1.2 实验概述

**本实验在阿里云PAI机器学习平台上，对某零售商的一款产品的销量进行了预测及分析。模型分为两个部分，一个部分是线性回归模型，选取了该产品的促销幅度、促销手段对实际销量进行回归；另一部分为时间序列分解建模，对线性回归模型的误差进行时间序列分解，从而可以在一定程度上解释线性模型的误差并提高整体预测的准确率。**

1.3 实验目标

1.掌握商品销量预测的基本概念。

2.了解常见的预测方法和价格弹性时间序列分解模型的优缺点和所适用的场景。

3.使用阿里云的相关产品来构建和部署销量预测的模型。

4.学会根据模型的结果进行商品定价与促销决策。

1.4 实验架构

1.MaxCompute

2.DataIDE

3.PAI

1.5 实验准备

**正确运营策略的制定取决于相关市场情况的准确预测**

企业营销的最终目的是为了获取利润，而要实现这个目标就需要在产品的定位、定价、分销渠道、促销方式、库存运输、生产计划和原料采购等方面都做出准确的预测，才能制定相应的策略，这些预测都是环环相扣并且互相依赖的，而他们的基础就是商品销量预测。

**定价策略的好坏会影响到利润的高低**

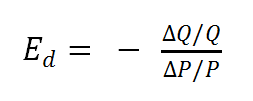
Profit = Price \* Quantity – Cost(total)

(利润 = 价格 \* 销量 - 总成本)

其中：价格Price的变动会引起销量Quantity的变化。

**需求价格弹性（Price elasticity of demand）**

简称为价格弹性或需求弹性，是指需求量对价格变动的反应程度，是需求量变化的百分比除以价格变化的百分比。

假设Q为某个[商品](http://wiki.mbalib.com/wiki/%E5%95%86%E5%93%81)的[需求](http://wiki.mbalib.com/wiki/%E9%9C%80%E6%B1%82)，P为该商品的价格，则计算需求的价格弹性Ed：

其中：Ed代表需求价格弹性系数，Q表示需求量，ΔQ是需求量的变化量，P表示价格，ΔP是价格的变化量。

**【注】**第一，将所有的百分比变动视为正数，从而去掉数值中的负号。这意味着尽管根据需求定律，需求量与价格反方向变动，但需求价格弹性均不小于零。当某产品的需求价格弹性大，即指其绝对值大。第二，注意弹性定义使用的相对变化的百分比，而非实际变化量。这意味着衡量单位的变化不会影响弹性的大小，无论以美元还是人民币衡量商品价格，需求价格弹性保持不变。第三，计算公式中的“价格”和“需求量”，通常采用平均值作为计算基础。

**对于商品的不同弹性程度，有不同的称呼**

当Ed = 0：完全无弹性（perfectly inelastic）

当0 < Ed < 1：缺乏弹性、不富弹性，或无弹性（relatively inelastic）

当Ed = 1：单位弹性、单弹性，或恒一弹性（unitary）

当1 < Ed < https://edu.aliyun.com/lab/files/courses/f3dd67c5dfc248cfad4ee1540ee61283/sections/3b072634eea24a3a97383004820a2800/content/images/course-f3dd67c5dfc248cfad4ee1540ee61283-section-3b072634eea24a3a97383004820a2800-content-image-1506652397109-SHaoZ3：富有弹性（relatively elastic）

当Ed = https://edu.aliyun.com/lab/files/courses/f3dd67c5dfc248cfad4ee1540ee61283/sections/3b072634eea24a3a97383004820a2800/content/images/course-f3dd67c5dfc248cfad4ee1540ee61283-section-3b072634eea24a3a97383004820a2800-content-image-1506652406308-stZ0EG：完全弹性或完全有弹性（perfectly elastic）

需求价格弹性是企业实行薄利多销策略的一个主要理论基础，一般都是富有弹性的商品，因为这种商品的价格下降时，需求量增加的幅度大于价格下降的幅度，所以总收益增加。

**影响商品需求价格弹性的因素**

假设大部分的人都是理性的，从整体销量的角度来看，影响需求弹性的因素有这么几个：

替代品的数量和相近程度：一种商品如果有许多相近的替代品（比如普通品牌矿泉水，替代品包括一些饮料、桶装水、自己烧水喝），这种商品的需求价格弹性就很大。一旦这种商品价格上涨，甚至是微小的上涨，消费者往往会舍弃它，而去选择它的替代品，从而引起需求量的变化。

