= (y\frac{1}{g(\theta^T x)} - (1-y)\frac{1}{1-g(\theta^T x)}) g(\theta^T x)(1-g(\theta^T x))\frac{\partial}{\partial\theta_j}\theta^T x \\ &= (y(1-g(\theta^T x)) - (1-y)g(\theta^T x))x_j \\ &= (y - h_\theta(x))x_j \end{aligned} \tag{20}$$

1.2.4 参数的迭代

第05课《回归》02:01:55

Logistic回归参数的学习规则

$$\theta_j := \theta_j + \alpha(y^{(i)} - h_\theta(x^{(i)}))x_j^{(i)} \tag{21}$$

比较上面的结果和“线性回归”的结论的差别：它们具有相同的形式，Logistic回归是用于做分类的，属于广义线性回归。

1.2.5 对数线性模型

一个事件的**几率**odds，是指该**事件发生的概率**与**事件不发生的概率**的比值。

对数几率：$\log it$函数

$$\begin{aligned} P(y=1|x;\theta) &= h_\theta(x) \\ P(y=0|x;\theta) &= 1-h_\theta(x) \end{aligned} \tag{22}$$

$$\begin{aligned} \log it(p) &= \log \frac{p}{1-p} \\ &= \log \frac{h_\theta(x)}{1-h_\theta(x)} \\ &= \log \frac{\frac{1}{1+e^{-\theta^T x}}}{\frac{e^{-\theta^T x}}{1+e^{-\theta^T x}}} \\ &= \theta^T x \end{aligned} \tag{23}$$

说一个事儿：如果有一个x_1和x_2，正常去求x_1占多大比例的方法是

$$\frac{x_1}{x_1+x_2} \tag{24}$$

如果不想这么算，而是想做一个指数的比例，那么x_1和x_2就变成了e^{x_1}和e^{x_2}，再求比例就是

$$\frac{e^{x_1}}{e^{x_1}+e^{x_2}} = \frac{1}{1+e^{x_2-x_1}} \tag{25}$$

如果令$x_2-x_1=-z$，就得到：

$$\frac{1}{1+e^{-z}} \tag{26}$$

这样就得到了sigmoid函数。

1.3 工具

梯度下降算法
极大似然估计

1.4 Softmax

1.5 聚类

按道理来说，“算法”和“模型”是两个完全不同的概念，算法一般用

于表示解决特定的数学问题，而模型则侧重于解决实际的问题，但在不同场景下，两者也会经常混用。

“聚类”是无监督的机器学习算法，而“线性回归、Logistic回归、Softmax、SVM、随机森林”都是有监督的机器学习算法。

“聚类”并不像“有监督分类”那样有y值，而只有$m \times n$维的向量数据，其中m表示m个样本，n表示有n个特征。“聚类”并不依赖于y值，而是依赖于$m \times n$维的数据，根据其内部之间的相似性，来做聚类。

一个$m \times n$维的数据，经过某种聚类算法之后，这m个样本聚类到k个簇当中，这样就把一个$m \times n$维矩阵转换成了一个$m \times k$维矩阵，这本质上就是一个聚类的过程。我们发现，这个数据是从一个$m \times n$维矩阵转换成了一个$m \times k$维矩阵，这本质上又是一个降维的过程。所以“聚类”这个词和“降维”这个词，两者本质上是一样的。

1.5.1 本次目标

掌握**K-means聚类**的思路和使用条件

了解**层次聚类**的思路和方法

理解**密度聚类**并能够应用于实践：（1）DBSCAN；（2）密度最大值聚类

掌握**谱聚类**的算法：考虑谱聚类和**PCA**的关系。

1.5.2 聚类的定义

聚类就是对大量未知标注的数据集，按数据的内在相似性将数据集划分为多个类别，使类别内的数据相似度较大而类别间的数据相似度较小。

而这种**内在相似性**，通常用**相似度**或**距离**来度量。往往，**距离**求出来之后，将距离的-1次方作为**相似度**。

1.5.3 相似度/距离计算方法总结

第10课聚类00:11:18

闵可夫斯基距离Minkowski

$$dist(X,Y) = \left(\sum_{i=1}^{n} |x_i - y_i|^p \right)^{\frac{1}{p}} \tag{27}$$

在上式中，当$p = 2$时，就是标准的“欧氏距离”；当$p = 1$时，就是“曼哈顿距离”；当$p = \infty$时，就相当于取$|x_i - y_i|$的最大距离，就称为“切比雪夫距离”。

杰卡德相似系数（Jaccard），从集合的角度来理解

$$J(A,B) = \frac{|A \bigcap B|}{|A \bigcup B|} \tag{28}$$

余弦相似度（cosine similarity），从角度来理解

$$\cos(\theta) = \frac{a^T b}{|a| \cdot |b|} \tag{29}$$

Pearson相关系数（Pearson CorrelationCoefficient）是用来衡量两个数据集合是否在一条线上面，它用来衡量定距变量间的线性关系。

$$\begin{aligned} \rho_{XY} &= \frac{cov(X,Y)}{\sigma_X \sigma_Y} \\ &= \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y} \\ &= \frac{\sum_{i=1}^{n}(X_i-\mu_X)(Y_i-\mu_Y)}{\sqrt{\sum_{i=1}^{n}(X_i-\mu_X)^2}\sqrt{(\sum_{i=1}^{n}Y_i-\mu_Y)^2}} \end{aligned} \tag{30}$$

假设$A = (A_1, A_2, ...A_n)$，有n个数，我们可以求它的平均值$\bar{A}$，那么$\frac{1}{n}(A_i - \bar{A})^2$就可以认为是随机变量A的均方差。同时，有$B = (B_1, B_2, ...B_n)$，也有$n$个数，它的均方差也可以表示为$\frac{1}{n}(B_i - \bar{B})^2$。也可以用$\frac{1}{n}(A_i - \bar{A})(B_i - \bar{B})$来度量$A$和$B$之间的协方差。这个$A$和$B$之间协方差，也可以除以它的标准差，就有了Pearson相关系数。

相对熵（K-L距离）

$$D(p \parallel q) = \sum_x p(x) \log \frac{p(x)}{q(x)} = E_{p(x)} \log \frac{p(x)}{q(x)} \tag{31}$$

Hellinger距离

$$D_\alpha(p \parallel q) = \frac{2}{1-a^2}\left(1 - \int p(x)^{\frac{1+\alpha}{2}} q(x)^{\frac{1-\alpha}{2}} dx\right) \tag{32}$$

注意：不要拘泥于某种相似性的计算方式，在实践当中，有些场景，适合用欧氏距离，如果是推荐系统，就习惯用杰卡德系数，如果是文本的相似度，就可能用余弦相似度，这都是有可能的。

我们要说明的就是，在“聚类”算法当中，可以任意挑选一个“相似度计算方法”，就能够算出样本i和样本j的相似性S_{ij}；如果有m个样本，就形成一个$m \times m$的相似度方阵，后面就对$m \times m$的方阵使用各种手段来去做聚类。至于说用哪一个“相似度”算法，则是一个相对独立的事情。

1.5.4 Hellinger distance

第10课聚类00:20:50

1.5.5 余弦相似度与Pearson相似系数

第10课聚类00:24:50

n维向量x和y的夹角记作θ，根据余弦定理，其余弦值为：

$$\cos(\theta) = \frac{x^T y}{|x| \cdot |y|} = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2}\sqrt{\sum_{i=1}^{n} y_i^2}} \tag{33}$$

这两个向量的Pearson相关系数是：

$$\begin{aligned} \rho_{XY} &= \frac{cov(X,Y)}{\sigma_X \sigma_Y} \\ &= \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y} \\ &= \frac{\sum_{i=1}^{n}(X_i-\mu_X)(Y_i-\mu_Y)}{\sqrt{\sum_{i=1}^{n}(X_i-\mu_X)^2}\sqrt{(\sum_{i=1}^{n} Y_i-\mu_Y)^2}} \end{aligned} \tag{34}$$

如果此时的μ_X和μ_Y都为0，则此时的Pearson相关系数恰好是“余弦相似度”。

相关系数即将x、y坐标向量各自平移到原点后的夹角余弦!

这即解释了为何文档间求距离使用夹角余弦——因为这一物理量表征了文档去均值化后随机向量间相关系数。

1.5.6 聚类的基本思想

第10课聚类00:27:32

给定一个有N个对象的数据集，构造数据的k个簇，$k \leq n$。满足下列条件：

（1）每一个簇至少包含一个对象

（2）每一个对象属于且仅属于一个簇

（3）将满足上述条件的k个簇称作一个合理划分。

基本思想：对于给定的类别数目k，首先给出初始划分，通过迭代改变**样本**和**簇**的隶属关系，使得每一次改进之后的划分方案都较前一次好。

1.5.7 K-means算法

K-means算法，也称为k-平均或k-均值，是一种广泛使用的聚类算法，或者成为其他聚类算法的基础。

假定输入样本为$S = x_1, x_2, ...x_m$，则算法步骤为：

（1）选择初始的k个类别中心$\mu_1, \mu_2, ...\mu_k$

（2）对于每个样本x_i，将其标记为距离类别中心最近的类别，即：

$$label_i = \arg\min_{1 \le j \le k} \parallel x_i - \mu_j \parallel \tag{35}$$

（3）将每个类别中心更新为隶属该类别的所有样本的均值

$$\mu_j = \frac{1}{|c_j|} \sum_{i \in c_j} x_j \tag{36}$$

（4）重复最后两步，直到类别中心的变化小于某阈值。

中止条件：迭代次数、簇中心变化率、最小平方误差MSE(Minimum Squared Error)

很好的一个问题就是：（1）K-means算法中的k如何选择？（2）k个中心如何初始化？

对于第一个问题，一般是使用两种方式。第一种，有的时候，可以通过“先验的知识”来确定的，比如说，抛骰子只能有6种可能的数值。第二种，就是交叉验证，不断的尝试k的大小，来看看最小平方误差是否会减小。当k没有更好的办法选择时，只能够通过相互交叉验证的方式帮助我们做。

第二个问题解答第10课聚类00:39:00

对于第二个问题，一种是随机的给定初始值。

1.5.8 K-means的公式化解释

第10课聚类00:55:32

记K个簇中心为$\mu_1, \mu_2, ...\mu_k$，每个簇的样本数目为$N_1, N_2, ..., N_k$

使用平方误差作为目标函数：

$$J(\mu_1,\mu_2,...\mu_k)=\frac{1}{2}\sum_{j=1}^{K}\sum_{i=1}^{N_j}(x_i-\mu_j)^2 \tag{37}$$

该函数为关于$\mu_1,\mu_2,...\mu_k$的凸函数，其驻点为：

$$\begin{aligned}\frac{\partial J}{\partial \mu_j}&=\sum_{x_i\in\{J\}}(x_i-\mu_j)=0\\ &\Rightarrow \mu_j=\frac{1}{N_j}\sum_{x_i\in\{J\}}x_i\end{aligned} \tag{38}$$

k-均值的聚类结果，一定是类“圆”的。

1.5.9 K-means聚类方法总结

第10课聚类01:01:32

优点：

（1）是解决聚类问题的一种经典算法，简单，快速

（2）对处理大数据集，该算法保持可伸缩性和高效率

（3）当簇近似为高斯分布时，它的效果较好

缺点：

（1）在簇的平均值可被定义的情况下才能使用，可能不适用于某些应用

（2）必须事先给出k（要生成的簇的数目），而且对初值敏感，对于不同的初始值，可能会导致不同结果

（3）不适合于发现非凸形状的簇或者大小差别很大的簇

（4）对噪声和孤立点数据敏感

可作为其他聚类方法的基础算法，如谱聚类。

1.5.10 对K-means的思考：K-Mediods聚类(K中值距离)

第10课聚类01:02:40

K-Means将簇中所有点的均值作为新质心，若簇中含有异常点，将导致均值偏离严重。以一维数据为例：

（1）数据1、2、3、4、100的均值为22，显然距离“大多数”数据1、2、3、4比较远

（2）改成求数组的中位数3，在该实例中更为稳妥。

（3）这种聚类方式即K-Mediods聚类（K中值距离）

初值的选择，对聚类结果有影响吗？如何避免？

1.5.11 轮廓系数(Silhouette)

第10课聚类01:07:37

Silhouette系数是对聚类结果有效性的解释和验证，由Peter J. Rousseeuw于1986年提出。

计算样本i到同簇其他样本的平均距离a_i。a_i越小，说明样本i越应该被聚类到该簇。将a_i称为样本i的簇内不相似度。簇C中所有样本的a_i均值称为簇C的簇不相似度。

计算样本i到其他某簇C_j的所有样本的平均距离b_{ij}，称为样本i与簇C_j的不相似度。定义为样本i的簇间不相似度：$b_i = \min\{b_{i1}, b_{i2}, ...b_{iK}\}$。其中，$b_i$越大，说明样本$i$越不属于其他簇。

根据样本i的簇内不相似度a_i和簇间不相似度b_i，定义样本i的轮廓系数：

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

s_i接近于1，则说明样本i聚类合理；s_i接近于-1，则说明样本i更应该聚类到另外的簇；若s_i近似为0，则说明样本i在两个簇的边界上。

所有的样本的s_i的均值，称为聚类结果的轮廓系数，是该聚类是否合理、有效的度量。

1.5.12 层次聚类方法

第10课聚类01:19:19

层次聚类方法对给定的数据集进行层次的分解，直到某种条件满足为止。具体又分为：（1）凝聚的层次聚类：AGNES算法，（2）分裂的层次聚类：DIANA算法。

凝聚的层次聚类（AGNES算法），是一种自底向上的策略，首先将每个对象作为一个簇，然后合并这些原子簇为越来越大的簇，直到某个终结条件被满足。

分裂的层次聚类（DIANA算法），采用自顶向下的策略，它首先将所有对象置于一个簇中，然后逐渐细分为越来越小的簇，直到达到了某个终

结条件。

1.5.13 密度聚类方法

第10课聚类01:22:39

密度聚类算法的指导思想是，只要样本点的密度大于某阈值，则将该样本添加到最近的簇中。

这类算法能克服基于距离的算法只能发现“类圆形”的聚类的缺点，可发现任意形状的聚类，且对噪声数据不敏感。但计算密度单元的计算复杂度大，需要建立空间索引来降低计算量。

常用的两个算法：DBSCAN算法和密度最大值算法。

DBSCAN算法（Density-Based Spatial Clustering of Applications with Noise）是一个比较有代表性的基于密度的聚类算法。与划分和层次聚类方法不同，它将簇定义为密度相连的点的最大集合，能够把具有足够高密度的区域划分为簇，并可在有“噪声”的数据中发现任意形状的聚类。

密度最大值聚类是一种简洁优美的聚类算法，可以识别各种形状的类簇，并且参数很容易确定。

第10课聚类01:40:39

1.5.14 谱和谱聚类

第10课聚类01:57:06

方阵作为线性算子，它的所有特征值的全体统称方阵的谱。

（1）方阵的谱半径为最大的特征值（2）矩阵A的谱半径：A^TA的最大特征值

谱聚类是一种基于图论的聚类方法，通过对样本数据的拉谱拉斯矩阵的特征向量进行聚类，从而达到对样本数据聚类的目的。

2 提升

bagging 并行RF

boosting 串行GDBT Adaboost

GBDT分类案例CART（Classification And Regression Tree）回归树

Adaboost Adaptive Boosting 自适应增加

问题：（1）选择哪个特征合适；（2）选取该特征的某个值进行切分呢？

3 支持向量机SVM

3.1 复习：对偶问题

第09课SVM 00:01:22 复习：对偶问题

一般优化问题的Lagrange乘子法

Lagrange函数

【lsieun】“仿射”和“线性变换”似乎是同一个意思。

第09课SVM 00:01:22 复习：Lagrange对偶函数(dual function)

概念：KKT条件

【lsieun】Lagrange函数的最大值，就等于“原函数”的最小值。这里主要是讲Lagrange乘子法作为一种工具，将原来求“最小值”的问题，转换为求“最大值”的问题。

第09课SVM 00:03:42 线性方程的最小二乘问题

此处是举例子，用于解释上面的“由最小值转换成求最大值的解决方法”。

第09课SVM 00:04:26 强对偶条件

若要对偶函数的最大值即为原问题的最小值，考察需要满足的条件

第09课SVM 00:05:37 强对偶KKT条件: Karush-Kuhn-Tucker

3.2 主要内容和目标

理解支持向量机SVM的原理和目标SVM核心的东西what.

掌握支持向量机的计算过程和算法步骤SVM核心的东西how.

