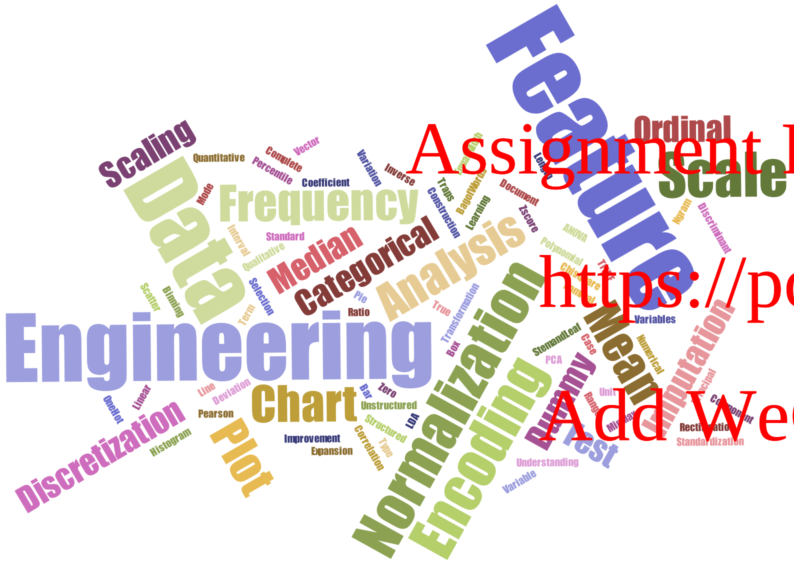
FEATURE ENGINEERING (CONCEPTS – PART 3)

Assignment Project Exam Help
https://powcoder.com
Add WeChat powcoder



Contents

Assignment Project Exam Help
https://powcoder.com
Add WeChat powcoder

Feature Selection

Assignment Project Exam Help
https://powcoder.com
Add WeChat powcoder

Feature selection selects the more relevant features and eliminate redundant, irrelevant, and noisy features

- Feature relevance is classified into three types: strong relevance, weak relevance, and irrelevant
- A feature which has an influence on the output and its role cannot be replaced by the rest is known as relevant feature and therefore cannot be removed
- A feature is said to be a weakly relevant if it is necessary for an optimal subset only at certain conditions
- An irrelevant feature is one which is not necessary at all because it does not contribute any information to the target and hence it should be removed
- A feature which takes the role of another is said to be redundant
- Removing irrelevant and redundant features will potentially give a better generalization, understanding and visualization with less training and testing time

Assignment Project Exam Help
https://powcoder.com
Add WeChat powcoder

Copyright (c) by Daniel K.C. Chan. All Rights Reserved.

Identifying which features are most relevant is particularly useful when there are only a few samples

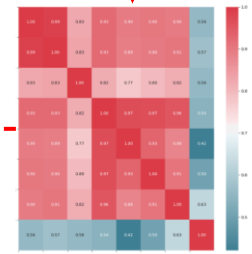
Name	Amount	Date	Issued In	Used In	Age	Education	Fraud?
Daniel	$2,600.45	1-Jul-2020	HK	HK	22	Secondary	No
Alex	$2,294.58	1-Oct-2020	HK	RUS	None	Postgraduate	Yes
Adrian	$1,003.30	3-Oct-2020	HK		25	Graduate	Yes
Vicky	$8,488.32	4-Oct-2020	JAPAN	HK	64	Graduate	No
Adams	¥20000	7-Oct-2020	AUS	JAP	58	Primary	No
...	...	...	...	...	...	...	...
Jones	₽3,250.11	Nov 1, 2020	HK	RUS	43	Graduate	No
Mary	₽8,156.20	Nov 1, 2020	HK	N/A	27	Graduate	Yes
Max	€7475,11	Nov 8, 2020	UK	GER	32	Primary	No
Peter	₽500.00	Nov 9, 2020	Hong Kong	RUS	0	Postgraduate	No
Anson	₽7,475.11	Nov 9, 2020	Hong Kong	RUS	20	Postgraduate	Yes

Observations

Feature

Target

feature selection



Assignment Project Exam Help
https://powcoder.com
Add WeChat powcoder

Copyright (c) by Daniel K.C. Chan. All Rights Reserved.

Feature selection is the process of selecting a subset of relevant features for use in model construction

- Reasons for doing feature selection include
  - to simplify models to make them easier to interpret by researchers / users
  - to shorten training time
  - to reduce the dimensionality of data involved
  - to enhance generalization by reducing overfitting
  - to reduce model scoring time (after model deployment)

Assignment Project Exam Help
https://powcoder.com
Add WeChat powcoder

Copyright (c) by Daniel K.C. Chan. All Rights Reserved.

The three main categories of supervised feature selection algorithms are filter, wrapper, and embedded methods

Assignment Project Exam Help

https://powcoder.com

Add WeChat powcoder

Filter Methods

- A proxy measure, often statistical, instead of the error rate is used to score a subset
- Computationally less expensive
- Selection is more general & with lower predictive performance

Wrapper Methods

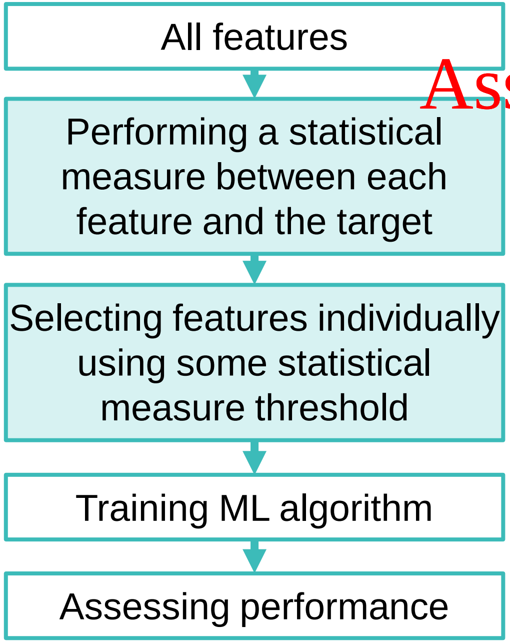
- Each subset is used to train a model and the model error rate provides the score for the subset
- Computationally very expensive
- Selection is usually good

Embedded Methods

- A catch-all group of techniques being part of the model construction process
- Computational complexity is between filters and wrappers

Copyright (c) by Daniel K.C. Chan. All Rights Reserved.

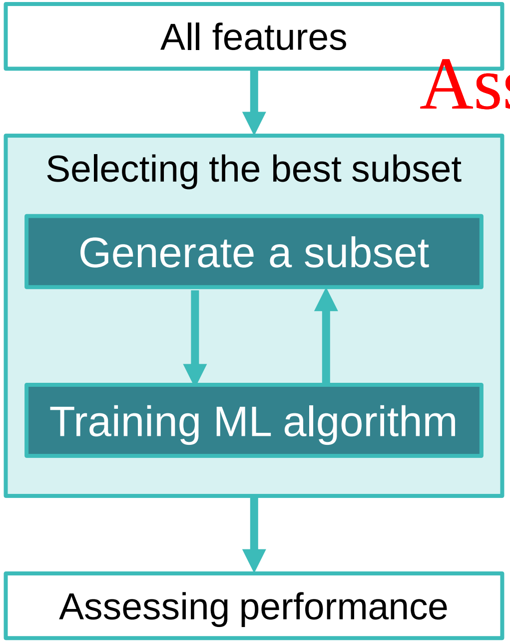
Filter Methods



- Apply a statistical measure (e.g. correlation with the target) to assign a score to each variable regardless of the ML model
- Variables are ranked by the score and either to be kept or removed from the dataset
- Often univariate and consider the feature independently
- Tend to select redundant variables as the relationships between variables are not considered
- No consideration is given to the ML model during the filtering process; hence, may not be able to select the right features for the model

Assignment Project Exam Help
https://powcoder.com
Add WeChat powcoder

Wrapper Methods

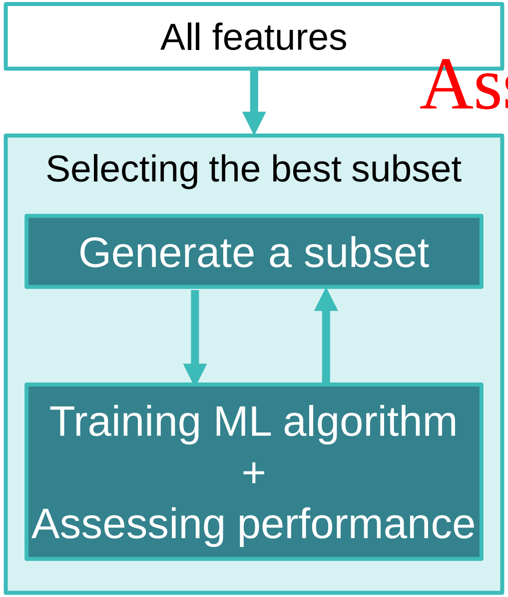


- Consider the selection as a search problem where different combinations are prepared, evaluated and compared to other combinations
- A predictive model is used to assign scores based on model accuracy
- Can detect possible interactions between variables
- Increase the overfitting risk when the number of observations is insufficient

Assignment Project Exam Help
https://powcoder.com
Add WeChat powcoder

Copyright (c) by Daniel K.C. Chan. All Rights Reserved.

