강의 계획 및 평가 방법

주차	날짜	강의 내용	과제	대면/비대면	평가
1	03/06	강의 소개		Online	
2	03/13	데이터 마이닝 절차	과제 1 (10%)	A704	
3	03/20	데이터 탐색 및 시각화		B224	
4	**03/27**	**차원 축소**		**Online**	
5	04/03	예측성능 평가		Online	
6	04/10	다중 선형 회귀분석		A704	
7	04/17	중간 프로젝트 발표		B224	30%
8	04/24	k-최근접이웃 알고리즘 나이브 베이즈 분류	과제 2 (10%)	Online	
9	05/01	분류와 회귀 나무		Online	
10	05/08	로지스틱 회귀분석		Online	
11	05/15	신경망		A704	
12	05/22	판별 분석		Online	
13	05/29	연관 규칙		Online	
14	06/05	군집 분석		A704	
15	06/12	기말 프로젝트 발표		B224	40%

Data Mining for Business Analytics

Ch. 04 Dimension Reduction

2023.03.27.

Contents

4.1 Introduction

차원 축소의 접근법

1. 범주를 제거하거나 결합하기 위해서는 주어진 자료와 관련된 특정 분야의 지식(domain knowledge)를 도입
2. 변수 간 중복되는 자료를 검출하기 위해서(그리고 불필요한 변수들이나 범주들을 제거하거나 합치기 위해서) 자료 요약을 사용
3. 범주형 변수를 수치형 변수로 변환하는 등 자료 변환 기술 사용
4. 주성분분석(PCA)과 같은 자동화된 차원 축소 기술 사용

차원 축소의 이유

- 데이터 준비 단계에서 변수의 개수 증가(범주형 변수에 대한 dummy 변수 작성, 이미 존재하는 변수들의 변환) → 변수들 간에 높은 상관관계가 존재할 가능성 증대 → 상관관계가 높은 변수 또는 결과변수와 관련 없는 변수들로 인한 과적합(overfitting) 가능성 증대
- 많은 변수로 인한 계산 문제 발생 가능
- 변수들을 생성하고 처리하기 위한 비용 발생

4.2 Curse of Dimensionality

차원의 저주

- 다변량 데이터 모델에 변수들을 추가함으로써 야기되는 문제
- 데이터 공간이 희박해지고, 분류와 예측 모델들을 구하기 어려움
- 변수를 추가함으로써 발생하는 어려움이 기하급수적으로 증가함
- 예) 체스판
  - ✓ 2차원, 64개의 정사각형 또는 선택지
  - ✓ 큐브로 확장: 차원 50% 증가(2차원 → 3차원), 선택지 800% 증가(512=8*8*8)
- 인공지능 분야: 요인선택(Factor Selection), 특성추출(Feature Extraction)

4.3 Practical Considerations

실질적인 고려사항

- 어떤 변수들이 가장 중요하고 어떤 것이 가장 쓸모없을 것 같은가?
- 어떤 변수들이 오차를 많이 가질 것 같은가?
- 다음에 분석이 반복된다면 어떤 변수들이 측정(이를 위한 비용 고려) 가능할 것인가?
- 결과값이 발생하기 전에 실제로 측정 가능한 것은 어떤 변수들인가?

Ex 1: Boston Housing Data

- 보스턴의 인구조사 구역에서 측정된 여러가지 정보(예: 범죄율, 학생-교사 비율 등)로 구성된 14개 변수
- CAT.MEDV: 주택 가격의 중앙값(MEDV)을 3만 달러 기준 “고”와 “저” 두 가지 범주로 변환한 변수

변수명	변수 내역
CRIM	범죄율
ZN	25,000 평방피트 기준 거주지 비율
INDUS	비소매업종 점유 구역 비율
CHAS	찰스강 인접 여부 (1: 인접, 0: 비인접)
NOX	10ppm 당 일산화 질소 농도
RM	거주자의 평균 방의 개수
AGE	1940년 이전에 건축된 주택에 사는 비율

변수명	변수 내역
DIS	보스톤 5대 상업지구와의 거리
RAD	고속도로 진입 용이성 정도
TAX	10,000달러 당 재산세율
PTRATIO	학생 대 교사 비율
LSTAT	저소득층 비율
MEDV	주택가격의 중앙값(단위: $1,000)
CAT.MEDV	주택가격의 중앙값이 3만 달러 이상 여부(1: 이상, 0: 이하)

4.4 Data Summaries

[실습] Table 4.3

Summary Statistics

mean(), std(), min(), max(), median(), len()

함수	활용
min, max	오류일 가능성이 있는 극단값 검출
mean, median	편차가 크다는 사실은 이 변수의 분포가 비대칭, 즉 한쪽으로 경사진 왜도(skewness)의 존재 파악
std	데이터의 분산 정도
isnull, sum	결측값에 대한 정보