理解软间隔最大化的含义：（1）对线性不可分的数据给出（略有错误）的分割面；（2）线性可分的数据需要使用“软间隔”目标函数吗？

了解核函数的思想

了解SMO算法的过程

核函数：SVM本身是个线性分类器，那么在原始的SVM上，可以加上一些核函数，来构造一个非线性的分隔面，来更好的解决分类问题。

3.3 各种概念

第09课SVM 00:09:32 各种概念

SVM进行简单的化分，可以分为三类：

第一类，**线性可分**支持向量机：（1）硬间隔最大化hard margin maximization；（2）硬间隔支持向量机

第二类，**线性**支持向量机：（1）软间隔最大化soft margin maximization；（2）软间隔支持向量机

第三类，**非线性**支持向量机：（1）核函数kernel function

从学习的角度来说，第一类（线性可分支持向量机）是最重要的，只要学会了第一类，稍微加一点东西就能变成第二类（线性支持向量机），对第二类稍微加一点东西就能变成第三类（非线性支持向量机）。[知识梯度lsieun]

在实际应用中，使用的较多的就是第二类和第三类了。

3.4 分隔超平面

第09课SVM 00:13:37 分隔超平面

[lsieun]什么是超平面？概念怎么理解，有时间查一查。之前查的时候，印象中是说，超过二维的平面，都叫超平面。

第09课SVM 00:13:37 分隔超平面的思考

如何定义两个集合的“最优”分割超平面？（1）找到集合“边界”上的若干点，以这些点为“基础”计算超平面的方向；以两个集合边界上的这些点的平均作为超平面的“截距”；（2）支持向量：support vector

若两个集合有部分相交，如何定义超平面，使得两个集合“尽量”分开？

第09课SVM 00:16:00 线性分类问题

假定一共有N个样本，最终可能只有n个样本参与到支持向量(support vector)里去了，一般而言，N是大于n的，甚至是N远大于n的，就比如说有10000个样本，有20个样本参与到支持向量里去了，相当于有9980都是零，只有20个非零，所以SVM是个稀疏的模型。在有些教材中，会特意将SVM放在“稀疏模型”中介绍。

第09课SVM 00:20:00 CNN 卷积神经网络也是个稀疏模型。在这一点上（都是稀疏模型），SVM和CNN是相似的。如果在面试中谈到了，或许是一个加分项呢，哈哈。。。

事实上，有很多很多条直线可以将两部分图形分开，但是谁是最优的直线呢？如果我们知道了“哪条直线是最优的”，我们又怎么把“目标函数”写出来呢？我们只有写出了“目标函数”，才能下一步去“如何优化它”。所以，首先要有目标函数。此处解决的是“有和无”（目标函数）的问题，之后才是“优化”的问题。

3.5 SVM的目标函数

第09课SVM 00:24:00 画图讲解“空间求距离”

在做SVM的时候，不要过多考虑究竟是多少维的问题，讲课的时候，画在纸上，用两维的数据是为了方便，其实它的算法/计算过程是一样的。推导的时候，是用二维来做，但真正做的时候，与N维是没有区别的。

sign函数：符号函数（一般用sign(x)表示）是很有用的一类函数，能够帮助我们在几何画板中实现一些直接实现有困难的构造。符号函数能够把函数的符号析离出来。在数学和计算机运算中，其功能是取某个数的符号（正或负）：当$x > 0$，$sign(x) = 1$;当$x = 0$，$sign(x) = 0$; 当$x < 0$，$sign(x) = -1$。

第09课SVM 00:32:00 很精彩的部分SVM的目标函数：求最小值的最大值，哈哈

3.6 求解SVM的目标函数

3.6.1 输入数据

第09课SVM 00:39:00 输入数据

假设给定一个特征空间上的训练数据集$T = \{(x_1, y_1), (x_2, y_2)...(x_N, y_N)\}$，其中，$x_i \in R^n$，$y_i \in \{+1, -1\}$，$i = 1, 2, ...N$。

x_i为第i个实例（若$n > 1$，则x_i为向量）。

y_i为x_i的类标记：（1）当$y_i = +1$时，称x_i为正例；（2）（1）当$y_i = -1$时，称x_i为负例。

(x_i, y_i)称为样本点。

[一个非常好的问题]为什么y_i的取值是+1和-1呢？

第09课SVM 00:44:00 在做logistic 回归的时候，y一个可以取1，一个可以取0。到了SVM里面呢，y一个取+1，一个取-1。为什么会这样呢？所有教科书中都不会讲为什么。事实上，logistic回归，也可以用+1和-1去推导，

是可以做的。在SVM里面用+1和-1是为了方便，仅此而已，方便推导。方便在哪儿了呢？如果y等于+1或-1，那么$y_i \cdot f(x) = \frac{f(x)}{y_i}$，这样的话，就方便我们做推导。

[问题]SVM中实际当中用的多吗？

第09课SVM 00:45:00

[问题]SVM和Spark？SVM的分类效果，真的是好，一般而言，SVM的分类结果优于logistic回归、优于随机森林（RF），但是SVM的计算速度慢。比如，几千个，几万个样本，SVM不做优化的情况下，用到分钟级（时间）能够把参数学出来，但是随机森林，秒级（时间）就能搞定。SVM往往是一个比较好的分类器。

[问题]SVM如何做多分类呢？

3.6.2 线性可分支持向量机

给定“线性可分训练集”，通过间隔最大化得到的分隔超平面为$y(x) = w^T\Phi(x) + b$，相应的分类决策函数$f(x) = sign(w^T\Phi(x) + b)$，该决策函数称为“线性可分支持向量机”。

其中，$\Phi(x)$是某个确定的特征空间转换函数，它的作用是将x映射到（更高的）维度。最简单直接的：$\Phi(x) = x$。

稍后会看到，求解分离超平面问题可以等价为求解相应的凸二次规划问题。

3.6.3 整理符号

分割平面：$y(x) = w^T\Phi(x) + b$

训练集：$x_1, x_2, \ldots, x_n$

目标值：$y_1, y_2, \ldots, y_n$

新数据的分类：$sign(y(x))$

[问题] 第09课SVM 00:48:00 分隔超平面，哪边是+1，哪边是-1，跟w方向有关系吗？是的。如果位于分隔超平面的法向量的正向，就是+1；位于法向量的负向，就是-1。

3.6.4 推导目标函数

第09课SVM 00:49:00 推导目标函数

3.6.5 最大间隔分离超平面

第09课SVM 00:52:00 最大间隔分离超平面

不要忘记了，虽然目标函数是我们的优化目标，但其实是想通过目标函数求解出其中w和b。

3.6.6 函数间隔和几何间隔

第09课SVM 00:54:00 函数间隔和几何间隔

3.6.7 建立新目标函数

第09课SVM 00:58:00 建立新目标函数

3.7 拉格朗日乘子法

第09课SVM 01:15:00 拉格朗日乘子法

目标函数

约束条件

原问题是极小极大问题

原始问题的对偶问题，是极大极小问题。

第09课SVM 01:30:00 整理目标函数：添加负号

3.8 线性可分支持向量机学习算法

第09课SVM 01:33:00 线性可分支持向量机学习算法

构造并求解约束最优化问题

$$\begin{aligned} \min_a \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j \left(\Phi(x_i)\cdot\Phi(x_y)\right) - \sum_{i=1}^{n} a_i \\ s.t. \sum_{i=1}^{n} a_i y_i = 0 \\ a \geq 0, i = 1, 2, \ldots, n \end{aligned} \tag{39}$$

求得最优解a^*，再将a^*代入到下面的方程中去

$$
\begin{aligned}
w^* &= \sum_{i=1}^{N} a_i y_i \Phi(x_i) \\
b^* &= y_i - \sum_{i=1}^{N} a_i^* \left(\Phi(x_i) \cdot \Phi(x_j)\right)
\end{aligned}
\tag{40}
$$

计算求得分离超平面

$$
w^* \Phi(x) + b^* = 0 \tag{41}
$$

分类决策函数

$$
f(x) = sign\left(w^* \Phi(x) + b^*\right) \tag{42}
$$

第09课SVM 01:36:00 将约束带入函数，化简计算

第09课SVM 01:37:00 现在要说一个重要的结论，只有α不为0的向量，才是支撑向量。

3.9 线性支持向量机

第09课SVM 01:39:00 线性支持向量机

不一定分类完全正确的超平面就是最好的。

样本数据本身线性不可分

若数据线性不可分，则增加松弛因子$\varepsilon_i \geq 0$，使函数间隔加上松弛变量大于等于1.这样约束条件变成

$$
y_i(w \cdot x_i + b) \geq 1 - \varepsilon_i \tag{43}
$$

目标函数：

$$
\min_{w,b} \frac{1}{2}||w||^2 + C\sum_{i=1}^{N} \varepsilon_i \tag{44}
$$

虽然可以这么做，但是松弛因子ε_i不能太过分（不能太大），因此将ε_i加起来求和并放到了目标函数当中（$\sum_{i=1}^{N} \varepsilon_i$），当作约束条件之一，让整体的值是最小的，这样就得到了线性的SVM。

现在需要来解释一下“目标函数”。首先说C，C是用来调节松弛因子ε_i和原始的函数$\frac{1}{2}||w||^2$之间的比例关系；如果C是无穷大的时候，那就

意味着：哪怕ε_i是一个很小的数，只要乘以C，就会使得最后的结果很大，此时就必须强制性的要求ε_i为0，如果不为0，就不可能取最小值了。因此，当C是无穷大的时候，ε_i为0，就相当于退化成了线性可分的SVM。“线性SVM”是“线性可分SVM”的一个推广、泛化。

3.10 线性SVM的目标函数

第09课SVM 01:48:00 线性SVM的目标函数

3.11 核函数

第09课SVM 02:07:00 核函数

可以使用核函数，将原始输入空间映射到新的特征空间，从而使原本线性不可分的样本可能在核空间可分。

多项式核函数$\kappa(x_1, x_2) = (||x_1 - x_2||^a + r)^b$，其中$a, b, r$为常数。

高斯核函数RBF $\kappa(x_1, x_2) = \exp\left(-\frac{||x_1-x_2||^2}{2\sigma^2}\right)$

字符串核函数，例如两字符串的满足某条件的子串的余弦相似度

在实际应用中，往往依赖先验领域知识/交叉验证等方案才能选择有效的核函数。如果没有更多先验信息，则使用高斯核函数。

3.12 SVM中系数的求解：SMO

第09课SVM 02:07:00 SVM中系数的求解：SMO

序列最小最优化Sequential Minimal Optimization

有多个拉格朗日乘子

每次只选择其中两个乘子做优化，其他因子认为是参数。将N个问题，转换成两个变量的求解问题，并且目标函数是凸的。

3.13 总结与思考

第09课SVM 02:26:00 总结与思考

SVM可以用来划分多分类别吗？（1）直播多分类；（2）1 vs rest / 1 vs 1

SVM和Logistic回归的比较：（1）经典的SVM，直接输出类别，不给出后验概率；（2）Logistic回归，会给出属于那哪个类别的后验概率。重点：两者目标函数的异同。

SVM框架下引入Logistic函数：输出条件后验概率

SVM用于回归问题：SVR

体会SVM的目标函数的建立过程：原始目标函数和Lagrange函数有什么关系。

4 贝叶斯网络

之前的线性回归、逻辑回归，都是已知X来求y，是建立X和y的一种直接关系。

SVM中也是已知X来求y，只不过多了一个$\kappa(x_1, x_2)$核函数，也是建立X和y的一种直接关系。

有一种情况是X和y之前并不是直接相关的，它们中间有一些我们看不到的其他东西，那这些东西我们就需要建立一个图模型来解释它，我们往往会把X和y看成某一个随机变量，那么X可能服从某个概率密度，y可能服从某个概率密度，如此一来，中间的每一个节点就都是概率上的东西，用概率的节点形成的这样一个图，就是概率图模型。如果模型稍微复杂一点，往往用概率图模型来解决问题是需要的，所以内容是非常大的。

4.1 主要内容

第14课贝叶斯网络00:02:30 主要内容

复习本次将用到知识：相对熵、互信息（信息增益）

朴素贝叶斯

贝叶斯网络的表达：（1）条件概率表示参数个数分析；（2）马尔科夫模型

D-separation：（1）条件相互独立的三种类型；（2）Markov Blanket

网络的构建流程：（1）混合（离散+连续）网络：线性高斯模型；（2）Chow-Liu算法：最大权生成树MSWT。

PLSA—CDA—DL

4.2 复习

4.2.1 复习：相对熵

第14课贝叶斯网络00:02:50 相对熵

相对熵，又称互信息，交叉熵，鉴别信息，Kullback熵，Kullback-Leible散度

设$p(x)$、$q(x)$是X中取值的两个概率分布，则p对q的相对熵是

$$D(p||q) = \sum_x p(x) \log \frac{p(x)}{q(x)} = E_{p(x)} \log \frac{p(x)}{q(x)} \tag{45}$$

说明：（1）相对熵可以度量两个随机变量的“距离”；（2）一般的$D(p||q) \neq D(q||p)$，只有当p和q相同的时候，等号才成立；（3）$D(p||q) \geq 0$、$D(q||p) \geq 0$

4.2.2 复习：互信息

第14课贝叶斯网络00:03:45 互信息

两个随机变量X, Y的互信息，定义为X，Y的“联合分布”和“独立分布乘积”的相对熵。

$$\begin{aligned} I(X,Y) &= D\left(P(X,Y)||P(X)P(Y)\right) \\ I(X,Y) &= \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \end{aligned} \tag{46}$$

4.2.3 复习：信息增益

第14课贝叶斯网络00:03:56 信息增益

信息增益，表示得知特征A的信息而使类X的信息的不确定性减少的程度。

定义：特征A对训练数据集D的信息增益$g(D, A)$，定义为集合D的经验熵$H(D)$[lsieun:经验熵，我理解为“不知道的东西的信息量”]与特征A给定条件下D的经验熵$H(D|A)$之差，即$g(D, A) = H(D) - H(D|A)$；显然，这即为训练集D和特征A的互信息，所以，“信息增益”和“互信息”本质上是一个东西。

4.3 概率

第14课贝叶斯网络00:04:32 概率

条件概率：$P(A|B) = \frac{P(AB)}{P(B)}$

全概率公式：$P(A) = \sum_i P(A|B_i)P(B_i)$

贝叶斯（Bayes）公式：$P(B_i|A)\frac{P(A|B_i)P(B_i)}{\sum_j P(A|B_j)P(B_j)}$

由“条件概率”和“全概率公式”可以推导出“贝叶斯公式”。

我感觉，自己好像懂了一点儿，但还是不那么特别懂，有时间百度一下。目前的理解是，“贝叶斯公式”是由“结果”求“原因”的概率，由“后”向“前”计算的概率，是“后验”概率。查一下“先验概率”和“后验概率”到底是什么意思。

第14课贝叶斯网络00:04:55 贝叶斯公式带的思考

4.4 朴素贝叶斯的假设

第14课贝叶斯网络00:13:33 朴素贝叶斯的假设

（1）特征独立性（概率）：一个特征出现的机率，与其他特征（条件）独立。其实是：对于给定分类的条件下，特征独立。

（2）特征均衡性（重要性）：每个特征同等重要。

4.4.1 以文本分类为例

第14课贝叶斯网络00:14:26 以文本分类为例

样本：10000封邮件，每个邮件被标记为垃圾邮件或非垃圾邮件

分类目标：给定第10001封邮件，确定它是垃圾邮件还是非垃圾邮件

方法：朴素贝叶斯

4.4.2 拉谱拉斯平滑

第14课贝叶斯网络00:34:41 拉谱拉斯平滑

4.5 贝叶斯网络

第14课贝叶斯网络00:45:08 贝叶斯网络

4.5.1 一个简单的贝叶斯网络

第14课贝叶斯网络00:50:39 一个简单的贝叶斯网络

4.5.2 全连接贝叶斯网络

第14课贝叶斯网络00:52:14 全连接贝叶斯网络

4.5.3 一个“正常”的贝叶斯网络

第14课贝叶斯网络00:55:12 一个“正常”的贝叶斯网络

4.5.4 贝叶斯网络的形式化定义

第14课贝叶斯网络01:13:33 贝叶斯网络的形式化定义

5 主题模型

5.1 主要内容

第15课主题模型00:00:58 主要内容

共轭先验分布

多项式分布-Dirichlet分布：二项式分布-Beta分布

LDA模型：Gibbs采样算法

多项分布，比如掷色子，有6个点，如果做N次就是6项分布，它的共轭分布就是Dirichlet分布。如果变成2点投硬币，投N次就是二项分布，二项分布的共轭分布叫Beta分布。

“共轭”在数学、物理、化学、地理等学科中都有出现。本意：两头牛背上的架子称为轭，轭使两头牛同步行走。共轭即为按一定的规律相配的一对。通俗点说就是孪生。在数学中有共轭复数、共轭根式、共轭双曲线、共轭矩阵等。

