08 机器学习 | 简约而不简单：线性回归

2017-12-26 王天一

人工智能基础课 进入课程 >

讲述：王天一
时长 11:54 大小 5.46M

数学中的线性模型可谓“简约而不简单”：它既能体现出重要的基本思想，又能构造出功能更加强大的非线性模型。在机器学习领域，线性回归就是这样一类基本的任务，它应用了一系列影响深远的数学工具。

在数理统计中，回归分析是确定多种变量间相互依赖的定量关系的方法。**线性回归假设输出变量是若干输入变量的线性组合，并根据这一关系求解线性组合中的最优系数。**在众多回归分析的方法里，线性回归模型最易于拟合，其估计结果的统计特性也更容易确定，因而得到广泛应用。而在机器学习中，回归问题隐含了输入变量和输出变量均可连续取值的前提，因而利用线性回归模型可以对任意输入给出对输出的估计。

1875 年，从事遗传问题研究的英国统计学家弗朗西斯·高尔顿正在寻找父代与子代身高之间的关系。在分析了 1078 对父子的身高数据后，他发现这些数据的散点图大致呈直线状态，

即父亲的身高和儿子的身高呈正相关关系。而在正相关关系背后还隐藏着另外一个现象：矮个子父亲的儿子更可能比父亲高；而高个子父亲的儿子更可能比父亲矮。

受表哥查尔斯·达尔文的影响，高尔顿将这种现象称为“**回归效应**”，即大自然将人类身高的分布约束在相对稳定而不产生两极分化的整体水平，并给出了历史上第一个线性回归的表达式：y = 0.516x + 33.73，式中的 y 和 x 分别代表以英寸为单位的子代和父代的身高。

高尔顿的思想在今天的机器学习中依然保持着旺盛的生命力。假定一个实例可以用列向量 $\mathbf{x} = (x_1; x_2; \cdots, x_n)$ 表示，每个 x_i 代表了实例在第 i 个属性上的取值，线性回归的作用就是习得一组参数 $w_i, i = 0, 1, \cdots, n$，使预测输出可以表示为以这组参数为权重的实例属性的线性组合。如果引入常量 $x_0 = 1$，线性回归试图学习的模型就是

$$f(\mathbf{x}) = \mathbf{w}^T\mathbf{x} = \sum_{i=0}^{n} w_i \cdot x_i$$

当实例只有一个属性时，输入和输出之间的关系就是二维平面上的一条直线；当实例的属性数目较多时，线性回归得到的就是 n 维空间上的一个超平面，对应一个维度等于 n - 1 的线性子空间。

在训练集上确定系数 w_i 时，预测输出 $f(\mathbf{x})$ 和真实输出 y 之间的误差是关注的核心指标。在线性回归中，这一误差是以**均方误差**来定义的。当线性回归的模型为二维平面上的直线时，均方误差就是预测输出和真实输出之间的**欧几里得距离**，也就是两点间向量的 L^2 范数。而以使均方误差取得最小值为目标的模型求解方法就是**最小二乘法**，其表达式可以写成

$$\mathbf{w}^* = \arg\min_{\mathbf{w}} \sum_{k=1} \left(\mathbf{w}^T\mathbf{x}_k - y_k\right)^2$$

$$= \arg\min_{\mathbf{w}} \sum_{k=1} ||y_k - \mathbf{w}^T\mathbf{x}_k||^2$$

式中每个 $\mathbf{x}_k$ 代表训练集中的一个样本。**在单变量线性回归任务中，最小二乘法的作用就是找到一条直线，使所有样本到直线的欧式距离之和最小。**

说到这里，问题就来了：凭什么使均方误差最小化的参数就是和训练样本匹配的最优模型呢？

这个问题可以从概率论的角度阐释。线性回归得到的是统计意义上的拟合结果，在单变量的情形下，可能每一个样本点都没有落在求得的直线上。

对这个现象的一种解释是回归结果可以完美匹配理想样本点的分布，但训练中使用的真实样本点是理想样本点和噪声叠加的结果，因而与回归模型之间产生了偏差，而每个样本点上噪声的取值就等于 $y_k - f(\mathbf{x}_k)$。

