INT104 ARTIFICIAL INTELLIGENCE

REVIEW – WEEK 14

Sichen Liu
Sichen.Liu@xjtlu.edu.cn
Xi'an Jiaotong-Liverpool University
西交利物浦大学

Data Collection

- Lots of places that host/share data online, or you can collect them yourself.
- Open data collections
- Social media data
- Multimodal data

Data Cleaning

- Handling Missing Data
  - Get rid of the corresponding instance.
  - Get rid of the whole column.
  - Set the values to some value (zero, the mean, the median, etc.).
- Smooth Noisy Data
  - Identify or remove the outliners
  - Try to resolve the inconsistent

  (there is no one way to remove noise, or smooth out the noisiness in the data)

Practice: Data Cleaning

#	Country	Alcohol (L/person)	Deaths (Per 100k)	Heart (Per 100k)	Liver (Per 100k)	Free healthcare
1	Australia	2.5	785	211	15.30000019	Y
2	Austria	3.000000095	863	167	45.59999847	Y
3	Belg/Lux	2.900000095	883	131	20.70000076	N
4	Canada	2.400000095	793	NA	16.39999962	Y
5	Denmark	2.900000095	971	220	23.89999962	Y
6	Finland	0.800000012	970	297	19	N
7	France	9.100000381	751	11	37.90000153	N
8	Iceland	-0.800000012	743	211	11.19999981	Y
9	Ireland	0.699999988	1000	300	6.5	Y
10	Israel	0.600000024	-834	183	13.69999981	Y
11	Italy	27.900000095	775	107	42.20000076	Y
12	Japan	1.5	680	36	23.20000076	N
13	Netherlands	1.799999952	773	167	9.199999809	N
14	New Zealand	1.899999976	916	266	7.699999809	Y
15	Norway	0.0800000012	806	227	12.19999981	N
16	Spain	6.5	724	NA	NA	Y
17	Sweden	1.600000024	743	207	11.19999981	N
18	Switzerland	5.800000191	693	115	20.29999924	N
19	UK	1.299999952	941	285	10.30000019	Y
20	US	1.200000048	926	199	22.10000038	N
21	West Germany	2.700000048	861	172	36.70000076	Y

Data Transformation

- Normalization
  - Min–max normalization.

$$x_{scaled} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

  - Z-score normalization.

    Normalizing every value in a dataset such that the mean of all of the values is 0 and the standard deviation is 1

$$x_{scaled} = \frac{x - mean}{sd}$$

  - Normalization by decimal scaling.

$$x_{scaled} = \frac{x}{10^j}$$

Data Reduction

Data reduction is a key process in which a reduced representation of a dataset that produces the same or similar analytical results is obtained.

	A1	A2	A3	....	A200
T1					
T2					
T3					
....					
T200					

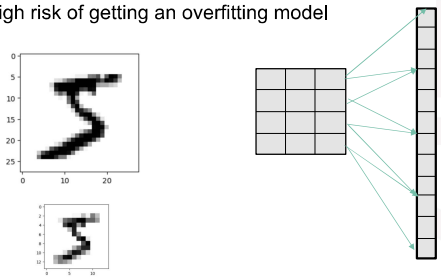
→

	A1	A2	A3	...	A120
T1					
T2					
T3					
....					
T150					

Dimensionality Reduction

Data with high dimensions:

- High computational complexity
- May contain many irrelevant or redundant features
- Difficulty in visualization
- With high risk of getting an overfitting model



Principal Component Analysis (PCA)

Preserving the Variance:

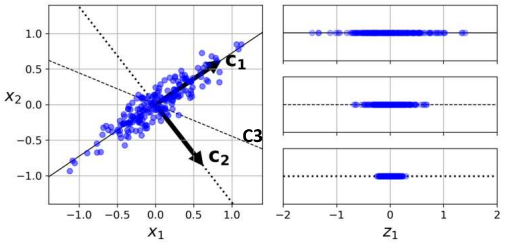
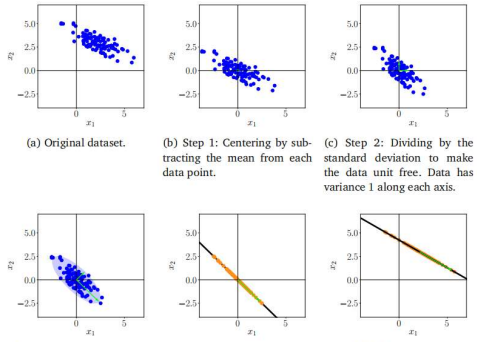


Figure 8-7. Selecting the subspace to project on

PCA identifies the axis that accounts for the largest amount of variance in the training set.

PCA

Key steps of PCA in practice



(a) Original dataset.

(b) Step 1: Centering by subtracting the mean from each data point.

(c) Step 2: Dividing by the standard deviation to make the data unit free. Data has variance 1 along each axis.

(d) Step 3: Compute eigenvalues and eigenvectors (arrows) of the data covariance matrix (ellipse).

(e) Step 4: Project data onto the principal subspace.

(f) Undo the standardization and move projected data back into the original data space from (a).

API Information

numpy.linalg.norm

linalg.norm(x, axis=None): Matrix or vector norm.

Parameters

x: array like

Input array. If axis is None, x must be 1-D or 2-D.

Returns

n: float or ndarray

Norm of the matrix or vector(s).

API Information

sklearn.cluster.KMeans

class sklearn.cluster.KMeans(n_clusters=8, random_state=None): KMeans clustering.

Parameters

n_clusters: int, default=8

The number of clusters to form as well as the number of centroids to generate

random_state: int, RandomState instance or None, default=None

Determines random number generation for centroid initialization. Use an int to make the randomness deterministic.

Attributes

labels_ : ndarray of shape (n samples,), Labels of each point.

Methods

fit(X), Compute k-means clustering.

Parameters

X: array-like, sparse matrix of shape (n samples, n features)

Training instances to cluster.

Returns

self: object

Fitted estimator.

Indentation matters!

- Code is grouped by its indentation
- Indentation is the number of whitespace or tab characters before the code.
- If you put code in the wrong block, then you will get unexpected behavior

```
Line 1  x = True
Line 2  if x:
Line 3      print("Executing if")
Line 4  else:
Line 5      print("Executing else")
Line 6  print("Prints regardless of the if-else block")
```

```
Executing if
Prints regardless of the if-else block
```