Embedded Methods



- Try to combine the advantages of both filter and wrapper methods
- A learning algorithm takes advantage of its own variable selection process and performs feature selection and assessment simultaneously

Assignment Project Exam Help
https://powcoder.com
Add WeChat powcoder

Copyright (c) by Daniel K.C. Chan. All Rights Reserved.

Filter-based Feature Selection

Assignment Project Exam Help
https://powcoder.com
Add WeChat powcoder

The choice of feature selection algorithm depends on the nature of the input features and output target

Assignment Project Exam Help
https://powcoder.com
Add WeChat powcoder

		Target	
		Categorical	Numerical
Features	Categorical	Chi-Squared Test (contingency table)	ANOVA Correlation Coefficient (linear)
		Mutual Information	Kendall's Rank Coefficient (non-linear)
	Numerical	ANOVA Correlation Coefficient (linear)	Pearson's Correlation Coefficient (linear)
		Kendall's Rank Coefficient (non-linear)	Spearman's Rank Correlation Coefficient (non-linear)

- Pearson's can be used on quantitative continuous variables
- Spearman's can be used on ordinal data when the ordered categories are replaced by their ranks
- Actually, mutual information is agnostic to data types

Copyright (c) by Daniel K.C. Chan. All Rights Reserved.

Feature Selection using Pearson's Correlation

Assignment Project Exam Help

https://powcoder.com

Add WeChat powcoder

Pearson's Correlation

- The Pearson's Correlation is a measure of the strength and direction of association that exists between two variables measured on at least an interval scale
- The coefficient measures the linear relationship between columns
- The coefficient value varies between -1 and +1
- The value 0 implies no correlation between columns
- Values closer to -1 or +1 imply an extremely strong linear relationship
- Pearson’s correlation coefficient generally requires that each column be normally distributed

Assignment Project Exam Help
https://powcoder.com
Add WeChat powcoder

Copyright (c) by Daniel K.C. Chan. All Rights Reserved.

Pearson's correlation calculates the effect of change in one variable when the other variable changes

$$r = \frac{N(\sum xy) - (\sum x)(\sum y)}{\sqrt{[N\sum x^2 - (\sum x)^2][N\sum y^2 - (\sum y)^2]}}$$

Assignment Project Exam Help

https://powcoder.com

Add WeChat powcoder

where

$N = \text{the number of pairs of scores}$

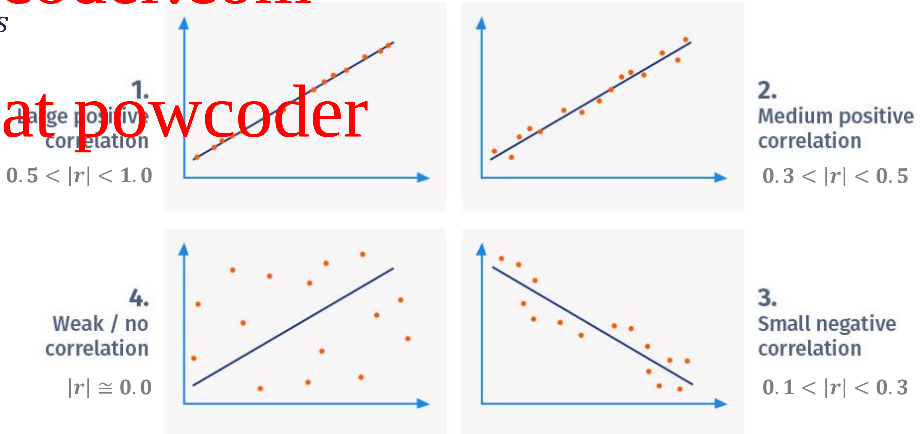
$\sum xy = \text{the sum of the products of paired scores}$

$\sum x = \text{the sum of } x \text{ scores}$

$\sum y = \text{the sum of } y \text{ scores}$

$\sum x^2 = \text{the sum of squared } x \text{ scores}$

$\sum y^2 = \text{the sum of squared } y \text{ scores}$



Default Credit Card Payments

- A dataset about customer default payments in Taiwan
- Number of observations = 30,000
- Number of features = 24
- From the perspective of risk management, the result of predictive accuracy of the estimated probability of default will be more valuable than the binary result of classification - credible or not credible clients

Assignment Project Exam Help
https://powcoder.com
Add WeChat powcoder

Copyright (c) by Daniel K.C. Chan. All Rights Reserved.

The credit card default payment dataset

#	Feature	Description
1	LIMIT_BAL	Credit amount in NT dollar
2	SEX	Gender: 1=male; 2=female
3	EDUCATION	Education: 1=postgraduate, 2=graduate, 3=secondary, 4=others
4	MARRIAGE	Marital status: 1=married; 2=single; 3=others
5	AGE	Age in year
6	PAY_0	Repayment status of September to April 2005: -1= paid duly, 1=1 month delay, … , 8=8 months' delay, 9=9 months or longer delay
7	PAY_2	
8	PAY_3	
9	PAY_4	
10	PAY_5	
11	PAY_6	

#	Feature	Description
12	BILL_AMT1	Bill statement amount in NT dollar from September to April 2005.
13	BILL_AMT2	
14	BILL_AMT3	
15	BILL_AMT4	
16	BILL_AMT5	
17	BILL_AMT6	
18	PAY_AMT1	Amount of previous payment in NT dollar from September to April 2005
19	PAY_AMT2	
20	PAY_AMT3	
21	PAY_AMT4	
22	PAY_AMT5	
23	PAY_AMT6	
24	default pay …	Default payment: yes=1, no=0

Assignment Project Exam Help
https://powcoder.com
Add WeChat powcoder

Copyright (c) by Daniel K.C. Chan. All Rights Reserved.

Python: Correlation-based Feature Selection (1)

load relevant packages

```
import matplotlib.pyplot as plt
import seaborn as sns
import pandas as pd
import numpy as np
```

Assignment Project Exam Help

load the credit card default dataset

https://powcoder.com

```
data = pd.read_csv('FIN7790-02-3-credit_card_default.csv', header=1, index_col=0)
```

Add WeChat powcoder

confirm the entire dataset is indeed loaded

```
data.shape
```

```
(30000, 24)
```

Copyright (c) by Daniel K.C. Chan. All Rights Reserved.

Python: Correlation-based Feature Selection (2)

examine the first 5 rows

```
data.head().T
```

ID	1	2	3	4	5
LIMIT_BAL	20000	120000	90000	50000	50000
SEX	2	2	2	2	1
EDUCATION	2	2	2	2	2
MARRIAGE	1	2	2	1	1
AGE	24	26	34	37	57
PAY_0	2	-1	0	0	-1
PAY_2	2	2	0	0	0
PAY_3	-1	0	0	0	-1
PAY_4	-1	0	0	0	0
PAY_5	-2	0	0	0	0
PAY_6	-2	2	0	0	0
BILL_AMT1	3913	2682	29239	46990	8617
BILL_AMT2	3102	1725	14027	48233	5670
BILL_AMT3	689	2682	13559	49291	35835
BILL_AMT4	0	3272	14331	28314	20940
BILL_AMT5	0	3455	14948	28959	19146
BILL_AMT6	0	3261	15549	29547	19131
PAY_AMT1	0	0	1518	2000	2000
PAY_AMT2	689	1000	1500	2019	36681
PAY_AMT3	0	1000	1000	1200	10000
PAY_AMT4	0	1000	1000	1100	9000
PAY_AMT5	0	0	1000	1069	689
PAY_AMT6	0	2000	5000	1000	679
default payment next month	1	1	0	0	0

Assignment Project Exam Help
https://powcoder.com
Add WeChat powcoder

Copyright (c) by Daniel K.C. Chan. All Rights Reserved.

