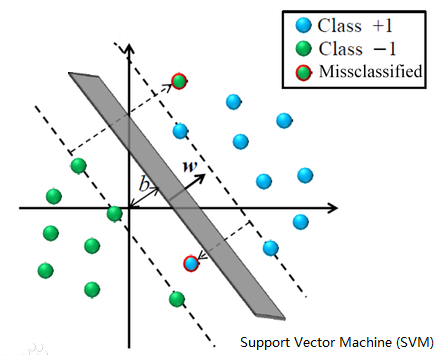
# 程序实现

## 模型选择

我们要根据数据判断出地产是否可以买，只有买和不买，二分的svm比较适合我们的数据



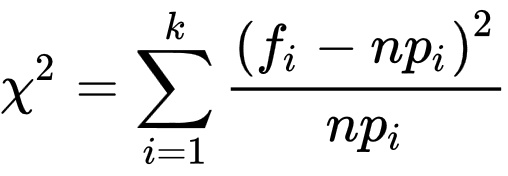
## 特征选择

**方差分析**

方差分析(Analysis of Variance，简称ANOVA)，又称“变异数分析”，是[R.A.Fisher](https://baike.baidu.com/item/R.A.Fisher/8606035)发明的，用于两个及两个以上[样本](https://baike.baidu.com/item/%E6%A0%B7%E6%9C%AC/5712132)均数差别的[显著性检验](https://baike.baidu.com/item/%E6%98%BE%E8%91%97%E6%80%A7%E6%A3%80%E9%AA%8C/8788495)。 由于各种因素的影响，研究所得的[数据](https://baike.baidu.com/item/%E6%95%B0%E6%8D%AE/33305)呈现波动状。造成波动的原因可分成两类，一是不可控的随机因素，另一是研究中施加的对结果形成影响的[可控因素](https://baike.baidu.com/item/%E5%8F%AF%E6%8E%A7%E5%9B%A0%E7%B4%A0/19223913)。

**卡方检验**

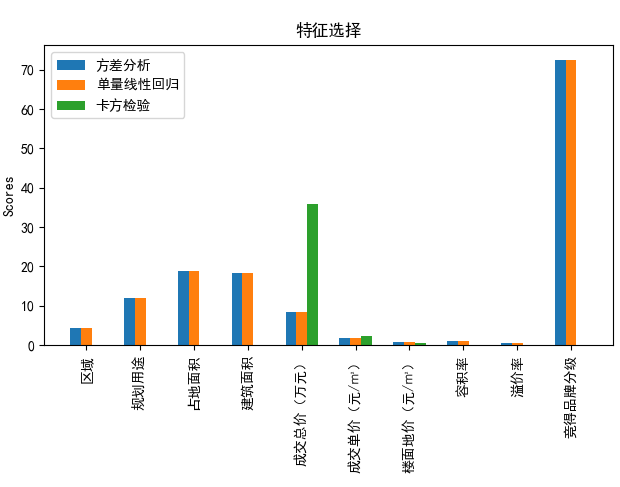
卡方检验就是统计[样本](https://baike.baidu.com/item/%E6%A0%B7%E6%9C%AC/5712132)的实际观测值与理论推断值之间的偏离程度，实际观测值与理论推断值之间的偏离程度就决定卡方值的大小，如果卡方值越大，二者偏差程度越大；反之，二者偏差越小；若两个值完全相等时，卡方值就为0，表明理论值完全符合。



**单量线性回归**

线性回归是利用数理统计中回归分析，来确定两种或两种以上变量间相互依赖的定量关系的一种统计分析方法，运用十分广泛。其表达形式为y = w'x+e，e为误差服从均值为0的正态分布。

