Dimension Reduction

Ling Xue

Dimension

- Dimension of a dataset refers to the number of variables in the dataset.
- Dimensionality of a model is the number of predictors (independent or input variables) used by the model
- Dimension reduction
  - Also known as factor selection, feature extraction.
  - Reduce the dimension of a dataset so that data mining algorithms can operate efficiently.

Dimension reduction approaches

1. Incorporating domain knowledge to remove or combine categories.
2. Using data summaries to detect information overlap between variables (and remove or combine redundant variables or categories).
3. Use data conversion techniques such as converting categorical variables into numerical variables
4. Employing automated reduction techniques such as **principal component analysis (PCA)**.
5. Using regression models, regression and classification trees to remove redundant variables or combine similar categories of categorical variables.

Principal Component Analysis

Principal Components Analysis

- Goal is to reduce a set of numerical variables
- **The idea**: remove the overlap of information between these variable.
  - "Information" is variability, measured by the sum of the variances of the variables.
- **Final product:** A smaller number of numerical variables (linear combination) that contain most of the information.

Principal Components Analysis

How does PCA do this?

- Create new variables that are ***linear combinations of original variables***
  - They are weighted averages of the original variables
- These linear combinations are uncorrelated (no information overlap), and only a few of them contain most of the original information.
- The new variables are called ***principal components***

Example – Breakfast Cereals

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

- **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

	calories	ratings
calories	379.63	-189.68
ratings	-189.68	197.32

- Total variance (=“information”) is sum of individual variances: 379.63 + 197.32=577
- Calories accounts for 379.63/577= 66%

Correlation between Calories and Rating

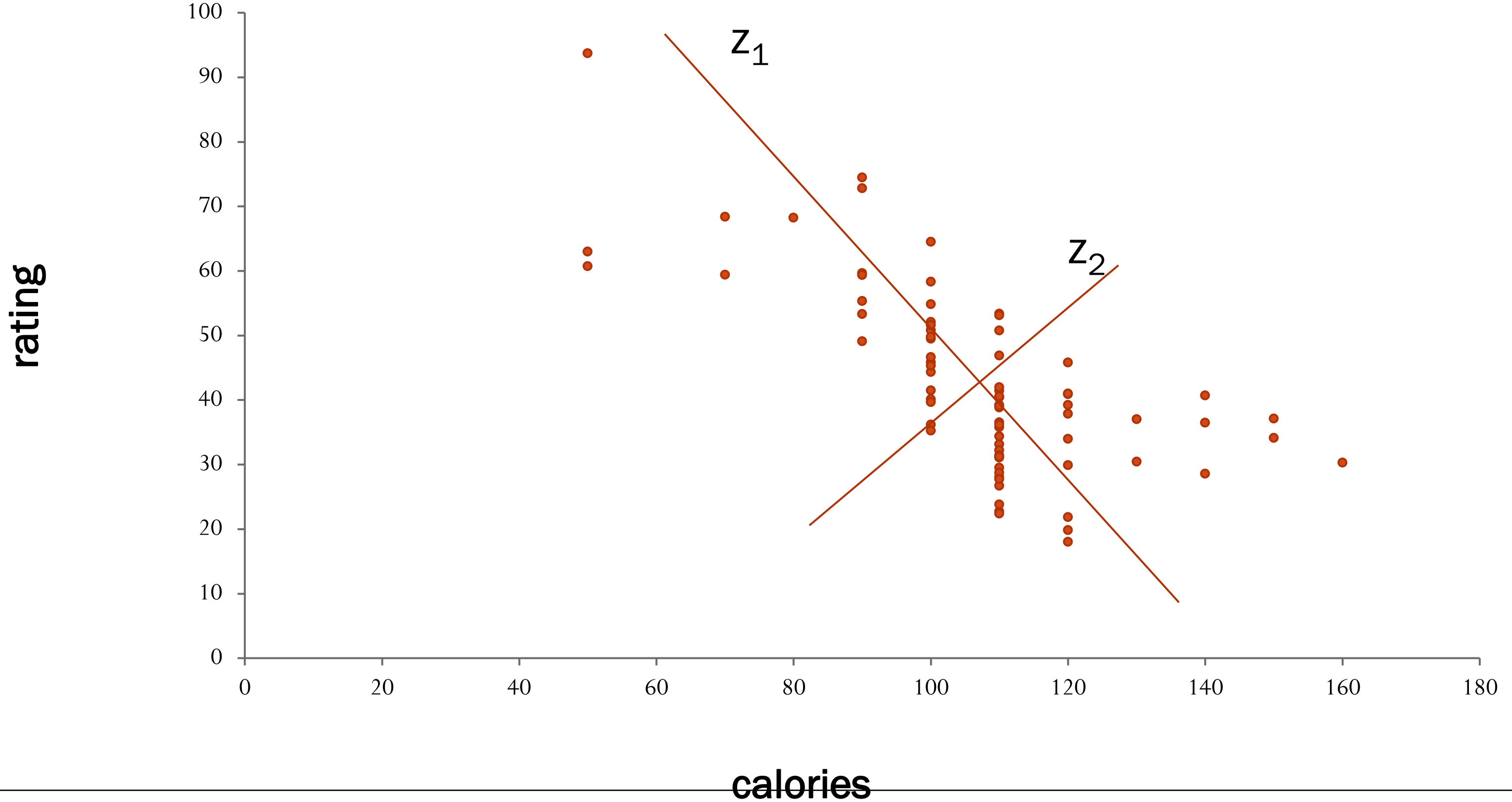
	Calories	Ratings
Calories	379.63	-188.68
Ratings	-188.68	197.32