Python: Correlation-based Feature Selection (3)

examine the statistics about the dataset

```
data.describe().T
```

Assignment Project Exam Help

	count	mean	std	min	25%	50%	75%	max
LIMIT_BAL	30000.0	167484.322667	129747.661567	10000.0	50000.00	140000.0	240000.00	1000000.0
SEX	30000.0	1.603733	0.489129	1.0	1.00	2.0	2.00	2.0
EDUCATION	30000.0	1.853133	0.790349	0.0	1.00	2.0	2.00	6.0
MARRIAGE	30000.0	1.551867	0.521970	0.0	1.00	2.0	2.00	3.0
AGE	30000.0	35.485500	9.217904	21.0	28.00	34.0	41.00	79.0
PAY_0	30000.0	-0.016700	1.123802	-2.0	-1.00	0.0	0.00	8.0
PAY_2	30000.0	-0.133767	1.197186	-2.0	-1.00	0.0	0.00	8.0
PAY_3	30000.0	-0.166200	1.196868	-2.0	-1.00	0.0	0.00	8.0
PAY_4	30000.0	-0.220667	1.169139	-2.0	-1.00	0.0	0.00	8.0
PAY_5	30000.0	-0.266200	1.133187	-2.0	-1.00	0.0	0.00	8.0
PAY_6	30000.0	-0.291100	1.149988	-2.0	-1.00	0.0	0.00	8.0

https://powcoder.com

Add WeChat powcoder

Copyright (c) by Daniel K.C. Chan. All Rights Reserved.

Python: Correlation-based Feature Selection (4)

check if there is any null values

```
data.isnull().sum()
```

```
LIMIT_BAL                     0
SEX                           0
EDUCATION                     0
MARRIAGE                      0
AGE                           0
PAY_0                         0
PAY_2                         0
PAY_3                         0
PAY_4                         0
PAY_5                         0
PAY_6                         0
BILL_AMT1                     0
BILL_AMT2                     0
BILL_AMT3                     0
BILL_AMT4                     0
BILL_AMT5                     0
BILL_AMT6                     0
PAY_AMT1                      0
PAY_AMT2                      0
PAY_AMT3                      0
PAY_AMT4                      0
PAY_AMT5                      0
PAY_AMT6                      0
default payment next month    0
dtype: int64
```

assumes no preprocessing is required
partition the dataset into features & target

Assignment Project Exam Help

```
target = 'default payment next month'
X = data.drop(target, axis = 1)
y = data[target]
```

https://powcoder.com

showed the normalized value counts

Add WeChat powcoder

```
y.value_counts(normalize=True)
```

```
0    0.7788
1    0.2212
Name: default payment next month, dtype: float64
```

Copyright (c) by Daniel K.C. Chan. All Rights Reserved.

Python: Correlation-based Feature Selection (5)

show the Pearson's correlation coefficients

```python
data.corr()
```

Assignment Project Exam Help

https://powcoder.com

Add WeChat powcoder

	LIMIT_BAL	SEX	EDUCATION	MARRIAGE	AGE	PAY_0	PAY_2	PAY_3	PAY_4	PAY_
LIMIT_BAL	1.000000	0.024755	-0.219161	-0.108139	0.144713	-0.271214	-0.296382	-0.286123	-0.267460	-0.2494
SEX	0.024755	1.000000	0.014232	-0.031389	-0.090874	-0.057643	-0.070771	-0.066096	-0.060173	-0.0550
EDUCATION	-0.219161	0.014232	1.000000	-0.143464	0.175061	0.105364	0.121566	0.114025	0.108793	0.0975
MARRIAGE	-0.108139	-0.031389	-0.143464	1.000000	-0.414170	0.019917	0.024199	0.032688	0.033122	0.0356
AGE	0.144713	-0.090874	0.175061	-0.414170	1.000000	-0.039447	-0.050148	-0.053048	-0.049722	-0.0538
PAY_0	-0.271214	-0.057643	0.105364	0.019917	-0.039447	1.000000	0.672164	0.574245	0.538841	0.5094
PAY_2	-0.296382	-0.070771	0.121566	0.024199	-0.050148	0.672164	1.000000	0.766552	0.662067	0.6227
PAY_3	-0.286123	-0.066096	0.114025	0.032688	-0.053048	0.574245	0.766552	1.000000	0.777359	0.6867
PAY_4	-0.267460	-0.060173	0.108793	0.033122	-0.049722	0.538841	0.662067	0.777359	1.000000	0.8198

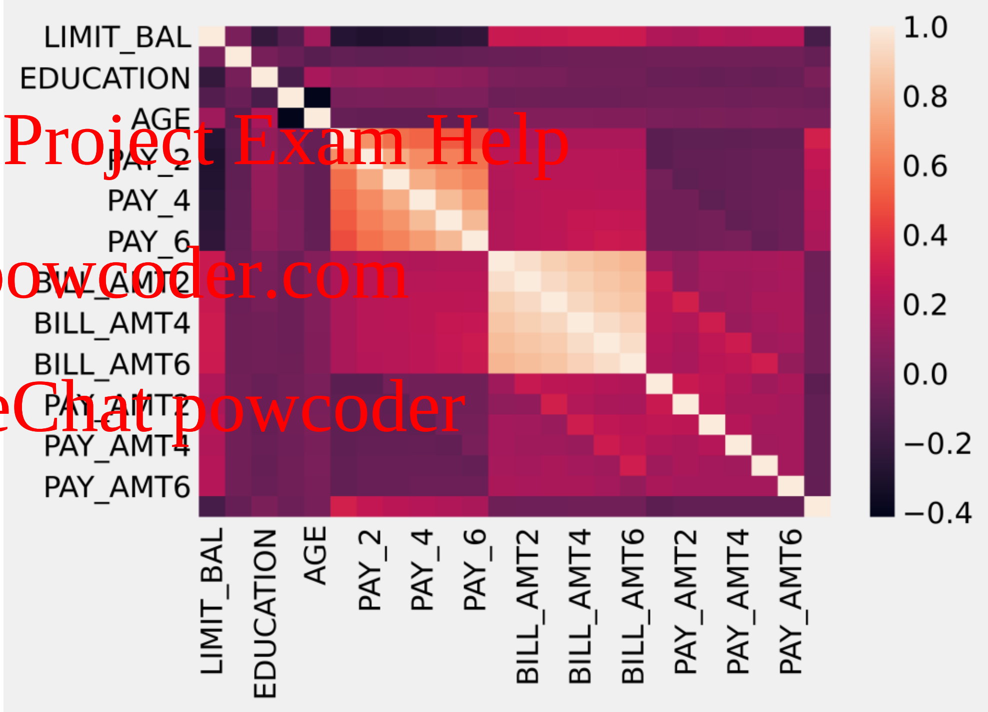
Copyright (c) by Daniel K.C. Chan. All Rights Reserved.

Python: Correlation-based Feature Selection (6)

```
# show Pearson's correlation
# coefficient as a heatmap
sns.heatmap(data.corr())
```

Note that the heatmap function automatically chose the most correlated features to show.

For simplicity no normalization is performed before computing the Pearson's coefficients.



Assignment Project Exam Help
https://powcoder.com
Add WeChat powcoder

Python: Correlation-based Feature Selection (7)

```
# list the coefficient against the target
data.corr()[target]
```

```
LIMIT_BAL                    -0.153520
SEX                          -0.039961
EDUCATION                     0.028006
MARRIAGE                     -0.024339
AGE                           0.013890
PAY_0                         0.324794
PAY_2                         0.263551
PAY_3                         0.235253
PAY_4                         0.216614
PAY_5                         0.204149
PAY_6                         0.186866
BILL_AMT1                    -0.019644
BILL_AMT2                    -0.014193
BILL_AMT3                    -0.014076
BILL_AMT4                    -0.010156
BILL_AMT5                    -0.006760
BILL_AMT6                    -0.005372
PAY_AMT1                     -0.072929
PAY_AMT2                     -0.058579
PAY_AMT3                     -0.056250
PAY_AMT4                     -0.056827
PAY_AMT5                     -0.055124
PAY_AMT6                     -0.053183
default payment next month    1.000000
Name: default payment next month, dtype: float64
```

```
# list the coefficient against the target
# only if the absolute value > 0.2
data.corr()[target].abs() > 0.2
```

```
LIMIT_BAL                     False
SEX                           False
EDUCATION                     False
MARRIAGE                      False
AGE                           False
PAY_0                          True
PAY_2                          True
PAY_3                          True
PAY_4                          True
PAY_5                          True
PAY_6                         False
BILL_AMT1                     False
BILL_AMT2                     False
BILL_AMT3                     False
BILL_AMT4                     False
BILL_AMT5                     False
BILL_AMT6                     False
PAY_AMT1                      False
PAY_AMT2                      False
PAY_AMT3                      False
PAY_AMT4                      False
PAY_AMT5                      False
PAY_AMT6                      False
default payment next month     True
Name: default payment next month, dtype: bool
```