商品的重要性：一种商品如果是人们基本生活的必需品，小量的价格变动是不会引起需求的剧烈变动，人们该买还得照样买（比如汽油），其需求弹性就比较小或者缺乏弹性；而一些奢侈品，比较像贵重首饰、高档服装等，只有当消费者购买力提高之后才买得起，其需求弹性就比较大。

时间：时间越短，商品的需求弹性就越小；时间越长，商品的需求弹性就越大。这是因为在较长的时间内，消费者就越有可能找到替代品，替代物品多了，它的需求弹性就会增加。

**【注】**在本实验中，该模型的建立需要假设这个商品的需求价格弹性不因时间而变化。

商品占总消费支出比例：一本杂志只售6-10元，同时替代品之间的差异比较大，而它的售价占消费者的支出比例不大，消费者不太会愿意因为价格的小小变动而购买其他杂志，更看中的是它本身提供的内容，所以它的需求弹性也是比较低的。

**目前常见的两种预测方法**

**第一种是因果预测法**

根据事物之间的因果关系来预测事物的发展和变化，通过分析那些对预测目标有直接或间接影响的因素，找出这些因素的变化规律来预测未来的值。

因果预测法也叫回归法，一般可以分为一元线性回归、多元线性回归和自回归。

其中一元线性回归只分析一个自变量与因变量之间的相关关系，然后用一元回归方程进行预测。例如，依据居民收入的变化预测某种日常消耗品的需求量。

多元线性回归法会分析因变量与若干个自变量之间的相关关系，运用多元回归方程从若干个自变量的变化去预测因变量的变化。例如，依据商品的促销手段、促销幅度等的变化，预测商品的销售量。

自回归法利用预测目标的历史数据在不同时期的取值之间的依存关系(即自身相关)，建立起回归方程进行预测。例如，根据消费者目前的食品消费水平，可以预测下一期的食品消费水平。其中因变量是观测值，而自变量则是因变量的滞后值（即之前的值）

**另一种方法是时间序列法**

**时间序列**

也叫时间数列、历史复数或动态数列。它是按时间的先后顺序，将某种统计指标的数值排列所形成的数列。

**时间序列预测法**

是通过编制和分析时间序列，根据它所反映出来的发展过程、方向和趋势，进行类推或延伸，借以预测下一段时间可能达到的水平。

从回归分析法的角度看，时间序列分析法实际上是一种特殊的回归分析法，因为时间序列不考虑事物之间的因果关系或其他相关关系，而仅研究对象与时间之间的相关关系。

**对时间序列进行预测的时候，有以下几种方法**

**简单平均法**

把N个历史数值作为观察值，求出算术平均数作为下期预测值。这种方法的假设是：“过去这样，今后也会这样”，把近期和远期数据看作是一样的，因此只适用于事物变化不大的趋势预测。

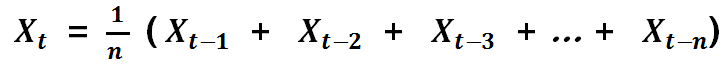
**指数平滑法**

根据历史数据的上期实际值和预测值，进行预测。

优点是只要有上期的实际值和上期预测值，用指数加权的办法就可计算下期的预测值，这样可以节省很多数据和处理数据的时间，减少数据的存储量，方法简便。缺点就是没有包含远期数据，可能忽略了很多信息，造成预测不够准确。

**移动平均法**

相继移动计算若干个时期的算术平均数，作为下期的预测值



https://edu.aliyun.com/lab/files/courses/f3dd67c5dfc248cfad4ee1540ee61283/sections/3b072634eea24a3a97383004820a2800/content/images/course-f3dd67c5dfc248cfad4ee1540ee61283-section-3b072634eea24a3a97383004820a2800-content-image-1506652641117-b4LxDh是当时间为t时的预测值，n是要计算的移动平均的周期数，https://edu.aliyun.com/lab/files/courses/f3dd67c5dfc248cfad4ee1540ee61283/sections/3b072634eea24a3a97383004820a2800/content/images/course-f3dd67c5dfc248cfad4ee1540ee61283-section-3b072634eea24a3a97383004820a2800-content-image-1506652655647-geMGob，https://edu.aliyun.com/lab/files/courses/f3dd67c5dfc248cfad4ee1540ee61283/sections/3b072634eea24a3a97383004820a2800/content/images/course-f3dd67c5dfc248cfad4ee1540ee61283-section-3b072634eea24a3a97383004820a2800-content-image-1506652668369-6qSboV ，… ，https://edu.aliyun.com/lab/files/courses/f3dd67c5dfc248cfad4ee1540ee61283/sections/3b072634eea24a3a97383004820a2800/content/images/course-f3dd67c5dfc248cfad4ee1540ee61283-section-3b072634eea24a3a97383004820a2800-content-image-1506652691852-y6owjX 表示前一期、前两期直到前n期的实际值。