“二项分布”就是重复n次独立的伯努利试验。在每次试验中只有两种可能的结果，而且两种结果发生与否互相对立，并且相互独立，与其它各次试验结果无关，事件发生与否的概率在每一次独立试验中都保持不变，则这一系列试验总称为n重伯努利实验，当试验次数为1时，二项分布服从0-1分布。

“多项式分布”（Multinomial Distribution）是二项式分布的推广。二项分布的典型例子是扔硬币，硬币正面朝上概率为p, 重复扔n次硬币，k次为正面的概率即为一个二项分布概率。把二项分布公式推广至多种状态，就得到了多项分布。

5.2 引：Γ函数

第15课主题模型00:01:50 Γ函数

Γ函数（读作Gamma函数，拼音“ga ma”函数）是阶乘在实数上的推广。是欧拉发现的。

5.3 Beta分布

第15课主题模型00:05:36 Beta分布

Beta分布的概率密度

其中系数B为

Gamma函数可以看成阶乘的实数域推广

5.3.1 Beta分布的期望

第15课主题模型00:08:26 Beta分布的期望

5.4 朴素贝叶斯的分析

第15课主题模型00:17:07 朴素贝叶斯的分析

可以胜任许多文本分类的问题

无法解决语料中“一词多义”和“多词一义”的问题，它更像是词法分析，而非语义分析

如果使用“词向量”作为文档的特征，“一词多义”和“多词一义”会造成计算文档间相似度的不准确性。

可以通过增加“主题”的方式，一定程度的解决上述问题：（1）一个词可能被映射到多个主题中，“一词多义”；（2）多个词可能被映射到某个主题的概率很高，“多词一义”。

5.5 文档和主题

第15课主题模型00:20:44 文档和主题

是一个标准的无监督学习

LDA是一个主题模型，也是无监督的东西。给了若干个文档，假设给定k个主题，其实就是讲若干个文档映射到了k个主题上，第1个文档跟第1个主题有多少相似度（相关度），第2个文档跟第1个主题有多少相似度（相关度）……每一个文档跟每一个主题都有一个相似度（相关度），就可以认为这些文档做了k个主题“概率化的聚类”，因此，LDA可以从一定程度上是一个“聚类”。这一段主要是讲“LDA”和“聚类（降维）”的关系。

LDA是一个无监督模型，因为只给定了文本X，没有告诉我们这些文档属于什么样的主题，只给了X，让我们去学那个y。我们经常讲，“无监督模型”，往往可以约等于“聚类”，往往可以约等于“降维”，本质上它们往往可以通用。如果你发现一个模型是“无监督的”，你就往里去套，往往你会发现，它就是“聚类”，就是“降维”。

D—Z—W：在文档（D）和词（W）之间加了一个隐变量（Z），有的时候说“主题”或者“话题”是一个意思。这个模型是02、03年才提出的。

这里边涉及到将“文档”进行“切词”的一个过程。注意，“切词”这个概念。

5.6 LDA涉及的主要问题

第15课主题模型00:29:10 LDA涉及的主要问题
共轭先验分布
Dirichlet分布
LDA模型：Gibbs采样算法学习参数

5.6.1 共轭先验分布

第15课主题模型00:29:22 共轭先验分布

5.6.2 复习：二项分布的最大似然估计

第15课主题模型00:38:41 二项分布的最大似然估计

5.6.3 共轭先验的直接推广

第15课主题模型00:51:22 共轭先验的直接推广
从2到K：（1）二项式到多项式；（2）Beta分布到Dirichlet分布

5.6.4 Dirichlet分布

第15课主题模型00:51:54 Dirichlet分布
Beta分布
Dirichlet分布

5.6.5 对称Dirichlet分布的参数分析

第15课主题模型01:06:54 对称Dirichlet分布的参数分析

我们从来没有说过“词频”满足Dirichlet分布，大家一定要清楚这个概念，这个“词频”就是一个非常普通的V点分布，“主题”也是一个标准的V项分布，而多项分布是需要有“参数”的，这个参数满足Dirichlet分布。“词”也好，“文档”也好，“主题”也好，它们满足的是正常多项式分布，这样是不是清楚一点。

5.6.6 Dirichlet分布分析

第15课主题模型01:08:17 Dirichlet分布分析

5.6.7 参数α对Dirichlet分布的影响

第15课主题模型01:16:17 参数α对Dirichlet分布的影响

主题的数目k是需要事先给定的，但是不需要指定“主题”是什么。主题模型里，相当于要指定什么，(α, k)指定α等于几，指定k等于几，其他的就没有了，其他的参数就不需要用户指定子。

5.6.8 对称Dirichlet分布

第15课主题模型01:20:17 对称Dirichlet分布

5.7 LDA的解释

第15课主题模型01:31:17 LDA的解释

共有m篇文章，一共涉及了K个主题；

每篇“文章”（长度为N_m，第1篇文章长度为N_1，第2篇文章长度为N_2，以此类推）都有各自的“主题分布”，“主题分布”是“多项分布”，该“多项分布”的参数服从“Dirichlet分布”，该“Dirichlet分布”的参数为α。

每个“主题”都有各自的“词分布”，“词分布”为“多项分布”，该“多项分布”的参数服从“Dirichlet分布”，该“Dirichlet分布”的参数为β。

对于某篇“文章”中的第n个“词”，首先从该“文章”的“主题分布”中采样一个“主题”，然后在这个“主题”对应的“词分布”中采样一个“词”。不断重复这个随机生成过程，直到m篇“文章”全部完成上述过程。

5.8 Gibbs updating rule

第15课主题模型01:50:17 Gibbs updating rule

5.9 代码实现

第15课主题模型02:06:17 代码实现

6 卷积神经网络

00:46:23

神经网络

隐藏层激活函数

input nodes / hidden nodes / output nodes

一般建议：隐藏层内的节点多一些，而不是隐藏层数多，谷歌的那个狗才5层而已。

relu

7 Tensorflow

7.1 安装Tensorflow

conda install tensorflow

pip install tensorflow

8 微积分与概率论基础

第01课数学分析与概率论00:00:00

能够在如何用“机器学习”算法的基础之上，再能够知道它为什么是起作用的。所以我们会给大家探讨算法的背后是什么，在“机器学习”的角度来看“数学到底是如何的”。

8.1 主要内容

本课程示例概述

机器学习的角度看数学：（1）复习数学分析（常数e、导数/梯度、Taylor展开式、凸函数）；（2）概率论基础（古典概型、贝叶斯公式、常见概率分布）

8.2 什么是机器学习

第01课数学分析与概率论00:01:21 什么是机器学习

Tom Michael Mitchell在1997年给出了“机器学习”的定义：对于某给定的任务T(task)，在合理的性能度量方案P的前提下，某计算机程序可以自主学习任务T的经验E；随着提供合适、优质、大量的经验E，该程序对于任务T的性能逐步提高。

这里最重要是机器学习的对象：（1）任务Task,T，一个或多个；（2）经验Experience, E；（3）性能Performance,P

“机器学习”的定义，这样一个描述看起来很严格，但有时候理解起来却不知所云，不知道它在说什么。我们可以简单的理解为：随着“任务”的不断的执行，“经验”的累积会带来计算机“性能”的提升。

8.2.1 换个表述

第01课数学分析与概率论00:02:59 换个表述

机器学习(Machine Learning)，是“人工智能(AI)”的一个分支。我们使用计算机设计一个系统，使它能够根据提供的训练数据/样本按照一定的方式来学习；随着训练次数的增加，该系统可以在性能上不断学习和改进；通过参数优化的学习模型，能够用于预测相关问题的输出。

现在我们来举一个例子，“机器学习”到底与我们传统的“算法”最大的区别在什么地方。我们举一个经常见的例子“无人驾驶汽车”，思考：如何设计无人驾驶机动车？

汽车的无人汽车模块已经成熟：全自动公共交通工具已经出现在了世界上的多个城市，Lutz探路者/CYCAB/Google。问题是，如何设计自动驾驶系统呢？把所有的交通规则录入到系统中去，人多的时候，人少的时候，应该以多大的速度来行驶，到路口应该怎么做，到了这种情况应该怎么做，到了那种情况应该怎么做，我们把所有的情况都要处理好，然后这个系统就做完了，这是一种思路。这种思路是传统的做法，却不是“机器学习”的算法。这种思路会带来的一个问题，我们不太方便去穷举所有的情况，那我们应该怎么考虑呢？我们就先去做一个简单的、刚刚能满足要

求的、能够在最简单的路况上进行自动行驶的系统，遇到某种情况（样本）之后，做参数调整，遇到新的情况（样本），再做参数调整，不停的做这样的事情，经过大量的样本迭代之后，就输出了我们想要的那个“模型参数”。这样一个过程，是通过我们的样本(sample)，我们采样，不停的让系统去学，这就是机器学习的基本想法，所以“机器学习”是一个非常拟人的说法，让机器来学习，不停的学，直到把“参数”学到手。这就是跟传统算法非常非常不一样的东西。“机器学习”是一个很“务实”的东西，它拿到“样本”，拿到“模型”之后，不断的训练“参数”。通过这样一个例子，就比只有一个“严格的定义”要更容易让大家理解到更多的东西。

因此，在机器学习里，我们需要有“样本”，需要建立我们的“模型”，然后需要做我们的“参数”，在后续的学习中，我们需要不停的做这个东西。

我们现在已经了解了“机器学习”的思路了，那么应该如何去做机器学习呢？如何去介入这个领域呢？遇到这种情况，我们会先想“人是怎么学习的呢”。

8.3 人类的学习？

如何从完全“无知”到掌握知识？语言、颜色、形状等特征统计

有监督学习：月亮

无监督学习：阅兵

增强学习：机器人走路、机器人踢球

8.4 很精彩

第01课数学分析与概率论00:18:32 很精彩

sample feature target

train test

training/labels/feature vectors/machine learning algorithm/model/expected label

sample分为train和test两部分，第一部分train是用“标记好”的数据做训练，第二部分test是用已经训练好的模型做预测。这是两个阶段，很显然的事情是：我们应该把重点放在前一部分，如何把“模型参数”给学出来，用什么样的算法、什么样的工具、什么样的优化手段，把这个参数学到手，

然后至于怎么测试它、怎么用它，相对而言是简单的，将未知的X带进去就可以了。

sample数据的格式，可以有text docs、images、sounds和transactions（交易数据）。

[问题]深度学习是什么概念呢？第01课数学分析与概率论00:19:57

8.5 机器学习方法

第01课数学分析与概率论00:22:04 机器学习方法

同一批“样本数据”，不同的“模型”做出来的结果是不一样的。因此，在机器学习中，第一步是确定用什么模型来为我们的数据做服务，不同的数据需要有不同的模型来做。这里面有一个认知问题：并不是复杂的模型，就一定是好的，一般同来而言，如果有两个模型可以胜任我们的工作，谁（“模型”）简单用谁做。真不是越复杂越好，而是越简单越好，一般我们把这种思路叫做“奥卡姆剃刀(Occam's Razor)”。

奥卡姆剃刀定律（Occam's Razor, Ockham's Razor）又称“奥康的剃刀”，它是由14世纪逻辑学家、圣方济各会修士奥卡姆的威廉（William of Occam，约1285年至1349年）提出。这个原理称为“如无必要，勿增实体”，即“简单有效原理”。正如他在《箴言书注》2卷15题说“切勿浪费较多东西去做，用较少的东西，同样可以做好的事情。”

奥卡姆剃刀原则是奥卡姆（全称是「奥卡姆的威廉」,「William of Ockham」）当年说过的某句话的前半部分。那一句话是：**Do not multiply entities beyond necessity**, but also do not reduce them beyond necessity.按照我的理解，这句话的前半句才是剃刀原则。

作者：知乎用户链接：https://www.zhihu.com/question/20159241/answer/14167100 来源：知乎著作权归作者所有。商业转载请联系作者获得授权，非商业转载请注明出处。

「问题」监督学习和非监督学习都有哪些？第01课数学分析与概率论00:25:04

监督学习：线性回归、K近邻、逻辑回归、Linear SVM、还有核函数的SVM、决策树、Naive Bayes、

非监督学习：聚类、EM算法、（推荐系统里的）协同过滤、关联分析

聚类是一个最重要的非监督学习方法，甚至所有非监督学习的算法都可以归并到“聚类”的概念里面去。

是否有y，是否有“标记/标签”是区分“监督学习”和“非监督学习”的重要标志。

「问题」机器学习和数据挖掘的区别？第01课数学分析与概率论00:28:23

8.6 思考：机器如何发现新词

第01课数学分析与概率论00:30:21 思考：机器如何发现新词

PPT当中描述了一种方法，但没有细讲；说是后面会讲到用HMM（隐马尔可夫模型）发现新词。

8.7 Python Code示例

第01课数学分析与概率论00:32:10 Python Code示例

第01课数学分析与概率论00:36:03 线性回归、rate、Loss

EM code

GMM（高斯混合模型）与图像

SVM：高斯核函数的影响1995-2006年非要重要的算法

贝叶斯网络

理解HMM框架

HMM分词

其他内容：最大熵模型（自然语言处理解决标记问题）、聚类（K-means、K-Mediods、密度聚类、谱聚类）、降维（PCA/SVD/ICA）、SVM（与核技术相结合）、主题模型PLSA/LDA（与聚类、标签传递算法相结合）、条件随机场（无向图模型、链式条件随机场解决标记问题）、变分推导Variation Interface（与EM、贝叶斯相结合、参数、隐变量的学习）、深度学习（大规模人工神经网络）

8.8 本课程参考书、文献

第01课数学分析与概率论00:38:18 本课程参考文献

8.9 回忆知识

第01课数学分析与概率论00:40:18 回忆知识

求S的值：

$$S=\frac{1}{0!}+\frac{1}{1!}+\frac{1}{2!}+\frac{1}{3!}+\frac{1}{4!}+\cdots+\frac{1}{n!}+\ldots \tag{47}$$

这个S会有极限吗？如果有，它的极限会是多少呢？

这个极限是存在的，极限等于e，为什么呢？

8.10 复习微积分：两边夹定理

第01课数学分析与概率论00:40:44 复习微积分：两边夹定理

当$x\in U(x_0,r)$时（x是在x_0的邻域是有定义的），有$g(x)\leq f(x)\leq h(x)$成立，并且$\lim\limits_{x\to x_0}g(x)=A$，$\lim\limits_{x\to x_0}h(x)=A$，那么：

$$\lim_{x\to x_0}f(x)=A \tag{48}$$

8.11 极限

第01课数学分析与概率论00:41:40 极限

8.12 复习微积分：极限存在定理

第01课数学分析与概率论00:48:33 复习微积分：极限存在定理

8.13 导数

第01课数学分析与概率论00:59:46 导数

简单的说，导数就是曲线的斜率，是曲线变化快慢的反映。

二阶导数是斜率变化快慢的反映，表征曲线凹凸性：（1）二阶导数连续的曲线，往往称之为“光滑”的；（2）还记得高中物理老师时常念叨的吗，加速度的方向总是指向轨迹曲线凹的一侧。

根据$\lim\limits_{x\to\infty}\left(1+\frac{1}{x}\right)^x=e$可以得到函数$f(x)=\ln x$的导数，进一步根据换底公式、反函数求导等，得到其他初等函数的导数。

$f(x)=\log_a x$，当$a=e$的时候，在$x=1$处点的导数为1；当$a=e$的时候，将$f(x)$写成$f(x)=\ln x$。

8.13.1 常用函数的导数

第01课数学分析与概率论01:01:46 常用函数的导数

$C' = 0$

$(x^n)' = nx^{n-1}$

$(\sin x)' = \cos x$

$(\cos x)' = -\sin x$

$(a^x)' = a^x \ln a$

$(e^x)' = e^x$

$(\log_a x)' = \frac{1}{x} \log_a e$

$(\ln x)' = \frac{1}{x}$

$(\mu + \nu)' = \mu' + \nu'$

$(\mu\nu)' = \mu'\nu + \mu\nu'$

8.13.2 应用1

第01课数学分析与概率论01:02:17 应用1

已知函数$f(x) = x^x$，其中$x > 0$，求$f(x)$的最小值。（1）领会幂指函数的一般处理思路；（2）在信息熵章节中将再次遇到它。

在算这种“任何函数”取“极值”的时候，往往是对它（该函数）进行“求导”；然后让“导数”等于0。导数为0的点，我们称为“驻点”；然后我们再通过别的方式来判断这个“驻点”是极大值，还是极小值；然后，就能从统一的概念上来看待求极值问题，就是这么一个过程。

幂指函数取“导数”是不方便的。在“常用函数的导数”中给出了“幂函数”的导数公式和“指数函数”的导数公式，但是却没有给出“幂指函数”的导数公式。这时候，我们应该怎么办呢？