Assignment Project Exam Help
https://powcoder.com
Add WeChat powcoder

Python: Correlation-based Feature Selection (8)

Retain the most correlated features

```
key_features = data.columns[data.corr()[target].abs() > 0.2]
key_features
```

```
Index(['PAY_0', 'PAY_2', 'PAY_3', 'PAY_4', 'PAY_5',
       'default payment next month'],
      dtype='object')
```

Assignment Project Exam Help

display the retained features of the dataset

```
data_trimmed = data[key_features]
data_trimmed
```

https://powcoder.com

Add WeChat powcoder

ID	PAY_0	PAY_2	PAY_3	PAY_4	PAY_5	default payment next month
1	2	2	-1	-1	-2	1
2	-1	2	0	0	0	1
3	0	0	[illegible]	0	0	0
4	0	0	0	0	0	0
5	-1	0	-1	0	0	0
...	...	...	...	...	...	...
29996	0	0	0	0	0	0
29997	-1	-1	-1	-1	0	0
29998	4	3	2	-1	0	1
29999	1	-1	0	0	0	1
30000	0	0	0	0	0	1

30000 rows × 6 columns

Copyright (c) by Daniel K.C. Chan. All Rights Reserved.

Prediction accuracy may suffer or improve as a result of feature selection depending of the choice of parameters

Before Feature Selection			
Model Name	Accuracy (%)	Fit Time (sec)	Predict Time (sec)
Decision Tree	0.8203	0.158	0.002

Assignment Project Exam Help
https://powcoder.com
Add WeChat powcoder

After Feature Selection					
Model Name	# of Features	Threshold	Accuracy (%)	Fit Time (sec)	Predict Time (sec)
Decision Tree	7	0.1	0.8206	0.105	0.003
Decision Tree	5	0.2	0.8197	0.010	0.002

Copyright (c) by Daniel K.C. Chan. All Rights Reserved.

Feature Selection using Hypothesis Testing

Assignment Project Exam Help

https://powcoder.com

Add WeChat powcoder

Hypothesis Testing

Assignment Project Exam Help

https://powcoder.com

Add WeChat powcoder

- Hypothesis testing is a method for testing a claim about a parameter in a population, using data measured in a sample
  1) State the hypothesis
  2) Set the criteria for a decision
  3) Compute the test statistic
  4) Make a decision
- The null hypothesis (H_0) is a statement about the population parameter that is assumed to be true
  - The reason of testing H_0 is because we think it is wrong!
- An alternative hypothesis (H_1) is a statement that directly contradicts H_0 by stating that the population parameter is different to what is stated in H_0

Copyright (c) by Daniel K.C. Chan. All Rights Reserved.

Presumption of Innocence 無罪推定原則

- The **presumption of innocence** is the legal principle that one is considered "**innocent until proven guilty**". Under the presumption of innocence, the legal **burden of proof** is thus on **the prosecution**, which must present compelling evidence to the trier of fact (a judge or a jury)
- **無罪推定原則**，意指一個人在法院上應該**先被假定為無罪**，除非被證實及判決有罪。 在這個原則下，提起公訴的**檢察官應負起舉證責任**，應負責收集足夠的可靠證據，以證明被告在事實上的確有罪；而若法院要判被告有罪，則所使用的證據必須符合法律限制，而且不能超越合理懷疑。

Assignment Project Exam Help
https://powcoder.com
Add WeChat powcoder

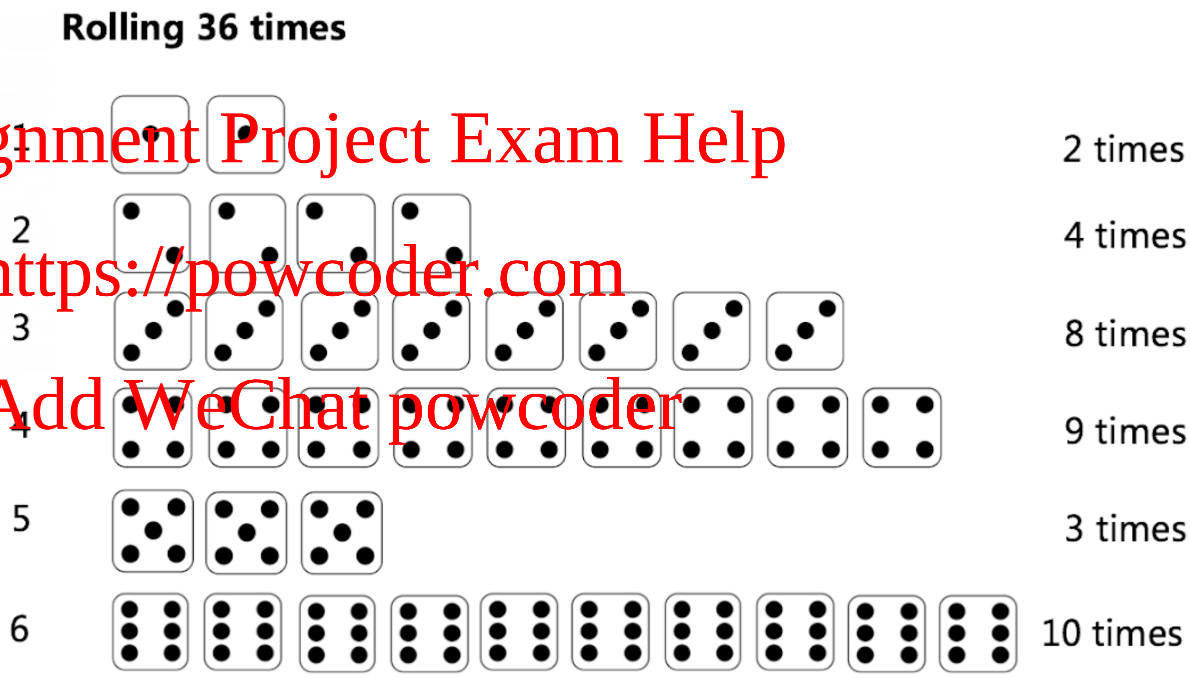
Copyright (c) by Daniel K.C. Chan. All Rights Reserved.

Chi-Squared Test

- The Chi-Squared test is used to determine whether a relationship between 2 categorical variables in a sample is likely to reflect a real association between these 2 variables in the population
  - In the case of 2 variables being compared, the test can be interpreted as determining if there is a difference between the 2 variables
- The sample data is used to calculate a single number, the test statistic
- The size of the test statistic reflects the probability that the observed relationship between 2 variables has occurred by chance

Assignment Project Exam Help
https://powcoder.com
Add WeChat powcoder

After rolling a dice 36 times, how can we determine if the dice is fair or unfair



Assignment Project Exam Help
https://powcoder.com
Add WeChat powcoder

How would you draw the conclusion?

Copyright (c) by Daniel K.C. Chan. All Rights Reserved.

Chi[2] test is used for categorical variables to reveal variance in observed & expected frequencies

Assignment Project Exam Help

$$Chi\!-\!Squared\ Score\ \chi^2 = \sum \frac{(Observed\ Frequency - Expected\ Frequency)^2}{Expected\ Frequency}$$

https://powcoder.com

Add WeChat powcoder

where, *Observed Frequency = Number of observations of class*

Expected Frequency = Number of expected observations of class if there was no relationship between the feature and the target

Copyright (c) by Daniel K.C. Chan. All Rights Reserved.

