Chapter 4 – Dimension Reduction

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

Exploring the data 探索数据

Statistical summary of data: common metrics
数据统计汇总：常用指标

- Average 均值
- Median 中位数
- Minimum 最小值
- Maximum 最大值
- Standard deviation 标准方差
- Counts & percentages 计数和百分比

Summary Statistics – Boston Housing
统计汇总 波士顿房屋

	Average	Median	Min	Max	Std	Count	Countblank
CRIM	3.61	0.26	0.01	88.98	8.60	506	0
ZN	11.36	0.00	0.00	100.00	23.32	506	0
INDUS	11.14	9.69	0.46	27.74	6.86	506	0
CHAS	0.07	0.00	0.00	1.00	0.25	506	0
NOX	0.55	0.54	0.39	0.87	0.12	506	0
RM	6.28	6.21	3.56	8.78	0.70	506	0
AGE	68.57	77.50	2.90	100.00	28.15	506	0
DIS	3.80	3.21	1.13	12.13	2.11	506	0
RAD	9.55	5.00	1.00	24.00	8.71	506	0
TAX	408.24	330.00	187.00	711.00	168.54	506	0
PTRATIO	18.46	19.05	12.60	22.00	2.16	506	0
B	356.67	391.44	0.32	396.90	91.29	506	0
LSTAT	12.65	11.36	1.73	37.97	7.14	506	0
MEDV	22.53	21.20	5.00	50.00	9.20	506	0

Correlations Between Pairs of Variables: Correlation Matrix from Excel 相关系数矩阵

	PTRATIO	*B*	*LSTAT*	*MEDV*
PTRATIO	1			
B	-0.17738	1		
LSTAT	0.374044	-0.36609	1	
MEDV	-0.50779	0.333461	-0.73766	1

Summarize Using Pivot Tables 数据透视表

Counts & percentages are useful for summarizing categorical data
计数和百分比用来总结类别型数据尤为有用

Boston Housing example:

- 471 neighborhoods border the Charles River (1)
- 35 neighborhoods do not (0)

Count of MEDV	
CHAS	Total
0	471
1	35
Grand Total	506

Pivot Tables 数据透视表

Averages are useful for summarizing grouped numerical data 均值用来汇总数值型数据尤为有用

Boston Housing example: Compare average home values in neighborhoods that border Charles River (1) and those that do not (0)

Average of MEDV	
CHAS	Total
0	22.09
1	28.44
Grand Total	22.53

Pivot Tables 数据透视表

Group by multiple criteria: 根据不同规则进行分组

- By # rooms and location 根据房间数和位置分组
- E.g., neighborhoods on the Charles with 6-7 rooms have average house value of 25.92 ($000)

Average of MEDV	CHAS		
RM	0	1	Grand Total
3-4	25.30		25.30
4-5	16.02		16.02
5-6	17.13	22.22	17.49
6-7	21.77	25.92	22.02
7-8	35.96	44.07	36.92
8-9	45.70	35.95	44.20
Grand Total	22.09	28.44	22.53

Correlation Analysis 相关性分析

Below: Correlation matrix for portion of Boston Housing data 部分波士顿房屋数据的相关系数矩阵
Shows correlation between variable pairs 展示变量对之间的相关系数

	CRIM	*ZN*	*INDUS*	*CHAS*	*NOX*	*RM*
CRIM	1					
ZN	-0.20047	1				
INDUS	0.406583	-0.53383	1			
CHAS	-0.05589	-0.0427	0.062938	1		
NOX	0.420972	-0.5166	0.763651	0.091203	1	
RM	-0.21925	0.311991	-0.39168	0.091251	-0.30219	1

Reducing Categories 减少类型数目

- A single categorical variable with *m* categories is typically transformed into *m-1* dummy variables 一个类别型变量如果有m个值，通常转换成m-1个虚拟变量。