**季节趋势分析法**

根据每年重复出现的周期性季节变动，预测未来的季节性变动趋势。一般使用的是季度(或月度)平均法。就是把各年度的数值按季度(或月度)求和之后取算术平均数。

例：假设总共有3年的完整数据，每年1月份的季节变动数值都等于将这三年的一月份数值加起来之后除以3（因为总共有3个一月份）。这样求得的每年相同月份的季节变动都是相同的。

**时间序列的变化，受到长期趋势、季节变动、循环变动和不规则变动这四个因素的影响**

**长期趋势（Trend）**反映了数据在一个较长时间段内的发展方向，表现为近似直线的持续向上或持续向下或平稳的趋势。长期趋势一般是受某种根本性因素作用而形成的，比如产品的生命周期或者社会总体的经济形势。

**季节变动（Seasonality）**反映了现象在一年内随着季节的变化而发生的有规律的周期性变动。比如毛衣在秋天和冬天卖的比较多，春夏卖的比较少。

**循环变动（Cycle）**是受各种因素影响而形成的上下起伏不定的波动。注意，季节变动和循环变动的区别是循环变动没有固定周期，可能是几年。但季节变动是有固定周期的，基本就是一年。

**不规则变动（Random）**又称随机变动，它是受各种偶然因素影响所形成的。

**本次实验使用的价格弹性时间序列分解模型，将促销幅度、促销手段、时间序列都考虑进来，进行多元线性回归并将时间序列进行分解的一个模型。**

**具体的操作可以分为下面几步：**

**1.数据预处理**

历史数据拆分成训练集和测试集，在训练集中，将促销幅度和促销类型作为自变量，将真实的销量（小写q）作为因变量进行多元线性回归后，就能得到回归方程式的系数。

**2.线性回归模型**

根据刚才算出来的回归系数和训练集里面的历史数据进行预测，得到一个预测的销量，记为大写的Q，然后计算实际销量小q和回归预测出来的销量大Q之间的误差（记做dQ），基本思想是将促销幅度和促销类型对销量造成的影响剥离出来，只留下时间造成的影响，即为真实值和回归预测值之间的差。

**3.时间序列分解模型**

对这个误差使用时间序列分解模型分解，得到时间序列分解模型的参数UTSCR。

时间序列https://edu.aliyun.com/lab/files/courses/f3dd67c5dfc248cfad4ee1540ee61283/sections/3b072634eea24a3a97383004820a2800/content/images/course-f3dd67c5dfc248cfad4ee1540ee61283-section-3b072634eea24a3a97383004820a2800-content-image-1506652830349-7agbRv可以表示为五个因素的函数https://edu.aliyun.com/lab/files/courses/f3dd67c5dfc248cfad4ee1540ee61283/sections/3b072634eea24a3a97383004820a2800/content/images/course-f3dd67c5dfc248cfad4ee1540ee61283-section-3b072634eea24a3a97383004820a2800-content-image-1506652809601-lfsfKc，其中U代表了序列的平均值。时间序列分解的方法有很多，较常用的模型有加法模型和乘法模型，这两个模型分别适用于不同的情形。

**加法模型**

https://edu.aliyun.com/lab/files/courses/f3dd67c5dfc248cfad4ee1540ee61283/sections/3b072634eea24a3a97383004820a2800/content/images/course-f3dd67c5dfc248cfad4ee1540ee61283-section-3b072634eea24a3a97383004820a2800-content-image-1506652875351-5Qe8rY 假定TSCR对现象的影响是[相互独立](https://www.baidu.com/s?wd=%E7%9B%B8%E4%BA%92%E7%8B%AC%E7%AB%8B&tn=44039180_cpr&fenlei=mv6quAkxTZn0IZRqIHckPjm4nH00T1d9mWbLmH01mHbdnW6Ln1Rd0ZwV5Hcvrjm3rH6sPfKWUMw85HfYnjn4nH6sgvPsT6KdThsqpZwYTjCEQLGCpyw9Uz4Bmy-bIi4WUvYETgN-TLwGUv3EnHTdn16znWDLnjT1njnzPjfvr0)的，每个成分均以绝对量表示。

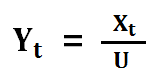
**乘法模型**

https://edu.aliyun.com/lab/files/courses/f3dd67c5dfc248cfad4ee1540ee61283/sections/3b072634eea24a3a97383004820a2800/content/images/course-f3dd67c5dfc248cfad4ee1540ee61283-section-3b072634eea24a3a97383004820a2800-content-image-1506652886546-05wwMk 假定四个因素的影响是相互作用的，以序列平均值的绝对量为基础，其余成分均以比率表示。

