Chapter 7 – K-Nearest-Neighbor

Instructor: Zach Zhizhong ZHOU,
Shanghai Jiao Tong University
主讲教师：周志中，上海交通大学

Data Mining for Business Intelligence

Shmueli, Patel & Bruce

Characteristics 特征

Data-driven, not model-driven 数据驱动而非模型驱动

Makes no assumptions about the data 对数据没有任何假设

Basic Idea 基本思路

For a given record to be classified, identify nearby records 对于一个需要进行分类的记录，找到附近的记录。

“Near” means records with similar predictor values $X_1, X_2, \cdots X_p$ “附近”指有类似预测因子取值的记录。

Classify the record as whatever the predominant class is among the nearby records (the “neighbors”) 把该记录归入附近记录中占优的分类当中。

How to measure “nearby”？如何衡量“附近”？

The most popular distance measure is
Euclidean distance 最常用的是“欧拉距离”

$$\sqrt{(x_1 - u_1)^2 + (x_2 - u_2)^2 + \cdots + (x_p - u_p)^2}$$

Choosing k 选择k

K is the number of nearby neighbors to be used to classify the new record　k是附近记录中用来对新纪录进行分类的记录数目。

K=1 means use the single nearest record

K=5 means use the 5 nearest records

Typically choose that value of *k* which has lowest error rate in validation data 通常是找到使得验证数据集的错误率最小的k。

Low *k* vs. High *k* 比较k的取值

Low values of *k* (1, 3, …) capture local structure in data (but also noise) 取值小的k可以捕捉到记录附近的分类结构（但同时也 捕捉到噪声）

High values of *k* provide more smoothing, less noise, but may miss local structure 取值高的k提供更为平滑更少噪声的估计但可能没有捕捉到记录附近的分类结构。

Note: the extreme case of k = n (i.e., the entire data set) is the same as the "naïve rule" (classify all records according to majority class) 极端情况是k=n（整个数据集）。这就得到与朴素贝叶斯准则相同的结果（把所有记录归入主要的分类中）。

Example: Riding Mowers 刈草车

Data: 24 households classified as owning or not owning riding mowers 24个家庭被划分为拥有刈草车或者未拥有刈草车。

Predictors: Income, Lot Size 预测因子：收入、房屋面积

Income	Lot_Size	Ownership
60.0	18.4	owner
85.5	16.8	owner
64.8	21.6	owner
61.5	20.8	owner
87.0	23.6	owner
110.1	19.2	owner
108.0	17.6	owner
82.8	22.4	owner
69.0	20.0	owner
93.0	20.8	owner
51.0	22.0	owner
81.0	20.0	owner
75.0	19.6	non-owner
52.8	20.8	non-owner
64.8	17.2	non-owner
43.2	20.4	non-owner
84.0	17.6	non-owner
49.2	17.6	non-owner
59.4	16.0	non-owner
66.0	18.4	non-owner
47.4	16.4	non-owner
33.0	18.8	non-owner
51.0	14.0	non-owner
63.0	14.8	non-owner

XLMiner Output

For each record in validation data (6 records) XLMiner finds neighbors amongst training data (18 records).

The record is scored for k=1, k=2, … k=18.

Best k appears to be k=8.

k = 9, k = 10, k=14 also share low error rate, but best to choose lowest k.

Value of k	% Error Training	% Error Validation	
1	0.00	33.33	
2	16.67	33.33	
3	11.11	33.33	
4	22.22	33.33	
5	11.11	33.33	
6	27.78	33.33	
7	22.22	33.33	
8	22.22	16.67	<--- Best k
9	22.22	16.67	
10	22.22	16.67	
11	16.67	33.33	
12	16.67	16.67	
13	11.11	33.33	
14	11.11	16.67	
15	5.56	33.33	
16	16.67	33.33	
17	11.11	33.33	
18	50.00	50.00	

Using K-NN for Prediction (for Numerical Outcome)
KNN用在数值型预测模型

- Instead of “majority vote determines class” use average of response values 不使用“多数投票决定分类”而是使用预测目标平均值。

- May be a weighted average, weight decreasing with distance 可以使用加权平均，权重随距离变小。

Advantages 好处

- Simple 简单

- No assumptions required about Normal distribution, etc. 不要求正态分布

- Effective at capturing complex interactions among variables without having to define a statistical model 可以有效捕捉到预测变量中复杂的相互关系而无需定义一个统计模型。

Shortcomings 坏处

- Required size of training set increases exponentially with # of predictors, p 训练集所要求的数据量随预测因子数指数增长。

  This is because expected distance to nearest neighbor increases with p (with large vector of predictors, all records end up “far away” from each other) 这是因为期望距离随预测因子数目增加而增加（如果有很多预测因子，最终所有记录彼此之间都相隔很远）
- In a large training set, it takes a long time to find distances to all the neighbors and then identify the nearest one(s) 大的训练数据集要花很长时间才能算出距离并找到最近的k个点。计算量大。
- These constitute “curse of dimensionality” 这就是所谓的“维数灾难”。

Dealing with the Curse 处理维数灾难

- Reduce dimension of predictors (e.g., with PCA)
  对预测因子进行降维（主成分分析）

- Computational shortcuts that settle for “almost nearest neighbors” 想办法降低运算量，可以接受“几乎是最近的临近点”而不一定去找到最近的临近点。

Summary 总结

- Find distance between record-to-be-classified and all other records 找到将被分类的记录与其他记录的距离。
- Select k-nearest records 选择k个距离被分类记录最接近的记录
  - Classify it according to majority vote of nearest neighbors
  - Or, for prediction, take the as average of the nearest neighbors 根据多数表决原则对记录进行分类，在预测数值时则使用临近点的平均值作为预测值。
- “Curse of dimensionality” – need to limit # of predictors 维数灾难 需要对预测因子数目进行控制。