- Each dummy variable takes the values 0 or 1 每个虚拟变量根据是否属于该类别取值0或者1

  0 = “no” for the category 如不在该类别取值为0

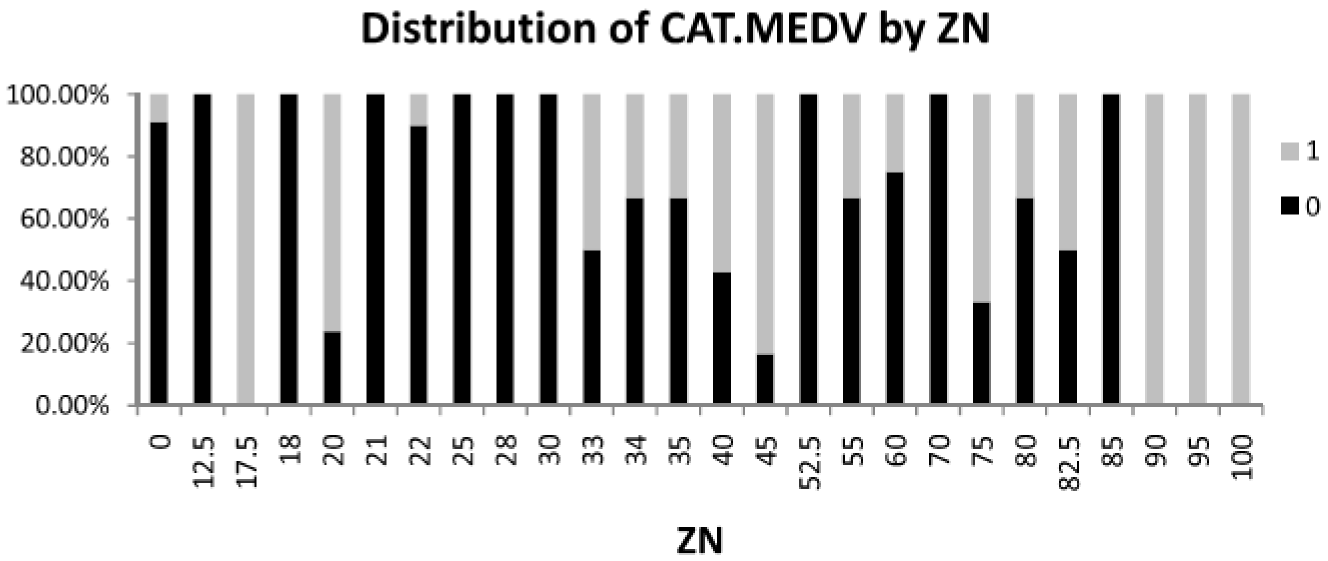
  1 = “yes” 反之取值为1

Reducing Categories 减少类型数目

- Problem: Can end up with too many variables 问题：是否会导致很多个变量
- Solution: Reduce by combining categories that are close to each other 解决：将多个类似的类型捆绑在一起组成一个新的类型。
- Use pivot tables to assess outcome variable sensitivity to the dummies 使用数据透视表评估新的类型对虚拟变量的敏感程度
- Exception: Naïve Bayes can handle categorical variables without transforming them into dummies 例外：朴素贝叶斯方法可以处理类别型数据，无需将它们转为虚拟变量。

Combining Categories 合并类型

Many zoning categories are the same or similar with respect to CATMEDV



Principal Components Analysis 主成分分析法

Goal: Reduce a set of numerical variables. 目标是减少一系列数值型变量的个数

The idea: Remove the overlap of information between these variable. ["Information" is measured by the sum of the variances of the variables.] 思想：移除去变量间重合的信息。（信息用变量的总方差度量）

Final product: A smaller number of numerical variables that contain most of the information 最终产品：更少数目的变量，但保留了原变量所包含的绝大多数信息。

How does PCA do this? 主成分分析法如何实现?

- Create new variables that are linear combinations of the original variables (i.e., they are weighted averages of the original variables). 建立新的变量，是原来变量的线性组合（例如，是原有变量的加权平均值）。

- These linear combinations are uncorrelated (no information overlap), and only a few of them contain most of the original information. 这些线性组合生成的新变量没有相关性（没有信息重合部分），所以少数的几个变量可以包含原有变量的大多数信息。