假定影响样本点的噪声满足参数为 $(0, \sigma^2)$ 的正态分布（还记得正态分布的概率密度公式吗？），这意味着噪声等于 0 的概率密度最大，幅度（无论正负）越大的噪声出现的概率越小。在这种情形下，对参数 $\mathbf{w}$ 的推导就可以用**最大似然的方式**进行，即在已知样本数据及其分布的条件下，找到使样本数据以最大概率出现的假设。

单个样本 $\mathbf{x}_k$ 出现的概率实际上就是噪声等于 $y_k - f(\mathbf{x}_k)$ 的概率，而相互独立的所有样本同时出现的概率则是每个样本出现概率的乘积，其表达式可以写成

$$p(\mathbf{x}_1, \mathbf{x}_2, \cdots \mathbf{x}_k, \cdots | \mathbf{w}) =$$

$$\prod_k \frac{1}{\sqrt{2\pi}\sigma} \exp[-\frac{1}{2\sigma^2}(y_k - \mathbf{w}^T \mathbf{x}_k)^2]$$

而最大似然估计的任务就是让以上表达式的取值最大化。出于计算简便的考虑，上面的乘积式可以通过取对数的方式转化成求和式，且取对数的操作并不会影响其单调性。经过一番运算后，上式的最大化就可以等效为 $\sum_k (y_k - \mathbf{w}^T \mathbf{x}_k)^2$ 的最小化。这不就是最小二乘法的结果么？

因此，**对于单变量线性回归而言，在误差函数服从正态分布的情况下，从几何意义出发的最小二乘法与从概率意义出发的最大似然估计是等价的。**

确定了最小二乘法的最优性，接下来的问题就是如何求解均方误差的最小值。在单变量线性回归中，其回归方程可以写成 $y = w_1 x + w_0$。根据最优化理论，将这一表达式代入均方误差的表达式中，并分别对 w_1 和 w_0 求偏导数，令两个偏导数均等于 0 的取值就是线性回归的最优解，其解析式可以写成

$$w_1 = \frac{\sum_{k=1}^{m} y_k(x_k - \frac{1}{m}\sum_{k=1}^{m} x_k)}{\sum_{k=1}^{m} x_k^2 - \frac{1}{m}(\sum_{k=1}^{m} x_k)^2}$$

$$w_0 = \frac{1}{m}\sum_{k=1}^{m}(y_k - w_1 x_k)$$

单变量线性回归只是一种最简单的特例。子代的身高并非仅仅由父母的遗传基因决定，营养条件、生活环境等因素都会产生影响。当样本的描述涉及多个属性时，这类问题就被称为**多元线性回归。**

多元线性回归中的参数 $\mathbf{w}$ 也可以用最小二乘法进行估计，其最优解同样用偏导数确定，但参与运算的元素从向量变成了矩阵。在理想的情况下，多元线性回归的最优参数为

$$\mathbf{w}^* = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

式中的 $\mathbf{X}$ 是由所有样本 $\mathbf{x} = (x_0; x_1; x_2; \cdots, x_n)$ 的转置共同构成的矩阵。但这一表达式只在矩阵 $(\mathbf{X}^T\mathbf{X})$ 的逆矩阵存在时成立。在大量复杂的实际任务中，每个样本中属性的数目甚至会超过训练集中的样本总数，此时求出的最优解 $\mathbf{w}^*$ 就不是唯一的，解的选择将依赖于学习算法的归纳偏好。

但不论采用怎样的选取标准，存在多个最优解都是无法改变的事实，这也意味着过拟合的产生。更重要的是，在过拟合的情形下，微小扰动给训练数据带来的毫厘之差可能会导致训练出的模型谬以千里，模型的稳定性也就无法保证。

要解决过拟合问题，常见的做法是正则化，即添加额外的惩罚项。**在线性回归中，正则化的方式根据其使用惩罚项的不同可以分为两种，分别是“岭回归”和“LASSO 回归”。**

在机器学习中，岭回归方法又被称为“参数衰减”，于 20 世纪 40 年代由前苏联学者安德烈·季霍诺夫提出。当然，彼时机器学习尚未诞生，季霍诺夫提出这一方法的主要目的是解决矩阵求逆的稳定性问题，其思想后来被应用到正则化中，形成了今天的岭回归。

岭回归实现正则化的方式是在原始均方误差项的基础上添加一个待求解参数的二范数项，即最小化的对象变为 $||y_k - \mathbf{w}^T\mathbf{x}_k||^2 + ||\Gamma\mathbf{w}||^2$，其中的 Γ 被称为**季霍诺夫矩阵**，通常可