✓ CRIM과 ZN(거주지 비율) 데이터의 왜도

CRIM	ZN
0.00632	18.0
0.02731	0.0
0.02729	0.0
0.03237	0.0
0.06905	0.0
0.02985	0.0
0.08829	12.5
0.14455	12.5
0.21124	12.5

	mean	sd	min	max	median	length	miss.val
CRIM	3.613524	8.601545	0.00632	88.9762	0.25651	506	0
ZN	11.363636	23.322453	0.00000	100.0000	0.00000	506	0
INDUS	11.136779	6.860353	0.46000	27.7400	9.69000	506	0
CHAS	0.069170	0.253994	0.00000	1.0000	0.00000	506	0
NOX	0.554695	0.115878	0.38500	0.8710	0.53800	506	0
RM	6.284634	0.702617	3.56100	8.7800	6.20850	506	0
AGE	68.574901	28.148861	2.90000	100.0000	77.50000	506	0
DIS	3.795043	2.105710	1.12960	12.1265	3.20745	506	0
RAD	9.549407	8.707259	1.00000	24.0000	5.00000	506	0
TAX	408.237154	168.537116	187.00000	711.0000	330.00000	506	0
PTRATIO	18.455534	2.164946	12.60000	22.0000	19.05000	506	0
LSTAT	12.653063	7.141062	1.73000	37.9700	11.36000	506	0
MEDV	22.532806	9.197104	5.00000	50.0000	21.20000	506	0
CAT_MEDV	0.166008	0.372456	0.00000	1.0000	0.00000	506	0

4.4 Data Summaries

Summary Statistics

Relationships: Correlation

INDUS	비소매업종 점유 구역 비율
CHAS	찰스강 인접 여부 (1: 인접, 0: 비인접)
NOX	10ppm 당 일산화 질소 농도
LSTAT	저소득층 비율
DIS	보스톤 5대 상업지구와의 거리

	CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PTRATIO	LSTAT	MEDV	CAT_MEDV
CRIM	1.00	-0.20	0.41	-0.06	0.42	-0.22	0.35	-0.38	0.63	0.58	0.29	0.46	-0.39	-0.15
ZN	-0.20	1.00	-0.53	-0.04	-0.52	0.31	-0.57	0.66	-0.31	-0.31	-0.39	-0.41	0.36	0.37
INDUS	0.41	-0.53	1.00	0.06	0.76	-0.39	0.64	-0.71	0.60	0.72	0.38	0.60	-0.48	-0.37
CHAS	-0.06	-0.04	0.06	1.00	0.09	0.09	0.09	-0.10	-0.01	-0.04	-0.12	-0.05	0.18	0.11
NOX	0.42	-0.52	0.76	0.09	1.00	-0.30	0.73	-0.77	0.61	0.67	0.19	0.59	-0.43	-0.23
RM	-0.22	0.31	-0.39	0.09	-0.30	1.00	-0.24	0.21	-0.21	-0.29	-0.36	-0.61	0.70	0.64
AGE	0.35	-0.57	0.64	0.09	0.73	-0.24	1.00	-0.75	0.46	0.51	0.26	0.60	-0.38	-0.19
DIS	-0.38	0.66	-0.71	-0.10	-0.77	0.21	-0.75	1.00	-0.49	-0.53	-0.23	-0.50	0.25	0.12
RAD	0.63	-0.31	0.60	-0.01	0.61	-0.21	0.46	-0.49	1.00	0.91	0.46	0.49	-0.38	-0.20
TAX	0.58	-0.31	0.72	-0.04	0.67	-0.29	0.51	-0.53	0.91	1.00	0.46	0.54	-0.47	-0.27
PTRATIO	0.29	-0.39	0.38	-0.12	0.19	-0.36	0.26	-0.23	0.46	0.46	1.00	0.37	-0.51	-0.44
LSTAT	0.46	-0.41	0.60	-0.05	0.59	-0.61	0.60	-0.50	0.49	0.54	0.37	1.00	-0.74	-0.47
MEDV	-0.39	0.36	-0.48	0.18	-0.43	0.70	-0.38	0.25	-0.38	-0.47	-0.51	-0.74	1.00	0.79
CAT_MEDV	-0.15	0.37	-0.37	0.11	-0.23	0.64	-0.19	0.12	-0.20	-0.27	-0.44	-0.47	0.79	1.00

4.4 Data Summaries

[실습] Table 4.5, 4.6, 4.7

Aggregation and Pivot Tables

value_count()

```
bostonHousing_df.CHAS.value_counts()
```

```
0    471
1     35
Name: CHAS, dtype: int64
```

범주의 수준별 레코드 결합

```
bostonHousing_df.groupby(['RM_bin', 'CHAS'])['MEDV'].mean()
```

```
RM_bin  CHAS
3       0       25.300000
4       0       15.407143
5       0       17.200000
        1       22.218182
6       0       21.769170
        1       25.918750
7       0       35.964444
        1       44.066667
8       0       45.700000
        1       35.950000
```

CHAS	찰스강 인접 여부 (1: 인접, 0: 비인접)
RM	거주자의 평균 방의 개수
MEDV	주택가격의 중앙값(단위: $1,000)

✓ RM: (6, 7] → RM_bin: 정수 6

Pivot table

```
pd.pivot_table(bostonHousing_df, values='MEDV', index=['RM_bin'], columns=['CHAS'], aggfunc=np.mean, margins=True)
```

CHAS	0	1	All
RM_bin			
3	25.300000	NaN	25.300000
4	15.407143	NaN	15.407143
5	17.200000	22.218182	17.551592
6	21.769170	25.918750	22.015985
7	35.964444	44.066667	36.917647
8	45.700000	35.950000	44.200000
All	22.093843	28.440000	22.532806

4.5 Correlation Analysis

상관관계 분석

- 많은 후보 예측변수들을 포함하고 있는 데이터 집합에서는 변수들이 담고 있는 정보가 많이 중복됨
- 상관관계 행렬을 통해 변수들 간의 중복성 탐색
- 매우 강한(양 또는 음) 상관관계를 갖는 한 쌍의 변수들은 서로 정보의 중복성이 크게 나타나기 때문에 이 중 한 변수를 제거하여 데이터 축소 수행

	CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS
CRIM	1.00	-0.20	0.41	-0.06	0.42	-0.22	0.35	-0.38
ZN	-0.20	1.00	-0.53	-0.04	-0.52	0.31	-0.57	0.66
INDUS	0.41	-0.53	1.00	0.06	0.76	-0.39	0.64	-0.71
CHAS	-0.06	-0.04	0.06	1.00	0.09	0.09	0.09	-0.10
NOX	0.42	-0.52	0.76	0.09	1.00	-0.30	0.73	-0.77
RM	-0.22	0.31	-0.39	0.09	-0.30	1.00	-0.24	0.21

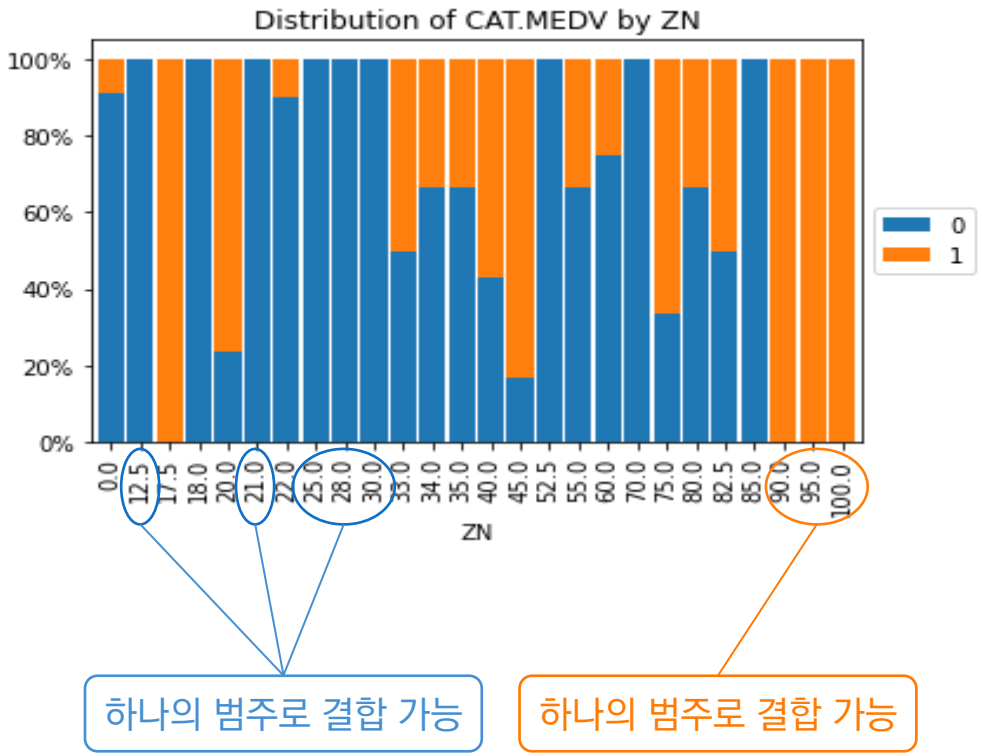
INDUS	비소매업종 점유 구역 비율
NOX	10ppm 당 일산화 질소 농도
DIS	보스톤 5대 상업지구와의 거리

4.6 Reducing the Number of Categories in Categorical Variables

[실습] Figure 4.1

범주형 변수의 범주 개수 축소

- 범주형 변수의 경우: 더미변수(Dummy variable) 사용 → 0 or 1
  - ✓ 0 = “no” (해당 범주에 속하지 않음)
  - ✓ 1 = “yes” (해당 범주에 속함)
- m개의 범주를 갖는 변수는 분석에 사용할 경우 m−1개의 더미변수로 변환
- 관측값의 수가 적은 것은 다른 것과 합침
- 의미있는 것들만 사용하고 나머지는 “기타(other)”로 처리



ZN	25,000 평방피트 기준 거주지 비율
CAT.MEDV	주택가격의 중앙값이 3만 달러 이상 여부 (1: 이상, 0: 이하)

4.7 Converting a Categorical Variable to a Numerical Variable

범주형 변수에서 수치형 변수로의 변환

- 범주 N = 구간 [n, m] → 범주형 변수 “N”을 “(n + m)/2” 로 변환
- 여러 개의 더미변수가 필요 없음
- 예) Category 2 = [20, 30] → 범주형 값 25 사용

4.8 Principal Components Analysis

주성분 분석

- 변수들의 개수가 클 때, 차원축소에 유용한 방법
- 데이터가 동일한 스케일로 측정되고 상관관계가 높은 측정값들을 포함할 때, 특히 유용
- 원래의 변수들을 가중선형결합(weighted linear combination)으로 재표현하여 작은 수(대개 3개 정도)의 변수들(principal component)을 생성 → 이 변수들이 데이터 셋 전체가 가지고 있는 정보의 대부분을 유지
- 주성분 분석은 양적변수(quantitative variable)에 대해 사용되는 분석기법
- 범주형 변수의 경우에는 대응분석(correspondence analysis)와 같은 다른 기법들이 더 적합

4.8 Principal Components Analysis

Example 2: Breakfast Cereals

- 77가지 아침식사용 시리얼제품의 영양정보와 소비자 평점 자료로 구성
- 소비자 평점: 소비자에게 정보를 제공하기 위한 시리얼의 “healthiness“ 평점(소비자에 의한 평점이 아님)
- 시리얼 데이터: 13개의 수치형 변수들로 구성
- 시리얼에 대한 정보는 섭취량(serving size)보다는 용기 단위의 시리얼을 기초로 함. (대부분의 사람들은 단순히 무게가 아닌 일정한 용량을 나타내는 시리얼 용기를 사용함)