大多数的情况下乘法模型都是更符合现实和更准确的，因此本次实验采用乘法模型。

**乘法模型的分解步骤如下**

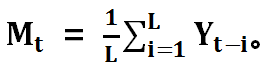
**a) 计算序列的平均值U**

然后将序列中所有个体的值除以U来移除平均值:   。

这样就创建了一个值在1左右的新序列Y(如果序列平均值的绝对值小于0.0000001，则不进行除法运算)。

**b) 计算移动平均值 M**

计算https://edu.aliyun.com/lab/files/courses/f3dd67c5dfc248cfad4ee1540ee61283/sections/3b072634eea24a3a97383004820a2800/content/images/course-f3dd67c5dfc248cfad4ee1540ee61283-section-3b072634eea24a3a97383004820a2800-content-image-1506652965325-QYeMKU的 L 步移动平均，其中L是季节变动 S 的周期长度（例如，数据是一个聚合到季度的序列，那么这个序列的季节变动周期就是4, L=4。如果是聚合到月度的序列，季节变动的周期就是12，L=12）

**M的公式是**

由于移动平均M算的是一年的平均值，季节性因素被移除。

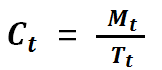
**c) 计算长期趋势T**

具体对移动平均https://edu.aliyun.com/lab/files/courses/f3dd67c5dfc248cfad4ee1540ee61283/sections/3b072634eea24a3a97383004820a2800/content/images/course-f3dd67c5dfc248cfad4ee1540ee61283-section-3b072634eea24a3a97383004820a2800-content-image-1506653020629-3mOvT3进行线性回归得到:https://edu.aliyun.com/lab/files/courses/f3dd67c5dfc248cfad4ee1540ee61283/sections/3b072634eea24a3a97383004820a2800/content/images/course-f3dd67c5dfc248cfad4ee1540ee61283-section-3b072634eea24a3a97383004820a2800-content-image-1506653032451-mGKtis（其中a是截距，b是斜率，是线性拟合的残差）。

**长期趋势T就是这个等式的线性部分：https://edu.aliyun.com/lab/files/courses/f3dd67c5dfc248cfad4ee1540ee61283/sections/3b072634eea24a3a97383004820a2800/content/images/course-f3dd67c5dfc248cfad4ee1540ee61283-section-3b072634eea24a3a97383004820a2800-content-image-1506653048880-xkboFa**

得到长期趋势的系数a和b。

**d) 计算循环变动C**

循环变动等于**移动平均**https://edu.aliyun.com/lab/files/courses/f3dd67c5dfc248cfad4ee1540ee61283/sections/3b072634eea24a3a97383004820a2800/content/images/course-f3dd67c5dfc248cfad4ee1540ee61283-section-3b072634eea24a3a97383004820a2800-content-image-1506653090946-q2nWy1除以**长期趋势**https://edu.aliyun.com/lab/files/courses/f3dd67c5dfc248cfad4ee1540ee61283/sections/3b072634eea24a3a97383004820a2800/content/images/course-f3dd67c5dfc248cfad4ee1540ee61283-section-3b072634eea24a3a97383004820a2800-content-image-1506653104498-v3IF8r ： 

预测时循环变动C的值可以根据商业知识手动输入，也可以输入一个固定的值。

**在实验中，我们计算了C序列之后，会取它的平均值作为预测用的C值。**

**e) 计算季节变动S**

**季节变动K等于 Y序列 除以 移动平均**

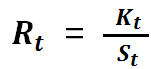
**https://edu.aliyun.com/lab/files/courses/f3dd67c5dfc248cfad4ee1540ee61283/sections/3b072634eea24a3a97383004820a2800/content/images/course-f3dd67c5dfc248cfad4ee1540ee61283-section-3b072634eea24a3a97383004820a2800-content-image-1506653271885-8QE6fO**

（注意，K序列是由季节变动和随机变动组成的）

**为了计算真正的季节变动S，取时间t所属“季节”的K序列平均值https://edu.aliyun.com/lab/files/courses/f3dd67c5dfc248cfad4ee1540ee61283/sections/3b072634eea24a3a97383004820a2800/content/images/course-f3dd67c5dfc248cfad4ee1540ee61283-section-3b072634eea24a3a97383004820a2800-content-image-1506653303312-YlPpic**