计算相关性



根据图表

方差分析和单量线性回归的表现看，占地面积，建筑面积，竞得品牌分级相关性比较大

卡方卡方检验中成交总价相关性比较大，其次是成交单价和楼面地价

实际测试一下准确率：

特征选择让准确率得到了比较明显的提高，均值提高一个百分点，众数更是提高9个百分点。

|  |  |  |  |
| --- | --- | --- | --- |
|  | 全部 | 选定特征 | 削峰 |
|  |  |  |  |
| 平均 | 0.793333 | 0.809524 | 0.797143 |
| 标准误差 | 0.011104 | 0.011464 | 0.011204 |
| 中位数 | 0.785714 | 0.809524 | 0.809524 |
| 众数 | 0.761905 | 0.857143 | 0.714286 |
| 标准差 | 0.078518 | 0.081064 | 0.079222 |
| 方差 | 0.006165 | 0.006571 | 0.006276 |
| 峰度 | -0.34656 | -0.52846 | 0.142382 |
| 偏度 | 0.174745 | -0.07757 | 0.016909 |
| 区域 | 0.333333 | 0.333333 | 0.380952 |
| 最小值 | 0.619048 | 0.619048 | 0.571429 |
| 最大值 | 0.952381 | 0.952381 | 0.952381 |
| 求和 | 39.66667 | 40.47619 | 39.85714 |
| 观测数 | 50 | 50 | 50 |
| 最大(1) | 0.952381 | 0.952381 | 0.952381 |
| 最小(1) | 0.619048 | 0.619048 | 0.571429 |
| 置信度(95.0%) | 0.022314 | 0.023038 | 0.022515 |

## 参数调整

调整惩罚系数，学习力度，过小会“欠学习”，过大会“过学习”

我们先从0.1到20之间取一下数据

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | 0.1 | 0.5 | 1 | 5 | 7 | 10 | 15 | 20 |
| 平均 | 0.705714 | 0.810476 | 0.809524 | 0.831429 | 0.827619 | 0.818095 | 0.812381 | 0.814286 |
| 标准误差 | 0.012668 | 0.010755 | 0.011464 | 0.010731 | 0.010617 | 0.009782 | 0.010128 | 0.009835 |
| 中位数 | 0.666667 | 0.809524 | 0.809524 | 0.833333 | 0.833333 | 0.809524 | 0.809524 | 0.809524 |
| 众数 | 0.666667 | 0.809524 | 0.857143 | 0.761905 | 0.761905 | 0.857143 | 0.857143 | 0.857143 |
| 标准差 | 0.089574 | 0.076051 | 0.081064 | 0.07588 | 0.075071 | 0.069167 | 0.071613 | 0.069541 |
| 方差 | 0.008024 | 0.005784 | 0.006571 | 0.005758 | 0.005636 | 0.004784 | 0.005128 | 0.004836 |
| 峰度 | 0.5248 | -0.14136 | -0.52846 | -1.13859 | -1.17987 | -1.21911 | -1.10875 | -0.52055 |
| 偏度 | 0.711985 | -0.2531 | -0.07757 | 0.138883 | 0.083705 | -0.07712 | -0.1435 | -0.13925 |
| 区域 | 0.428571 | 0.333333 | 0.333333 | 0.238095 | 0.238095 | 0.238095 | 0.285714 | 0.285714 |
| 最小值 | 0.52381 | 0.619048 | 0.619048 | 0.714286 | 0.714286 | 0.714286 | 0.666667 | 0.666667 |
| 最大值 | 0.952381 | 0.952381 | 0.952381 | 0.952381 | 0.952381 | 0.952381 | 0.952381 | 0.952381 |
| 求和 | 35.28571 | 40.52381 | 40.47619 | 41.57143 | 41.38095 | 40.90476 | 40.61905 | 40.71429 |
| 观测数 | 50 | 50 | 50 | 50 | 50 | 50 | 50 | 50 |
| 最大(1) | 0.952381 | 0.952381 | 0.952381 | 0.952381 | 0.952381 | 0.952381 | 0.952381 | 0.952381 |
| 最小(1) | 0.52381 | 0.619048 | 0.619048 | 0.714286 | 0.714286 | 0.714286 | 0.666667 | 0.666667 |
| 置信度(95.0%) | 0.025457 | 0.021613 | 0.023038 | 0.021565 | 0.021335 | 0.019657 | 0.020352 | 0.019763 |

根据数据，查看，惩罚系数设定在10的时候，平均值，众数，最小值比较高，标准差，方差比较低，是比较合适的值

## 检验：

用今年的数据测试，准确率有所降低

大概是因为疫情造成的大企业没有拿地

|  |  |  |
| --- | --- | --- |
| **准确度** | **精确度** | **召回率** |
| 0.684211 | 0.69697 | 0.92 |

准确度接近70%