Cereal name	mfr	Type	Calories	Protein	Fat	Sodium	Fiber	Carbo	Sugars	Potass	Vitamins	Rating
100%_Bran	N	C	70	4	1	130	10	5	6	280	25	68
100%_Natural_Bran	Q	C	120	3	5	15	2	8	8	135	0	34
All-Bran	K	C	70	4	1	260	9	7	5	320	25	59
All-Bran_with_Extra_Fiber	K	C	50	4	0	140	14	8	0	330	25	94
Almond_Delight	R	C	110	2	2	200	1	14	8		25	34
Apple_Cinnamon_Cheerios	G	C	110	2	2	180	1.5	10.5	10	70	25	30
Apple_Jacks	K	C	110	2	0	125	1	11	14	30	25	33
Basic_4	G	C	130	3	2	210	2	18	8	100	25	37
Bran_Chex	R	C	90	2	1	200	4	15	6	125	25	49
Bran_Flakes	P	C	90	3	0	210	5	13	5	190	25	53
Cap'n'Crunch	Q	C	120	1	2	220	0	12	12	35	25	18

4.8 Principal Components Analysis

Example 2: Breakfast Cereals

변수명	변수 내역
mfr	시리얼의 제조업체명(American Home Food Products, General Mills, Kellogg 등)
type	저온용 또는 고온용
calories	1회분에 대한 칼로리
protein	단백질(g)
fat	지방(g)
sodium	나트륨(mg)
fiber	식이섬유(g)
carbo	복합 탄수화물(g)
sugars	설탕(g)
potass	칼륨(mg)
vitamins	비타민과 미네랄: 0, 25, 또는 100(FDA 권장 비율)
shelf	디스플레이 선반(바닥으로부터 1, 2, 3으로 번호 부여)
weight	1회분의 무게(ounces)
cups	1회분에 제공되는 컵의 수
rating	소비자 보고서(Customer Report)에 의한 시리얼 평점

Cereal	Calories	Rating
100% Bran	70	68.40297
100% Natural Bran	120	33.98368
All-Bran	70	59.42551
Wheat Chex	100	49.787445
Wheaties	100	51.592193
Wheaties Honey Gold	110	36.187559

4.8 Principal Components Analysis

Example 2: Breakfast Cereals

Calories(X_1) and Consumer Rating(X_2)

- 77개 시리얼 평균 칼로리 = 106.86
- 평균 소비자 평점 = 42.67
- Var(X1) = 379.63, Var(X2) = 197.32
- Cov(X1, X2)= −188.68, Corr(X1, X2)=−0.69
- 공분산 행렬 $S = \begin{bmatrix} 379.63 & -188.68 \\ -188.68 & 197.32 \end{bmatrix}$
- 칼로리와 소비자 평점의 상관관계: 강한 음의 상관관계

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

- 두 변수의 전체 변동 중 69%는 실질적으로 두 변수가 공유하고 있는 변동으로, 중복되었다고 볼 수 있음 → 변수 축소의 가능성

Cereal	Calories	Rating
100% Bran	70	68.40297
100% Natural Bran	120	33.98368
All-Bran	70	59.42551
Wheat Chex	100	49.787445
Wheaties	100	51.592193
Wheaties Honey Gold	110	36.187559

- 확률벡터 X의 공분산 행렬
  - ✓ 확률벡터 X의 Xi번째 원소와 Xj번째 원소 사이의 공분산(Covariance)을 i행, j열의 원소로 갖는 행렬
  - ✓ n개의 확률변수를 원소로 갖는 확률벡터 X의 공분산 행렬: nxn의 정방행렬(Square matrix), 대각선 원소들은 Xi번째 원소의 분산인 대칭행렬(symmetric matrix)

$$Cov(X,X) = \begin{bmatrix} Var(X_1) & Cov(X_1,X_2) & \cdots & Cov(X_1,X_n) \\ Cov(X_2,X_1) & Var(X_2) & \cdots & Cov(X_2,X_n) \\ \vdots & \vdots & \ddots & \vdots \\ Cov(X_n,X_1) & Cov(X_n,X_2) & \cdots & Var(X_n) \end{bmatrix}$$

4.8 Principal Components Analysis

Example 2: Breakfast Cereals

Calories(X_1) and Consumer Rating(X_2)