Chi^2 test calculates the variances in frequency and compares the sum with the Chi^2 distribution

Rolling 36 times

Table : 2rows*6 columns	1	2	3	4	5	6
Expected(E)	6	6	6	6	6	6
Observed(O)	2	4	8	9	3	10

Assignment Project Exam Help
https://powcoder.com
Add WeChat powcoder

$$\frac{(O-E)^2}{E} + \frac{(O-E)^2}{E} + \frac{(O-E)^2}{E} + \frac{(O-E)^2}{E} + \frac{(O-E)^2}{E} + \frac{(O-E)^2}{E}$$

$$\frac{(2-6)^2}{6} + \frac{(4-6)^2}{6} + \frac{(8-6)^2}{6} + \frac{(9-6)^2}{6} + \frac{(3-6)^2}{6} + \frac{(10-6)^2}{6}$$

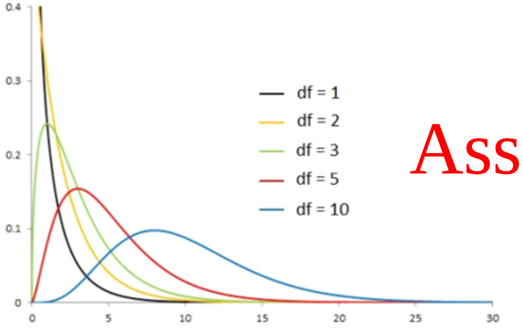
①Chi-Squared = **9.6**

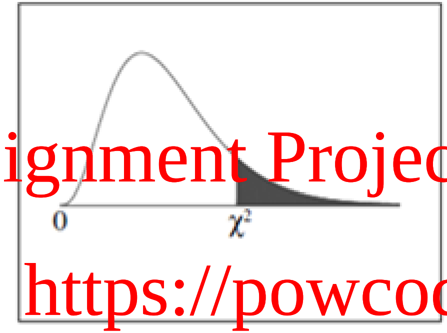
②Degree of freedom (#rows- 1) * (#columns– 1) = (2-1) * (6-1) = **5**

③ Significant level **90%** level of significance, typically set at 95%

Copyright (c) by Daniel K.C. Chan. All Rights Reserved.

The threshold in the Chi^2 distribution for the corresponding degree of freedom determines H_0's acceptance or rejection





The shaded area is equal to α for $\chi^2 = \chi^2_\alpha$.

Conclusion

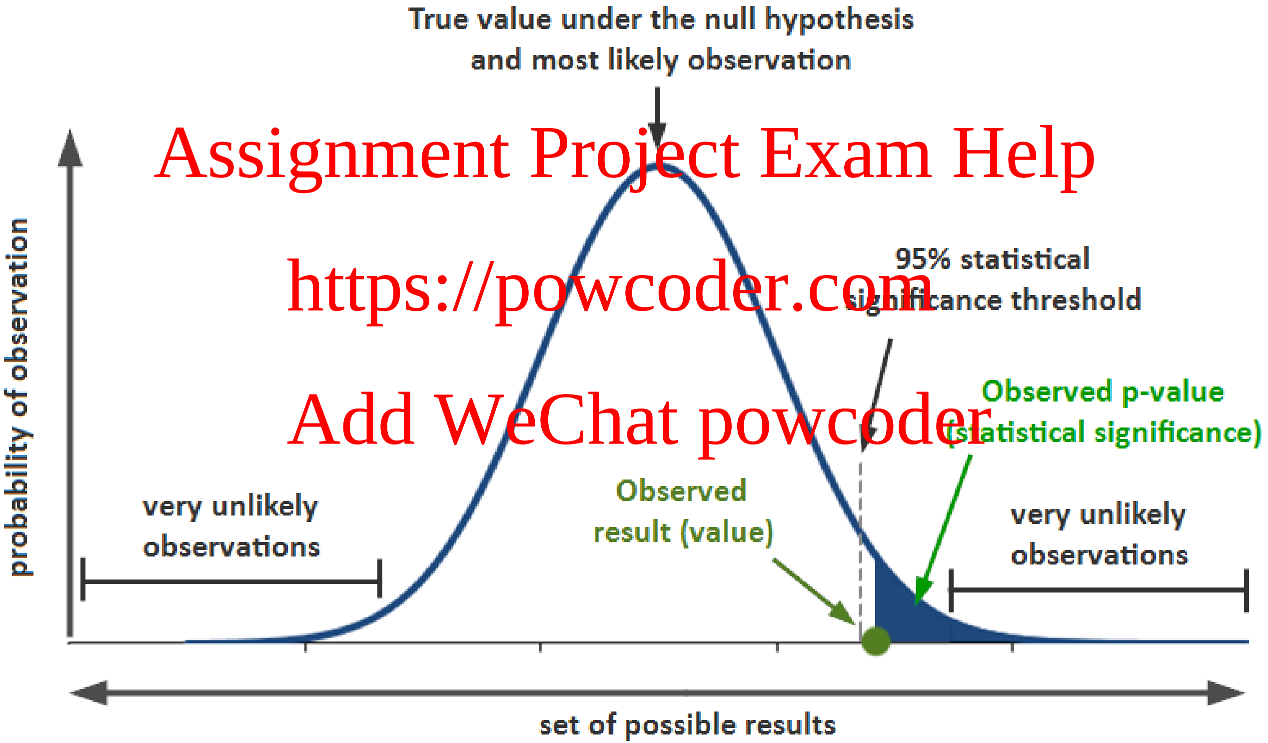
- test statistic (9.6) > threshold (9.236)
- suggests the dice is unbalanced
- reject the H_0 hypothesis

Assignment Project Exam Help
https://powcoder.com
Add WeChat powcoder

df	$\chi^2_{.995}$	$\chi^2_{.990}$	$\chi^2_{.975}$	$\chi^2_{.950}$	$\chi^2_{.900}$	$\chi^2_{.100}$	$\chi^2_{.050}$	$\chi^2_{.025}$	$\chi^2_{.010}$	$\chi^2_{.005}$
1	0.000	0.000	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.070	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090	21.955

Copyright (c) by Daniel K.C. Chan. All Rights Reserved.

Probability & Statistical Significance Explained



Assignment Project Exam Help
https://powcoder.com
Add WeChat powcoder

Dataset

Online grocery purchase	gender
1	Male
1	Male
1	Female
0	Male
0	Male
1	Female
1	Female
0	Female
...	...

Assignment Project Exam Help

https://powcoder.com

Add WeChat powcoder

Observed

	Male	Female	Total
Do not purchase grocery online	527	72	$\boxed{599}$
purchase grocery online	206	102	308
Total	733	174	$\boxed{907}$

Copyright (c) by Daniel K.C. Chan. All Rights Reserved.

Observed Table:

	Male	Female	Total
Do not purchase	527	72	599
purchase	206	102	308
Total	733	174	907

Assignment Project Exam Help

We found 66% of people don’ t purchase grocery food online, and 34% purchase from above table.

https://powcoder.com

If there are 733 male , 174 female, we can generate the following table by calculating the expected value with these ratio.

Add WeChat powcoder

Expected Table :

	Male	Female	Total
Do not purchase	$\boxed{484}$	115	599 (66%)
purchase	249	59	308 (34%)
Total	733	174	907

733 male * 66% don’ t purchase = 484

Copyright (c) by Daniel K.C. Chan. All Rights Reserved.

$$\frac{(O-E)^2}{E} + \frac{(O-E)^2}{E} + \frac{(O-E)^2}{E} + \frac{(O-E)^2}{E}$$

$$\frac{(527-484)^2}{484} + \frac{(72-115)^2}{115} + \frac{(206-249)^2}{249} + \frac{(102-59)^2}{59}$$

Assignment Project Exam Help

① Chi-Squared = **58.4**

https://powcoder.com

② Degree of freedom (rows − 1) * (column − 1) = (2-1) * (2-1) = **1**

Add WeChat powcoder

③ Significant level **90%** threshold = 2.706

Yes! Correlation!

conclusion : There is a correlation between gender and online purchase decision.

Copyright (c) by Daniel K.C. Chan. All Rights Reserved.

Chi-Squared based Feature Selection

- Chi^2 measures the distance between observed and expected frequencies
- The null hypothesis (H_0) is that the observed frequencies for a **categorical** variable match the expected frequencies for the **categorical** variable
- If **Score >= Threshold**: target depends on the feature, significant result, reject the null hypothesis (H_0), feature is to be retained
- If **Statistic < Threshold** : target does not depend on the feature, not significant result, fail to reject the null hypothesis (H_0), feature should be removed

Assignment Project Exam Help
https://powcoder.com
Add WeChat powcoder

Copyright (c) by Daniel K.C. Chan. All Rights Reserved.

