# 关于第二题成绩预测的一种做法（1）

数据挖掘里面的一大方法是机器学习，而机器学习从抽象的层面来看可以看作是函数拟合的过程。预测，是机器学习算法最为关键的评价标准之一。也就是说，一个好的机器学习算法，应该能根据少量的已知信息来预测未知的、或者缺失的数据。为什么说线性回归其实可以看作机器学习的一种，就是因为它可以根据少量的离散数据点，给出一个关于这些数据内在规律的假设，也就是空间中的一条直线。这样的话，即使我们只掌握的少量的数据，但我们可以自认为掌握了这些数据生成的内在规律，从而可以用来预测未知数据点更有可能的分布情况。

而第二题，由于已经提供所有学生在过去两个学期的相关数据，如借书、图书馆、消费以及相应学期的成绩，且只提供这些数据，因此该问题就可以看作是在已知某个学生在过去两个学期的学习以及消费情况的前提下，对该学生在第三个学期的学习成绩作出相应的预测的问题。

预测的第一步当然是对数据进行抽象分析并建模。在这个应用里面，我们得到的数据全部是离散且数值化的（不存在纯文本值），这就给我们的抽象分析提供了一定的便利。比如，我们可以这样做：针对数据给出的每一个学生，假设其对应的借书、图书馆门禁、消费以及学期成绩的相应记录分别为 B、L、C、G，其中 B 包含借书的日期、书的 ID，L 包括进出图书馆门禁的时间而 C 则包含消费日期以及消费的数量等，则该学生可以被唯一表征为向量 (B,L,C,G)。当然，由于没一个学生的 B、L、C、G 的取值并不一定是对齐的（有多有少，有长有短），对其进行统一分析是很难的，所以这种表示方法只有象征性的意义，对我们的具体分析用处不大。

现在，如果我们假设这样一个前提：**学生在第三个学期的学业成绩是一个函数，这个函数的自变量是该学生在前两个学期中的学习和消费情况[1]**。也就是说，在这里，我们构建了一个关于学生学业成绩的数学模型，模型的参数是学生的历史行为记录，而模型的产出就是学生的学业成绩。这样的话，我们就可以以抽象公式来进一步细化在上文提到的表征方式,即:

$G' = f(B, L, C, G)$ ，其中 f 是一未知函数。

那么，这种建模方法有什么用处呢？仔细想想，如果我们能找出这样一种关系，那如果现在来了一个新同学 x，我们想根据这个同学以往的行为以及学业成绩记录来预测一下他在第三个学期所能够取得的成绩，我们就可以这样做：计算 $G'(x)=f(B(x),L(x),C(x),G(x))$ ，然后以 G'(x)的值作为对该学生成绩的预测值。这样，我们就把预测问题转换成了纯粹的数学计算了。

当然，上面的假设还不太具体，参数也没有对齐，因此还不太可能形成有用的算法。所以接下来我们需要考虑怎么样把 B,L,C,G 这些参数进行对齐处理。

首先，我们看看 B 里面有什么东西。B 是图书馆的借阅记录，对于某一个特定的学生 x 来说，B 应该包含三个字段：学期、书号以及借阅的日期。因为每个学生有不同长度的借阅记录，所以问题的关键就在于，如何把这些不同长度的序列都映射到一个等长的序列里面，从而完成对 B 的对齐呢？有一种笨方法是这样的：把数据集中出现所有的书目排列起来作为横轴，把借阅历史数据中的日期以一天为间隔排列起来作为纵轴，形成一个 M*N 维的矩阵 Bm，其中 M 是书的数量，N 是借阅历史中不同的日期数，Bm(i,j)表示该学生在第 j 天对书籍 i 的借阅情况（可以假设 Bm(i,j)=1 当且仅当有借阅，否则为 0)。这样的话，我们就可以把每个学生的借阅历史都表示成一个 M*N 维的二值矩阵了。

上述方法是有其合理性的，比如，它对书的不同类型都进行了记录，保证原数据的各种细节都能在对齐后的表示方法里得到体现，从而保证了精确度。但是问题在于，这种对齐方法其实是没必要的，因为它很有可能把一些并不是那么重要的、以至于足以影响结果的因素也考虑在内了。举例说，大多数情况下，学生借书的具体日期可能并不能真正影响到他的学业成绩，因此对日期的精确转换其实是没必要的。所以，基于上文提到的那个假设[1]，我们可以进行更强化的假设，从而简化数据冗余，消除数据噪音。对于借阅记录，我们可以有下述的强化假设：**学生的学业成绩仅仅与其在一个学期中借阅书的数量相关，而与借书的日期和具体是哪本书无关[2]。**这个假设基于这样的简单判断：**借书的数量足以表明学生的努力程度，而具体日期和到底借的是什么书我们并不关心**。这样的话，借阅记录就可以表示为这样一组数据：

$Bv=(bn_1,bn_2,....,bn_k)$ ，其中 $bn_i$ 表示第 i 个学期该学生借书的数量。

这样我们就完成了对学生借阅记录的对齐。基于类似的考虑，我们也可以假设：（1）**图书馆门禁记录中，我们只关心进出的次数而不考虑具体时间[3]**。（2）**消费记录中，我们只考虑消费的总额而不关心消费的时间和地点[4]**。

假设对每个学生 x，其对齐后的借阅记录、图书馆门禁以及消费记录、前两个学期的学业成绩分别为 Bv(x)、Lv(x)、Cv(x)、G(x)，则参照上面的表示方法，该学生可被标识为向量 $H(x)=(Bv(x),Lv(x),Cv(x),G(x))$ 。更进一步，对历史记录到第三学期学业成绩 G'的映射函数 f，有 $G'(x)=f(H(x))=f(Bv(x),Lv(x),Cv(x),G(x))$ 。由于 Bv,Lv,Cv,G 都已知且对齐，我们就完成了基于假设前提[1]、[2]、[3]、[4]的对该问题的建模过程。

**未完待续。。。**