“幂函数”是基本初等函数之一。一般地，形如$y = x^a$(a为有理数）的函数，即以“底数”为自变量，“幂”为因变量，“指数”为常数的函数称为**幂函数**。例如函数$y = x^0$、$y = x^1$、$y = x^2$、$y = x^{-1}$等都是幂函数。

“指数函数”是重要的基本初等函数之一。一般地，$y = a^x$函数(a为常数且以$a > 0$，$a \neq 1$)叫做**指数函数**，函数的定义域是R。

幂指函数既像“幂函数”，又像“指数函数”，二者的特点兼而有之。作为“幂函数”，其“幂指数”确定不变，而“幂底数”为自变量；相反地，“指数函数”却是“底数”确定不变，而“指数”为自变量。**幂指函数**

就是“幂底数”和“幂指数”同时都为“自变量”的函数。这种函数的推广，就是广义幂指函数。

附：$N^{\frac{1}{\log N}}$ =?（1）在计算机算法跳跃表Skip List的分析中，用到了该常数。（2）背景：跳表是支持增删改查的动态数据结构，能够达到与平衡二叉树、红黑树近似的效率，而代码实现简单。

有时间百度一下“平衡二叉树”和“红黑树”吧。

8.13.3 求解x^x

第01课数学分析与概率论01:04:19 求解x^x

令$t = x^x$，然后取“对数”可以得到$\ln t = x \ln x$；两边对x求导，得到$\frac{1}{t}t' = \ln x + 1$；令$t' = 0$，得到$\ln x + 1 = 0$；再求得$x = e^{-1}$；再求得$t = e^{-\frac{1}{e}}$。

8.13.4 积分应用2

第01课数学分析与概率论01:06:21 积分应用2

当$N \to \infty$的时候（当N趋向于无穷大的时候），$\ln N!$大致等于$N(\ln N - 1)$，即$\ln N! \to N(\ln N - 1)$。

细节，我在这里先省略了，后续再看。

$$
\begin{aligned}
\ln N! &= \sum_{i=1}^{n} \ln i \\
&\approx \int_1^N \ln x dx \\
&= x \ln x \mid_1^N - \int_1^N x d(\ln x) \\
&= N \ln N - \int_1^N x \cdot \frac{1}{x} dx \\
&= N \ln N - x \mid_1^N \\
&= N \ln N - N + 1 \\
&\to N \ln N - N
\end{aligned}
\tag{49}
$$

8.13.5 Taylor公式-Maclaurin公式

第01课数学分析与概率论01:10:46 Taylor公式-Maclaurin公式

Taylor公式

$f(x)=f(x_0)+f'(x_0)(x-x_0)+\frac{f''(x_0)}{2!}(x-x_0)^2+\cdots+\frac{f^{(n)}(x_0)}{n!}(x-x_0)^n+R_n(x)$

Maclaurin公式

$f(x)=f(0)+f'(0)x+\frac{f''(0)}{2!}x^2+\cdots+\frac{f^{(n)}(0)}{n!}x^n+o(x^n)$

Taylor公式和Maclaurin公式的区别在于：Taylor公式是在“任意点”展开，而Maclaurin公式是在0点展开，应该说Taylor公式是Maclaurin公式的范化。

我感觉，经常用的就是“Maclaurin公式”，用它来求解一些函数的函数值。下面进行举例。

8.13.6 Taylor公式的应用1

第01课数学分析与概率论01:11:40 Taylor公式的应用1

数值计算：初等函数值的计算（在原点展开）

$$\sin x=x-\frac{x^3}{3!}+\frac{x^5}{5!}-\frac{x^7}{7!}+\frac{x^9}{9!}+\cdots+(-1)^{m-1}\frac{x^{2m-1}}{(2m-1)!}+R_{2m} \tag{50}$$

$$e^x=1+x+\frac{x^2}{2!}+\frac{x^3}{3!}+\cdots+\frac{x^n}{n!}+R_n \tag{51}$$

这两个都是在0点展开，使用的是“Maclaurin公式”。

在实践中，往往需要做一定程度的变换。如果自己实现一种新的语言，类似于C语言或Java语言，我们可以实现一些数据类库的功能，例如计算$\sin x$或e^x，按道理上来说，是可以用“Maclaurin公式”来计算的；但是，有一个问题，“Maclaurin公式”展开之后可以有无穷多项（n可以取很大的值），而我们往往取前10项，它的精度就足够了。这里的描述可能有点问题。

8.13.7 Taylor公式的应用1：计算e^x

第01课数学分析与概率论01:13:40 Taylor公式的应用1：计算e^x

给定正实数x，计算$e^x=?$

一种可行的思路：

求整数k和小数r，使得$x=k\cdot\ln 2+r$，其中$|r|\le 0.5\cdot\ln 2$

从而

$$\begin{aligned} e^x &= e^{k\cdot\ln 2+r} \\ &= e^{k\cdot\ln 2}\cdot e^r \\ &= 2^k\cdot e^r \end{aligned} \tag{52}$$

8.13.8 Taylor公式的应用2

第01课数学分析与概率论01:14:53 Taylor公式的应用2

考察Gini系数的图像、熵、分类误差率三者之间的关系：将$f(x)=-\ln x$在$x=1$处一阶展开，忽略高阶无穷小，得到$f(x)\approx 1-x$。

上述结论，在决策树章节会进一步讨论。

8.14 方向导数

第01课数学分析与概率论01:18:49 方向导数

上面的部分讲的是“一元的导数”，如果是“二元”或“多元”的呢？

如果函数$z=f(x,y)$在点$P(x,y)$是可微分的，那么函数在该点任一方向L的方向导数都存在，且有：

$$\begin{aligned} \frac{\partial f}{\partial l} &= \left(\frac{\partial f}{\partial x},\frac{\partial f}{\partial y}\right)\cdot\begin{pmatrix}\cos\varphi\\ \sin\varphi\end{pmatrix} \\ &= \frac{\partial f}{\partial x}\cos\varphi+\frac{\partial f}{\partial y}\sin\varphi \end{aligned} \tag{53}$$

其中，φ为x轴到方向L的转角，$\left(\frac{\partial f}{\partial x},\frac{\partial f}{\partial y}\right)$是和“方向”**无关**的，而$\binom{\cos\varphi}{\sin\varphi}$是和“方向”**有关**的。

我的问题：“可微分”的概念怎么理解呢？

8.14.1 梯度

第01课数学分析与概率论01:20:56 梯度

设函数$z=f(x,y)$在平面区域D内具有一阶连续偏导数，则对于每一个点$P(x,y)\in D$，向量

$$\nabla_{(x,y)}=\left(\frac{\partial f}{\partial x},\frac{\partial f}{\partial y}\right) \tag{54}$$

为函数$z=f(x,y)$在点P的梯度，记做$gradf(x,y)$

梯度的方向是函数在该点变化最快的方向：考虑一座山，假如它的解析式为$z=H(x,y)$，在(x_0,y_0)的梯度是在该点坡度变化最快的方向。

梯度下降法。思考：若下山方向和梯度方向呈θ角，下降速度是多少？

“导数”这个概念，把它从“一元”扩展到“n元”就得到了“梯度”。“梯度”这个概念，把它从“n元”降到“一元”就是“导数”。在后续的讨论中，这两个词（“导数”和“梯度”）往往是不做严格区分的。

8.15 概率论

第01课数学分析与概率论01:23:18 概率论

对概率的认识：$P(x)\in[0,1]$：（1）若$P=0$，则事件出现的概率为0，但并不表示事件一定不会发生（一根针，扎到一个平面上，它的每一点的概率都为0，但是它还是会发生的）。（2）若x为“离散”或“连续”变量，则$P(x=x_0)$表示x_0发生的“概率”或“概率密度”。

累计分布函数$\Phi(x)=P(0\le x_0)$：（1）$\Phi(x)$一定为单增函数；（2）$\min(\Phi(x))=0$，$\max(\Phi(x))=1$

思考：将值域为$[0,1]$的某单增函数$y=F(x)$看成X事件的累积概率函数：若$y=F(x)$可导，则$f(x)=F'(x)$为某概率密度函数。

cumulative distribution function, CDF 累积分布函数

Probability Density Function, PDF 概率密度函数

8.15.1 古典概型

第01课数学分析与概率论01:33:36 古典概型

“古曲概型”，我的理解是“古典的概率模型”

举例：将n个不同的球放入N（$N\ge n$）个盒子中，假设盒子容量无限，求事件A=｛每个盒子至多有1个球｝的概率。

只要是“古典概型”，要解决它，直接三步走：第一步，算一下所有的“基本事件总数”；第二步，算一下有效事件的总数；第三显，二者相除就可以了。

解：$P(A)=\frac{P_N^n}{N^n}$

基本事件总数：第1个球，有N种放法；第2个球，有N种放法；……共N^n种放法。

每个盒子至多放1个球的事件数：第1个球，有N种放法；第2个球，有$N-1$种放法；第3个球，有有$N-2$种放法；……共$N(N-1)(N-2)\dots(N-n+1)=P_N^n$

8.15.2 生日悖论

第01课数学分析与概率论01:36:41 生日悖论

假定班里有50位同学，则至少有2个生日相同的概率是多少？

答案是：$1-\frac{P_N^n}{N^n}=1-\frac{P_{365}^{50}}{365^{50}}$

之所以称为生日悖论，是因为它和我们的直觉是相违背的。

8.15.3 装箱问题

第01课数学分析与概率论01:38:55 装箱问题

将12件正品和3件次品随机装在3个箱子中，每箱装5件，则每箱中恰好有1件次品的概率是多少？

我现在的问题是：无法解决每箱放5件的条件。

解：

将15件产品装入3个箱子，每箱装5件，共有$\frac{15!}{5!\cdot5!\cdot5!}$种装法

先把3件次品放入3个箱子，有3!种装法。对于这样的每一种装法，把其余12件产品装入3个箱子，每箱装4件，共有$\frac{12!}{4!\cdot4!\cdot4!}$种装法

$$P(A)=\frac{\frac{3!\cdot12!}{4!\cdot4!\cdot4!}}{\frac{15!}{5!\cdot5!\cdot5!}}=\frac{25}{91} \tag{55}$$

8.15.4 与组合数的关系

第01课数学分析与概率论01:41:23 与组合数的关系

把n个物品分成k组，使得每组物品的个数分别为$n_1,n_2,\dots,n_k$，$(n=n_1+n_2+\cdots+n_k)$，则不同的分组方法有$\frac{n!}{n_1!n_2!n_3!\dots n_k!}$种。

上述问题的简化版本，即n个物品分成2组，第一组m个，第二组$n-m$个，则分组方法有$\frac{n!}{m!(n-m)!}$，即C_n^m。

8.15.5 组合数背后的秘密

第01课数学分析与概率论01:42:24 组合数背后的秘密

$N\to\infty\Rightarrow\ln N!\to N(\ln N-1)$

$$
\begin{aligned}
H &= \frac{1}{N}\ln\frac{N!}{\prod_{i=1}^{k} n_i!} = \frac{1}{N}\ln(N!) - \frac{1}{N}\sum_{i=1}^{k}\ln(n_i!) \\
&= (\ln N - 1) - \frac{1}{N}\sum_{i=1}^{k} n_i(\ln n_i - 1) \\
&= \ln N - \frac{1}{N}\sum_{i=1}^{k} n_i \ln n_i = -\frac{1}{N}\left(\left(\sum_{i=1}^{k} n_i \ln n_i\right) - N\ln N\right) \\
&= -\frac{1}{N}\left(\left(\sum_{i=1}^{k} n_i \ln n_i\right) - \sum_{i=1}^{k} n_i \ln N\right) \\
&= -\frac{1}{N}\sum_{i=1}^{k}(n_i \ln n_i - n_i \ln N) = -\frac{1}{N}\sum_{i=1}^{k}\left(n_i \ln\frac{n_i}{N}\right) \\
&= -\sum_{i=1}^{k}\left(\frac{n_i}{N}\ln\frac{n_i}{N}\right) \to -\sum_{i=1}^{k}(p_i \ln p_i)
\end{aligned}
\tag{56}
$$

这个H叫作“熵”。

8.15.6 商品推荐

第01课数学分析与概率论01:45:46 商品推荐

商品推荐场景中“过于聚焦的商品推荐”往往会损害用户的购物体验，在有些场景中，系统会通过一定程度的随机性给用户带来发现的惊喜感。

假设某推荐场景中，经计算A和B两个商品与当前访问用户的匹配度分别为0.8分和0.2分，系统将随机为A商品生成一个均匀分布于0到0.8的最终得分，为B商品生成一个均匀分布于0到0.2的最终得分，试计算最终B的分数大于A的分数的概率。

这里没有听明白。这里的概念大概是“几何概型”。

古典概型的基本事件都是有限的，概率为事件所包含的基本事件除以总基本事件个数。**几何概型**的基本事件通常不可计数，只能通过一定的测度，像长度，面积，体积的的比值来表示

8.16 概率公式

第01课数学分析与概率论01:49:46 概率公式

条件概率：

$$P(A|B) = \frac{P(AB)}{P(B)} \tag{57}$$

全概率公式：

$$P(A) = \sum_i P(AB_i) = \sum_i P(A|B_i)P(B_i) \tag{58}$$

贝叶斯（Bayes）公式：

$$P(B_i|A) = \frac{P(AB_i)}{P(A)} = \frac{P(A|B_i)P(B_i)}{\sum_j P(A|B_j)P(B_j)} \tag{59}$$

思考题：8支步枪中有5支校准过，3支未校准。一名射手用校准过的枪射击，中靶概率为0.8；用未校准的枪射击，中靶概率为0.3；现从8支枪中随机取1支，结果中靶。求该枪是已校准过的概率。

8.16.1 两种认识下的两个学派

第01课数学分析与概率论01:53:35 两种认识下的两个学派

给定某系统的若干样本，求该系统的参数。

矩估计/MLE(最大似然估计)/MaxEnt/EM等：（1）假定**参数是**某个/某些未知的**定值**，求这些参数如何取值，能够使得某目标函数取极大/极小；（2）频率学派。

贝叶斯模型：（1）假定**参数本身是变化的**，服从某个分布，求在这个分布约束下使得目标函数极大或极小；（2）贝叶斯学派。

8.16.2 频率学派和贝叶斯学派

第01课数学分析与概率论01:56:03 频率学派和贝叶斯学派

无高低好坏之分，只是认识自然的手段。只是在当前人们掌握的数学工具和需要解决的实践问题中，贝叶斯学派的理论体系往往能够比较好的解释目标函数、分析相互关系等。

课程的前半段的内容，大多是频率学派的思想；后半段的内容，使用贝叶斯学派的观点。

思考：大数据。频率学派对于贝叶斯学派一次强有力逆袭。

我要查一下“频率学派”和“贝叶斯学派”两者的区别。

8.16.3 贝叶斯公式

第01课数学分析与概率论01:57:26 贝叶斯公式

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{60}$$

给定某系统的若干样本x，计算该系统的参数，即：

$$P(\theta|x) = \frac{P(x|\theta)P(\theta)}{P(x)} \tag{61}$$

$P(\theta)$：没有数据支持下θ发生的概率：先验概率。

$P(\theta|x)$：在数据x的支持下，θ发生的概率：后验概率。

$P(x|\theta)$：给定某参数θ的概率分布：似然函数。

例如：（1）在没有任何信息的前提下，猜测某个姓氏：先猜“李、王、张、刘……”猜对的概率相对较大：先验概率；（2）若知道某人来自“牛家村”，则他姓“牛”的概率很大：后验概率——但并不排除他姓“郭、杨”等情况。

百度一下“先验概率、后验概率、似然函数”究竟是什么意思。

我目前的理解是这样子的，“先验概率”和“后验概率”的中的“先”是指“拿到样本之前，并不知道该样本的一些特殊值信息，根据经验计算它属于某一种情况的概率”；“后”是指“拿到样本之后，知道了样本的特征值信息，这个时候再根据经验计算它属于某一种情况的概率”。

Probability distribution

In probability theory and statistics, a probability distribution is a mathematical function that can be thought of as providing the probabilities of occurrence of different possible outcomes in an experiment. For instance, if the random variable X is used to denote the outcome of a coin toss (“the experiment”), then the probability distribution of X would take the value 0.5 for $X = heads$, and 0.5 for $X = tails$ (assuming the coin is fair). 从概率论和统计的角度来说（第一个视角），“概率分布”是一个mathematical function，用于给定different possible outcomes各自发生的概率。

In more technical terms, the probability distribution is a description of a random phenomenon in terms of the probabilities of events. Examples of random phenomena can include the results of an experiment or survey. A probability distribution is defined in terms of an underlying sample space,

which is the set of all possible outcomes of the random phenomenon being observed. The sample space may be the set of real numbers or a higher-dimensional vector space, or it may be a list of non-numerical values; for example, the sample space of a coin flip would be {heads, tails} . 从技术的角度来说（第二个视角），“概率分布”更多的描述一种随机的现象；由“概率分布”引出新的概率“sample space（样本空间）”。

