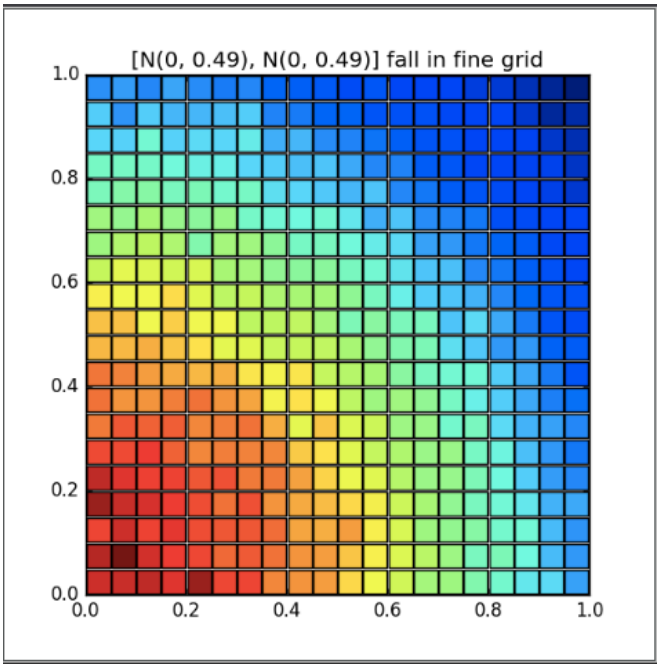
仿真实验报告

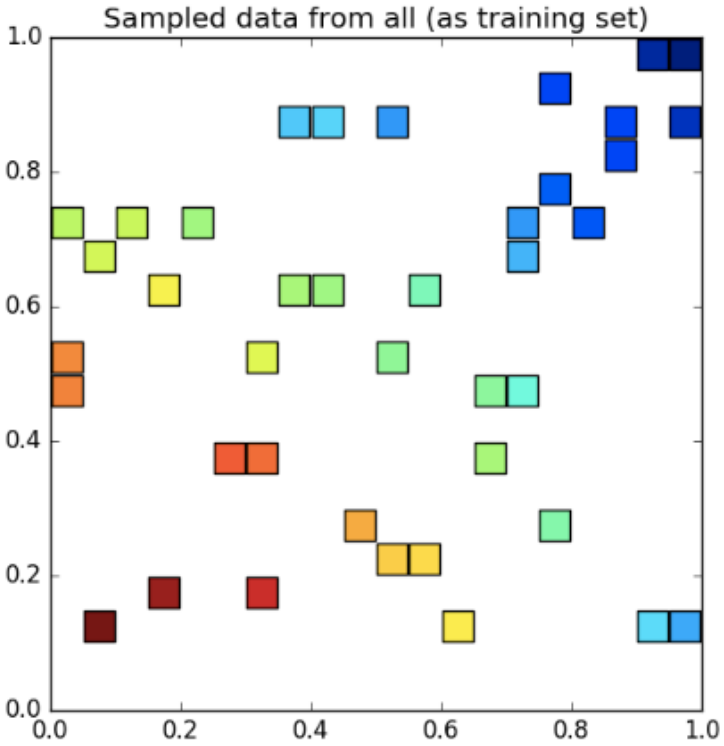
Jinwei Han, jinwei.han93@gmail.com, han550@purdue.edu

1. 数据

在0-1grid上，分成M*M个区域（以下M取20）。使用二元标准正态分布，生成大量点，使其落入M*M个区域中，统计每个小格子里的点数量。红色区域表示点的数量多，蓝色表示少。



然后为了模拟仿真的预测问题，随机抽出上面20*20个点中的10%.



2. 模型参数

我们试图使用上述的训练集预测剩余的空白格子（测试集），然后将预测值和测试集中的真实值进行比对，使用一个类似R^2的概念，去衡量模型的预测能力。

我们会分别将两个模型分别去施加在上述数据集上。

涉及到的模型参数：

参数	取值	参数意义	备注
M	20	0-1方格横竖分割出的数量	规模可调
N	M*M	总的小格子数量，400	
data_sample_count	500000	生成的仿真正态分布数据量，会被取绝对值之后构建为仿真数据	
data_var	0.49	二元分布的方差，（协方差矩阵的对角线，非对角线为0）	
phi	0.16	covariance function $\text{cov}(s_i, s_j) = e^{-\frac{d_{ij}}{\phi}}$	可调整，似乎有比0.16更好的值
sigma2	1	covariance function中使用	
training_testing_split_ratio	0.1	使用10%的数据做训练集，剩下是测试集，400*0.1=40	可调整
neighbor_relative_ratio	0.2	生成近似V^-1y矩阵时，用测试集中最近（欧式距离）的20%的测试集	可调整

3. 实验

基于上述实验数据和参数，我们对不同M进行测试，对预测能力和性能做评估。

3.1 实验1 （参数设定M=20）

验证集上的SST=180720990.889

- 模型1: 使用全体预测集中的点
  R^2 = 0.90186（SSE=17734642.9）
  V矩阵的Conditional number=24.1（40维）

- 模型2: 使用近邻的点去近似（近邻策略是固定比例的近邻预测集）
  R^2 = 0.91280（SSE=15757420.5）
  40次V矩阵的计算，平均Conditional number=13.26523（8维）

在这组参数设定下，模型2貌似更好一些。由于低维数据，运算时长基本一致（都小于1s）。我们用更高维数据来验证性能问题。

3.1 实验1 （参数设定M=60）

验证集上的SST=21301995.4

- 模型1: 使用全体预测集中的点

  R^2 = 0.912192（SSE=1870484.3）

  V矩阵的Conditional number=703.61（360维）

  耗时：29s

- 模型2: 使用近邻的点去近似（近邻策略是固定比例的近邻预测集）

  R^2 = 0.911353（SSE=1888355.4）

  360次V矩阵的计算，平均Conditional number=350.22（72维）

  耗时：18s

R^2反映的预测能力差不多，但是耗时确实是有大幅度的减少