- The new variables are called *principal components*. 新的变量被称为主成分。

Example – Breakfast Cereals 早餐麦片

name	mfr	type	calories	protein	…	rating
100%_Bran	N	C	70	4	…	68
100%_Natural_Bran	Q	C	120	3	…	34
All-Bran	K	C	70	4	…	59
All-Bran_with_Extra_Fiber	K	C	50	4	…	94
Almond_Delight	R	C	110	2	…	34
Apple_Cinnamon_Cheerios	G	C	110	2	…	30
Apple_Jacks	K	C	110	2	…	33
Basic_4	G	C	130	3	…	37
Bran_Chex	R	C	90	2	…	49
Bran_Flakes	P	C	90	3	…	53
Cap'n'Crunch	Q	C	120	1	…	18
Cheerios	G	C	110	6	…	51
Cinnamon_Toast_Crunch	G	C	120	1	…	20

Description of Variables 变量描述

Name: name of cereal

mfr: manufacturer

type: cold or hot

calories: calories per serving

protein: grams

fat: grams

sodium: mg.

fiber: grams

carbo: grams complex carbohydrates

sugars: grams

potass: mg.

vitamins: % FDA rec

shelf: display shelf

weight: oz. 1 serving

cups: in one serving

rating: consumer reports

Consider calories & ratings 考虑卡路里和评级

- Total variance (= "information") is sum of individual variances: 379.63 + 197.32 总方差

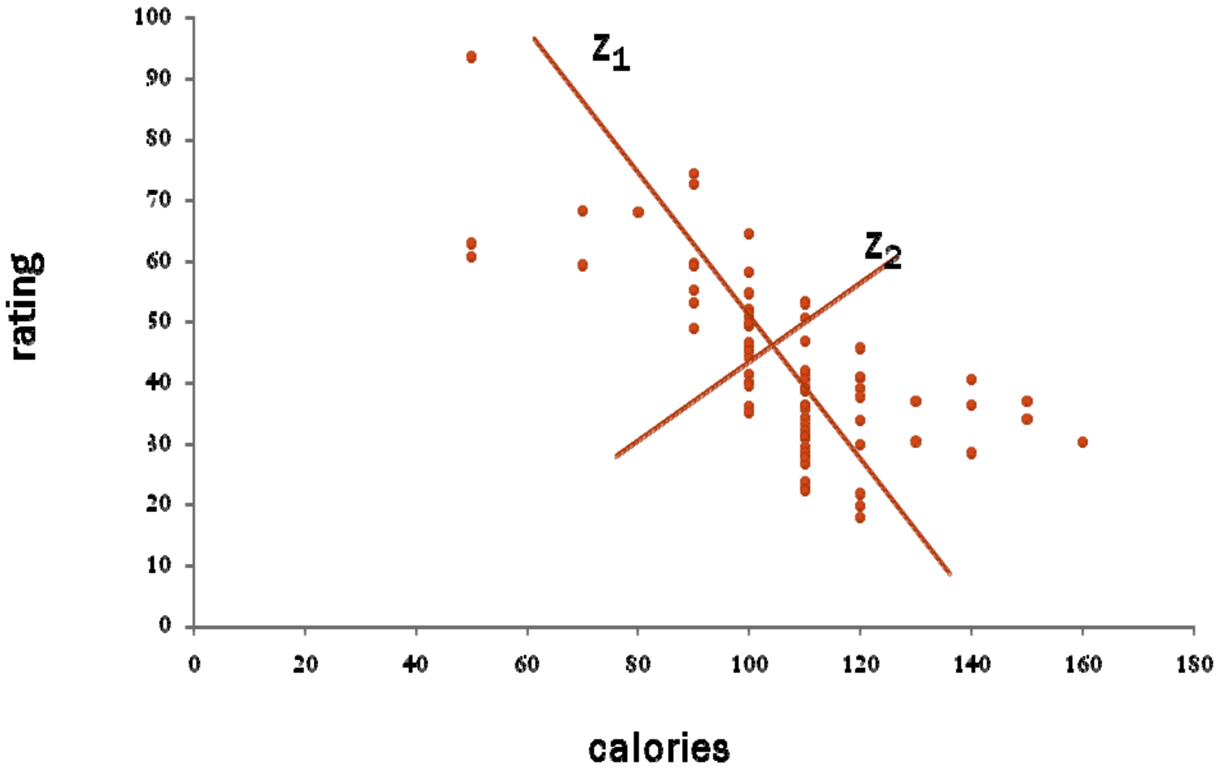
	calories	ratings
calories	379.63	-189.68
ratings	-189.68	197.32