Probability distributions are generally divided into two classes. A discrete probability distribution (applicable to the scenarios where the set of possible outcomes is discrete, such as a coin toss or a roll of dice骰子) can be encoded by a discrete list of the probabilities of the outcomes. On the other hand, a continuous probability distribution (applicable to the scenarios where the set of possible outcomes can take on values in a continuous range (e.g. real numbers), such as the temperature on a given day) is typically described by probability density functions (with the probability of any individual outcome actually being 0). The normal distribution is a commonly encountered continuous probability distribution. lsieun: （1）概率分布分成两类，即“离散分布”和“连续分布”；（2）“正态分布”是“连续分布”的举例。

In Bayesian statistical inference, a prior probability distribution, often simply called “the prior”, of an uncertain quantity is the probability distribution that would express one’s beliefs about this quantity before some evidence is taken into account. For example, the prior could be the probability distribution representing the relative proportions of voters who will vote for a particular politician in a future election. The unknown quantity may be a parameter of the model or a latent variable rather than an observable variable. lsieun: 先验概率，是在some evidence is taken into account之前，来表达人的相信程度。

In Bayesian statistics, the posterior probability of a random event or an uncertain proposition is the conditional probability that is assigned after the relevant evidence or background is taken into account. Similarly, the posterior probability distribution is the probability distribution of an unknown quantity, treated as a random variable, conditional on the evidence obtained from an experiment or survey. “Posterior”, in this context,

means after taking into account the relevant evidence related to the particular case being examined. lsieun: 后验概率，是在the relevant evidence or background is taken into account之后的conditional probability。

贝叶斯学派和频率学派的最大区别并不在于信息的利用和整合上。虽然贝叶斯方法可以用先验分布来引入以往的信息，但是频率学派也有方法来整合各种domain knowledge，比如在最优化likelihood的时候加入各种constrain。从这个意义上来说两者其实差别并不大。

频率学派和贝叶斯学派最大的差别其实产生于对**参数空间**的认知上。**所谓参数空间，就是你关心的那个参数可能的取值范围。频率学派**（其实就是当年的Fisher）并不关心参数空间的所有细节，他们相信数据都是在这个空间里的“某个参数值”下产生的（虽然你不知道那个值是啥），所以他们的方法论一开始就是从“哪个值最有可能是真实值”这个角度出发的。于是就有了最大似然（maximum likelihood）以及置信区间（confidence interval）这样的东西，你从名字就可以看出来他们关心的就是我有多大把握去圈出那个唯一的真实参数。而**贝叶斯学派**恰恰相反，他们关心参数空间里的每一个值，因为他们觉得我们又没有上帝视角，怎么可能知道哪个值是真的呢？所以**参数空间**里的每个值都有可能是真实模型使用的值，区别只是概率不同而已。于是他们才会引入先验分布（prior distribution）和后验分布（posterior distribution）这样的概念来设法找出参数空间上的每个值的概率。最好诠释这种差别的例子就是想象如果你的后验分布是双峰的，频率学派的方法会去选这两个峰当中较高的那一个对应的值作为他们的最好猜测，而贝叶斯学派则会同时报告这两个值，并给出对应的概率。

如果从概率的角度看，贝叶斯学派的想法其实更为自然，这也是为什么贝叶斯学派的产生远早于频率学派（去年是贝叶斯250周年）。但是贝叶斯方法本身有很多问题，比如当先验选的不好或者模型不好的时候你后验分布的具体形式可能都写不出来，跟别说做统计推断了。在当年电子计算机还没发展出来的时候，对这些情况做分析几乎是不可能的，这也就大大限制了贝叶斯方法的发展。而频率学派主要使用最优化的方法，在很多时候处理起来要方便很多。所以在频率学派产生后就快速地占领了整个统计领域。直到上世纪90年代依靠电子计算机的迅速发展，以及抽样算法的进步（Metropolis-hastings, Gibbs sampling）使得对于任何模型任何先验分布都可以有效地求出后验分布，贝叶斯学派才重新回到人们的视线当中。就现在而言，贝叶斯学派日益受到重视当然是有诸多原因的，所以这并不

意味这频率学派就不好或者不对。两个学派除了在**参数空间**的认知上有区别以外，方法论上都是互相借鉴也可以相互转化的。

作者：Xiangyu Wang 链接：https://www.zhihu.com/question/20587681/answer/41436978 来源：知乎著作权归作者所有。商业转载请联系作者获得授权，非商业转载请注明出处。

贝叶斯分析的思路对于由证据的积累来推测一个事物发生的概率具有重大作用，它告诉我们当我们要预测一个事物，我们需要的是首先根据已有的经验和知识推断一个先验概率，然后在新证据不断积累的情况下调整这个概率。整个通过积累证据来得到一个事件发生概率的过程我们称为贝叶斯分析。

8.17 分布

第01课数学分析与概率论01:58:44 分布

复习各种常见分布本身的统计量

在复习各种分布的同时，重温积分、Taylor展开式等前序知识

常见分布是可以完美统一为“一类分布”

8.17.1 两点分布

第01课数学分析与概率论01:58:51 两点分布

0-1分布

已知随机变量X的分布律为：当X取1时的概率为p，X取0时的概率为$1-p$，则有期望$E(X)=1\cdot p+0\cdot q=p$；方差（这里是方差吗？？？）$D(X)=E(X^2)-[E(X)]^2=1^2\cdot p+0^2\cdot(1-p)-p^2=p(1-p)=pq$。

8.17.2 二项分布Bernolli distribution

第01课数学分析与概率论01:58:59 二项分布Bernolli distribution

“两点分布”和“二项分布”的区别是，“两点分布”是做1次实验，而“二项分布”重复做n次实验。

设随机变量X服从参数为n,p二项分布，

（法一）设X_i为第i次试验中事件A发生的次数，$i=1,2,\ldots,n$，则$X=\sum_{i=1}^{n}X_i$。

显然，X_i相互独立均服从参数为p的0−−1分布，所以$E(X)=\sum_{i=1}^{n}E(X_i)=np$；$D(X)=\sum_{i=1}^{n}D(X_i)=np(1-p)$。

（法二）X的分布律为

第01课数学分析与概率论01:59:20

8.17.3 考察Taylor展开式

第01课数学分析与概率论01:59:32 考察Taylor展开式

泊松分布的概念。

8.17.4 泊松分布

第01课数学分析与概率论02:01:32 泊松分布

8.17.5 均匀分布

第01课数学分析与概率论02:02:06 均匀分布

“二项分布”和“泊松分布”是离散的；“均匀分布”是连续的

8.17.6 指数分布

第01课数学分析与概率论02:03:09 指数分布

8.17.7 正态分布

第01课数学分析与概率论02:05:09 正态分布

8.17.8 指数族

第01课数学分析与概率论02:06:39 指数族

9 数理统计与参数估计

“概率论”和“数理统计”有区别吗？它们不一样吗？

事实上，一般来讲，“概率论”和“数理统计”是两门课程，只不过，如果大家不是数学专业的话，一般是由“概率论”和“数理统计”合成一门课程，由一个老师，在一个学期讲完，所以大家会觉得是同一个东西，但本质是两者是不同的。在“数理统计”这个词中，更多的偏重于“统计”；

在机器学习当中的“统计”往往做的是“参数估计”，所以标题就是“数理统计与参数估计”。

9.1 主要内容

第02课数理统计与参数估计00:01:14 主要内容

统计量：（1）期望/方差/偏度/峰度；（2）协方差和相关系数；（3）独立和不相关

矩估计：中心矩和原点矩

最大似然估计：深刻理解

首先，“统计量”，大家应该是清楚的，只不过我们再说一遍这个事情。一般来说，大家对于“期望”和“方差”是比较理解的，而对于“偏度”和“峰度”就不那么熟悉；由“方差”这个概念，我们会提到“协方差”，而“协方差”要谈两件事情，第一件就是“独立”和“不相关”的区别和联系，第二件就是说一下“协方差”之间做一个“相关系数”，而这个“相关系数”我们后面还会用到。

关于“参数估计”，我们会说两个方案，第一个是“矩估计”，第二个是“最大似然估计”。“矩估计”比较简单，因此并不是重点，而“最大似然估计”才是重点，并且是我们以后的绝对重点。

我猜想，上面讲的“相关系数”和下面讲的“两个参数估计方案”应该是有联系的。

9.2 事件的独立性

第02课数理统计与参数估计00:02:36 事件的独立性

给定A和B是两个事件，若有$P(AB) = P(A)P(B)$，则称事件A和事件B相互独立。有的时候，我们称$P(AB)$为“联合概率”，而称$P(A)$、$P(B)$为“边缘概率”，以后遇到的“联合概率”和“边缘概率”概念的时候，要知道是什么意思。$P(AB) = P(A)P(B)$就可以表示成：**联合概率**，就等于各自**边缘概率**的乘积。

说明：（1）A和B独立，则$P(A|B) = P(A)$；（2）实践中，并不是根据$P(AB) = P(A)P(B)$这个公式来判断两个事件的独立性，而是往往根据两个事件是否相互影响而判断独立性，如果给定M个样本、若干次采样等情形，往往**假定**它们相互独立。换句话说，在实践中，我们往往先假定两个事件是独立的，然后再用$P(AB) = P(A)P(B)$这个公式。

思考：试给出A、B相互包含的信息量的定义$I(A,B)$，要求：如果A、B独立，则$I(A,B)=0$。

9.3 期望

第02课数理统计与参数估计00:09:20 期望

离散型，例如掷色子

$$E(X)=\sum_i x_i p_i \tag{62}$$

连续型

$$E(X)=\int_{-\infty}^{+\infty} xf(x)dx \tag{63}$$

即：概率加权下的“平均值”。

9.3.1 期望的性质

第02课数理统计与参数估计00:10:07 期望的性质

无条件成立

$$\begin{aligned} E(kX)&=kE(X)\\ E(X+Y)&=E(X)+E(Y)\end{aligned} \tag{64}$$

若X和Y相互独立（下面的式子，有点不太明白）

$$E(XY)=E(X)E(Y) \tag{65}$$

注意：（1）反之不成立。事实上，若$E(XY)=E(X)E(Y)$，只能说明X和Y“不相关”，并不能说X和Y相互独立。（2）关于“不相关”和“独立”的区别，稍后马上给出。

9.3.2 例1:计算期望

第02课数理统计与参数估计00:11:44 例1:计算期望

从$1,2,3,\ldots,99,2015$这100个数中任意选择若干个数（可能为0个数）求异或，试求异或的期望值。

这个我不懂哎。。。

9.3.3 计算每一位的期望

第02课数理统计与参数估计00:13:18 例1:计算期望

针对任何一个二进制位：取奇数个1异或后会得到1，取偶数个1异或后会得到0；与取0的个数无关。

给定最大数$2015 = (11111011111)_2$，共11位

针对每一位分别计算，考虑第i位X_i，假定给定的100个数中第i位一共有N个1，M个0，某次采样取到的1的个数为k。则有

$$P\{X_i = 1\} = \frac{2^m \cdot \sum\limits_{k\in odd} C_n^k}{2^{m+n}} = \frac{\sum\limits_{k\in odd} C_n^k}{2^n} = \frac{1}{2} \tag{66}$$

解决这个问题的思路，背后的思想是“事件加和的期望，等于各个事件期望的加和”$E(X+Y) = E(X) + E(Y)$。这样，当前题目本身听着很复杂，但使用这种思路后，就极大的简化了。

9.3.4 例2:集合Hash问题

第02课数理统计与参数估计00:14:20 例2:集合Hash问题

某Hash函数将任一字符串非均匀映射到正整数k，概率为2^{-k}，如下所示。现有字符串集合S，其元素经映射后，得到的最大整数为10，试估计S的元素个数。

9.4 方差

第02课数理统计与参数估计00:16:01 方差

定义$Var(X) = E\{[X - E(X)]^2\} = E(X^2) - E^2(X)$。（这个可以自己手动推导一下）一个随机变量X的方差，就是“变量平方的期望”减去“变量期望的平方”。

$E\{[X - E(X)]^2\} \geq 0 \Rightarrow E(X^2) \geq E^2(X)$，当X为定值时，取等号。

无条件成立

$$\begin{aligned} &Var(c) = 0 \\ &Var(X + c) = Var(X) \\ &Var(kX) = k^2 Var(X) \end{aligned} \tag{67}$$

X和Y独立

$$Var(X+Y) = Var(X) + Var(Y) \tag{68}$$

此外，方差的平方根，称为“标准差”。

我的理解“方差”，“方差”就表示：对随机变量X，先做“差”，再做“平方”。

9.5 协方差

第02课数理统计与参数估计00:19:08 协方差

刚刚的“方差”是针对1个随机变量计算方差，而“协方差”是对两个随机变量之间计算协方差。

定义：$Cov(X,Y) = E\{[X-E(X)][Y-E(Y)]\}$

性质：

$$\begin{aligned} Cov(X,Y) &= Cov(Y,X) \\ Cov(aX+b,cY+d) &= acCov(X,Y) \\ Cov(X_1+X_2,Y) &= Cov(X_1,Y) + Cov(X_2,Y) \\ Cov(X,Y) &= E(XY) - E(X)E(Y) \end{aligned} \tag{69}$$

9.5.1 协方差和独立、不相关

第02课数理统计与参数估计00:20:16 协方差和独立、不相关

X和Y独立时，$E(XY) = E(X)E(Y)$，而$Cov(X,Y) = E(XY) - E(X)E(Y)$，从而：当X和Y独立时，$Cov(X,Y) = 0$。

但X和Y独立这个前提太强，我们定义：若$Cov(X,Y) = 0$，称X和Y不相关。

9.5.2 协方差的意义

第02课数理统计与参数估计00:21:31 协方差的意义

协方差是两个随机变量具有**相同方向变化趋势**的度量：（1）若$Cov(X,Y) > 0$，它们的变化趋势相同；（2）若$Cov(X,Y) < 0$，它们的变化趋势相反；（3）若$Cov(X,Y) = 0$，称X和Y不相关。

思考：两个随机变量的协方差，是否有上界？

9.5.3 协方差的上界

第02课数理统计与参数估计00:24:46 协方差的上界

若X的方差$Var(X) = \sigma_1^2$，Y的方差$Var(Y) = \sigma_2^2$，则$|Cov(X,Y)| \leq \sigma_1\sigma_2$。当且仅当X和Y之间有线性关系时，等号成立。

也就是说，两个随机变量X和Y的协方差，能够被两个随机变量X和Y的方差限制住。

试分析该证明过程？

第02课数理统计与参数估计00:29:42 试分析该证明过程？

9.5.4 再谈独立与不相关

第02课数理统计与参数估计00:33:10 再谈独立与不相关

因为上述定理的保证，使得“不相关”事实上即“二阶独立”。

即：若X与Y不相关，说明X与Y之间没有线性关系（但有可能存在其他函数关系），不能保证X和Y相互独立。

但对于二维正态随机变量，X与Y不相关等价于X与Y相互独立。

9.5.5 Pearson相关系数

第02课数理统计与参数估计00:34:07 Pearson相关系数

定义：

$$\rho_{XY} = \frac{Cov(X,Y)}{\sqrt{Var(x)Var(Y)}} \tag{70}$$

我的理解，这个“相关系数”就是“协方差”与“各自标准差乘积”的商。

由协方差上界定理可知，$|\rho| \leq 1$

当且仅当X与Y有线性关系时，等号成立

容易看到，相关系数是标准尺度下的协方差（这里的“标准尺度”可能是这么理解的，根据上面的式子，它的分母是$\sqrt{Var(x)Var(Y)}$，从而使最终的$|\rho| \leq 1$）。上面关于协方差与XY相互关系的结论，完全适用于相关系数和XY的相互关系。

9.5.6 协方差矩阵

第02课数理统计与参数估计00:35:46 协方差矩阵

对于n个随机向量$(X_1, X_2, \dots X_n)$，任意两个元素X_i和X_j都可以得到一个协方差，从而形成$n \times n$的矩阵；协方差矩阵是对称阵。

$$c_{ij} = E\{[X_i - E(X_i)][X_j - E(X_j)]\} = Cov(X_i, X_j) \tag{71}$$

$$C = \begin{bmatrix} c_{11} & c_{12} & \cdots & c_{1n} \\ c_{21} & c_{22} & \cdots & c_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ c_{n1} & c_{n2} & \cdots & c_{nn} \end{bmatrix} \tag{72}$$

