LO
GO

K 近邻法

王晨超 2019/10/25

CONTENTS

目录

01

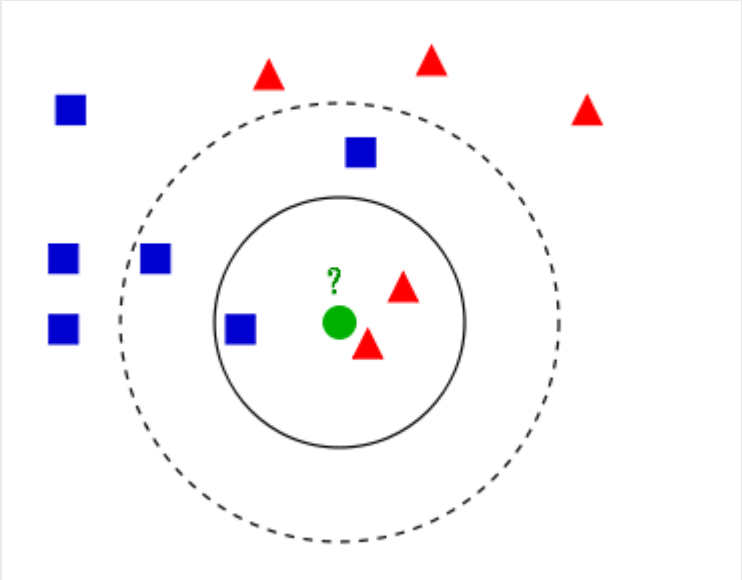
PART ONE

KNN的描述和模型

K近邻法（k-nearest neighbor，k-NN）是一种基本的分类与回归方法。《统计学习方法》只讨论了分类问题中的k近邻法。

K近邻的描述

PART ONE



K近邻的模型

PART ONE

1.距离度量

3.分类决策规则

2.K值的选择

距离度量

PART ONE

设特征空间 $\mathcal{X}$ 是 n 维实数向量空间 $\mathbf{R}^n$，$x_i, x_j \in \mathcal{X}$，$x_i = (x_i^{(1)}, x_i^{(2)}, \cdots, x_i^{(n)})^{\mathrm{T}}$，$x_j = (x_j^{(1)}, x_j^{(2)}, \cdots, x_j^{(n)})^{\mathrm{T}}$，$x_i, x_j$ 的 L_p 距离定义为

$$L_p(x_i, x_j) = \left(\sum_{l=1}^{n} | x_i^{(l)} - x_j^{(l)} |^p \right)^{\frac{1}{p}} \tag{3.2}$$

这里 $p \geqslant 1$．当 $p = 2$ 时，称为欧氏距离(Euclidean distance)，即

$$L_2(x_i, x_j) = \left(\sum_{l=1}^{n} | x_i^{(l)} - x_j^{(l)} |^2 \right)^{\frac{1}{2}} \tag{3.3}$$