Feature Transformation

Assignment Project Exam Help
https://powcoder.com
Add WeChat powcoder

Feature transformation creates new columns that are fundamentally different from the original dataset

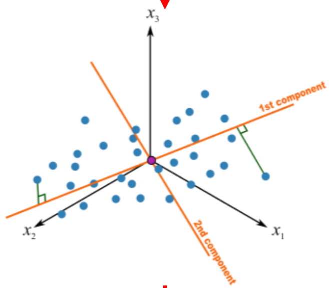
Name	Amount	Date	Issued In	Used In	Age	Education	Fraud?
Daniel	$2,600.45	1-Jul-2020	HK	HK	22	Secondary	No
Alex	$2,294.58	1-Oct-2020	HK	RUS	None	Postgraduate	Yes
Adrian	$1,003.30	3-Oct-2020	HK		25	Graduate	Yes
Vicky	$8,488.32	4-Oct-2020	JAPAN	HK	64	Graduate	No
Adams	¥20000	7-Oct-2020	AUS	JAP	58	Primary	No
…	…	…	…	…	…	…	…
Jones	₽3,250.11	Nov 1, 2020	HK	RUS	43	Graduate	No
Mary	₽8,156.20	Nov 1, 2020	HK	N/A	27	Graduate	Yes
Max	€7475,11	Nov 8, 2020	UK	GER	32	Primary	No
Peter	₽500.00	Nov 9, 2020	Hong Kong	RUS	0	Postgraduate	No
Anson	₽7,475.11	Nov 9, 2020	Hong Kong	RUS	20	Postgraduate	Yes

Observations

Feature

Target

feature transformation



…	…	…
…	…	…
…	…	…
…	…	…
…	…	…
…	…	…
…	…	…
…	…	…
…	…	…
…	…	…
…	…	…

Assignment Project Exam Help
https://powcoder.com
Add WeChat powcoder

Feature transformation creates an entirely new, structurally different dataset from the original dataset

- Feature selection processes are limited to only being able to select features from the original set of columns
- Feature transformation uses the original columns and combines them in useful ways to create new columns that are better at describing the data than any single column from the original dataset
- These algorithms create brand new columns that are so powerful that we only need a few of them to explain the entire dataset accurately

Assignment Project Exam Help
https://powcoder.com
Add WeChat powcoder

Copyright (c) by Daniel K.C. Chan. All Rights Reserved.

Feature transformation relies on matrix algorithms whereas feature learning relies on deep learning

- Feature transformation deploys a suite of algorithms designed to alter the internal structure of data to produce mathematically superior columns

Assignment Project Exam Help

- Feature learning will focus on using non-parametric algorithms (those that do not depend on the shape of the data) to automatically learn new features

https://powcoder.com

- Feature transformation uses a set of matrix algorithms that will structurally alter the dataset and produce what is essentially a brand new matrix of data

Add WeChat powcoder

  - The basic idea is that
    - the original features of a dataset are the descriptors / characteristics of data points and
    - it should be able to create a new set of features that explain the data-points just as well, perhaps even better, with fewer columns

Copyright (c) by Daniel K.C. Chan. All Rights Reserved.

Feature Transformation using Principal Component Analysis

Assignment Project Exam Help

https://powcoder.com

Add WeChat powcoder

Principal Component Analysis (PCA)

Assignment Project Exam Help

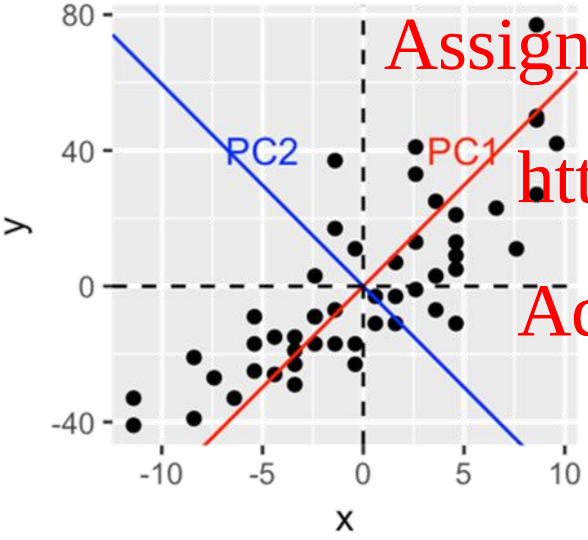
https://powcoder.com

Add WeChat powcoder

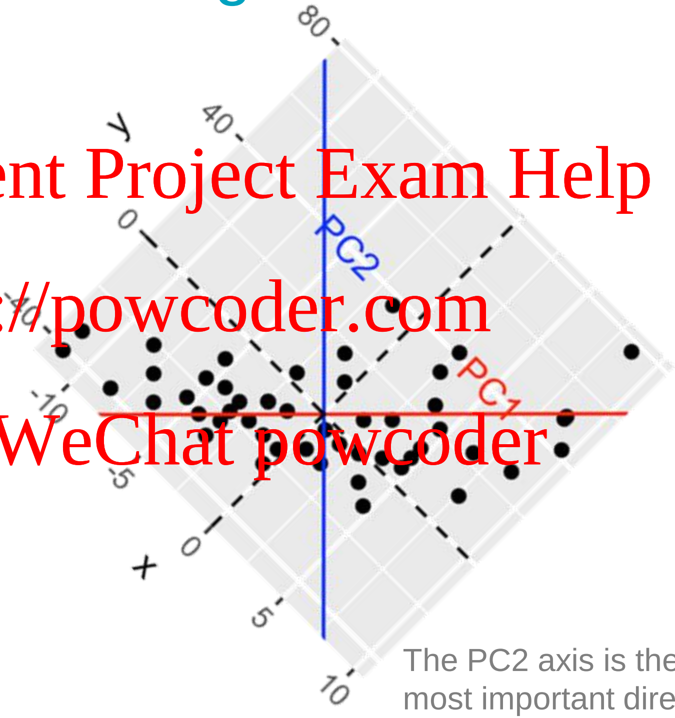
- PCA is used to extract the important information from a multivariate dataset and to express this information as a set of few new variables called principal components
- The principal components explain most of the patterns & latent structures observed in the original dataset
  - Often possible with only a few principal components
- An unsupervised dimension reduction technique providing a new lower-dimensional variable space to project the dataset on
- A linear static transformation using matrix multiplication

Copyright (c) by Daniel K.C. Chan. All Rights Reserved.

Graphically, PCA finds new orthogonal important dimensions that capture the largest variances



The dataset is represented in the X-Y coordinate system

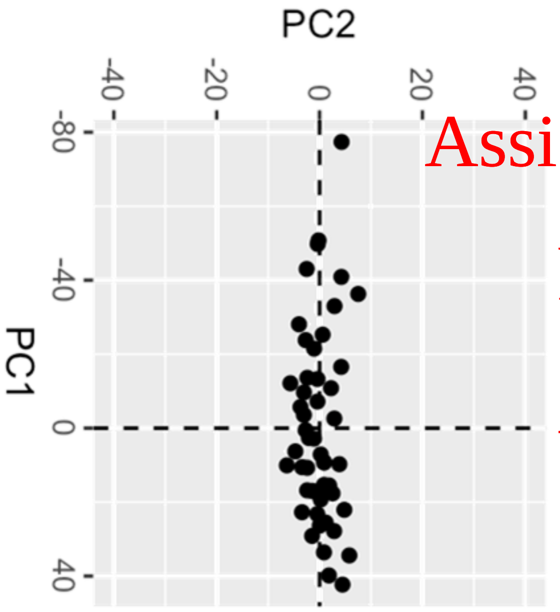


The PC1 axis is the first principal direction giving the largest sample variation

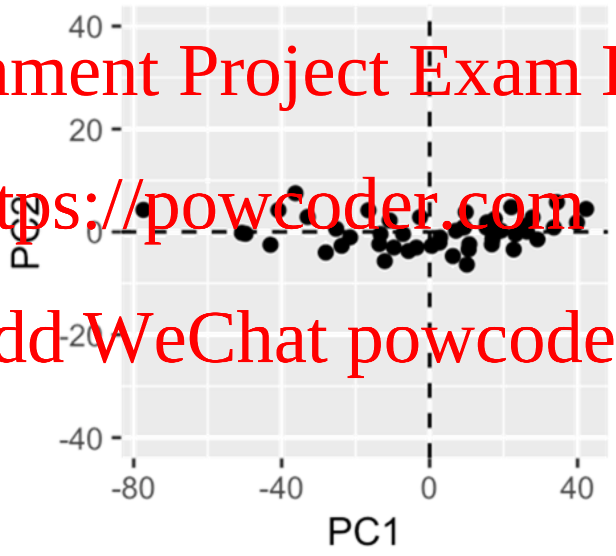
The PC2 axis is the second most important direction, orthogonal to the PC1 axis

Assignment Project Exam Help
https://powcoder.com
Add WeChat powcoder

The original data in a 2-dimension space can be effectively represented in a 1-dimension space



The PC2 axis is the second most important direction, orthogonal to the PC1 axis



The PC1 axis is the first principal direction giving the largest sample variation

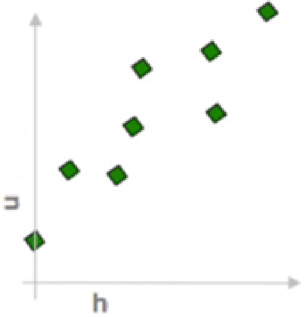
Assignment Project Exam Help
https://powcoder.com
Add WeChat powcoder

- The dimension reduction is achieved by identifying the principal directions, called principal components
- PCA assumes that the directions with the largest variances are the most important
- In this example, the two-dimensional data can be reduced to a single dimension by projecting each data onto the first principal component

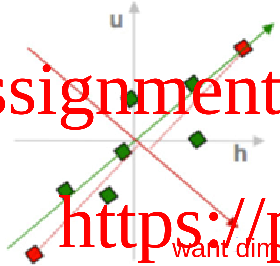
Copyright (c) by Daniel K.C. Chan. All Rights Reserved.