9.5.7 联想与思考

第02课数理统计与参数估计00:37:36 联想与思考

若X、Y独立，则$Var(XY) = Var(X)Var(Y)+Var(X)E^2(Y)+E(Y)E^2(X)$。

对称阵的不同特征值对应的特征向量，是否一定正交？对称阵和正交阵是否能够建立联系？

我现在不理解“正交阵”的概念。

9.6 矩

第02课数理统计与参数估计00:38:40 矩

对于随机变量X，X的k阶原点矩为$E(X^k)$。这里是X的k次幂求期望$E(X^k)$，把这个称之为**X的k阶原点矩**。当$k = 1$时，就是期望。

X的k阶中心矩为$E\{[X - E(X)]^k\}$。当$k = 2$时，就是方差；换句话说，2阶的中心矩就是方差。

这里是讲“期望”和“方差”向k阶的推广，其实没有新内容的。

9.7 统计参数的总结

第02课数理统计与参数估计00:40:09 统计参数的总结

期望（一阶原点矩）

方差（标准差，二阶中心矩）

变异系数（Coefficient of Variation）：标准差与均值（期望）的比值称为变异系数，记为$C \cdot V$。

偏度Skewness（三阶）
峰度Kurtosis（四阶）

9.8 偏度

第02课数理统计与参数估计00:40:43 偏度

偏度衡量随机变量概率分布的不对称性，是相对于**平均值**不对称程度的度量。（1）偏度的值可以为正，可以为负，或者无定义；（2）偏度为负（负偏）/正（正偏）表示在概率密度函数左侧/右侧的尾部比右侧的长，长尾在左侧/右侧；（3）偏度为零表示数值相对均匀地分布在平均值的两侧，但不一定意味着一定是对称分布。

9.8.1 偏度公式

第02课数理统计与参数估计00:42:10 偏度公式

三阶累积量与二阶累积量的1.5次方的比率。

听不懂了。

9.9 峰度

第02课数理统计与参数估计00:44:20 峰度

峰度是概率密度在均值处峰值高低的特征，通常定义四阶中心矩除以方差的平方减3。

看不懂了。

9.10 思考

1、给定两个随机变量X和Y，如何度量这两个随机变量的“距离”？

2、设随机变量X和期望为μ，方差为σ^2，对于任意正数ε，试估计概率$P\{|X-\mu|<\varepsilon\}$的下限。即：随机变量的变化值落在期望值附近的概率。

第2个问题与“切比雪夫不等式”和“大数定理”有关系。

解：第02课数理统计与参数估计01:00:20 解

9.10.1 切比雪夫不等式

第02课数理统计与参数估计01:02:10 切比雪夫不等式

设随机变量X的期望是μ，方差为σ^2，对于任意正数ϵ，有：

$$P\{|X-\mu| \geq \epsilon\} \leq \frac{\sigma^2}{\epsilon^2} \tag{73}$$

切比雪夫不等式说明，X的方差越小，事件$|X-\mu| \geq \epsilon$发生的概率越大。即：X取的值基本上集中在期望μ附近。（1）该不等式进一步说明了方差的含义；（2）该不等式可证明大数定理。

9.10.2 大数定理

第02课数理统计与参数估计01:03:34 大数定理

设随机变量$X_1, X_2, \ldots, X_n \ldots$互相独立，并且具有相同的期望$\mu$和方差$\sigma^2$。作为前$n$个随机变量的平均$Y_n = \frac{1}{n}\sum_{i=1}^{n} X_i$，则对任意正数$\epsilon$，有

$$\lim_{n\to\infty} P\{|Y_n - \mu| < \varepsilon\} = 1 \tag{74}$$

9.10.3 大数定理的意义

第02课数理统计与参数估计01:05:18 大数定理的意义

当n很大时，随机变量$X_1, X_2, \ldots, X_n$的平均值Y_n在概率意义下无限接近期望μ。出现偏离是可能的，但这种可能性很小，当n无限大时，这种可能性的概率为0。

9.10.4 重要推论

第02课数理统计与参数估计01:06:02 重要推论

一次试验中事件A发生的概率为p；重复n次独立实验中，事件A发生了n_A次，则p、n、n_A的关系满足：

对于任意正数ϵ，

$$\lim_{n\to\infty} P\left\{|\frac{n_A}{n} - p| < \epsilon\right\} = 1 \tag{75}$$

换句话说，当n趋向于无穷时，频率$\frac{n_A}{n}$会无限趋近于概率p。

9.10.5 伯努利定理

第02课数理统计与参数估计01:09:02 伯努利定理

上述推论是最早的大数定理的形式，称为伯努利定理。该定理表明事件A发生的频率$\frac{n_A}{n}$以概率收敛于事件A的概率p，以严格的数学形式表达了频率的稳定性。

上述事实为我们在实际应用中用频率来估计概率提供了一个理论依据。（1）正态分布的参数估计；（2）朴素贝叶斯做垃圾邮件分类；（3）隐马尔可夫模型有监督参数学习。

9.10.6 中心极限定理

第02课数理统计与参数估计01:09:22 中心极限定理

Central Limit Theorem

设随机变量$X_1, X_2, \ldots, X_n$互相独立，服从同一分布，并且具有相同的期望μ和方差σ^2，则随机变量

$$Y_n = \frac{\sum_{i=1}^{n} X_i - n\mu}{\sqrt{n}\sigma} \tag{76}$$

的分布收敛到标准正态分布。

容易得到：$\sum_{i=1}^{n} X_i$收敛到正态分布$N(n\mu, n\sigma^2)$。

9.10.7 中心极限定理的意义

第02课数理统计与参数估计01:10:48 中心极限定理的意义

实际问题中，很多随机现象可以看做许多因素的独立影响的综合反应，往往近似服从正态分布。

城市耗电量：大量用户的耗电总和。

测量误差：许多观察不到的、微小误差的总和。注意：是多个随机变量的“和”才可以，有些问题是乘性误差，则需要鉴别或者取对数后再使用。

线性回归中，将使用该定理论证最小二乘法的合理性。

9.10.8 例：标准的中心极限定理的问题

第02课数理统计与参数估计01:10:54 例：标准的中心极限定理的问题

有一批样本（字符串），其中a-z开头的比例是固定的，但是量很大，需要从中随机抽样。样本量n，总体中a开头的字符串占比1%，需要每次投到的a开头的字符串占比（0.99%,+1.01%），样本量n至少是多少？

问题可以重新表达一下：大量存在的两点分布$B_i(1,p)$，其中，B_i发生的概率为0.01，即$p=0.01$。取其中的n个，使得发生的个数除以总数的比例落在区间(0.009,0.0101)，则n至少是多少？

9.11 样本的统计量

第02课数理统计与参数估计01:12:17 样本的统计量

事实上，要拿到随机变量的“全体数据”是做不到的，而只能取其中“一些样本”来进行统计。

设随机变量X的N个样本为$X_1,X_2,\ldots,X_n$，则样本均值为：

$$\bar{X}=\frac{1}{n}\sum_{i=1}^{n}X_i \tag{77}$$

样本方差为：

$$S^2=\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2 \tag{78}$$

样本方差的分母使用$n-1$而非n，是为了无偏。思考：如何证明？

9.12 样本的矩

第02课数理统计与参数估计01:14:53 样本的矩

k阶样本原点矩：

$$A_k=\frac{1}{n}\sum_{i=1}^{n}X_i^k \tag{79}$$

k阶样本中心矩

$$M_k=\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^k \tag{80}$$

9.13 思考

第02课数理统计与参数估计01:15:53 思考

随机变量的矩和样本的矩，有什么关系？

换个想法：

（1）假设总体服从某参数θ（存在且未知，有可能是值或者向量）的分布，从总体中抽出一组样本$X_1, X_2, \ldots, X_n$，如何估计参数θ？（这绝对是最大的重点）

（2）样本是独立同分布的

（3）可以通过$X_1, X_2, \ldots, X_n$方便的计算样本的k阶矩

（4）假设样本的k阶矩等于总体的k阶矩，可估计出总体的参数。

这段讲的很精彩

9.14 矩估计

第02课数理统计与参数估计01:19:53 矩估计

设总体的均值为μ，方差σ^2，（μ和σ未知，待求）则有原点距离表达式：

$$\begin{aligned} E(X) &= \mu \\ E(X^2) &= Var(X) + [E(X)]^2 = \sigma^2 + \mu^2 \end{aligned} \tag{81}$$

根据该总体的一组样本，求得原点距：

$$\begin{aligned} A_1 &= \frac{1}{n}\sum_{i=1}^{n} X_i \\ A_2 &= \frac{1}{n}\sum_{i=1}^{n} X_i^2 \end{aligned} \tag{82}$$

9.14.1 矩估计的结论

第02课数理统计与参数估计01:21:53 矩估计的结论

根据各自阶的中心矩相等，计算得到：

$$\begin{aligned} \mu &= \bar{X} \\ \sigma^2 &= \frac{1}{n}\sum_{i=1}^{n} X_i - \bar{X}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 \end{aligned} \tag{83}$$

由于是根据样本求得的估计结果，根据记号习惯，写作：

$$\begin{aligned} \hat{\mu} &= \bar{X} \\ \hat{\sigma}^2 &= \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 \end{aligned} \tag{84}$$

注：“估计”的值往往加一个“hat”

9.14.2 例：正态分布的矩估计

第02课数理统计与参数估计01:23:34 例：正态分布的矩估计

在正态分布的总体中采样得到n个样本：$X_1, X_2, \ldots, X_n$，估计该总体的均值和方差。

解：直接使用矩估计的结论：

$$\begin{aligned} \hat{\mu} &= \bar{X} \\ \hat{\sigma}^2 &= \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 \end{aligned} \tag{85}$$

9.14.3 例：均匀分布的矩估计

第02课数理统计与参数估计01:24:58 例：均匀分布的矩估计

设$X_1, X_2, \ldots, X_n$为定义在$[a, b]$上的均匀分布的总体采样得到的样本，求a, b。

解：

均匀分布的均值和方差为：

$$\begin{aligned} E(X) &= \frac{a+b}{2} \\ Var(X) &= \frac{(b-a)^2}{12} \end{aligned} \tag{86}$$

矩估计要求满足：

$$\begin{aligned} \hat{\mu} &= \bar{X} \\ \hat{\sigma}^2 &= \frac{1}{n}\sum_{i=1}^{n}(X_i = \bar{X})^2 \end{aligned} \tag{87}$$

从而：

这里省略了。。。

9.15 贝叶斯公式带来的思考

第02课数理统计与参数估计01:26:28 贝叶斯公式带来的思考

这里与“最大似然估计”做了关联

这个我见到过，暂不记了

9.16 最大似然估计

第02课数理统计与参数估计01:36:03 最大似然估计

设总体分布为$f(x,\theta)$，$X_1,X_2,\ldots,X_n$为总体采样得到的样本。因为$X_1,X_2,\ldots,X_n$独立同分布，于是，它们的联合密度函数为：

查一下“独立同分布”的意思

$$L(x_1,x_2,\ldots,x_n;\theta_1,\theta_2,\ldots,\theta_k)=\prod_{i=1}^{n}f(x_i;\theta_1,\theta_2,\ldots,\theta_k); \tag{88}$$

这里，θ被看做固定但未知的参数；反过来，因为样本已存在，可以看成$x_1,x_2,\ldots,x_n$是固定的，$L(x,\theta)$是关于θ的函数，即似然函数。

求参数θ的值，使得似然函数取最大值，这种方法就是最大似然估计。

这里讲述了“what”，即“到底是什么”；后面讲如何做，如何做是技术手段问题。

9.17 最大似然估计的具体实践操作

第02课数理统计与参数估计01:42:55 最大似然估计的具体实践操作

在实践中，由于求导数的需要，往往将似然函数取对数，得到对数似然函数；若对数似然函数可导，可通过求导的方式，解下列方程组，得到驻点，然后分析该驻点是极大值点。

$$\log L(\theta_1,\theta_2,\ldots,\theta_k)=\sum_{i=1}^{n}\log f(x_i;\theta_1,\theta_2,\ldots,\theta_k) \tag{89}$$

$\frac{\partial L(\theta)}{\partial\theta}=0 \quad i=1,2,\ldots,k$

9.17.1 举例：抛硬币最大似然估计

第02课数理统计与参数估计01:48:28 举例：抛硬币最大似然估计

找出与样本的分布最接近的概率分布模型

简单的例子，10次抛硬币的结果是：正正反正下正反反正正

假设p是每次抛硬币结果为正的概率，则得到这样的实验结果的概率是：

$$P = pp(1-p)ppp(1-p)(1-p)pp = p^7(1-p)^3 \tag{90}$$

最优解是：$p = 0.7$

9.17.2 二项分布的最大似然值估计

第02课数理统计与参数估计01:48:57 二项分布的最大似然值估计

投硬币试验中，进行N次独立试验，n次朝上，$N-n$次朝下。

假定朝上的概率为p，使用对数似然函数作为目标函数：

$$\begin{aligned} h(p) = \log f(n|p) &= \log(p^n(1-p)^{N-n}) \\ \frac{\partial h(p)}{\partial p} &= \frac{n}{p} - \frac{N-n}{1-p} = 0 \\ &\Rightarrow p = \frac{n}{N} \end{aligned} \tag{91}$$

求导过程

$$\begin{aligned} h(p) = \log f(n|p) &= \log(p^n(1-p)^{N-n}) \\ &= \log p^n + \log(1-p)^{N-n} \\ &= n \cdot \log p + (N-n) \cdot \log(1-p) \end{aligned} \tag{92}$$

$$\begin{aligned} \frac{\partial h(p)}{\partial p} &= \frac{n}{p} - \frac{N-n}{1-p} = 0 \\ &\Rightarrow p = \frac{n}{N} \end{aligned}$$

到这个时候，我们发现，通过“最大似然估计”这个估计得到的概率p就是频率$\frac{n}{N}$。

9.17.3 正态分布的最大似然值估计

第02课数理统计与参数估计01:51:25 正态分布的最大似然值估计

若给定一组样本$X_1, X_2, \ldots, X_n$，已知它们来自于高斯分布$N(\mu, \sigma)$，试估计参数μ, σ。

按照MLE的过程分析

高斯分布的概率密度函数：

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{93}$$

将X_i的样本值带入xi带入（注意：X_i是“样本”，而xi是“样本值”。），得到：

$$L(X) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} \tag{94}$$

化简对数似然函数：

$$\begin{aligned} l(x) &= \log \prod_i \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} \\ &= \sum_i \log \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} \\ &= \left(\sum_i \log \frac{1}{\sqrt{2\pi}\sigma} \right) + \left(\sum_i -\frac{(x_i-\mu)^2}{2\sigma^2} \right) \\ &= \sum_i \log(2\pi\sigma^2)^{-\frac{1}{2}} - \sum_i \frac{(x_i-\mu)^2}{2\sigma^2} \\ &= -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_i (x_i-\mu)^2 \end{aligned} \tag{95}$$

9.17.4 参数估计的结论

第02课数理统计与参数估计01:56:10 参数估计的结论

目标函数

$$l(x) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_i (x_i-\mu)^2 \tag{96}$$

将目标函数对参数μ, σ分别求偏导，很容易得到μ, σ的式子：

$$\begin{aligned}\mu &= \frac{1}{n}\sum_i x_i \\ \sigma^2 &= \frac{1}{n}\sum_i (x_i-\mu)^2\end{aligned} \tag{97}$$

9.17.5 符合直观想像

第02课数理统计与参数估计01:56:32 符合直观想像

$$\begin{aligned}\mu &= \frac{1}{n}\sum_i x_i \\ \sigma^2 &= \frac{1}{n}\sum_i (x_i-\mu)^2\end{aligned} \tag{98}$$

上述结论和“矩估计”的结果是一致的，并且意义非常直观；“样本的均值”即高斯分布的均值，“样本的伪方差”即高斯分布的方差。注：经典意义下的方差，分母是$n-1$；在似然估计的方法中，求的方差的分母是n。

该结论将在EM（期望最大化算法）、GMM高斯混合模型中将继续使用。

9.17.6 最大似然估计与过拟合

第02课数理统计与参数估计01:57:09 最大似然估计与过拟合

10 矩阵和线性代数

第03课矩阵和线性代数00:01:40

我们来破个题，第一个是“数”，但我们谈的不是“数”，而是“代数”，是用“某些符号”来替代真正的“数”来做的运算。我们主要的方案是“线性”的，而“线性代数”里面最重要的一个算子就是“矩阵”，尤其是“矩阵的乘法”，所以我们题目叫作“矩阵和线性代数”。