例： 假设一共有3年数据的月度数据，那么 L=12，https://edu.aliyun.com/lab/files/courses/f3dd67c5dfc248cfad4ee1540ee61283/sections/3b072634eea24a3a97383004820a2800/content/images/course-f3dd67c5dfc248cfad4ee1540ee61283-section-3b072634eea24a3a97383004820a2800-content-image-1506656618981-kQXvxf，https://edu.aliyun.com/lab/files/courses/f3dd67c5dfc248cfad4ee1540ee61283/sections/3b072634eea24a3a97383004820a2800/content/images/course-f3dd67c5dfc248cfad4ee1540ee61283-section-3b072634eea24a3a97383004820a2800-content-image-1506656545655-86lQyL，依次类推，到https://edu.aliyun.com/lab/files/courses/f3dd67c5dfc248cfad4ee1540ee61283/sections/3b072634eea24a3a97383004820a2800/content/images/course-f3dd67c5dfc248cfad4ee1540ee61283-section-3b072634eea24a3a97383004820a2800-content-image-1506656679239-oeGcxx。

**f) 计算不规则变动R**

将**K序列**除以**S序列**： 

由于预测时假设R值为1，所以R序列不计算也是可以的。

**g) 时间序列预测**

得到所有的参数之后，就可以预测时间为t时的值：

https://edu.aliyun.com/lab/files/courses/f3dd67c5dfc248cfad4ee1540ee61283/sections/3b072634eea24a3a97383004820a2800/content/images/course-f3dd67c5dfc248cfad4ee1540ee61283-section-3b072634eea24a3a97383004820a2800-content-image-1506653386524-GKNWxc

其中：

长期趋势https://edu.aliyun.com/lab/files/courses/f3dd67c5dfc248cfad4ee1540ee61283/sections/3b072634eea24a3a97383004820a2800/content/images/course-f3dd67c5dfc248cfad4ee1540ee61283-section-3b072634eea24a3a97383004820a2800-content-image-1506653402959-xXNHyZ

季节变动 https://edu.aliyun.com/lab/files/courses/f3dd67c5dfc248cfad4ee1540ee61283/sections/3b072634eea24a3a97383004820a2800/content/images/course-f3dd67c5dfc248cfad4ee1540ee61283-section-3b072634eea24a3a97383004820a2800-content-image-1506653420074-xMF2Iv

循环变动https://edu.aliyun.com/lab/files/courses/f3dd67c5dfc248cfad4ee1540ee61283/sections/3b072634eea24a3a97383004820a2800/content/images/course-f3dd67c5dfc248cfad4ee1540ee61283-section-3b072634eea24a3a97383004820a2800-content-image-1506653518398-e2hshr取训练集的平均值

不规则变动https://edu.aliyun.com/lab/files/courses/f3dd67c5dfc248cfad4ee1540ee61283/sections/3b072634eea24a3a97383004820a2800/content/images/course-f3dd67c5dfc248cfad4ee1540ee61283-section-3b072634eea24a3a97383004820a2800-content-image-1506653529744-fJuqHz则设为1

**4.销量预测与准确率评估**

根据上面算出来的系数和测试集中的自变量（包括时间t,促销幅度和促销类型），分别计算时间序列分解模型和回归模型的预测值，将这两个预测值加起来就是我们最终的预测值。

**实验资源**

本实验需要使用阿里云的MaxCompute与PAI机器学习平台，具体步骤如下：

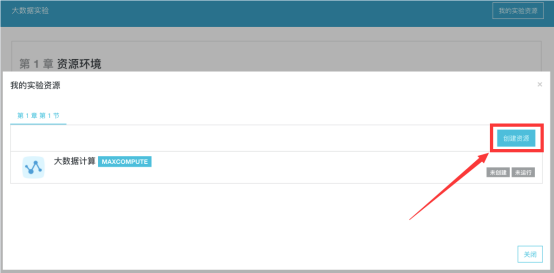
**【注】**本实验均在Google Chrome浏览器下测试运行，为了达到最大兼容，推荐使用Windows7以上的操作系统以及Chrome浏览器进行实验。

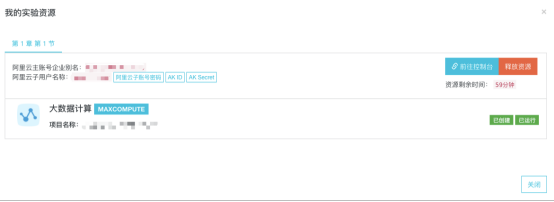
**【注】**一旦开始创建资源，该实验就开始计时，并在到达实验规定的时长时，将自动结束实验并清除资源。

