

跟我学人工智能

刘森

2018 年 2 月 4 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 σ^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。

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

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

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

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:

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

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

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

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：

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

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

$$\theta = (X^T X)^{-1} \cdot X^T y \quad (5)$$

1.1.4 最小二乘法意义下的参数最优解

参数的解析解

$$\theta = (X^T X)^{-1} X^T y \quad (6)$$

若 $X^T X$ 不可逆或防止过拟合，增加 λ 扰动

$$\theta = (X^T X + \lambda I)^{-1} X^T y \quad (7)$$

第05课《回归》00:38:36

“简单”方法记忆结论

$$\begin{aligned} X\theta &= y \\ \Rightarrow X^T X \theta &= X^T y \\ \Rightarrow \theta &= (X^T X)^{-1} X^T y \end{aligned} \quad (8)$$

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 $\theta = (X^T X)^{-1} X^T y$ 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

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

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

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

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

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

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

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

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

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

1.2.2 Logistic回归参数估计

假定

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

上面的两个式子，可以用一个式子来表示：

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

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：logit函数

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

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

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

$$\frac{x_1}{x_1 + x_2} \quad (24)$$

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

$$\frac{e^{x_1}}{e^{x_1} + e^{x_2}} = \frac{1}{1 + e^{x_2 - x_1}} \quad (25)$$

如果令 $x_2 - x_1 = -z$ ，就得到：

$$\frac{1}{1 + e^{-z}} \quad (26)$$

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

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

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

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} \quad (28)$$

余弦相似度（cosine similarity），从角度来理解

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

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

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

假设 $A = (A_1, A_2, \dots, A_n)$ ，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 $B = (B_1, B_2, \dots, B_n)$ ，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）

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

Hellinger距离

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

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：

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

这两个向量的Pearson相关系数是：

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

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 $S = x_1, x_2, \dots, x_m$ ，则算法步骤为：

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

(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：

$$label_i = \arg \min_{1 \leq j \leq k} \|x_i - \mu_j\| \quad (35)$$

(3) 将每个类别中心更新为隶属该类别的所有样本的均值

$$\mu_j = \frac{1}{|c_j|} \sum_{i \in c_j} x_j \quad (36)$$

(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 $\mu_1, \mu_2, \dots, \mu_k$ ，每个簇的样本数目为 N_1, N_2, \dots, N_k

使用平方误差作为目标函数：

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

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

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

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-Medoids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-Medoids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 的簇内不相相似度。簇 C 中所有样本的 a_i 均值称为簇 C 的簇不相相似度。

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 的簇间不相相似度: $b_i = \min\{b_{i1}, b_{i2}, \dots, b_{iK}\}$ 。其中, b_i 越大, 说明样本 i 越不属于其他簇。

根据样本 i 的簇内不相相似度 a_i 和簇间不相相似度 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一类簇，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**： $A^T A$ 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**拉普拉斯矩阵**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 $\text{sign}(x)$ 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 $x > 0$ ， $\text{sign}(x) = 1$ ；当 $x = 0$ ， $\text{sign}(x) = 0$ ；当 $x < 0$ ， $\text{sign}(x) = -1$ 。

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 $T = \{(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N)\}$ ，其中， $x_i \in R^n$ ， $y_i \in \{+1, -1\}$ ， $i = 1, 2, \dots, N$ 。

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 $f(x) = \text{sign}(w^T \Phi(x) + b)$ ，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于求解相应的**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, \dots, x_n

目标值： y_1, y_2, \dots, y_n

新数据的分类： $\text{sign}(y(x))$

[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

$$\begin{aligned} \min_a \quad & \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j y_i y_j (\Phi(x_i) \cdot \Phi(x_j)) - \sum_{i=1}^n a_i \\ \text{s.t.} \quad & \sum_{i=1}^n a_i y_i = 0 \\ & a_i \geq 0, i = 1, 2, \dots, n \end{aligned} \quad (39)$$

求得最优解 a^* ，再将 a^* 代入到下面的方程中去

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

计算求得分离超平面

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

分类决策函数

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

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-Leibler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

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

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。

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

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

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 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。

8.0.1 Probability Density Functions

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

9 Vocabulary

generalized linear model (GLM)

10 论算法之间的区别

各自想法的出处

自己的总结

11 面试

11.1 监督分类和非监督分类都有哪些呢？

11.2 监督分类

线性回归

逻辑回归

11.3 非监督分类

聚类：K-means

PCA

LDA（文档主题模型）

12 数学家的故事

12.1 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Aeneid，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数联结起来。

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，俄国的圣彼得堡科学院刚建立不久，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，研究和发展的弹性理论、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。

13 其它知识

13.1 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.]]
```

13.2 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/36814916>

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。

13.3 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。