The computation of PCA is typically done using linear algebra and the identification of eigenvectors & eigenvalues

(1) correlated data of n dimensions



(2) center (don't scale) the data



want dimension of
the highest variance

(3) compute the covariance matrix

$$\begin{bmatrix} cov(h,h) & cov(h,u) \\ cov(u,h) & cov(u,u) \end{bmatrix}$$

$$= \begin{bmatrix} 2.0 & 0.8 \\ 0.8 & 0.6 \end{bmatrix}$$

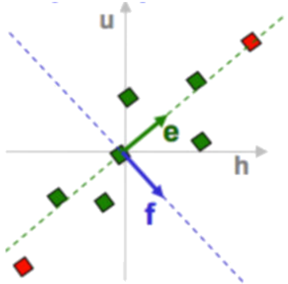
(4) compute the eigenvector and eigenvalues of the covariance matrix

$$\begin{bmatrix} 2.0 & 0.8 \\ 0.8 & 0.6 \end{bmatrix} \begin{bmatrix} e_h \\ e_u \end{bmatrix} = \lambda_e \begin{bmatrix} e_h \\ e_u \end{bmatrix}$$

$$\begin{bmatrix} 2.0 & 0.8 \\ 0.8 & 0.6 \end{bmatrix} \begin{bmatrix} f_h \\ f_u \end{bmatrix} = \lambda_f \begin{bmatrix} f_h \\ f_u \end{bmatrix}$$

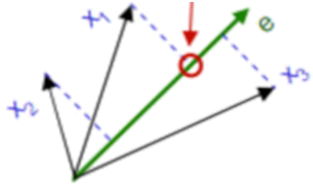
n orthogonal eigenvectors
for data of n dimensions

(5) keep the top k eigenvalues
(sorted by descending order)

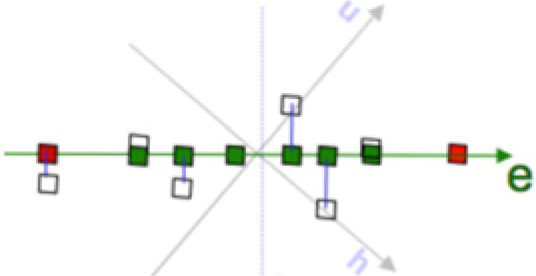


(6) project data multiplying the transpose of the feature vector with the eigenvectors corresponding to the top eigenvalues

$$x_e = x^{\tau} e$$



(6) uncorrelated data of lower dimensionality



Assignment Project Exam Help
https://powcoder.com
Add WeChat powcoder

Copyright (c) by Daniel K.C. Chan. All Rights Reserved.

Some reminders on linear algebra

- Variance computes the variation of the data distributed across the dimensionality graph

$$var(x) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}$$

- Covariance identifies the dependencies and relationships between the characteristics of datasets

$$cov(x, y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n}$$

- The sign of $cov(x, y)$ is the key
  - Positive: both dimensions increase together
  - Negative: one dimension increases, the other dimension decreases
  - 0: two dimensions are independent of each other

- An eigenvector ($\boldsymbol{v}$) of a linear transformation (A) is a non-zero vector (typically, a unity vector) that changes by a scalar factor ($\boldsymbol{\lambda}$) when that linear transformation is applied

$$A\boldsymbol{v} = \boldsymbol{\lambda}\boldsymbol{v}$$

$$\begin{bmatrix} t_{11} & \cdots & t_{1n} \\ \vdots & \ddots & \vdots \\ t_{m1} & \cdots & t_{mn} \end{bmatrix} \begin{bmatrix} v_1 \\ \vdots \\ v_n \end{bmatrix} = \boldsymbol{\lambda} \begin{bmatrix} v_1 \\ \vdots \\ v_n \end{bmatrix}$$

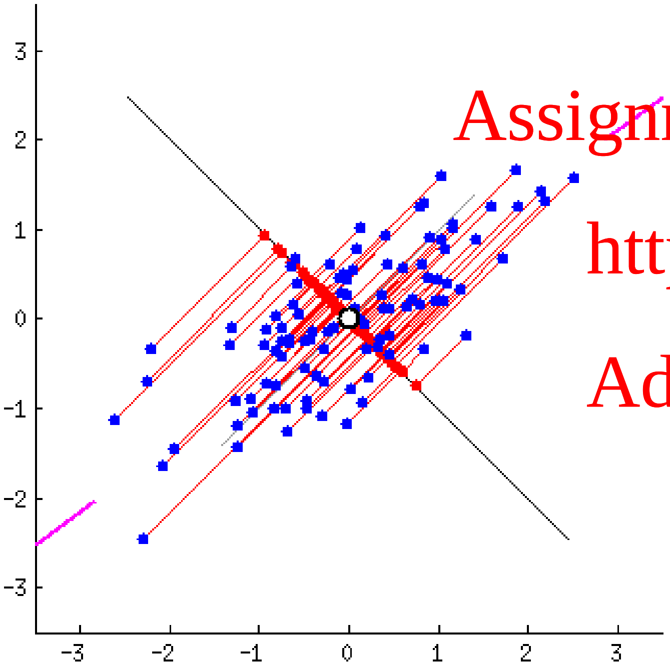
- The corresponding eigenvalue is the factor by which the eigenvector is scaled

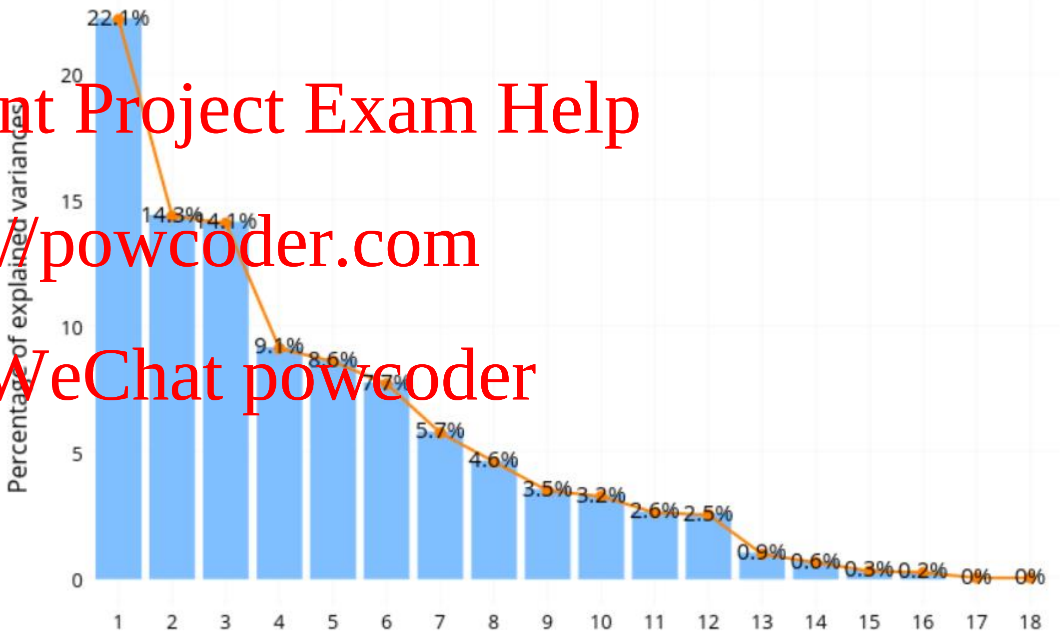
- Eigenvector and eigenvalue come in pair for a given linear transformation

Assignment Project Exam Help
https://powcoder.com
Add WeChat powcoder

Copyright (c) by Daniel K.C. Chan. All Rights Reserved.