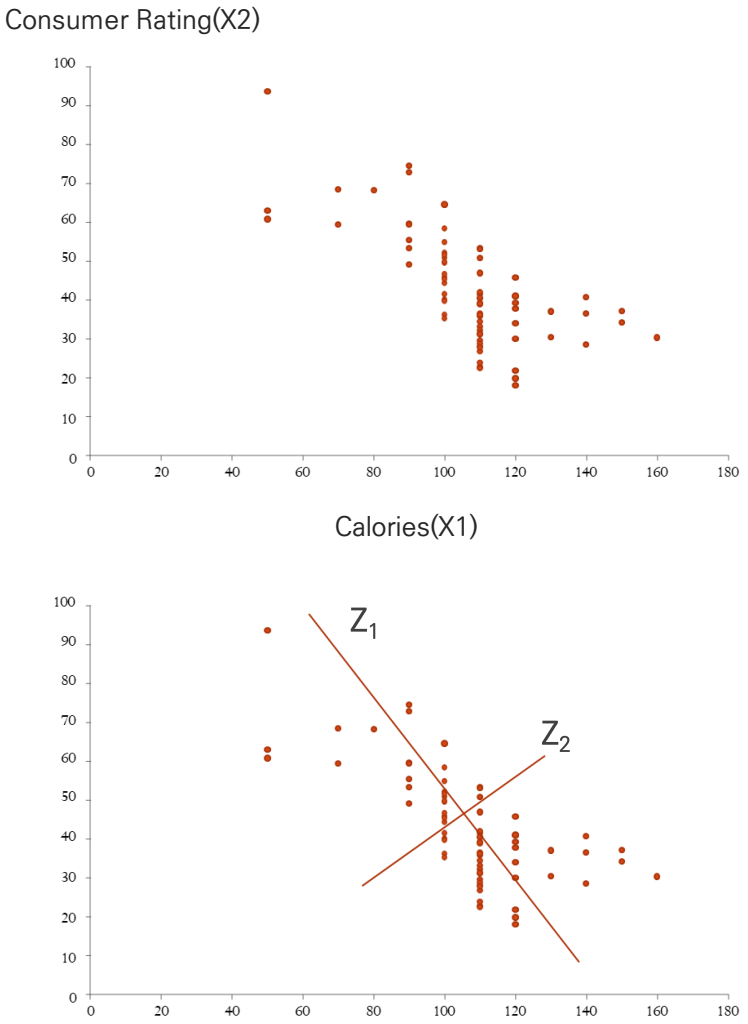
- PCA의 아이디어: 비록 두 개의 변수에 있는 정보(두 변수의 변동성)를 모두 포함하지 못하더라도 대부분은 포함될 수 있는 선형결합을 찾음
- 총 변동(Total variability) = Var(X1) + Var(X2) = 379.63 + 197.32 = 577
- 각 변수의 총 변동에서의 설명력 비율(Portions of variability):
  - ✓ X1 : 379.63/577=66% vs. X2 : 197.32/577=34%
  - → Dropping X2 will lose 34% of information
  - → 두 변수의 선형결합으로 생성된 새로운 변수에 두 변수 사이의 총 변동을 재분배

4.8 Principal Components Analysis

Example 2: Breakfast Cereals

Calories(X_1) and Consumer Rating(X_2)

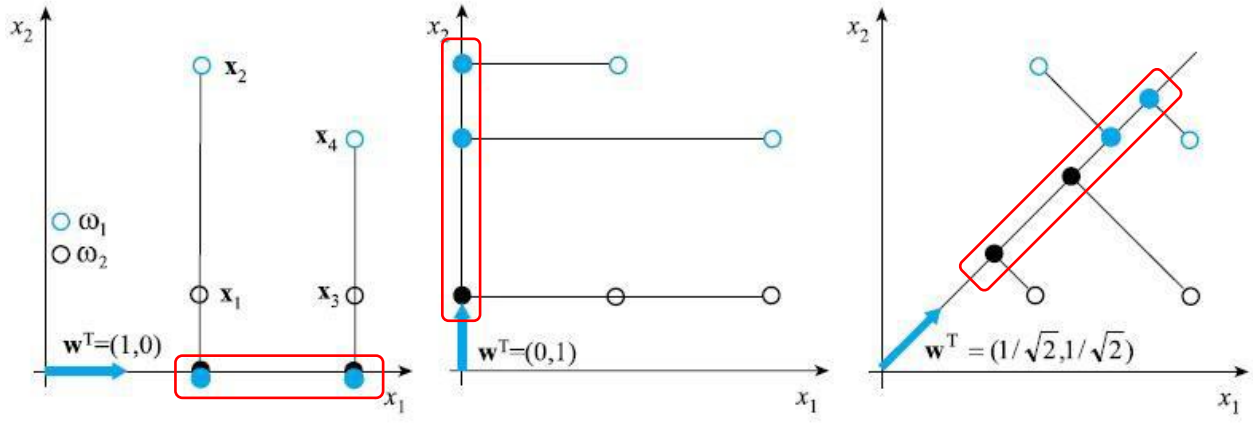
- 새로운 두 변수 생성: Z_1 and Z_2
- Z_1 : 1차 주성분(First principal component)
  - ✓ 데이터의 차원을 2개에서 1개로 축소하려고 할 때 데이터의 변동성이 가장 큰 부분을 보여주는 선
  - ✓ 모든 가능한 직선들 중 77개의 일차원 값들로 구성된 집합을 얻기 위하여 데이터셋에 있는 점들을 직각으로 교차하여 투영한 것 (Z_1의 분산은 최대가 됨)
  - ✓ 이 직선과 점들 간의 수직선 거리의 제곱합을 최소화하는 선
- Z_2 : 2차 주성분(Second principal component)
  - ✓ Z_1 축과 수직이면서 두 번째로 큰 변동성을 갖는 선



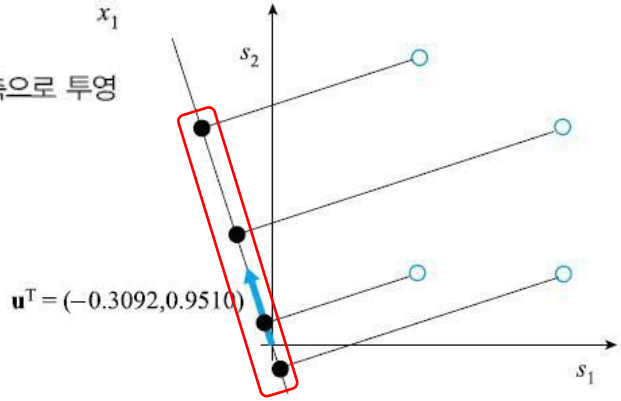
4.8 Principal Components Analysis

Example 2: Breakfast Cereals

Calories(X_1) and Consumer Rating(X_2)