10.1 主要内容

第03课矩阵和线性代数00:02:26 主要内容

矩阵：（1）线性代数是有用的，以SVD为例；（2）矩阵的乘法/状态转移矩阵；（3）矩阵和向量组

特征值和特征微量：（1）对称阵、正交阵、正定阵；（2）数据白化

矩阵求导：（1）向量对向量求导；（2）标量对向量求导；（3）标量对矩阵求导。

大多数人都了解过线性代数，但是我想说的第一个话题就是，线性代数其实是非常有用的，而不是像我们在上学时候学的、看起来很无聊的一种变换，我们举一个SVD应用的例子来说明线性代数的有用性。然后我们重点探索一件事情“矩阵的乘法”，如果矩阵A是$m \times n$阶的$A_{m\times n}$，矩阵B是$n \times r$阶的$B_{n\times r}$，A的列数和B的行数是一样的，则这两个矩阵是可乘的。两个矩阵乘法如何计算，大家是知道的；我们要说的是，矩阵的乘法为什么要这么计算呢？大家不觉得第一次看到矩阵乘法的时候很奇怪吗？我们举一个例子来看一下有意思的东西：状态转移矩阵。另外，我们说说“向量组、矩阵、特征值、特征向量”，说说各种变换（对称阵、正交阵、正定阵）、数据的白化、矩阵求导等事情。

10.2 SVD的提法（奇异值分解）

第03课矩阵和线性代数00:03:58 SVD的提法

此处主要是说明“线性代数是有用的”，举了一个图片的例子，很有趣，但我现在不是特别懂，只听懂了一个大概意思。将推荐系统的时候，还会再说SVD。

奇异值分解（Singular Value Decomposition）是一种重要的矩阵分解方法，可以看做对称方阵在任意矩阵上的推广。Singular并不是指“奇怪的”，而是指“突出的、奇特的、非凡的”，似乎更应该称之为“优值分解”。

假设A是一个$m \times n$阶实矩阵，则存在一个分解使得：

$$A_{m\times n} = U_{m\times m} \sum_{m\times n} V^T_{n\times n} \tag{99}$$

其中，$U_{m\times m}$和$V^T_{n\times n}$是正交的，而$\sum_{m\times n}$称为奇异值，是一个对角阵，即只有对角线上有值。通常将奇异值由大而排列。这样，$\sum$便能由A唯一确定了。（第03课矩阵和线性代数00:06:32 这里讲的很精彩）

什么是正交阵？如果U是一个$m \times m$的方阵，并且$U^TU = I$(U的转置乘以U正好是个单位阵)，则U称为“正交阵”。单位阵的意思是，只有对角线上的元素是1，其它的元素都是0。

什么是实矩阵？实矩阵指的是矩阵中所有的数都是实数的矩阵。如果一个矩阵中含有除实数以外的数，那么这个矩阵就不是实矩阵。

与特征值、特征向量的概念相对应：

- $\sum$对角线上的元素σ称为矩阵A的奇异值
- U的第i列称为A的关于σ_i的左奇异向量
- V的第i列称为A的关于σ_i的右奇异向量

10.2.1 SVD举例$A_{m\times n}=U_{m\times m}\sum_{m\times n}V^T_{n\times n}$

第03课矩阵和线性代数00:07:15 SVD举例

已知4×5阶实矩阵A，求A的SVD分解：

$$A_{4\times 5}=\begin{bmatrix}1&0&0&0&2\\0&0&3&0&0\\0&0&0&0&0\\0&4&0&0&0\end{bmatrix} \tag{100}$$

分解为：

$$U_{4\times 4}=\begin{bmatrix}0&0&1&0\\0&1&0&0\\0&0&0&1\\1&0&0&0\end{bmatrix}\quad \sum_{4\times 5}=\begin{bmatrix}4&0&0&0&0\\0&3&0&0&0\\0&0&\sqrt{5}&0&0\\0&0&0&0&0\end{bmatrix}\quad V^T_{5\times 5}=\begin{bmatrix}0&1&0&0&0\\0&0&1&0&0\\\sqrt{0.2}&0&0&0&\sqrt{0.8}\\0&0&0&1&0\\\sqrt{0.8}&0&0&0&-\sqrt{0.2}\end{bmatrix}$$

矩阵U和V都单位正交方阵：$U^TU=I, V^TV=I$

10.2.2 奇异值分解-效果

第03课矩阵和线性代数00:07:51 奇异值分解-效果

可以做图像处理。这里讲图像处理的过程，很有趣。

$\sum$是个对角阵，里面的元素σ是从大到小排列的$(\sigma_1>\sigma_2>\dots)$，这时如果取前$k$个$\sigma$值，就相当于做了特征的提取。

一般处理的灰度图，因为有RGB三个通道，分别处理3个通道，然后再进行合成彩色就好了。讲算法的时候，一般就用灰度来做例子。

这一部分的目的，就是要讲：线性代数是有用的。

10.3 线性代数

第03课矩阵和线性代数00:16:58 线性代数

10.3.1 方阵的行列式

第03课矩阵和线性代数00:16:58 方阵的行列式

定义：方阵的行列式

- 1阶方阵的行列式为该元素本身
- n阶方阵的行列式等于它的任一行（或列）的各元素与其对应的代数余子式乘积之和。

什么是行列式?（百度一下）

什么是代数余子式?（百度一下）

1×1的方阵，其行列式等于该元素本身。

$A = \begin{pmatrix} a_{11} \end{pmatrix} \quad |A| = a_{11}$

2×2的方阵，其行列式用“主对角线元素乘积”减去“次对角线元素的乘积”。

$A = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \quad |A| = a_{11}a_{22} - a_{12}a_{21}$

3×3的方阵，

$$A = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix} \tag{101}$$

根据“主对角线元素乘积减去次对角线元素的乘积”的原则，得：

$$\begin{aligned} |A| = a_{11}a_{12}a_{13} + a_{12}a_{23}a_{31} + a_{13}a_{21}a_{32} \\ -a_{11}a_{23}a_{32} - a_{12}a_{21}a_{33} - a_{13}a_{22}a_{31} \end{aligned} \tag{102}$$

10.3.2 代数余子式

第03课矩阵和线性代数00:18:53 代数余子式

在一个n阶行列式A中，把(i,j)元素a_{ij}所在的第i行和第j列划去后，留下的$n-1$阶方阵的行列式叫做元素a_{ij}的余子式，记作M_{ij}。

代数余子式：$A_{ij} = (-1)^{i+j} M_{ij}$

$$A = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix} \tag{103}$$

$\forall 1 \le j \le n, |A| = \sum_{i=1}^{n} a_{ij} \cdot (-1)^{i+j} M_{ij}$
$\forall 1 \le i \le n, |A| = \sum_{j=1}^{n} a_{ij} \cdot (-1)^{i+j} M_{ij}$

$$\begin{aligned} |A| = a_{11}a_{12}a_{13} + a_{12}a_{23}a_{31} + a_{13}a_{21}a_{32} \\ -a_{11}a_{23}a_{32} - a_{12}a_{21}a_{33} - a_{13}a_{22}a_{31} \end{aligned} \tag{104}$$

10.3.3 伴随矩阵

第03课矩阵和线性代数00:19:13 伴随矩阵

对于$n \times n$方阵的任意元素a_{ij}都有各自的代数余子式$A_{ij} = (-1)^{i+j} M_{ij}$，构造$n \times n$的方阵$A^*$：

$$A^* = \begin{pmatrix} A_{11} & A_{21} & \cdots & A_{n1} \\ A_{12} & A_{22} & \cdots & A_{n2} \\ \vdots & \vdots & \ddots & \vdots \\ A_{1n} & A_{2n} & \cdots & A_{nn} \end{pmatrix} \tag{105}$$

A^*称为A的伴随矩阵。注意：A_{ij}位于A^*的第j行第i列。

10.3.4 方阵的逆$A \cdot A^* = |A| \cdot I$

第03课矩阵和线性代数00:20:53 方阵的逆

$A \cdot A^* = |A| \cdot I$，我们发现，任何一个方阵$A$和它的伴随矩阵$A^*$的乘积$A \cdot A^*$，正好是行列式$|A|$的若干倍。

由前述结论：$\forall 1 \le i \le n, |A| = \sum_{j=1}^{n} a_{ij} \cdot (-1)^{i+j} M_{ij}$

根据：

$$A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{pmatrix}$$

$$A^* = \begin{pmatrix} A_{11} & A_{21} & \cdots & A_{n1} \\ A_{12} & A_{22} & \cdots & A_{n2} \\ \vdots & \vdots & \ddots & \vdots \\ A_{1n} & A_{2n} & \cdots & A_{nn} \end{pmatrix}$$

计算：

$$A \cdot A^* = \begin{pmatrix} |A| & 0 & \cdots & 0 \\ 0 & |A| & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & |A| \end{pmatrix} = |A| \cdot I \Rightarrow A^{-1} = \frac{1}{|A|} A^*$$

思考：该等式有什么用？在后续求偏导的时候会用到。

10.3.5 范德蒙行列式Vandermonde

第03课矩阵和线性代数00:22:03 范德蒙行列式Vandermonde

证明范德蒙行列式Vandermonde：

$$D_n = \begin{vmatrix} 1 & 1 & 1 & \cdots & 1 \\ x_1 & x_2 & x_3 & \cdots & x_n \\ x_1^2 & x_2^2 & x_3^2 & \cdots & x_n^2 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_1^{n-1} & x_2^{n-1} & x_3^{n-1} & \cdots & x_n^{n-1} \end{vmatrix} = \prod_{i,j(n \geq i \geq j \geq 1)} (x_i - x_j)$$

在上面的行列式D_n中，我们保证$x_1 \neq x_2 \neq x_3 \ldots$，即$x_i$互不相等。

提示：数学归纳法

注：参考Lagrange/Newton插值法

这部分讲的很精彩。

10.4 矩阵乘法

第03课矩阵和线性代数00:29:00 矩阵乘法

A为$m \times s$阶的矩阵，B为$s \times n$阶的矩阵，那么，$C = A \times B$是$m \times n$阶的矩阵，其中，

$$c_{ij} = \sum_{k=1}^{s} a_{ik} b_{kj} \tag{106}$$

10.4.1 矩阵模型

第03课矩阵和线性代数00:30:31 矩阵模型

考虑某随机过程π，它的状态有n个，用1 n表示。记在当前时刻t时位于i状态，它在$t+1$时刻位于j状态的概率为$P(i,j)=P(j|i)$：即“状态转移的概率”只依赖于“前一个状态”。

第03课矩阵和线性代数00:36:50 概率转移矩阵

第$n+1$代中处于第j个阶层的概率为：

$\pi(X_{n+1}=j)=\sum_{i=1}^{K}\pi(X_n=i)\cdot P(X_{n+1}=j|X_n=i)$

$\Rightarrow \pi^{(n+1)}=\pi^{(n)}\cdot P$

因此，矩阵P即为（条件）概率转移矩阵。第i行元素表示：在上一个状态为i时的分布概率，即每一行元素的和为1。

思考：初始概率分布π对最终分布的影响？

第03课矩阵和线性代数00:44:36 平稳分布

初始概率不同，但经过若干次迭代，π最终稳定收敛在某个分布上。

从而，这是转移概率矩阵P的性质，而非初始分布的性质。事实上，上述矩阵P的n次幂，每行都是$(0.286,0.489,0.225), n>20$。

如果一个非周期马尔科夫随机过程具有转移概率矩阵P，且它的任意两个状态都是连通的，则$\lim\limits_{n\to\infty}P_{ij}^{n}$存在，记做$\lim\limits_{n\to\infty}P_{ij}^{n}=\pi(j)$。

10.5 矩阵和向量的乘法

第03课矩阵和线性代数00:51:56 矩阵和向量的乘法

10.6 矩阵的秩

第03课矩阵和线性代数00:57:33 矩阵的秩

如果把“数学”看成是“机器学习”的工具，那么“线性代数”是数学的“工具”。

10.7 向量组等价

第03课矩阵和线性代数01:00:51 向量组等价

10.8 系数矩阵

第03课矩阵和线性代数01:06:38 系数矩阵

10.9 正交阵

第03课矩阵和线性代数01:09:52 正交阵

10.10 特征值和特征向量

第03课矩阵和线性代数01:15:18 特征值和特征向量

10.11 特征值的性质

第03课矩阵和线性代数01:27:32 特征值的性质

11 马同学线性代数

11.1 向量

11.1.1 物理中的向量

将有向线段的起点与终点分别表示为字母A, B，则向量表示为$\overrightarrow{AB}$。

为了书写方便，向量也可以用一个字母来代替，比如$\overrightarrow{AB}$也可以用$\vec{u}$来表示。

既然向量是具有方向和大小的几何对象。那么只要大小相等，方向相同，向量自然也就相等

11.1.2 数学中的向量

物理是物理，一向讲究差不多就行，甚至认为“近似”是物理的精华。

但是数学不能这么干，说向量就是一个具有方向和大小的几何对象，这事情数学干不出来。

“严格性虽然不是数学的一切，但是没有了严格性数学就没有了一切”。

从向量的物理概念出发，是一个具有方向和大小的几何对象。

比如，这么一个向量$\vec{u} = \overrightarrow{OP}$，我们把它的起点放在原点$O$点，终点放在$P$点，就可以画出这个向量。

在数学中，我们始终遵循向量的起点在原点O，那么我们就可以用终点的坐标来表示向量，即上面的向量可以表示为：$\vec{u} = (x, y)$。这样，向量和空间中的点就建立了一一映射的关系。往更高维度走也是一样的，比如三维空间中的一个点就对应了一个三维向量。

这个时候，我们就需要向量的形式化定义了，如下：

n个有序的数$a_1, a_2, \ldots, a_n$所组成的数组称为n维向量，这n个数称为该向量的n个分量，第i个数a_i称为第i个分量。n维向量可以写成一行，也可以写成一列，分别称为“行向量”和“列向量”。

n维列向量

$$\begin{pmatrix} a_1 \\ a_2 \\ \cdots \\ a_n \end{pmatrix} \tag{107}$$

n维行向量

$$\left(a_1, a_2, \cdots, a_n\right) \tag{108}$$

上面两种写法都代表同一个向量，这两种写法到后面矩阵出现了才有区别。

长度

向量是有向线段，它的大小就是线段的长度。

对于向量$\vec{u}=\overrightarrow{OP}$的长度被记做$||\vec{u}||$或者$||\overrightarrow{OP}||$。

在二维中，对于向量$\vec{u}=\begin{pmatrix} x \\ y \end{pmatrix}$而言，这个长度可以根据毕达哥拉斯定理计算，等于$\sqrt{x^2+y^2}$。扩展到三维就是$||\vec{v}||=\sqrt{x^2+y^2+z^2}$。

更一般的，对于：

$$\vec{w}=\begin{pmatrix} a_1 \\ a_2 \\ \cdots \\ a_n \end{pmatrix} \tag{109}$$

的长度为：

$$||\vec{w}||=\sqrt{a_1^2+a_2^2+\cdots+a_n^2} \tag{110}$$

注意：向量长度是一个数，不是一个向量或点。

对于向量$\overrightarrow{AB}$来讲，如果交换起点和终点的顺序，就会得到另一个向量$\overrightarrow{BA}$。它们的长度相等，方向相反，即：$||\overrightarrow{AB}||=||\overrightarrow{BA}||$。

零向量

$\overrightarrow{AA}$也是一个向量，它被称为零向量。

起点与终点为同一个点的向量为零向量,被记做$\vec{0}$。零向量的长度为零。

平行的向量

向量不仅有长度还有方向，不过我们这里暂时没有办法给出方向的数学定义，这得等后面介绍了点积才能严格定义。

两个具有相同或相反方向的向量平行

零向量与任意向量平行。

11.1.3 总结

向量是有序对
向量大小的标记方法、计算方法（长度）
平行的向量

11.2 向量基本操作（数乘和加法）

11.2.1 加法

三角法则

物理中，我们学过的力的合成其实就是向量的加法。

代数定义

对于：

$$\vec{a} = \begin{pmatrix} a_1 \\ a_2 \\ \cdots \\ a_n \end{pmatrix} \qquad \vec{b} = \begin{pmatrix} b_1 \\ b_2 \\ \cdots \\ b_n \end{pmatrix}$$

它的加法定义为：

$$\vec{a} + \vec{b} = \begin{pmatrix} a_1 \\ a_2 \\ \cdots \\ a_n \end{pmatrix} + \begin{pmatrix} b_1 \\ b_2 \\ \cdots \\ b_n \end{pmatrix} = \begin{pmatrix} a_1 + b_1 \\ a_2 + b_2 \\ \cdots \\ a_n + b_n \end{pmatrix}$$

当然也可以看作是行向量：

$$\vec{a}+\vec{b} = \Big(a_1, a_2, \cdots, a_n\Big)+\Big(b_1, b_2, \cdots, b_n\Big) = \Big(a_1 + b_1, a_2 + b_2, \cdots, a_n + b_n\Big)$$

推论

根据三角形原则，我们可以得到一个推论：$||\vec{u}|| + ||\vec{v}|| \geq ||\vec{u} + \vec{v}||$，其中$\vec{u}, \vec{v}$共线的时候取等号。