以简化为一个常数。

从最优化的角度看，二范数惩罚项的作用在于优先选择范数较小的 $\mathbf{w}$，这相当于在最小均方误差之外额外添加了一重关于最优解特性的约束条件，将最优解限制在高维空间内的一个球里。岭回归的作用相当于在原始最小二乘的结果上做了缩放，虽然最优解中每个参数的贡献被削弱了，但参数的数目并没有变少。

LASSO 回归的全称是“**最小绝对缩减和选择算子**”（Least Absolute Shrinkage and Selection Operator），由加拿大学者罗伯特·提布什拉尼于 1996 年提出。与岭回归不同的是，LASSO 回归选择了待求解参数的一范数项作为惩罚项，即最小化的对象变为 $||y_k - \mathbf{w}^T\mathbf{x}_k||^2 + \lambda||\mathbf{w}||_1$，其中的 λ 是一个常数。

与岭回归相比，LASSO 回归的特点在于稀疏性的引入。它降低了最优解 $\mathbf{w}$ 的维度，也就是将一部分参数的贡献削弱为 0，这就使得 $\mathbf{w}$ 中元素的数目大大小于原始特征的数目。

这或多或少可以看作奥卡姆剃刀原理的一种实现：当主要矛盾和次要矛盾同时存在时，优先考虑的必然是主要矛盾。虽然饮食、环境、运动等因素都会影响身高的变化，但决定性因素显然只存在在染色体上。值得一提的是，**引入稀疏性是简化复杂问题的一种常用方法，在数据压缩、信号处理等其他领域中亦有广泛应用。**

从概率的角度来看，最小二乘法的解析解可以利用正态分布以及最大似然估计求得，这在前文已有说明。岭回归和 LASSO 回归也可以从概率的视角进行阐释：岭回归是在 w_i 满足正态先验分布的条件下，用最大后验概率进行估计得到的结果；LASSO 回归是在 w_i 满足拉普拉斯先验分布的条件下，用最大后验概率进行估计得到的结果。

但无论岭回归还是 LASSO 回归，其作用都是通过惩罚项的引入抑制过拟合现象，以训练误差的上升为代价，换取测试误差的下降。将以上两种方法的思想结合可以得到新的优化方法，在此就不做赘述了。