当 $p = 1$ 时，称为曼哈顿距离（Manhattan distance），即

$$L_1(x_i, x_j) = \sum_{l=1}^{n} | x_i^{(l)} - x_j^{(l)} | \tag{3.4}$$

当 $p = \infty$ 时，它是各个坐标距离的最大值，即

$$L_\infty(x_i, x_j) = \max_l | x_i^{(l)} - x_j^{(l)} | \tag{3.5}$$

距离度量

PART ONE

线性比例变换法：

$$y_i = \frac{x_i}{\max(x)}$$

极差变换法：

$$y_i = \frac{x_i - \min(x)}{\max(x) - \min(x)}$$

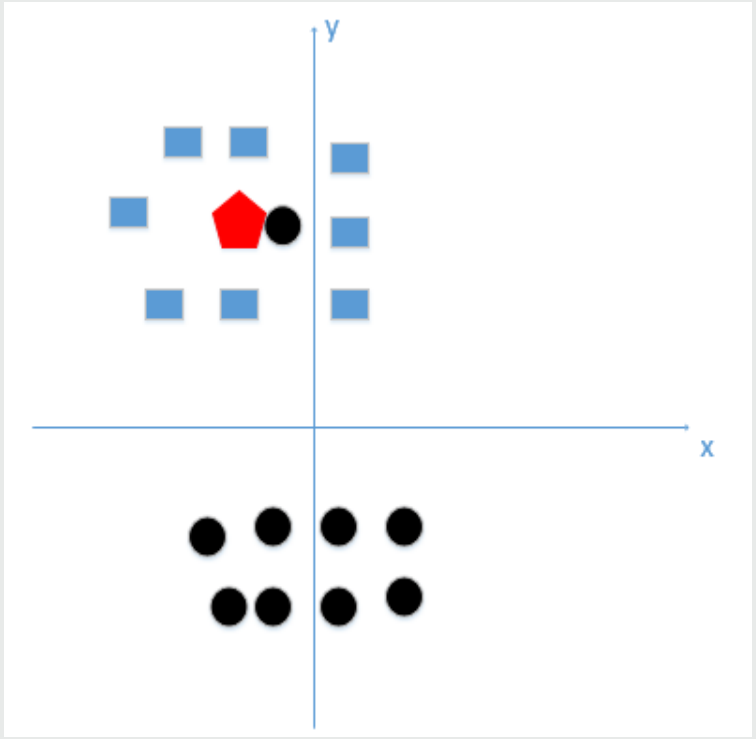
0 均值标准化（Z-score 方法）：

$$y_i = \frac{x_i - mean(x)}{\sigma}$$

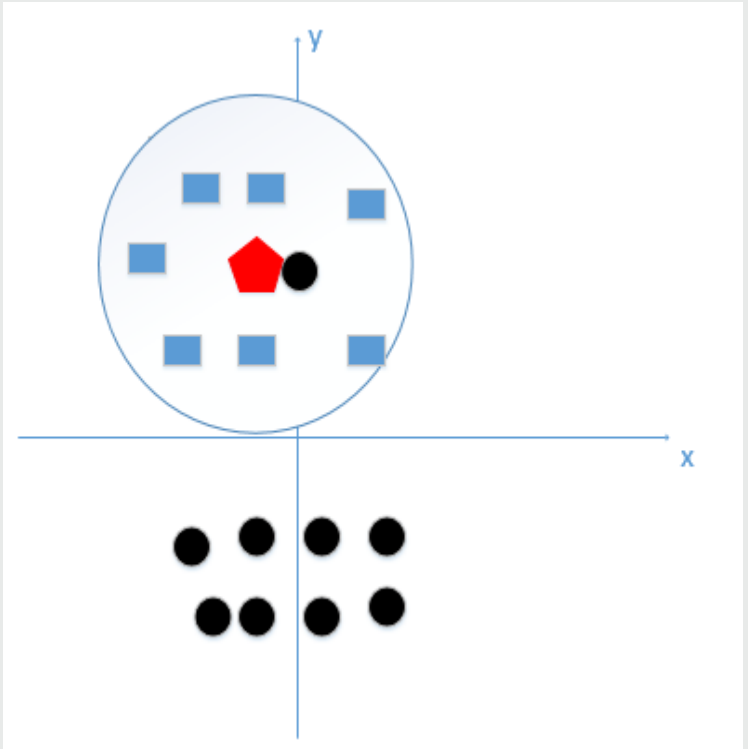
K值选择

PART ONE

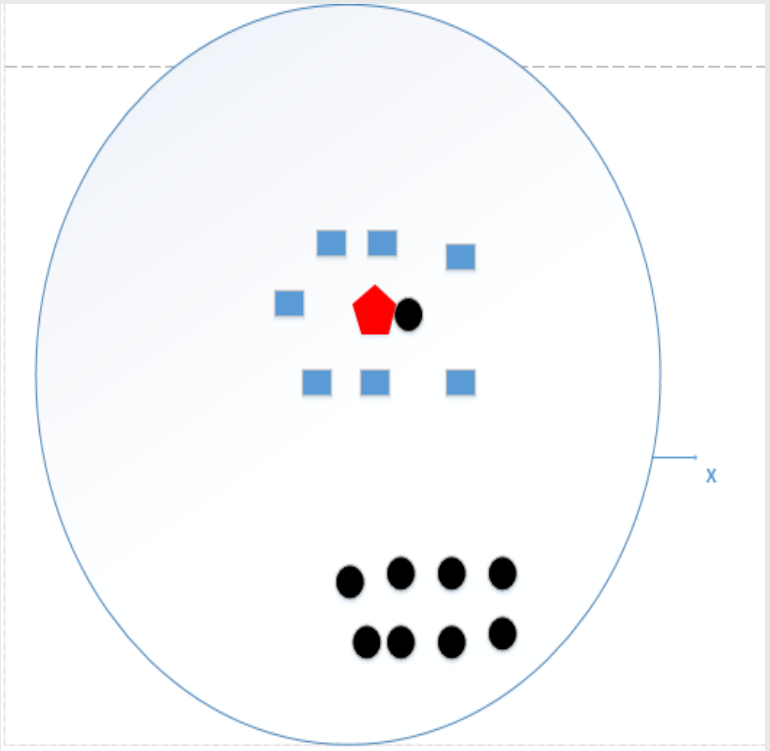
K值太小，模型复杂，过拟合



K的取值，比较合适



K值太大，模型简单，欠拟合



分类决策规则

PART ONE

对给定的实例 $x \in \chi$，其最近邻的 k 个训练实例点构成集合 $N_k(x)$。如果涵盖 $N_k(x)$ 的区域的类别是 c_j，那么误分类是

$$\frac{1}{k}\sum_{x_i \in N_k(x)} I(y_i \neq c_j) = 1 - \frac{1}{k}\sum_{x_i \in N_k(x)} I(y_i = c_j)$$

要使误分类率最小即经验风险最小，就要使 $\sum_{x_i \in N_k(x)} I(y_i = c_j)$ 最大，所以**多数表决规则等价于经验风险最小化**。