$$-0.69 = \frac{-188.68}{\sqrt{(379.63)(197.32)}}$$

- 69% of the total variation in both variables is actually covariation.

First & Second Principal Components

- Z_1 and Z_2 are two linear combinations.
  - Z_1 has the highest variation (spread of values)
  - Z_2 has the lowest variation



PCA output for these 2 variables

Top: weights to project original data onto z_1 & z_2

e.g. (-0.847, 0.532) are weights for z_1

	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

Bottom: reallocated variance for new variables

z_1 : 86% of total variance

z_2 : 14%

Principal Component 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

New distribution of information:

- New variances = 498 (for z_1) and 79 (for z_2)
- <u>Sum</u> of variances = sum of variances for original variables *calories* and *ratings*
- New variable z_1 has most of the total variance, might be used as proxy for both *calories* and *ratings*
- z_1 and z_2 have correlation of zero (no information overlap)

Generalization

X_1, X_2, X_3, … X_p, original *p* variables

Z_1, Z_2, Z_3, … Z_p, weighted averages of original variables

All pairs of Z variables have 0 correlation

Order Z's by variance (z_1 largest, z_p smallest)

Usually the first few Z variables contain most of the information, and so the rest can be dropped.

PCA on full data set

Feature\Co	1	2	3	4	5	6
calories	-0.077984	0.0093116	0.6292058	0.6010215	0.4549585	0.1188478
protein	0.0007568	-0.008801	0.0010261	-0.0032	0.056176	0.112745
fat	0.0001018	-0.002699	0.0161958	0.0252622	-0.016098	-0.131816
sodium	-0.980215	-0.140896	-0.135902	0.0009681	0.0139481	0.022793
fiber	0.0054128	-0.030681	-0.018191	-0.020472	0.013605	0.2628413
carbo	-0.017246	0.0167833	0.01737	-0.025948	0.349267	-0.537837
sugars	-0.002989	0.0002535	0.097705	0.1154809	-0.299066	0.6479234
potass	0.1349	-0.986562	0.0367825	0.0421757	-0.047151	-0.049999
vitamins	-0.094293	-0.016729	0.6919778	-0.714118	-0.037009	0.0157572
shelf	0.0015414	-0.00436	0.0124888	-0.005647	-0.007876	-0.059901
weight	-0.000512	-0.000999	0.003806	0.0025464	0.0030221	0.0090516
cups	-0.00051	0.001591	0.0006943	-0.000985	0.0021485	-0.010305
rating	0.0752963	-0.071742	-0.307947	-0.334534	0.757708	0.4130206