今天我和你分享了机器学习基本算法之一的线性回归的基本原理，其要点如下：

线性回归假设输出变量是若干输入变量的线性组合，并根据这一关系求解线性组合中的最优系数；

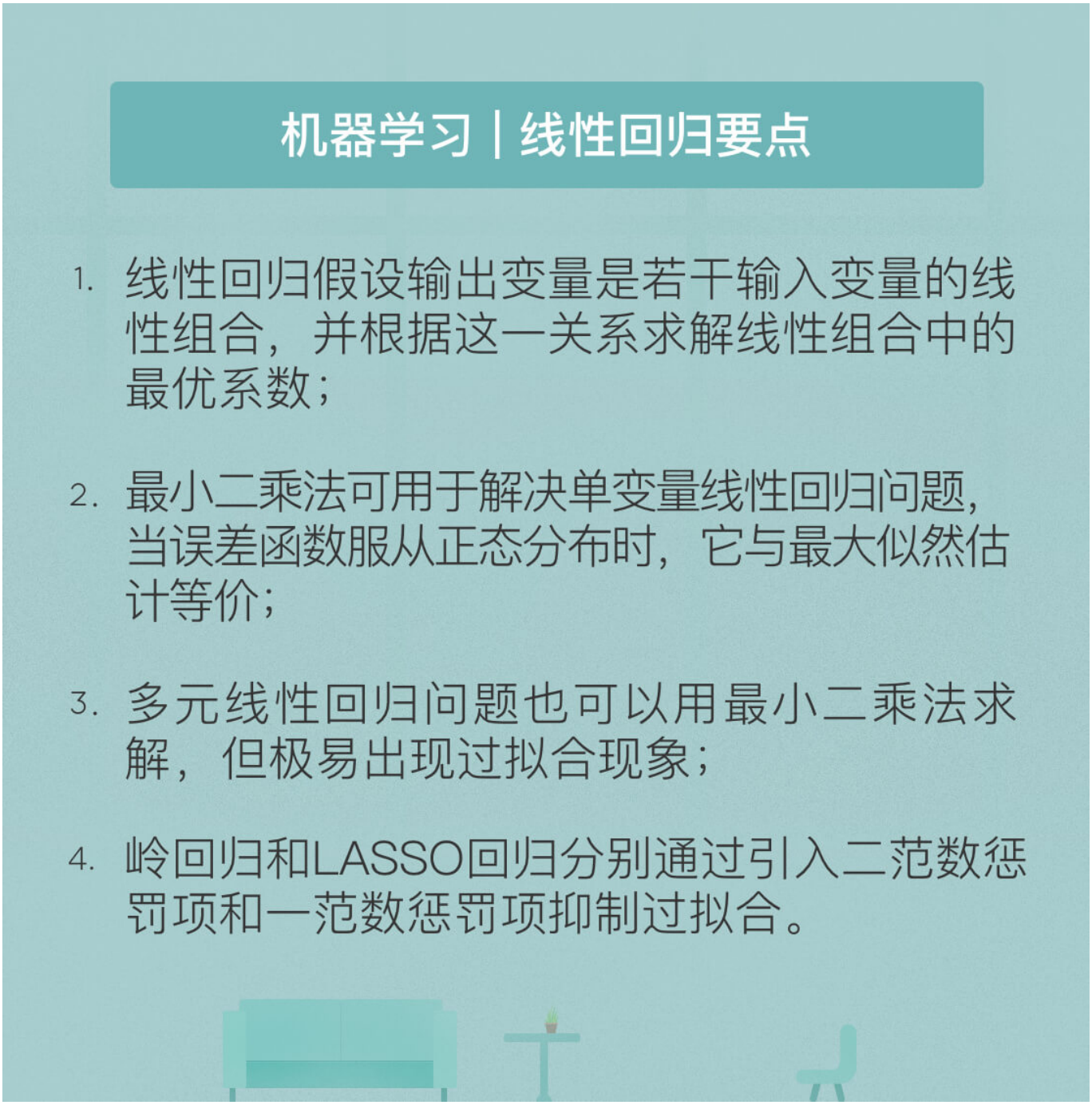
最小二乘法可用于解决单变量线性回归问题，当误差函数服从正态分布时，它与最大似然估计等价；

多元线性回归问题也可以用最小二乘法求解，但极易出现过拟合现象；

岭回归和 LASSO 回归分别通过引入二范数惩罚项和一范数惩罚项抑制过拟合。

在深度学习大行其道的今天，巨量的参数已经成为常态。在参数越来越多，模型越来越复杂的趋势下，线性回归还能发挥什么样的作用呢？

欢迎发表你的观点。



上一篇　07 机器学习 | 数山有路，学海无涯：机器学习概论

下一篇　09 机器学习 | 大道至简：朴素贝叶斯方法

精选留言 (19)

写留言

Maiza
2018-01-06

12

老师 每次看到公式的地方就跪了...
能麻烦给每个公式标明下出处，方便理解吗?
老师认为理所当然的事情对小白来说就是天书啊 。。。。。😂😂😂

展开

作者回复: 公式的出处都是前辈们的推导啊……后面我会再过一遍公式，看看有没有符号解释不清的地方

Haley_Hu
2018-02-24

2

2范数是不是就指L2正则 1范数就指L1正则

展开

作者回复: 范数本身是个数学指标，用这个指标做正则化就是对应的正则化方法

秦龙君
2017-12-29

2

学习了。

展开

wdf
2018-07-07

1

为什么说ringe对应的是正态，lasso是拉普拉斯分布

展开

作者回复: 这是从贝叶斯角度出发的。假定参数本身已经服从正态/拉普拉斯分布，那么在高斯噪声之下，用参数的似然概率乘以先验就可以得到后验。对后验概率取对数得到的结果和正则化的损失函数形式一致，所以对后验概率的最大化就是对正则化损失函数的最小化。
这相当于先假定参数符合特定条件，在此基础上再来计算最优参数。

全全
2019-03-07

天一老师，讲的真好！有些内容您可以加上一些图片帮助理解，比如讲特征值分解那里，变换是由矩阵引起的，特征值和特征向量，您用文字解释的非常明白，我又看维基百科里给了一个动图，觉得豁然开朗。还有这次课里那几个范数应用在回归正则化里，也有图片可以帮助理解。有时候学习这个东西，在我完全不懂的时候，解释的多明白都是天书的感觉。只有懂了，再看的时候就有共鸣，看出您文字里背后的意思。这是我的体会。...