02

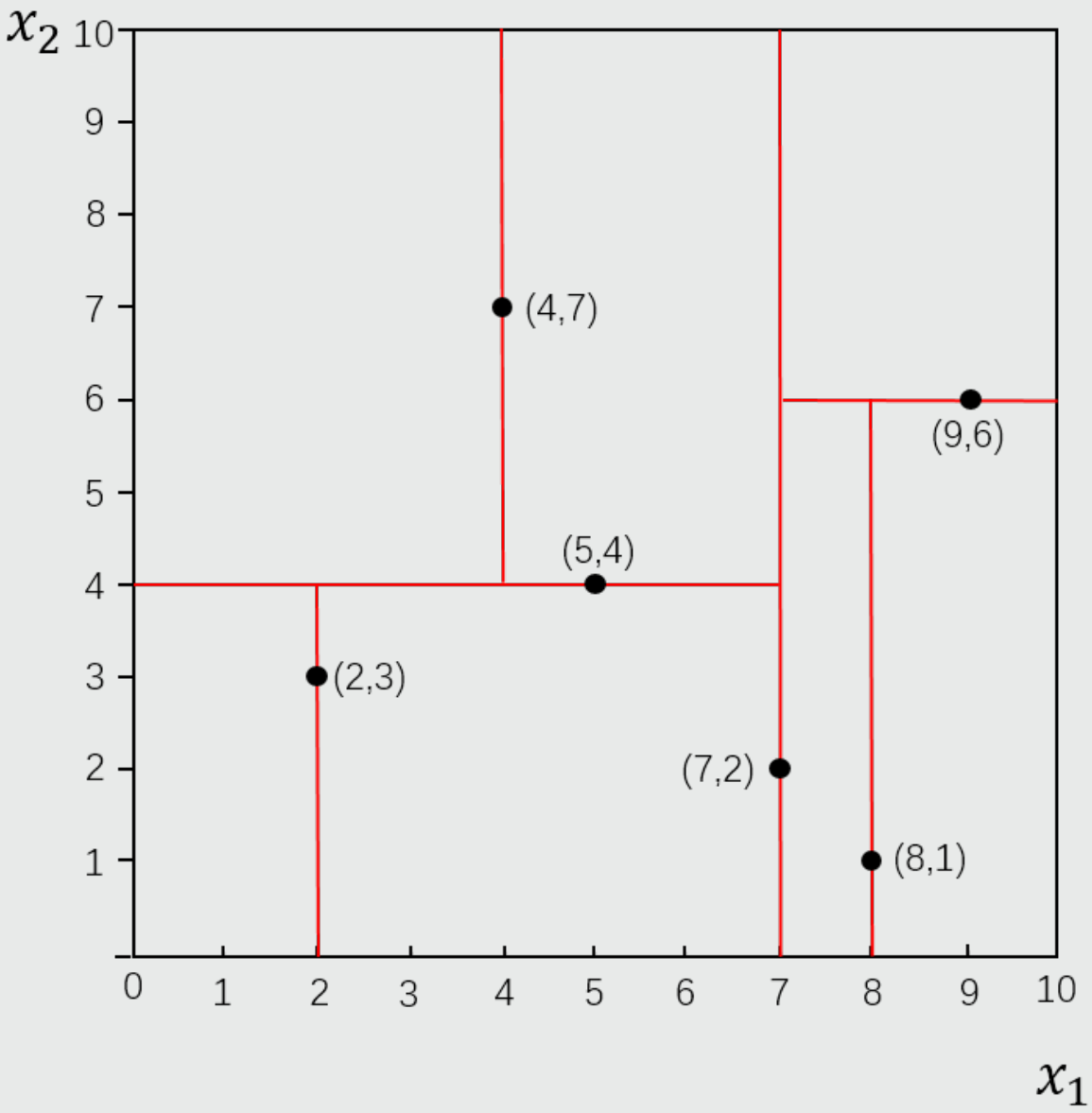
PART TWO

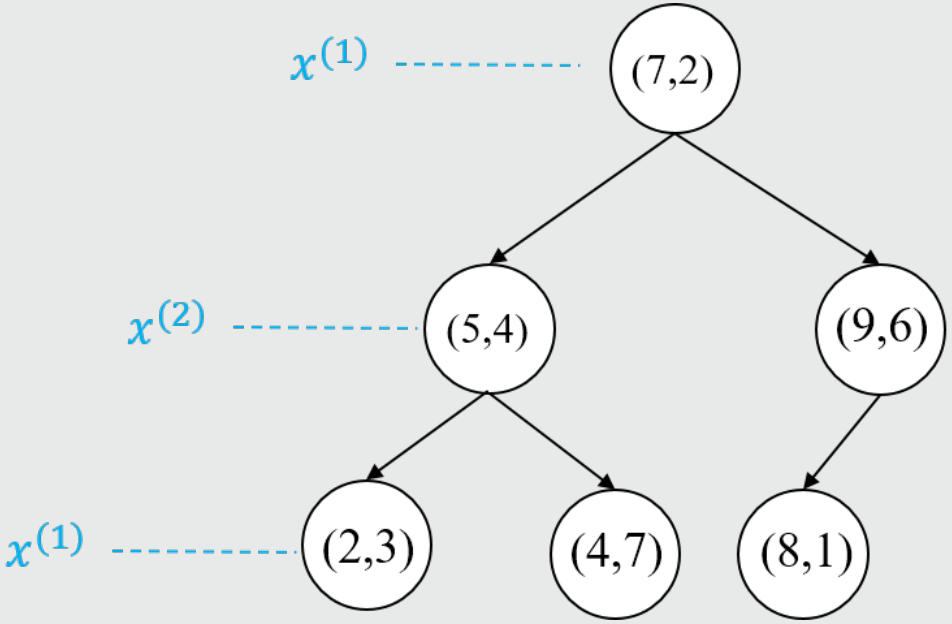
KNN的实现（KD树）

K近邻最简单的实现方法就是线性扫描，计算输入实例与每一个训练实例的距离，挑出最大的k个，时间复杂度是O(n)，当训练集很大的时候，计算是十分耗时的。而用kd树可以效率