1.在实验的的第一章第一节，可以看到如下画面。点击右上方的我的实验资源按钮来创建MaxCompute资源。



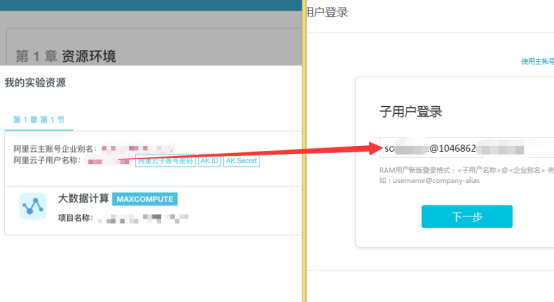
2.在弹出的界面上点击创建资源按钮，等待片刻，系统会自动创建好数加子帐号资源。





在该页面上可以看到类似于上图的信息，并且页面上会有接下来需要使用的子用户名称与子用户密码。

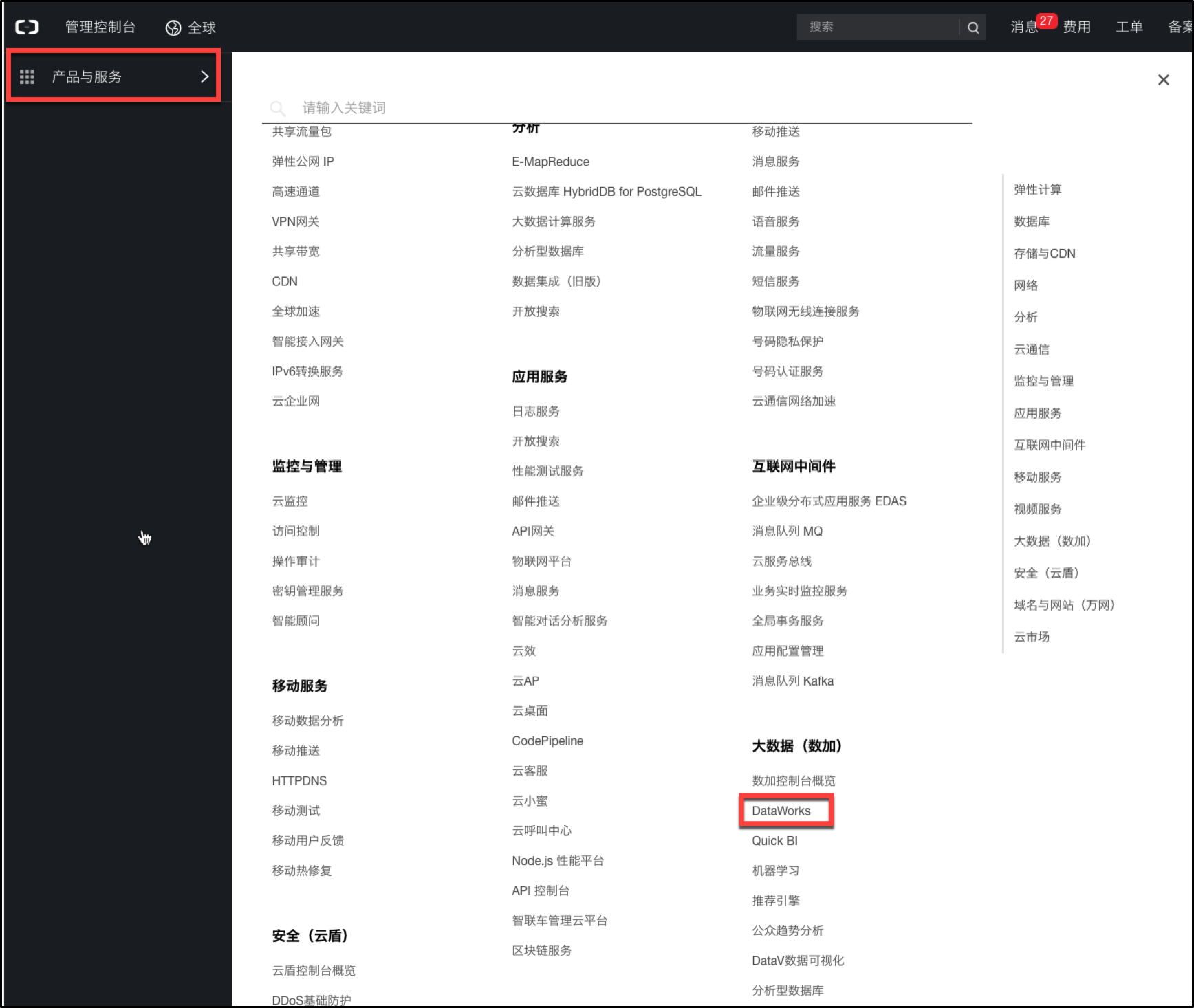
3.其中点击前往控制台按钮后将在一个新页面中连接到阿里云登录界面，其中，登陆时需要子用户名称以及子用户密码。如图所示，将子用户名称复制粘贴在登陆框中的@符号前，点击下一步。