(a) $\mathbf{w}^T=(1,0)$ 축으로 투영 (b) $\mathbf{w}^T=(0,1)$ 축으로 투영 (c) $\mathbf{w}^T=(1/\sqrt{2},1/\sqrt{2})$ 축으로 투영



PCA로 구한 최적의 축

4.8 Principal Components Analysis

[실습] Table 4.10

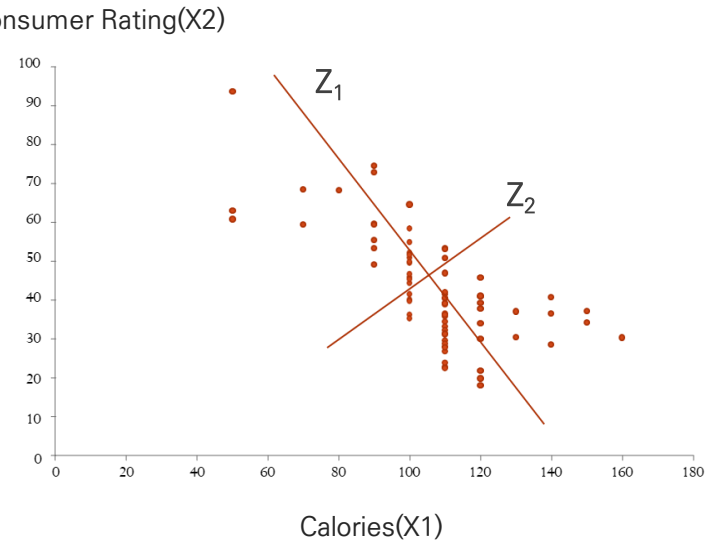
Example 2: Breakfast Cereals

Calories(X_1) and Consumer Rating(X_2)

[PCS Summary]

	PC1	PC2
Standard deviation	22.3165	8.8844
Proportion of variance	0.8632	0.1368
Cumulative proportion	0.8632	1.0000

- PCS Summary
  - ✓ Z_1 : 총 변동의 86% 설명
  - ✓ Z_2 : 총 변동의 14% 설명
- Drop Z_2



[PCS Components]

	PC1	PC2
calories	-0.847053	0.531508
rating	0.531508	0.847053

- PCS Components
- 원래 점을 두 가지 방향으로 투영하는데 사용되는 가중치 제공
  - ✓ Z_1 : (−0.847, 0.532)
  - ✓ Z_2 : (0.532, 0.847)
- 가중치: PCS Components Score를 계산하는 데 사용

✓ Notice that
$||(-0.847, 0.532)||=1$

4.8 Principal Components Analysis

[실습] Table 4.10

Example 2: Breakfast Cereals

Calories(X_1) and Consumer Rating(X_2)

- Principal Component Score
  - ✓ X_1 (칼로리)과 X_2 (소비자 평점)을 (평균을 뺀 후) 새로운 축에 투영한 값 → **새로운 축에서의 변동성**
  - ✓ PC1 열: 가중치 (−0.847, 0.532)을 사용하여 Z_1 에 투영한 값
  - ✓ PC2 열: 가중치 (0.532, 0.847)을 사용하여 Z_2 에 투영한 값

[PCS Components]

	PC1	PC2
calories	-0.847053	0.531508
rating	0.531508	0.847053

[Principal Component Score]

	PC1	PC2
100% Bran	44.921528	2.197183
100% Natural Bran	-15.725265	-0.382416
All-Bran	40.149935	-5.407212
All-Bran with Extra Fiber	75.310772	12.999126
Almond Delight	-7.041508	-5.357686

Mean of X_1 =106.88, mean of X_2 = 42.67

"100% Bran (cal=70, rating=68.4)"

→ 평균조정 : cal=70−106.88, rating=68.4−42.67

PC1(First principal components score) = (−0.847)(70−106.88) + (0.532)(68.4−42.67) = 44.92

PC2(Second principal components score) = (0.532)(70−106.88) + (0.847)(68.4−42.67) = 2.197

4.8 Principal Components Analysis

Example 2: Breakfast Cereals

Calories(X_1) and Consumer Rating(X_2)

- Properties of the resulting variables
  - ✓ Z_1 과 Z_2 의 평균 = 0 (평균을 뺀 값이므로)
  - ✓ Z_1 과 Z_2 는 관련이 없음 (수직 방향이므로): corr = 0 (no information overlap)
  - ✓ 분산의 합: var(X_1) + var(X_2) = var(Z_1) + var(Z_2)
  - ✓ var(Z_1) = 498, var(Z_2) = 79
  - ✓ 첫번째 주성분인 Z_1 은 총 변동의 86%를 설명 → Drop Z_2

Variances	**Var(X_1) = 379.63** **Var(X_2) = 197.32**	**Var(Z_1) = 498.02** **Var(Z_2) = 78.93**
Proportion of Variances (%)	X_1 : 66% X_2 : 34%	Z_1 : 86% Z_2 :14%
Corr	Corr (X_1, X_2) = -0.69	0 (uncorrelated)

4.8 Principal Components Analysis

Principal Components

p 〉2 인 경우

- 다중공선성(multicollinearity) 문제
  - ✓ 통계학의 회귀분석에서 독립변수들 간에 강한 상관관계가 나타나는 문제
  - ✓ 해당 변수들의 설명력이 약해지고, 변수들의 표준오차 증가