KD树的构造

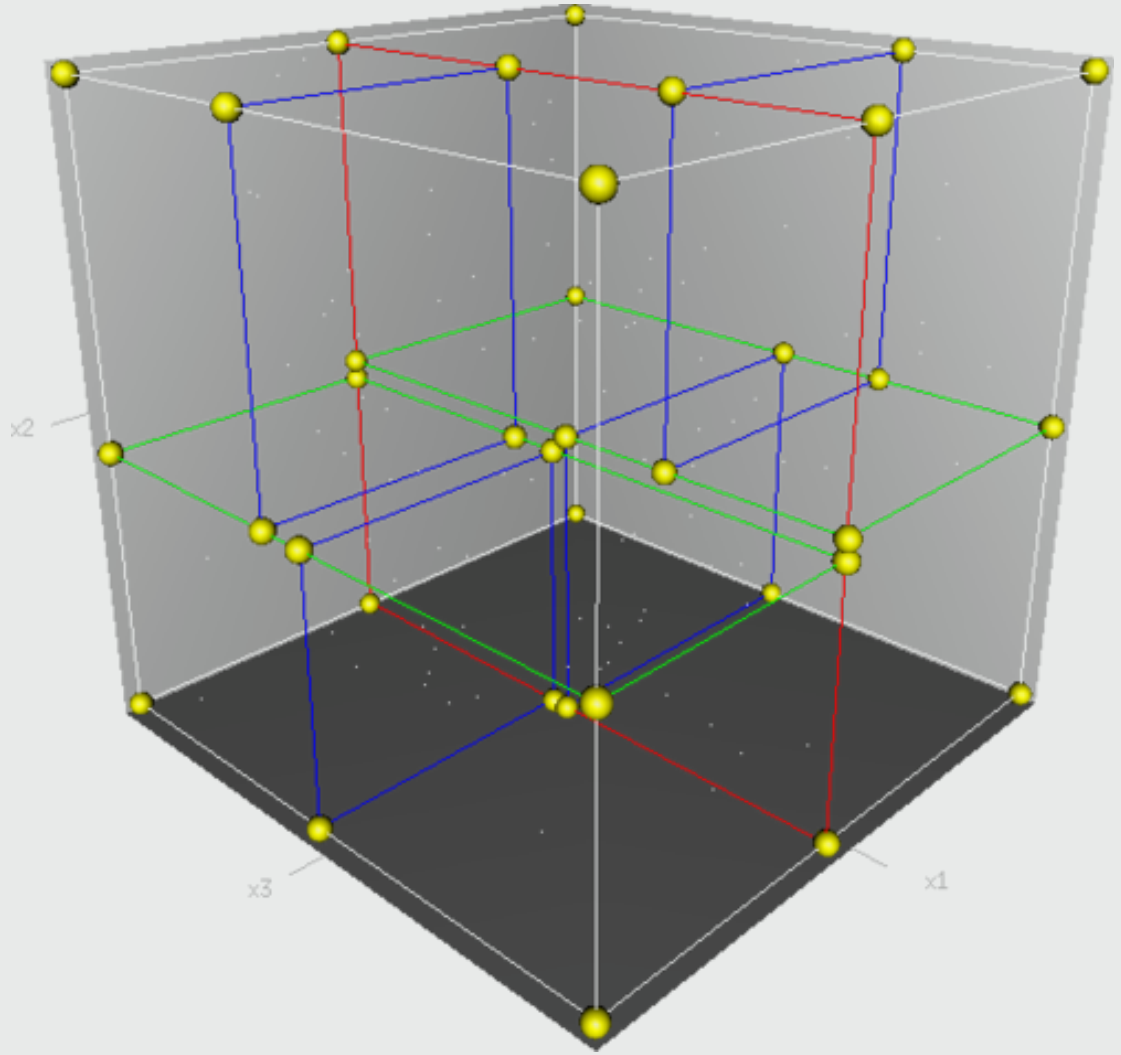
PART TWO





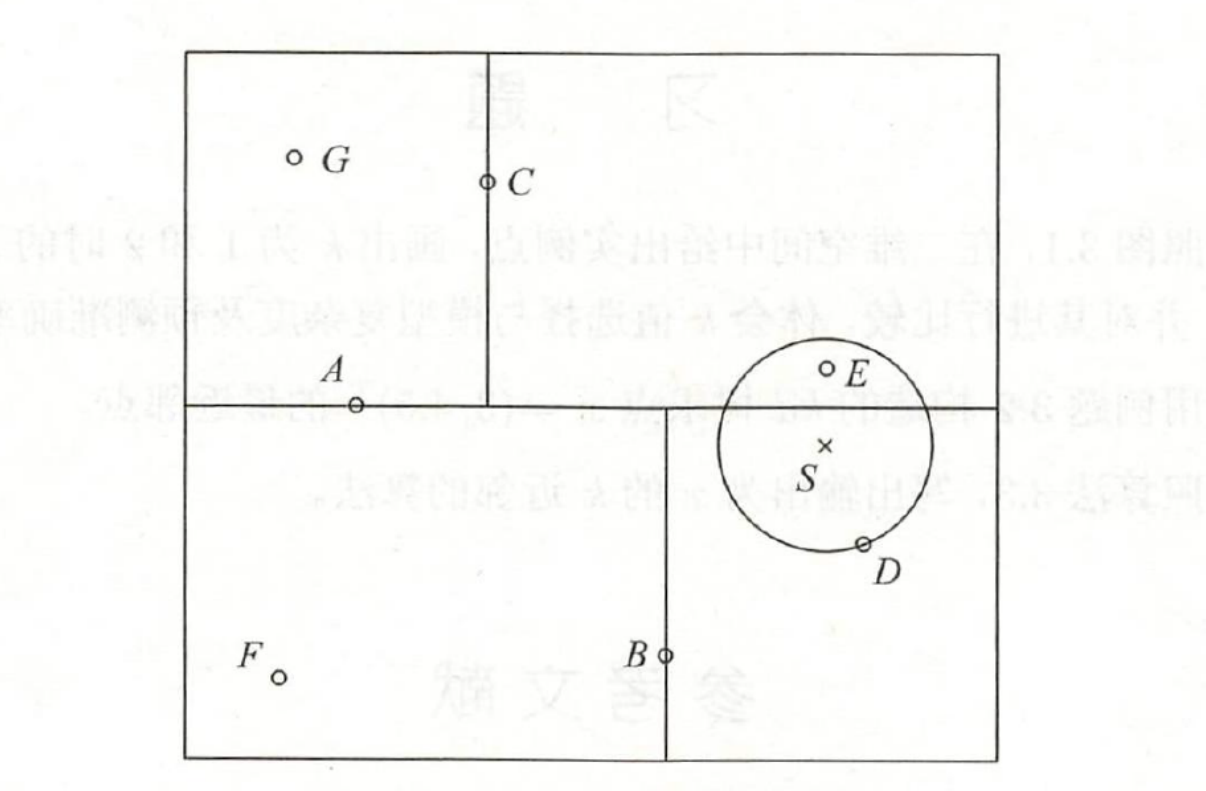
KD树的构造

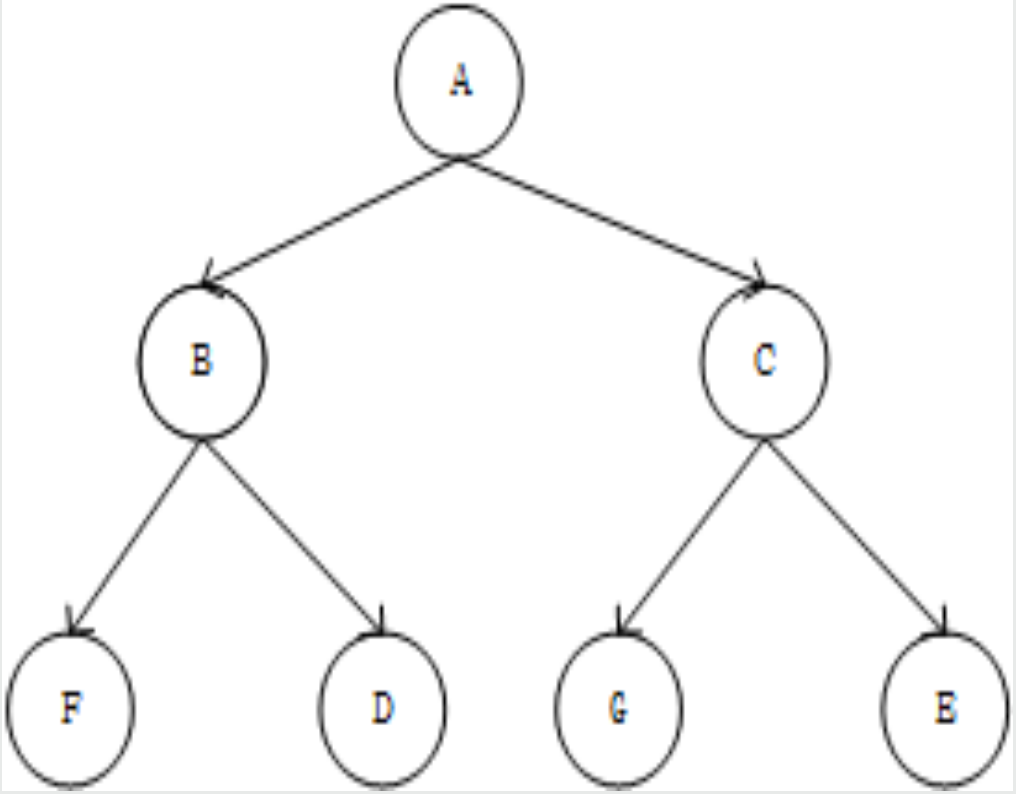
PART TWO



KD树的搜索（最近邻）

PART TWO





KD树的搜索（K近邻）

PART TWO

03

PART THREE

用K近邻对约会对象分类

有一位女士一直使用在线约会网站寻找适合自己的约会对象。尽管约会网站会推荐不同的任选，但她并不是喜欢每一个人。现根据她以前约会过的男士的数据，来判断新的以为男士是否她喜欢