Variances

	1	2	3	4	5	6
Variance	7016.4202	5028.8316	512.73921	367.92924	70.950765	4.3750841
Variance P	53.950258	38.667405	3.942525	2.8290605	0.5455506	0.0336406
Cumulative	53.950258	92.617663	96.560188	99.389248	99.934799	99.968439

- First 6 components shown
- First 2 capture 93% of the total variation

Normalizing data

- In these results, sodium dominates first PC
- Just because of the way it is measured (mg), its scale is greater than almost all other variables
- Hence its variance will be a dominant component of the total variance
- Normalize each variable to remove scale effect
  - Divide by std. deviation (may subtract mean first)
- Normalization (= standardization) is usually performed in PCA; otherwise measurement units affect results
- Use correlation matrix option to use normalized variables

PCA using standardized variables

Feature\Co	1	2	3	4	5	6	7	8
calories	-0.299542	0.3931479	0.1148575	-0.204359	0.2038989	-0.255906	0.0255955	0.0024775
protein	0.3073563	0.1653233	0.277282	-0.300743	0.319749	0.1207519	-0.282705	0.4266319
fat	-0.039915	0.3457243	-0.20489	-0.186833	0.5868933	0.3479673	0.0511547	-0.06305
sodium	-0.183397	0.1372205	0.389431	-0.120337	-0.338364	0.6643722	0.2837032	-0.17672
fiber	0.4534904	0.1798119	0.0697661	-0.039174	-0.255119	0.0642436	-0.112325	-0.216216
carbo	-0.192449	-0.149448	0.5624525	-0.087835	0.1827425	-0.326393	0.260468	-0.167436
sugars	-0.228068	0.3514345	-0.355405	0.0227072	-0.314872	-0.152082	-0.227985	0.0630881
potass	0.4019643	0.3005442	0.0676202	-0.090878	-0.14836	0.0251538	-0.148808	-0.262222
vitamins	-0.11598	0.1729092	0.3878587	0.6041106	-0.049287	0.1294858	-0.294276	0.4570409
shelf	0.1712634	0.2650503	-0.001531	0.6388786	0.3291013	-0.052044	0.1748344	-0.414146
weight	-0.050299	0.4503085	0.2471383	-0.153429	-0.221283	-0.398774	-0.013921	-0.075248
cups	-0.294636	-0.212248	0.1399997	-0.047489	0.1208164	0.0994609	-0.748567	-0.498959
rating	0.4383784	-0.251539	0.1818424	-0.038316	0.0575842	-0.186145	-0.063445	-0.014945

Variances								
	1	2	3	4	5	6	7	8
Variance	3.6336058	3.1480547	1.9093495	1.0194762	0.9893597	0.7220618	0.6715164	0.4162229
Variance Pe	27.950814	24.215805	14.687304	7.8421244	7.6104593	5.5543213	5.1655109	3.2017146
Cumulative	27.950814	52.166619	66.853923	74.696047	82.306507	87.860828	93.026339	96.228053

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

When Should We Normalize Data?

- When units of measurement are different so it is unclear how to compare the variability of different variables
  - Yes
- When units of measurement are different so scale does not reflect importance
  - Yes
- When units of measurement are common
  - No
- Scale reflects importance
  - No

PCA in Classification/Prediction

- Apply PCA to training data
- Decide how many PC's to use
- Use variable weights in those PC's with validation/new data
- This creates a new reduced set of predictors in validation/new data

Regression-Based Dimension Reduction

- Multiple Linear Regression or Logistic Regression
  - Use subset selection procedures to choose a subset of variables
- Combine similar categories
  - Having coefficients that are not statistically significant
  - Having similar coefficient values and same sign
- Classification and regression trees
  - Determine the important predictors

Summary

- **Dimension reduction** is useful for compressing the information in the data into a smaller subset.
  - Categorical variables can be reduced by combining similar categories.
  - 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.