- p개의 변수들: $X_1, X_2, \ldots, X_p$
- 원래 변수들의 (각 평균을 뺀 후) 가중평균인 새로운 변수 집합 찾음: $Z_1, Z_2, \ldots, Z_p$

$$Z_i = a_{i,1}(X_1 - \overline{X_1}) + a_{i,2}(X_2 - \overline{X_2}) + \cdots + a_{i,p}(X_p - \overline{X_p}) \quad i = 1, \cdots, p$$

- Z 변수들의 각 쌍에 관한 상관계수 = 0 → 주성분을 독립변수로 하여 회귀모델을 세우는 경우 다중공선성(multicollinearity) 문제에서 벗어남
- Z_1 은 가장 큰 분산, Z_p 는 가장 작은 분산
- 가중치 $a_{i,j}$ 계산

$a_{1,1}$ $a_{1,2}$

PC1(First principal components score) = (−0.847)(70−106.88) + (0.532)(68.4−42.67) = 44.92 → Z_1

PC2(Second principal components score) = (0.532)(70−106.88) + (0.847)(68.4−42.67) = 2.197 → Z_2

4.8 Principal Components Analysis

[실습] Table 4.11

Principal Components

Ex. Breakfast Cereals

- 처음 3개의 성분이 원래 변수 13개 모두와 관련된 총 변동의 96% 이상 차지
- 데이터의 원래 차원 수의 25% 미만의 차원으로 데이터의 변동을 대부분 파악 가능
- 첫 2개의 성분 만으로도 총 변동의 92.6% 차지

[PCS Components]

	PC1	PC2	PC3	PC4	PC5
calories	-0.077984	-0.009312	0.629206	-0.601021	0.454959
protein	0.000757	0.008801	0.001026	0.003200	0.056176
fat	0.000102	0.002699	0.016196	-0.025262	-0.016098
sodium	-0.980215	0.140896	-0.135902	-0.000968	0.013948
fiber	0.005413	0.030681	-0.018191	0.020472	0.013605
carbo	-0.017246	-0.016783	0.017370	0.025948	0.349267
sugars	-0.002989	-0.000253	0.097705	-0.115481	-0.299066
potass	0.134900	0.986562	0.036782	-0.042176	-0.047151
vitamins	-0.094293	0.016729	0.691978	0.714118	-0.037009
shelf	0.001541	0.004360	0.012489	0.005647	-0.007876
weight	-0.000512	0.000999	0.003806	-0.002546	0.003022
cups	-0.000510	-0.001591	0.000694	0.000985	0.002148
rating	0.075296	0.071742	-0.307947	0.334534	0.757708

[PCS Summary]

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11	PC12	PC13
Standard deviation	83.7641	70.9143	22.6437	19.1815	8.4232	2.0917	1.6994	0.7796	0.6578	0.3704	0.1864	0.063	0.0
Proportion of variance	0.5395	0.3867	0.0394	0.0283	0.0055	0.0003	0.0002	0.0000	0.0000	0.0000	0.0000	0.000	0.0
Cumulative proportion	0.5395	0.9262	0.9656	0.9939	0.9993	0.9997	0.9999	1.0000	1.0000	1.0000	1.0000	1.000	1.0

4.8 Principal Components Analysis

[실습] Table 4.11

Normalizing the Data

- PC1에 지배적인 변수: 나트륨(sodium)
- PC2에 지배적인 변수: 칼륨(potass)

protein	단백질(g)
fat	지방(g)
sodium	나트륨(mg)
fiber	식이섬유(g)
carbo	복합 탄수화물(g)
sugars	설탕(g)
potass	칼륨(mg)

- 나트륨과 칼륨: mg / 다른 변수: g
- 다른 변수들의 분산보다 나트륨과 칼륨의 분산이 훨씬 큼 → 총 분산을 독차지
- 분산이 1인 표준화된 변수로 정규화(표준화) 할 필요 있음

[PCS Components]

	PC1	PC2	PC3	PC4	PC5
calories	-0.077984	-0.009312	0.629206	-0.601021	0.454959
protein	0.000757	0.008801	0.001026	0.003200	0.056176
fat	0.000102	0.002699	0.016196	-0.025262	-0.016098
sodium	-0.980215	0.140896	-0.135902	-0.000968	0.013948
fiber	0.005413	0.030681	-0.018191	0.020472	0.013605
carbo	-0.017246	-0.016783	0.017370	0.025948	0.349267
sugars	-0.002989	-0.000253	0.097705	-0.115481	-0.299066
potass	0.134900	0.986562	0.036782	-0.042176	-0.047151
vitamins	-0.094293	0.016729	0.691978	0.714118	-0.037009
shelf	0.001541	0.004360	0.012489	0.005647	-0.007876
weight	-0.000512	0.000999	0.003806	-0.002546	0.003022
cups	-0.000510	-0.001591	0.000694	0.000985	0.002148
rating	0.075296	0.071742	-0.307947	0.334534	0.757708

4.8 Principal Components Analysis

Normalizing the Data

- 정규화 또는 표준화가 필요한 경우
  - ✓ 변수들이 다른 단위로 측정되어 변수 간의 변동성을 비교하기 불분명한 경우

    ex) 일부는 달러 단위, 다른 변수들은 parts per million
  - ✓ 변수들이 동일한 단위로 측정되었지만 변수의 스케일이 중요성을 반영하지 않는 경우

    ex) 주당 순이익, 총수익
- 정규화 또는 표준화를 하지 않아야 하는 경우
  - ✓ 변수의 스케일이 변수의 중요성을 반영하는 경우: 스케일 조정하지 않음

    ex) 제트연료 판매액, 난방기름의 판매액

4.8 Principal Components Analysis

[실습] Table 4.12

Normalizing the Data

Ex. Breakfast Cereals

[PCS Summary before Standardization]