1	40920	8.326976	0.953952	largeDoses
2	14488	7.153469	1.673904	smallDoses
3	26052	1.441871	0.805124	didntLike 不喜欢
4	75136	13.147394	0.428964	didntLike
5	38344	1.669788	0.134296	didntLike
6	72993	10.141740	1.032955	didntLike
7	35948	6.830792	1.213192	largeDoses 极具魅力
8	42666	13.276369	0.543880	largeDoses
9	67497	8.631577	0.749278	didntLike
10	35483	12.273169	1.508053	largeDoses
11	50242	3.723498	0.831917	didntLike
12	63275	8.385879	1.669485	didntLike
13	5569	4.875435	0.728658	smallDoses 魅力一般
14	51052	4.680098	0.625224	didntLike
15	77372	15.299570	0.331351	didntLike
16	43673	1.889461	0.191283	didntLike
17	61364	7.516754	1.269164	didntLike
18	69673	14.239195	0.261333	didntLike
19	15669	0.000000	1.250185	smallDoses
20	28488	10.528555	1.304844	largeDoses
21	6487	3.540265	0.822483	smallDoses
22	37708	2.991551	0.833920	didntLike
23	22620	5.297865	0.638306	smallDoses
24	28782	6.593803	0.187108	largeDoses
25	19739	2.816760	1.686209	smallDoses
26	36788	12.458258	0.649617	largeDoses
27	5741	0.000000	1.656418	smallDoses
28	28567	9.968648	0.731232	largeDoses