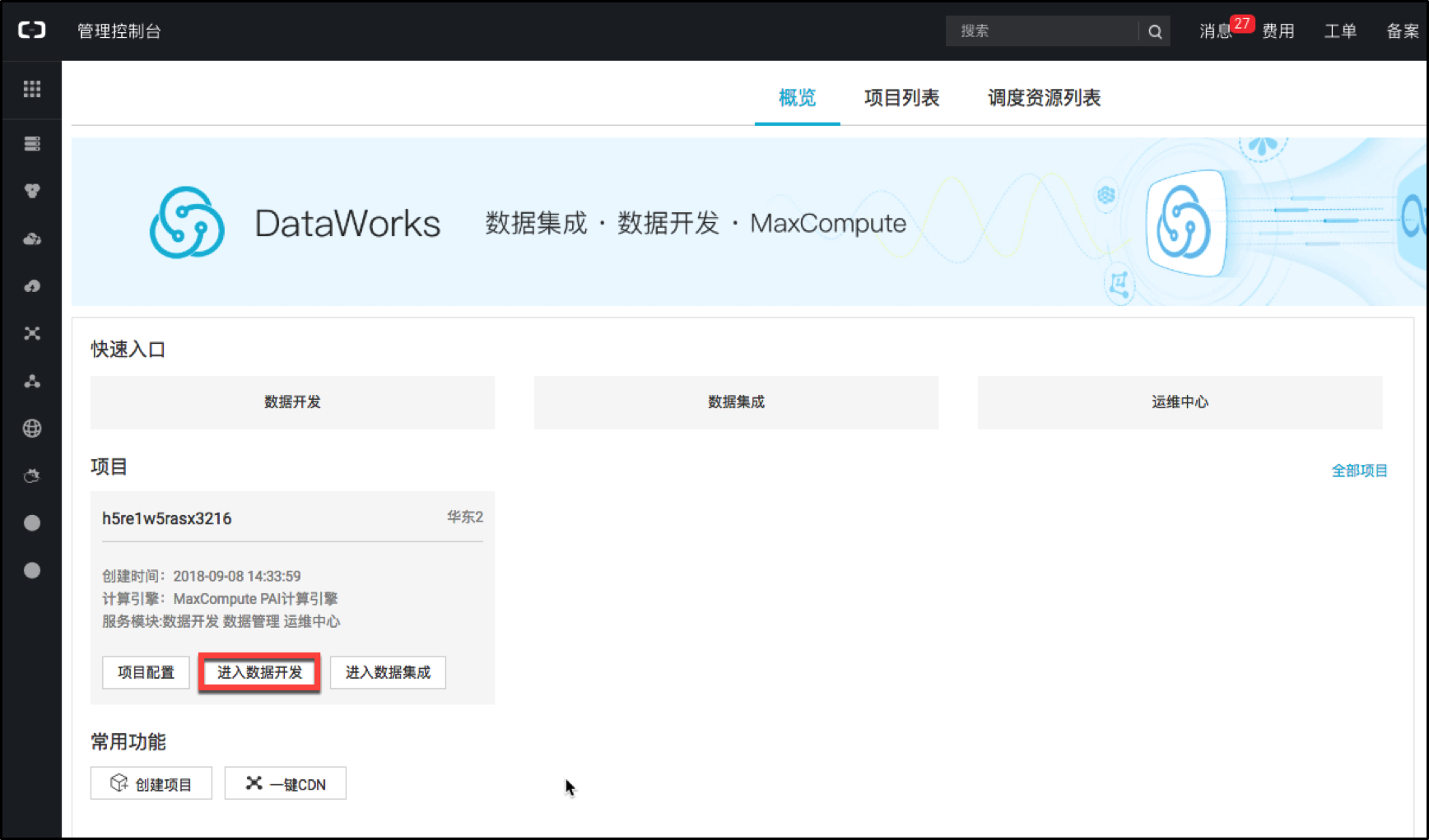
4.点击子用户密码显示按钮，将登陆密码填充进密码框中，点击登陆。



5.进入管理控制台后，本实验需要使用阿里云的DataWorks产品。在管理控制台左侧导航栏的“产品与服务”中的弹出菜单中，找到大数据（数加）下的“DataWorks”，点击它