- Calories accounts for 379.63/197.32 = 66% 卡路里占总方差66%。

First & Second Principal Components 第一和第二个主成分

Z_1 and Z_2 are two linear combinations. Z1和Z2是2个线性组合。

- Z_1 has the highest variation (spread of values) Z1有最大的方差
- Z_2 has the lowest variation Z2有最小的方差



PCA output for these 2 variables
对2个变量进行PCA的输出结果

Top: weights to project original data onto Z_1 & Z_2 上面的表：显示原来变量投影到Z1和Z2上使用的权重。

e.g. (-0.847, 0.532) are weights for Z_1

Bottom: reallocated variance for new variables 下面的表：显示新变量分配到的总方差百分比

Z_1 : 86% of total variance

Z_2 : 14%

	Components	
Variable	1	2
calories	-0.84705347	0.53150767
rating	0.53150767	0.84705347

Variance	498.0244751	78.932724
Variance%	86.31913757	13.68086338
Cum%	86.31913757	100
P-value	0	1

Principal Component Scores 主成分取值

XLMiner : Principal Components Analysis - Scores

Row Id.	1	2
100%_Bran	44.92	2.20
100%_Natural_Bran	-15.73	-0.38
All-Bran	40.15	-5.41
All-Bran_with_Extra_Fiber	75.31	13.00
Almond_Delight	-7.04	-5.36
Apple_Cinnamon_Cheerios	-9.63	-9.49
Apple_Jacks	-7.69	-6.38
Basic_4	-22.57	7.52
Bran_Chex	17.73	-3.51

Weights are used to compute the above scores

- e.g., col. 1 scores are computed z_1 scores using weights (-0.847, 0.532)

Properties of the resulting variables 新变量的性质

New distribution of information 信息新的分布:

- New variances = 498 (for z_1) and 79 (for z_2) 新变量的方差
- Sum of variances = sum of variances for original variables *calories* and *ratings* 总方差不变。
- New variable z_1 has most of the total variance, might be used as proxy for both *calories* and *ratings* 新变量Z1的方差占总方差比最高，接下来是Z2。
- z_1 and z_2 have correlation of zero (no information overlap) 两个变量的相关系数为0，无信息重合。

Generalization 推广

X_1, X_2, X_3, … X_p, original *p* variables 初始变量有p个

Z_1, Z_2, Z_3, … Z_p, weighted averages of original variables 新的变量是初始变量的加权平均值。

All pairs of Z variables have 0 correlation 新变量两两之间相关系数为0。

Order Z’s by variance (z_1 largest, Z_p smallest) 新的变量按方差大小从大到小进行排序。

Usually the first few Z variables contain most of the information, and so the rest can be dropped. 通常前面几个新变量包含了原有变量的绝大多数信息，所以剩下的新变量可以被剔除不予考虑。

PCA on full data set 对整个数据集进行PCA

Variable	1	2	3	4	5	6
calories	0.07624155	-0.01066097	0.61074823	-0.61706442	0.45754826	0.12601775
protein	-0.00146212	0.00873588	0.00050506	0.0019389	0.05533375	0.10379469
fat	-0.00013779	0.00271266	0.01596125	-0.02595884	-0.01839438	-0.12500292
sodium	0.98165619	0.12513085	-0.14073193	-0.00293341	0.01588042	0.02245871
fiber	-0.00479783	0.03077993	-0.01684542	0.02145976	0.00872434	0.271184
carbo	0.01486445	-0.01731863	0.01272501	0.02175146	0.35580006	-0.56089228
sugars	0.00398314	-0.00013545	0.09870714	-0.11555841	-0.29906386	0.62323487
potass	-0.119053	0.98861349	0.03619435	-0.042696	-0.04644227	-0.05091622
vitamins	0.10149482	0.01598651	0.7074821	0.69835609	-0.02556211	0.01341988
shelf	-0.00093911	0.00443601	0.01267395	0.00574066	-0.00823057	-0.05412053
weight	0.0005016	0.00098829	0.00369807	-0.0026621	0.00318591	0.00817035
cups	0.00047302	-0.00160279	0.00060208	0.00095916	0.00280366	-0.01087413
rating	-0.07615706	0.07254035	-0.30776858	0.33866307	0.75365263	0.41805118
Variance	7204.161133	4833.050293	498.4260864	357.2174377	72.47863007	4.33980322
Variance%	55.52834702	37.25226212	3.84177661	2.75336623	0.55865192	0.0334504
Cum%	55.52834702	92.78060913	96.62238312	99.37575531	99.93440247	99.96785736