展开

Snail@AI_M...
2019-01-18

和学习的课程相互印证之后，发现居然更糊涂了:这是上面的章节没有的现象，总结来说，本文更详细并做了一些拓展，比如线性回归的来历，正则化的应用等，课程只是告诉我们正则化的特点和应用，简单粗暴呢

展开

历尽千帆

2018-12-27

老师~我不太明白，为什么说LASSO回归的特点在于稀疏性的引入？我不太懂这里说的稀疏性是指的什么~

历尽千帆

2018-12-26

您好，我没有明白，为什么引入常量 x0=1，后面的y=wx才成立呢？

展开

十八哥

2018-12-01

假定给出美女的标准，数据化。问题提出如何确定美女在泳池中出现的概率？模型输入，地区、区域房价、游泳单价、时间维度、年龄参数等。此时我们可以用线性回归找到这些模型中那个变量是最优的，并且能给出排序。我的理解对吗？

作者回复: 实际上线性回归只能给出每个属性的权重，当然可以人为地认定权重越大，属性越优。

haiker

2018-10-14

引入抑制过拟合现象，以训练误差的上升为代价，换取测试误差的下降，训练误差和测试误差有时候是不是鱼和熊掌，不可兼得，训练误差太低可能就过拟合了，在测试集上效果就不好了。有些学者建议在训练集上训练的时候要等到稍微过拟合了再结束，因为提前结束的话，可能模型还没训练到位。

作者回复: 抑制过拟合几乎成为机器学习的核心问题了。

haiker
2018-10-14

LASSO 回归感觉也是在做特征降维，会不会和做完特征降维之后再做线性回归效果差不多呢?

作者回复: LASSO的降维应该说是无心插柳，不是以降维为目的，但起到降维的效果。它和PCA这种直接降维的效果还是有区别的。

haiker
2018-10-14

季霍诺夫矩阵是超参数要提前指定，还是参数，在训练过程中获得呢?

展开

作者回复: 超参数，需要提前确定，在不同的正则化超参数下计算出的模型参数也是不一样的。

Howard.Wu...
2018-09-22

老师的数学公式是用什么工具写的，可以具体分享一下吗?

展开

作者回复: Latex，学术写作排版神器

wdf
2018-07-07

请问老师，如果是多元回归，假定噪声服从高斯分布极大似然估计和最小二乘法等价吗

作者回复: 是的，文中考虑的就是多元的情形。

duchao_hit
2018-05-25

老师，对误差的概率就等于在参数w下样本的条件概率觉得不是很理解

展开

作者回复: 这个表达式写的有问题，等式左侧应该是在给定参数w和输入x的条件下，输出y的概率，也就是p(y | x, w)

duchao_hit

2018-05-25

老师，关于样本x的概率就等于误差的概率不是很理解。疑惑是y-wx=e，只能说p(y-wx)=p(e)，但不能说p(y-wx)=p(x|w)吧?

刘滨

2018-01-27

老师，请问最小二乘法跟梯度下降方法有什么区别？这里可以用梯度下降方法吗

作者回复: 最小二乘定义了最优化的目标函数，梯度下降要找到最优化问题的最优解，两者大致是目的和手段的关系。
最小二乘是有解析解的，如果解析解难以求解，就可以用梯度下降这些数值方法。

wolfog

2018-01-17

天一老师有段，开头是“LASSO回归的全称最小绝对缩减和选择算子”这一段的倒数第二行的2泛数项和一泛数项写法是否应该统一，要么数字都在右下角，要么都在右上角。今天后面这个惩罚项目不太了解为什么惩罚项目一加，就可以降低了他的过拟合。

作者回复: 感谢你的细心，但两个标号意思是不一样的哈：下标代表的是范数维度，上标表示的是平方操作，意思是2范数的平方。所以在上标下面再加一个表示2范数的下标2，是标准的写法。
所谓过拟合呢，就是用来描述模型的参数数目太多了，正则化项的作用就是通过不同范数控制参数的取值大小，如果把某些参数抑制为0，这个参数就消失了，也就起到了通过减少参数数目抑制过拟合的作用。

Andy
2017-12-29

最小二乘的形式，为何跟极大似然估计后的形式一致呢？是巧合吗？

展开

作者回复: 这是由正态分布的特性决定的