11.2.2 数乘

向量的基本操作除了加法，还有数乘。

几何意义

数乘$k\vec{u}, k \in R$，就是对$\vec{u}$进行缩放。

$|k|$为缩放比例，$k < 0$时，$k\vec{u}$与$\vec{u}$方向相反；$k = 0$时，$k\vec{u}$为零向量。

代数

对于

$$\vec{a} = \begin{pmatrix} a_1 \\ a_2 \\ \cdots \\ a_n \end{pmatrix}$$

数乘的代数表达为：

$$k\vec{a} = k \begin{pmatrix} a_1 \\ a_2 \\ \cdots \\ a_n \end{pmatrix} = \begin{pmatrix} ka_1 \\ ka_2 \\ \cdots \\ ka_n \end{pmatrix}, k \in R$$

当然也可以写成行向量：

$$k\vec{a} = k\left(a_1, a_2, \cdots, a_n\right) = \left(ka_1, ka_2, \cdots, ka_n\right), k \in R$$

11.2.3 运算规则

我们不难看出，向量**加法**满足交换律，结合律：

交换律：$\vec{v} + \vec{u} = \vec{u} + \vec{v}$

结合律：$\vec{u} + \vec{v} + \vec{w} = \vec{u} + (\vec{v} + \vec{w})$

数乘满足交换律，结合律和分配率的：

交换律：$k \cdot \vec{u} = \vec{u} \cdot k$

结合律：$k \cdot m \cdot \vec{u} = k \cdot (m \cdot \vec{u})$

分配律：$k(\vec{u}+\vec{v})=k\vec{u}+k\vec{v}$

11.3 线性表示与线性相关

11.3.1 线性表示

调色

人眼大致有三种感光细胞：红色、绿色、蓝色的感光细胞。如果通过特定的光线，单独“激活”这三种感光细胞，我们分别看到红色、绿色、蓝色，这些光线也就是红光、绿光、蓝光：这三种颜色的光线同时作用这三种感光细胞上，混合在一起，就得到了我们所看到的颜色。因此，这三种颜色我们也称为“三原色”，简写为RGB（红色、绿色、蓝色英文首字母的缩写）。

白色的太阳光，就是按照$R:G:B=1:1:1$的比例合成的；如果调整比例，我们可以得到海棠红$\approx R+\frac{1}{5}G+\frac{1}{3}B$。

线性表示or线性组合

上面啰嗦了那么多，我们终于可以开始数学化，首先把RGB表示为向量：

$$R:\begin{pmatrix}255\\0\\0\end{pmatrix}\qquad G:\begin{pmatrix}0\\255\\0\end{pmatrix}\qquad B:\begin{pmatrix}0\\0\\255\end{pmatrix}$$

这样之前的海棠红，就可以写成这样了：

$$\begin{pmatrix}255\\52\\85\end{pmatrix}\approx\begin{pmatrix}255\\0\\0\end{pmatrix}+\frac{1}{5}\begin{pmatrix}0\\255\\0\end{pmatrix}+\frac{1}{3}\begin{pmatrix}0\\0\\255\end{pmatrix}\tag{111}$$

为了严格进行数学定义，我们先给出向量组的定义：

若干同维数的列向量（或者同维数的行向量）所组成的集合，叫做**向量组**。比如，之前的RGB放在一个集合里，$\{R,G,B\}$就是一个向量组。

然后定义“线性表示”或“线性组合”：

给定向量组$A:\vec{a_1},\vec{a_2},\ldots,\vec{a_m}$和向量$\vec{b}$，如果存在一组实数$k_1,k_2,\ldots,k_m$，使：

$$\vec{b} = k_1\vec{a_1} + k_2\vec{a_2} + \cdots + k_m\vec{a_m} \tag{112}$$

则称向量$\vec{b}$能由向量组A线性表示，也可以说向量$\vec{b}$是向量组A的线性组合。

11.3.2 线性相关

海棠红可以被RGB线性表示，我们就说海棠红和RGB线性相关。

红色不能用绿色和蓝色调出来，绿色不能用红色和蓝色调出来，蓝色不能用红色和绿色调出来，我们就说RGB线性无关。

一个向量能被某向量组线性表示，则由它们组成的向量组也被称为线性相关的。

线性相关的定义

严格定义如下：

给定向量组$A:\vec{a_1},\vec{a_2},\ldots,\vec{a_m}$，如果存在**不全为零**的实数$k_1,k_2,\ldots,k_m$，使：

$$k_1\vec{a_1} + k_2\vec{a_2} + \cdots + k_m\vec{a_m} = \vec{0} \tag{113}$$

则称向量组A是线性相关的，否则称它为线性无关。

线性相关与线性表示的关系

因为$k_1,k_2,\ldots,k_m$不全为零，不妨假设$k_1 \neq 0$

$$k_1\vec{a_1} + k_2\vec{a_2} + \cdots + k_m\vec{a_m} = \vec{0} \Rightarrow -k_1\vec{a_1} = k_2\vec{a_2} + \cdots + k_m\vec{a_m} \tag{114}$$

则：

$$k_1\vec{a_1} + k_2\vec{a_2} + \cdots + k_m\vec{a_m} = \vec{0} \Rightarrow \vec{a_1} = -\frac{k_2\vec{a_2} + \cdots + k_m\vec{a_m}}{k_1} \tag{115}$$

$\vec{a_1}$是$\vec{a_2},\vec{a_3},\ldots,\vec{a_m}$的线性表示，因此，$\vec{a_1},\vec{a_2},\ldots,\vec{a_m}$线性相关。

12 数学知识

1.LATEX控制序列的概念（类似于函数）
控制序列可以是作为命令：以“\”开头，参数：必须参数和可选参数。

12.0.3 Probability Density Functions

Probability Density Functions https://onlinecourses.science.psu.edu/stat414/node/97

13 Vocabulary

过拟合
样本
重采样
标记、打标签
逻辑回归、SVM、随机森林、决策树
generalized linear model (GLM)

14 论算法之间的区别

各自想法的出处
自己的总结

15 面试

15.1 监督分类和非监督分类都有哪些呢？

15.2 监督分类

线性回归
逻辑回归

15.3 非监督分类

聚类：K-means
PCA

LDA（文档主题模型）

16 数学家的故事

16.1 欧拉

欧拉（L.Euler,1707.4.15-1783.9.18）是瑞士数学家。生于瑞士的巴塞尔（Basel），卒于彼得堡（Petepbypt）。父亲保罗·欧拉是位牧师，喜欢数学，所以欧拉从小就受到这方面的熏陶。但父亲却执意让他攻读神学，以便将来接他的班。幸运的是，欧拉并没有走父亲为他安排的路。父亲曾在巴塞尔大学上过学，与当时著名数学家约翰·伯努利（Johann Bernoulli,1667.8.6-1748.1.1）及雅各布·伯努利（Jacob Bernoulli,1654.12.27-1705.8.16）有几分情谊。由于这种关系，欧拉结识了约翰的两个儿子：擅长数学的尼古拉（Nicolaus Bernoulli,1695-1726）及丹尼尔（Daniel Bernoulli,1700.2.9-1782.3.17）兄弟二人，（这二人后来都成为数学家）。他俩经常给小欧拉讲生动的数学故事和有趣的数学知识。这些都使欧拉受益匪浅。1720年，由约翰保举，才13岁的欧拉成了巴塞尔大学的学生，而且约翰精心培育着聪明伶俐的欧拉。当约翰发现课堂上的知识已满足不了欧拉的求知欲望时，就决定每周六下午单独给他辅导、答题和授课。约翰的心血没有白费，在他的严格训练下，欧拉终于成长起来。他17岁的时候，成为巴塞尔有史以来的第一个年轻的硕士，并成为约翰的助手。在约翰的指导下，欧拉从一开始就选择通过解决实际问题进行数学研究的道路。1726年，19岁的欧拉由于撰写了《论桅杆配置的船舶问题》而荣获巴黎科学院的资金。这标志着欧拉的羽毛已丰满，从此可以展翅飞翔。

欧拉的成长与他这段历史是分不开的。当然，欧拉的成才还有另一个重要的因素，就是他那惊人的记忆力!，他能背诵前一百个质数的前十次幂，能背诵罗马诗人维吉尔（Virgil）的史诗Aeneil，能背诵全部的数学公式。直至晚年，他还能复述年轻时的笔记的全部内容。高等数学的计算他可以用心算来完成。

尽管他的天赋很高，但如果没有约翰的教育，结果也很难想象。由于约翰·伯努利以其丰富的阅历和对数学发展状况的深刻的了解，能给欧拉以重要的指点，使欧拉一开始就学习那些虽然难学却十分必要的书，少走了不少弯路。这段历史对欧拉的影响极大，以至于欧拉成为大科学家之后仍不忘记育新人，这主要体现在编写教科书和直接培养有才华的数学

工作者，其中包括后来成为大数学家的拉格朗日（J.L.Lagrange,1736.1.25-1813.4.10）。

欧拉本人虽不是教师，但他对教学的影响超过任何人。他身为世界上第一流的学者、教授，肩负着解决高深课题的重担，但却能无视"名流"的非议，热心于数学的普及工作。他编写的《无穷小分析引论》、《微分法》和《积分法》产生了深远的影响。有的学者认为，自从1784年以后，初等微积分和高等微积分教科书基本上都抄袭欧拉的书，或者抄袭那些抄袭欧拉的书。欧拉在这方面与其它数学家如高斯（C.F.Gauss,1777.4.30-1855.2.23）、牛顿（I.Newton,1643.1.4-1727.3.31）等都不同，他们所写的书一是数量少，二是艰涩难明，别人很难读懂。而欧拉的文字既轻松易懂，堪称这方面的典范。他从来不压缩字句，总是津津有味地把他那丰富的思想和广泛的兴趣写得有声有色。他用德、俄、英文发表过大量的通俗文章，还编写过大量中小学教科书。他编写的初等代数和算术的教科书考虑细致，叙述有条有理。他用许多新的思想的叙述方法，使得这些书既严密又易于理解。欧拉最先把对数定义为乘方的逆运算，并且最先发现了对数是无穷多值的。他证明了任一非零实数R有无穷多个对数。欧拉使三角学成为一门系统的科学，他首先用比值来给出三角函数的定义，而在他以前是一直以线段的长作为定义的。欧拉的定义使三角学跳出只研究三角表这个圈子。欧拉对整个三角学作了分析性的研究。在这以前，每个公式仅从图中推出，大部分以叙述表达。欧拉却从最初几个公式解析地推导出了全部三角公式，还获得了许多新的公式。欧拉用a 、b 、c 表示三角形的三条边，用A、B、C表示第个边所对的角，从而使叙述大大地简化。欧拉得到的著名的公式：

又把三角函数与指数函联结起来。

在普及教育和科研中，欧拉意识到符号的简化和规则化既有有助于学生的学习，又有助于数学的发展，所以欧拉创立了许多新的符号。如用sin、cos 等表示三角函数，用e 表示自然对数的底，用f(x) 表示函数，用Σ表示求和，用i表示虚数等。圆周率π虽然不是欧拉首创，但却是经过欧拉的倡导才得以广泛流行。而且，欧拉还把e 、π、i 统一在一个令人叫绝的关系式中。欧拉在研究级数时引入欧拉常数C，这是继π、e 之后的又一个重要的数。

欧拉不但重视教育，而且重视人才。当时法国的拉格朗日只有19岁，而欧拉已48岁。拉格朗日与欧拉通信讨论"等周问题"，欧拉也在研究这个

问题。后来拉格朗日获得成果，欧拉就压下自己的论文，让拉格朗日首先发表，使他一举成名。

欧拉19岁大学毕业时，在瑞士没有找到合适的工作。1727年春，在巴塞尔他试图担任空缺的教研室主任职务，但没有成功。这时候，俄国的圣彼得堡科院刚建立不久，正在全国各地招聘科学家，广泛地搜罗人才。已经应聘在彼得堡工作的丹尔·伯努利深知欧拉的才能，因此，他竭力聘请欧拉去俄罗斯。在这种情况下，欧拉离开了自己的祖国。由于丹尼尔的推荐，1727年，欧拉应邀到圣彼得堡做丹尼尔的助手。在圣彼得堡科学院，他顺利地获得了高等数学副教授的职位。1731年，又被委任领导理论物理和实验物理教研室的工作。1733年，年仅26岁的欧拉接替回瑞士的丹尼尔，成为数学教授及彼得堡科学院数学部的领导人。

在这期间，欧拉勤奋地工作，发表了大量优秀的数学论文，以及其它方面的论文、著作。

古典力学的基础是牛顿奠定的，而欧拉则是其主要建筑师。1736年，欧拉出版了《力学，或解析地叙述运动的理论》，在这里他最早明确地提出质点或粒子的概念，最早研究质点沿任意一曲线运动时的速度，并在有关速度与加速度问题上应用矢量的概念。

同时，他创立了分析力学、刚体力学，研究和发展了弹性理论、振动理论以及材料力学。并且他把振动理论应用到音乐的理论中去，1739年，出版了一部音乐理论的著作。1738年，法国科学院设立了回答热本质问题征文的奖金，欧拉的《论火》一文获奖。在这篇文章中，欧拉把热本质看成是分子的振动。

欧拉研究问题最鲜明的特点是：他把数学研究之手深入到自然与社会的深层。他不仅是位杰出的数学家，而且也是位理论联系实际的巨匠，应用数学大师。他喜欢搞特定的具体问题，而不象现代某些数学家那样，热衷于搞一般理论。

17 其它知识

17.1 数据处理——One-Hot Encoding

原文地址：http://blog.csdn.net/google19890102/article/details/44039761

在实际的机器学习的应用任务中，**特征**有时候并不总是连续值，有可能是一些分类值，如性别可分为“male”和“female”。在机器学习任务中，

对于这样的特征，通常我们需要对其进行特征数字化，如下面的例子：有如下三个特征属性：

性别：["male"，"female"]

地区：["Europe"，"US"，"Asia"]

浏览器：["Firefox"，"Chrome"，"Safari"，"Internet Explorer"]

不合适的方法。对于某一个样本，如["male"，"US"，"Internet Explorer"]，我们需要将这个分类值的特征数字化，最直接的方法，我们可以采用序列化的方式：[0,1,3]。但是这样的特征处理并不能直接放入机器学习算法中。

合适的方法。对于上述的问题，性别的属性是二维的，同理，地区是三维的，浏览器则是四维的，这样，我们可以采用One-Hot编码的方式对上述的样本“["male"，"US"，"Internet Explorer"]”编码，“male”则对应着[1，0]，同理“US”对应着[0，1，0]，“Internet Explorer”对应着[0,0,0,1]。则完整的特征数字化的结果为：[1,0,0,1,0,0,0,0,1]。这样导致的一个结果就是数据会变得非常的稀疏。

实际的Python代码如下：

```
from sklearn import preprocessing
enc = preprocessing.OneHotEncoder()
enc.fit([[0,0,3],[1,1,0],[0,2,1],[1,0,2]])
array = enc.transform([[0,1,3]]).toarray()
print array
```

结果：[[1. 0. 0. 1. 0. 0. 0. 0. 1.]]

17.2 什么是张量(tensor)？

什么是张量(tensor)？ https://www.zhihu.com/question/20695804

张量的数学与物理意义是什么，张量的特性与优势是什么？ https://www.zhihu.com/question/368149

怎么通俗地理解张量? https://www.zhihu.com/question/23720923

What is a tensor? https://www.quora.com/What-is-a-tensor

小学课本上画杨桃的故事每个人都听过，一个杨桃在不同角度看，就会呈现不同的样子。有些物理量也是一样的，它在不同的角度看就会有不同的数值。比如对于一个矢量，你的基底变化了，矢量的表示也会变化。但是矢量的长度永远不变。杨桃还是那个杨桃，物理量也还是那个物理量，但是一旦你换了个角度看，杨桃的形状就变了，物理量的数值也就变了。

那么如果一个物理系统没有一个更好的观察方向，或者说我们需要频繁的变换我们的视角的时候，应该怎么把握一个胡乱变化的东西呢？你要记住，杨桃和物理量本身都是不变的，变的只是它在你眼中的形象。于是张量就出现了，它将视角变换时候的变换关系作为张量的定义，看似在乱七八糟变，实际上只有满足这样的变换关系，它才是不变的！研究一个看似乱七八糟变，实际上不变的东西，就是张量分析。

作者：大野喵渣链接：https://www.zhihu.com/question/36814916/answer/69248640来源：知乎著作权归作者所有。商业转载请联系作者获得授权，非商业转载请注明出处。

17.3 数学符号“s.t.”的意义

在优化问题的求解中，如线性规划、非线性规划问题等，经常会遇到数学符号“s.t.”，它的意思是什么呢？

“s.t.”，指subject to，受限制于...。

例如：

目标函数：min x+2

约束条件：s.t. x=1,2,3

其题意为，求x+2的最小值以使得x的取值为1、2、3时。

或者理解为，x的取值为1、2、3时，求x+2的最小值。