- First 6 components shown
- First 2 capture 93% of the total variation
- Note: data differ slightly from text

Normalizing data 对数据进行标准化处理

- In these results, sodium dominates first PC 结果显示，变量“钠”在第一个主成分中权重很高。
- Just because of the way it is measured (mg), its scale is greater than almost all other variables 仅是因为它使用了mg的量纲，导致它的取值范围超过其他所有变量。
- Hence its variance will be a dominant component of the total variance 所以它变成权重高的指标。
- Normalize each variable to remove scale effect 需要对数据进行标准化处理，移除量纲的影响。

  Divide by std. deviation (may subtract mean first)
- Normalization (= standardization) is usually performed in PCA; otherwise measurement units affect results 通常需要在PCA中进行标准化处理，移除量纲对结果的影响。

PCA using standardized variables

Variable	1	2	3	4	5	6
calories	0.32422706	0.36006299	0.13210163	0.30780381	0.08924425	-0.20683768
protein	-0.30220962	0.16462311	0.2609871	0.43252215	0.14542894	0.15786675
fat	0.05846959	0.34051308	-0.21144024	0.37964511	0.44644874	0.40349057
sodium	0.20198308	0.12548573	0.37701431	-0.16090299	-0.33231756	0.6789462
fiber	-0.43971062	0.21760374	0.07857864	-0.10126047	-0.24595702	0.06016004
carbo	0.17192839	-0.18648526	0.56368077	0.20293142	0.12910619	-0.25979191
sugars	0.25019819	0.3434512	-0.34577203	-0.10401795	-0.27725372	-0.20437138
potass	-0.3834067	0.32790738	0.08459517	0.00463834	-0.16622125	0.022951
vitamins	0.13955688	0.16689315	0.38407779	-0.52358848	0.21541923	0.03514972
shelf	-0.13469705	0.27544045	0.01791886	-0.4340663	0.59693497	-0.12134896
weight	0.07780685	0.43545634	0.27536476	0.10600897	-0.26767638	-0.38367996
cups	0.27874646	-0.24295618	0.14065795	0.08945525	0.06306333	0.06609894
rating	-0.45326898	-0.22710647	0.18307236	0.06392702	0.03328028	-0.16606605
Variance	3.59530377	3.16411042	1.86585701	1.09171081	0.96962351	0.72342771
Variance%	27.65618324	24.3393116	14.35274601	8.39777565	7.45864248	5.5648284
Cum%	27.65618324	51.99549484	66.34824371	74.74601746	82.20465851	87.76948547

- First component accounts for smaller part of variance
- Need to use more components to capture same amount of information

- Apply PCA to training data 对训练数据集使用PCA
- Decide how many PC’s to use 决定应该使用多少个主成分变量
- Use variable weights in those PC’s with validation/new data 在验证数据集或者新数据集中针对原有变量使用主成分中的权重值。
- This creates a new reduced set of predictors in validation/new data 这样就减少了新数据集或者验证数据集中的预测因子个数。

Regression-Based Dimension Reduction

- Multiple Linear Regression or Logistic Regression
- Use subset selection
- Algorithm chooses a subset of variables
- This procedure is integrated directly into the predictive task

Summary

- Data summarization is an important for data exploration 数据汇总是探索数据过程中的一个重要方法。
- Data summaries include numerical metrics (average, median, etc.) and graphical summaries 汇总数据包括数值矩阵（平均值，中位数等）和可视化汇总。
- Data reduction is useful for compressing the information in the data into a smaller subset 数据减缩对于把数据中的信息压缩到一个小一点的数据集是有用的。
  - Categorical variables can be reduced by combining similar categories 可以通过合并类似的类别减少类别的个数。
  - Principal components analysis transforms an original set of numerical data into a smaller set of weighted averages of the original data that contain most of the original information in less variables. 主成分分析法将原始数值型数据转换成数量更少的变量，这些变量包含原始数据的绝大多数信息，是原始数据的变量的线性组合。