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11	PC12	PC13
Standard deviation	83.7641	70.9143	22.6437	19.1815	8.4232	2.0917	1.6994	0.7796	0.6578	0.3704	0.1864	0.063	0.0
Proportion of variance	0.5395	0.3867	0.0394	0.0283	0.0055	0.0003	0.0002	0.0000	0.0000	0.0000	0.0000	0.000	0.0
Cumulative proportion	0.5395	0.9262	0.9656	0.9939	0.9993	0.9997	0.9999	1.0000	1.0000	1.0000	1.0000	1.000	1.0

[PCS Summary after Standardization]

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11	PC12	PC13
Standard deviation	1.9192	1.7864	1.3912	1.0166	1.0015	0.8555	0.8251	0.6496	0.5658	0.3051	0.2537	0.1399	0.0
Proportion of variance	0.2795	0.2422	0.1469	0.0784	0.0761	0.0555	0.0517	0.0320	0.0243	0.0071	0.0049	0.0015	0.0
Cumulative proportion	0.2795	0.5217	0.6685	0.7470	0.8231	0.8786	0.9303	0.9623	0.9866	0.9936	0.9985	1.0000	1.0

4.8 Principal Components Analysis

[실습] Table 4.12

Normalizing the Data

Ex. Breakfast Cereals

- PC1
  - ✓ Negative weights: 칼로리, 1회분 분량(cup)
  - ✓ Positive weights: 단백질, 식이섬유, 칼륨(potass)
- 칼로리와 1회분 분량이 작고, 단배질과 칼륨이 많다면
  → 높은 소비자 평점
- 데이터의 구조를 이해할 수 있는 레이블링 가능

[PCS Components after Standardization]

	PC1	PC2	PC3	PC4	PC5
calories	-0.299542	-0.393148	0.114857	-0.204359	0.203899
protein	0.307356	-0.165323	0.277282	-0.300743	0.319749
fat	-0.039915	-0.345724	-0.204890	-0.186833	0.586893
sodium	-0.183397	-0.137221	0.389431	-0.120337	-0.338364
fiber	0.453490	-0.179812	0.069766	-0.039174	-0.255119
carbo	-0.192449	0.149448	0.562452	-0.087835	0.182743
sugars	-0.228068	-0.351434	-0.355405	0.022707	-0.314872
potass	0.401964	-0.300544	0.067620	-0.090878	-0.148360
vitamins	-0.115980	-0.172909	0.387859	0.604111	-0.049287
shelf	0.171263	-0.265050	-0.001531	0.638879	0.329101
weight	-0.050299	-0.450309	0.247138	-0.153429	-0.221283
cups	-0.294636	0.212248	0.140000	-0.047489	0.120816
rating	0.438378	0.251539	0.181842	-0.038316	0.057584

[PCS Summary after Standardization]

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11	PC12	PC13
Standard deviation	1.9192	1.7864	1.3912	1.0166	1.0015	0.8555	0.8251	0.6496	0.5658	0.3051	0.2537	0.1399	0.0
Proportion of variance	0.2795	0.2422	0.1469	0.0784	0.0761	0.0555	0.0517	0.0320	0.0243	0.0071	0.0049	0.0015	0.0
Cumulative proportion	0.2795	0.5217	0.6685	0.7470	0.8231	0.8786	0.9303	0.9623	0.9866	0.9936	0.9985	1.0000	1.0

4.8 Principal Components Analysis

Using Principal Components for Classification and Prediction

- 학습 데이터를 사용하여 PCA 적용
- 예측변수로 사용할 주성분 개수 결정
- 모델에서 축소된 개수의 주성분 점수를 예측변수로 사용
- 검증용 데이터에 대해서는 학습용 데이터로부터 계산된 가중치를 적용하여 주성분 점수 계산
- 새롭게 계산된 이들 주성분 점수가 검증용 데이터의 예측변수로 사용
- 지도학습에서 예측변수로 주성분 사용의 단점: PCA가 선형 변환을 생성하여 원래 변수들 간의 선형관계를 보존하므로 비선형적인 예측 정보를 잃을 수 있음

4.9 Dimension Reduction Using Regression Models

- 탐색적 방법을 통한 차원 축소: 요약 통계량, 시각화 방법, PCA
  - ✓ PCA: 결과변수를 무시
  - ✓ y값 관점에서 유사한 범주 결합: 예측변수와 결과변수 사이의 관계를 비공식적으로 포함

- 예측이나 분류 작업을 직접 고려하여 회귀 모델을 적합시키는 방법
  - ✓ 예측변수의 수를 줄이기 위해 변수 선택 절차 사용
  - ✓ 예측: 선형회귀 모델(ch. 06)
  - ✓ 분류: 로지스틱 회귀 모델(ch. 10)

4.10 Dimension Reduction Using Classification and Regression Trees

- 분류와 회귀나무(CART: Classification And Regression Tree) 사용
  - ✓ 예측: 회귀 나무
  - ✓ 분류: 분류 나무
  - ✓ 결과변수를 가장 잘 분류 또는 예측하도록 예측변수들에 관하여, 이분할(binary splits) 가지 생성
  - ✓ 결과로 나타난 tree를 사용하여 예측변수 결정 및 제거