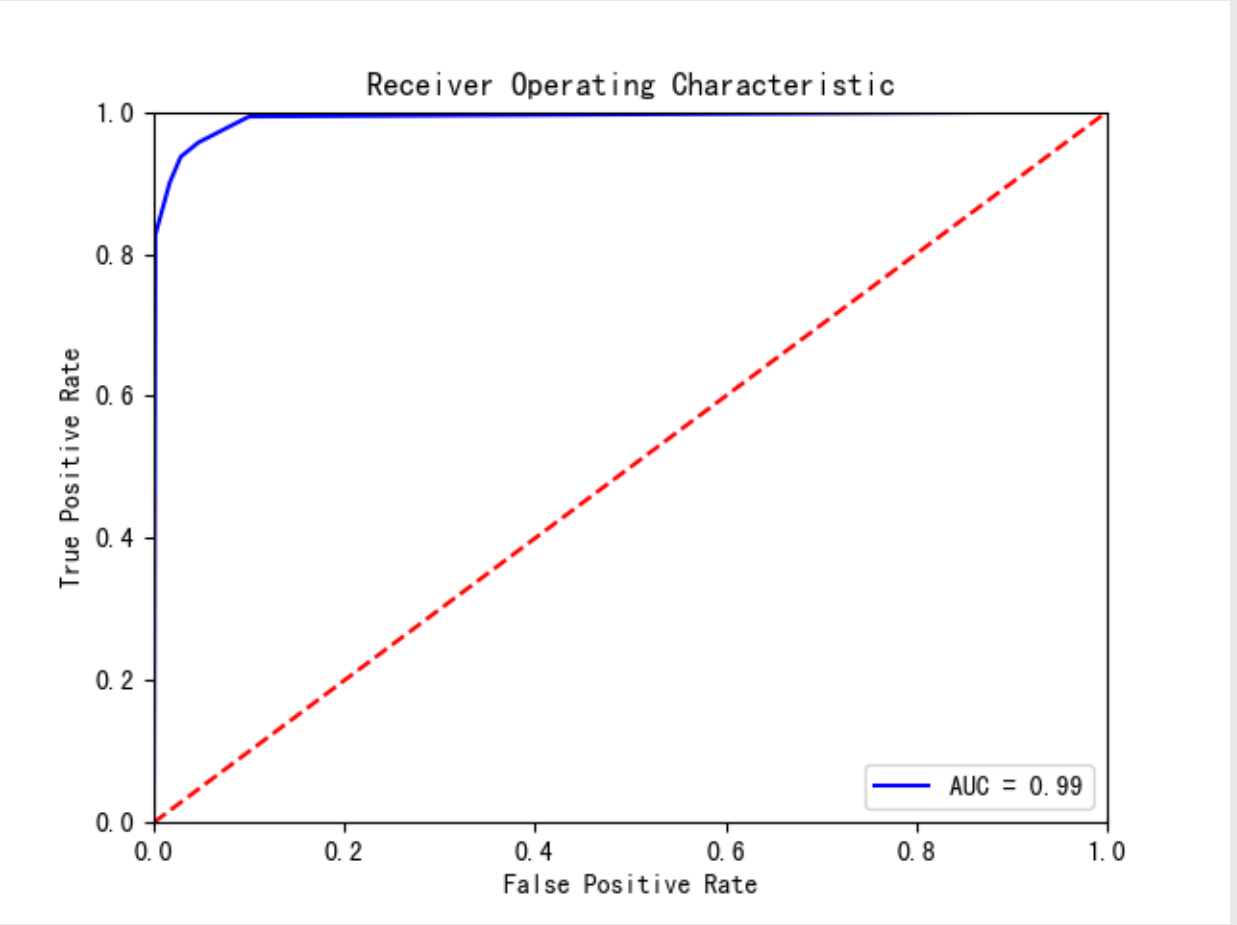
数据集

收集的样本数据主要包含以下3种特征：
1.每年获得的飞行常客里程数
2.玩视频游戏所消耗时间百分比
3.每周消费的冰淇淋公升数

根据文本中标记的喜欢的程度进行处理,0代表不喜欢,1代表魅力一般,2代表极具魅力

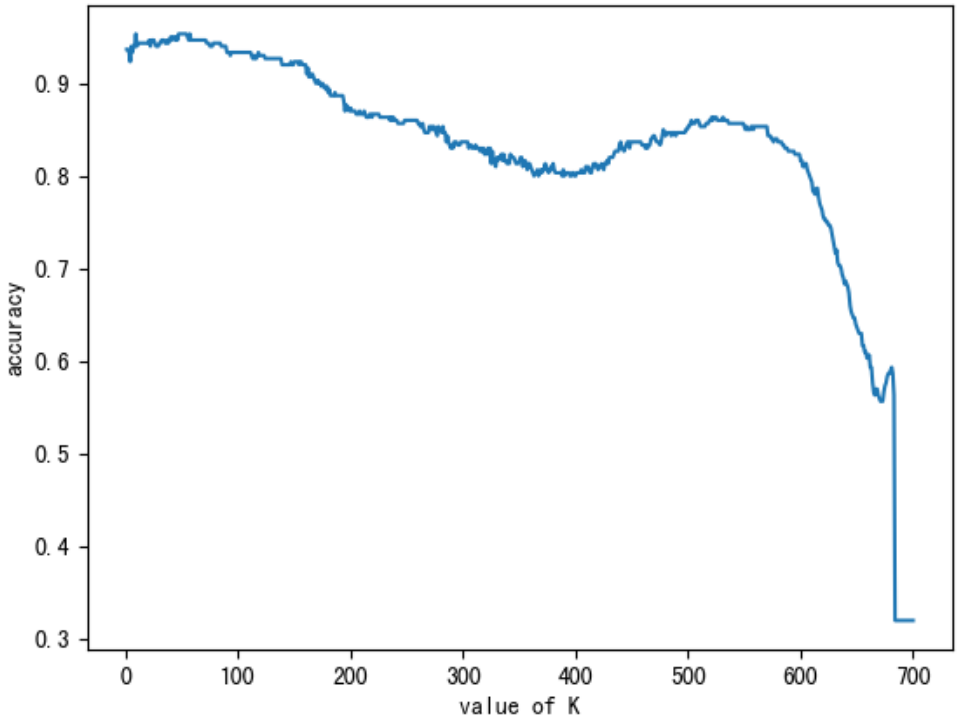