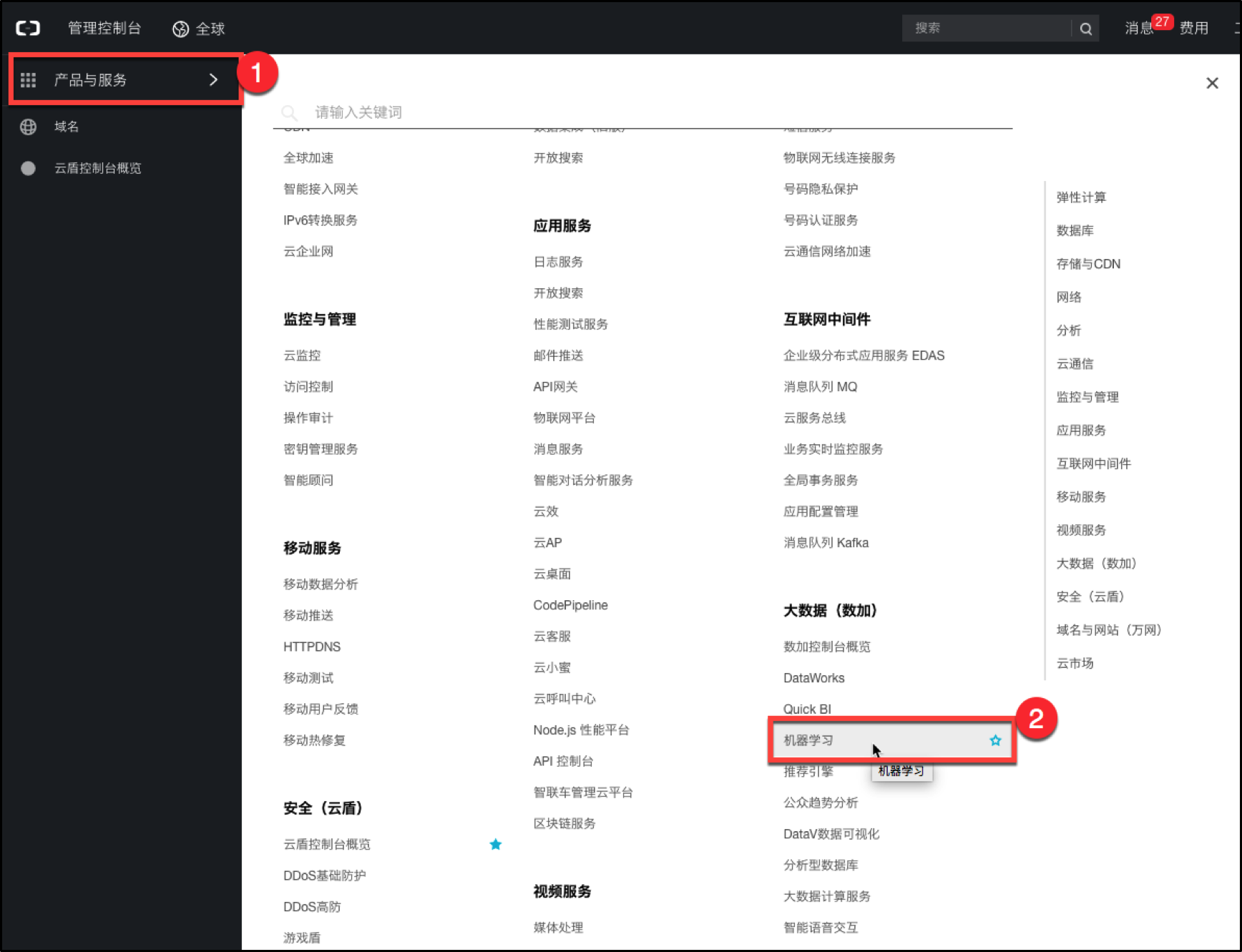
6.在此时进入到DataWorks控制台界面, 可以看到系统已经自动创建了一个项目（项目名称是自动生成的随机值），点击”进入数据开发”即可进入到在线IDE环境中。





本实验还需用到机器学习平台PAI，进入步骤如下：

7.进入控制台的界面后，接下来的实验需要进入机器学习平台。在管理控制台左侧导航栏的“产品与服务”中的弹出菜单中，找到大数据（数加）下的“机器学习”，点击它



8.可以看到系统已经自动创建了一个项目（项目名称是自动生成的随机值），单击【进入机器学习】，如下所示