ROC图

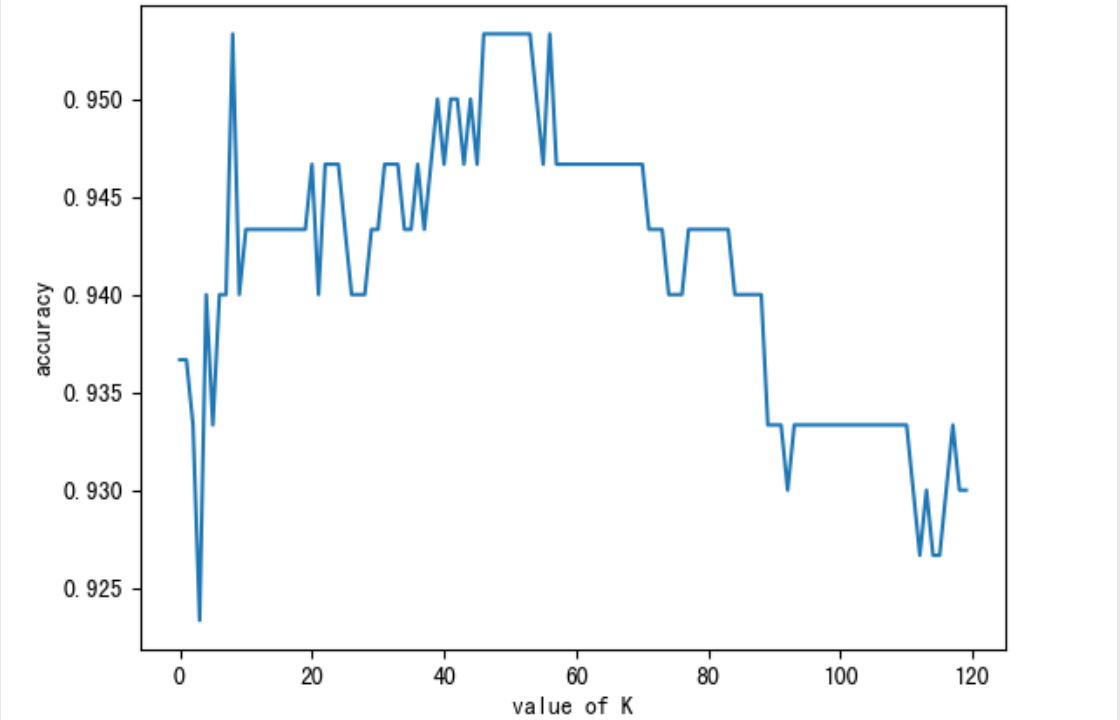
PART THREE



不同K对于准确率的影响

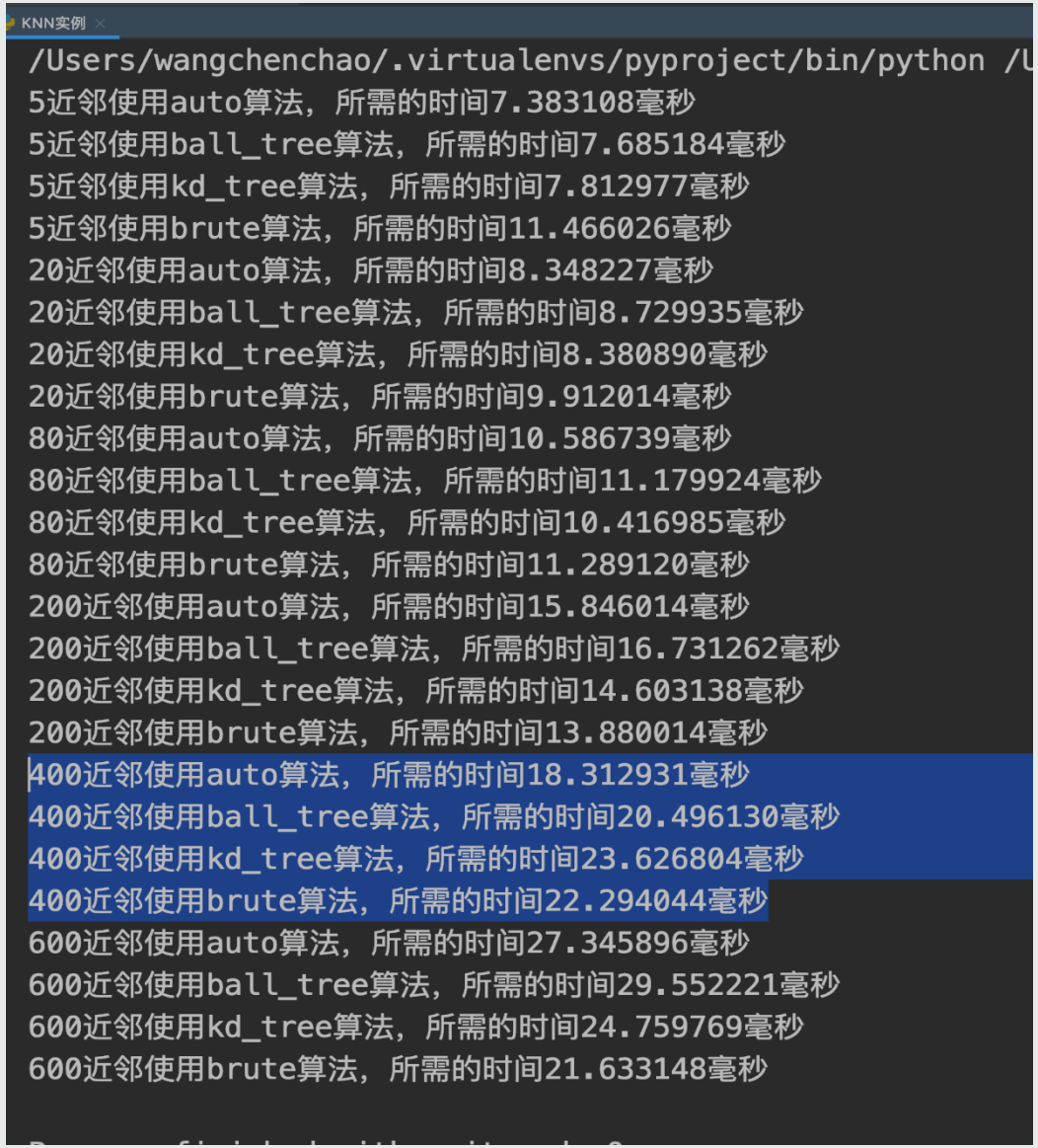
PART THREE





不同选择K近邻的算法对于时间的影响

PART THREE



THANKS

谢　谢　聆　听