Majority of the variance in the original dataset can be effectively explained by a few principal components





Assignment Project Exam Help

https://powcoder.com

Add WeChat powcoder

Understanding the Mathematics behind Principal Component Analysis (https://heartbeat.fritz.ai/understanding-the-mathematics-behind-principal-component-analysis-efd7c9ff0bb3)

Copyright (c) by Daniel K.C. Chan. All Rights Reserved.

Python: Using Principal Component Analysis (1)

```
# load relevant packages and data
import pandas as pd
import matplotlib.pyplot as plt
from sklearn.datasets import load_boston
from sklearn.model_selection import train_test_split
from sklearn.decomposition import PCA

boston_dataset = load_boston()
data = pd.DataFrame(boston_dataset.data, columns = boston_dataset.feature_names)
data['MEDV'] = boston_dataset.target
```

Assignment Project Exam Help
https://powcoder.com
Add WeChat powcoder

```
# separate the features from the target
X = data.drop('MEDV', axis = 1)
y = data['MEDV']
```

Copyright (c) by Daniel K.C. Chan. All Rights Reserved.

Python: Using Principal Component Analysis (2)

show the first 5 observations

```
data.head()
```

Assignment Project Exam Help
https://powcoder.com
Add WeChat powcoder

	CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PTRATIO	B	LSTAT	MEDV
0	0.00632	18.0	2.31	0.0	0.538	6.575	65.2	4.0900	1.0	296.0	15.3	396.90	4.98	24.0
1	0.02731	0.0	7.07	0.0	0.469	6.421	78.9	4.9671	2.0	242.0	17.8	396.90	9.14	21.6
2	0.02729	0.0	7.07	0.0	0.469	7.185	61.1	4.9671	2.0	242.0	17.8	392.83	4.03	34.7
3	0.03237	0.0	2.18	0.0	0.458	6.998	45.8	6.0622	3.0	222.0	18.7	394.63	2.94	33.4
4	0.06905	0.0	2.18	0.0	0.458	7.147	54.2	6.0622	3.0	222.0	18.7	396.90	5.33	36.2

show the number of rows and number of columns of the dataset

```
data.shape
```

```
(506, 14)
```

Python: Using Principal Component Analysis (3)

split the dataset into training dataset (70%) and testing dataset (30%)

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=0)
X_train.shape
```

```
(354, 13)
```

Assignment Project Exam Help

set up the PCA
n_components=None will keep all features in the original dataset
features will be ranked and selected in subsequent steps

https://powcoder.com

```
pca = PCA(n_components = None)
```

Add WeChat powcoder

train the PCA with the training dataset

```
pca.fit(X_train)
```

Copyright (c) by Daniel K.C. Chan. All Rights Reserved.

Python: Using Principal Component Analysis (4)

a few of the components will capture most of the variance of the original dataset
to identify how many components capture most of the variability,
we can plot the percentage of variance explained (by each component)
versus the component number

Assignment Project Exam Help

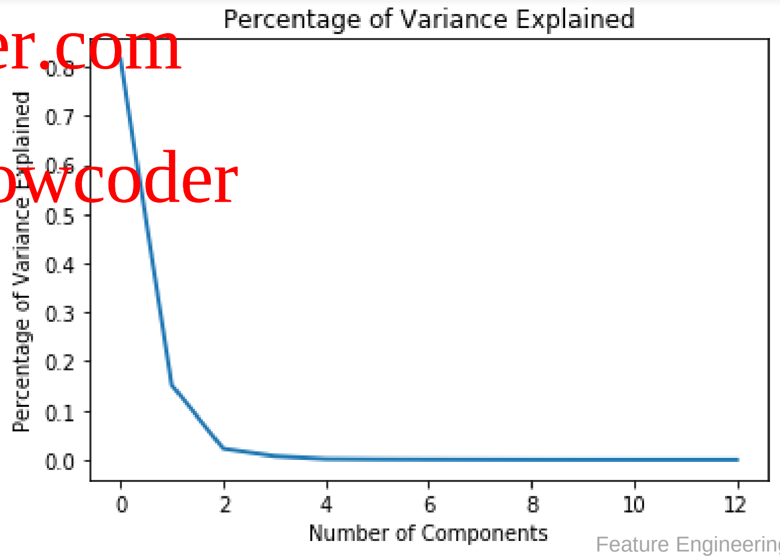
plot the percentage of the total variance
explained by each component

https://powcoder.com

```
plt.plot(pca.explained_variance_ratio_,
          linewidth =  2)
plt.title('Percentage of Variance Explained')
plt.xlabel('Number of Components')
plt.ylabel('Percentage of Variance Explained')
```

Add WeChat powcoder

the plot indicates that we can use the first
two components to train our machine learning
models using a linear model



Copyright (c) by Daniel K.C. Chan. All Rights Reserved.

Python: Using Principal Component Analysis (5)

transform the training and testing datasets

```
X_train_transformed = pca.transform(X_train)
X_test_transformed = pca.transform(X_test)
print(X_train_transformed)
```

Assignment Project Exam Help
https://powcoder.com
Add WeChat powcoder

```
[[ 2.84963123e+01 -4.38499065e+01 -3.14502960e+01 ... -3.84593774e-01
   5.86474148e-[illegible] -1.[illegible]e-02]
 [-1.82330673e+02  1.13476100e+01 -3.80261965e+00 ...  1.17281942e-01
  -8.48270610e-02 -2.37081258e-02]
 [ 2.84897466e+01 -4.1[illegible]e+01 -2.83140568e+01 ... -5.64773056e-01
  -6.28850879e-02  1.35[illegible]e-02]
 ...
 [ 2.20048157e+01 -4.04332881e+01 -1.53608650e+01 ...  4.35759491e-01
  -4.95804979e-02 -3.45554384e-02]
 [-1.68810736e+02  1.09954558e+01 -3.66885623e+01 ... -2.26660389e-01
  -6.91370706e-02 -6.74739271e-02]
 [-1.09061632e+02 -8.78956198e+00 -3.25570942e+01 ...  8.51171902e-01
  -6.56124719e-02 -6.40805251e-02]]
```

Copyright (c) by Daniel K.C. Chan. All Rights Reserved.

Python: Using Principal Component Analysis (6)

reduce the dimensionality of the dataset based on the result of PCA

```
X_train_trimmed = pd.DataFrame(X_train_transformed[:,0:2])
X_train_trimmed.head()
```

Assignment Project Exam Help

	0	1
0	28.496312	-43.849907
1	-182.350673	11.147610
2	28.489747	-41.863175
3	-51.448908	3.772947
4	-207.354549	36.293186

https://powcoder.com

Add WeChat powcoder

show the number of rows and columns of the reduced dataset

```
X_train_trimmed.shape
```

(354, 2)

Copyright (c) by Daniel K.C. Chan. All Rights Reserved.

Feature Learning

Assignment Project Exam Help

https://powcoder.com

Add WeChat powcoder

Feature learning relieves the restriction on the original dataset and uses deep learning to create new columns

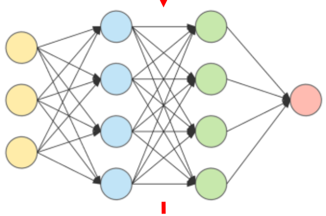
feature learning

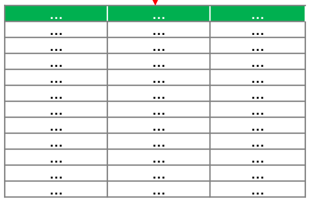
Name	Amount	Date	Issued In	Used In	Age	Education	Fraud?
Daniel	$2,600.45	1-Jul-2020	HK	HK	22	Secondary	No
Alex	$2,294.58	1-Oct-2020	HK	RUS	None	Postgraduate	Yes
Adrian	$1,003.30	3-Oct-2020	HK		25	Graduate	Yes
Vicky	$8,488.32	4-Oct-2020	JAPAN	HK	64	Graduate	No
Adams	¥20000	7-Oct-2020	AUS	JAP	58	Primary	No
...	...	...	...	...	...	...	...
Jones	₽3,250.11	Nov 1, 2020	HK	RUS	43	Graduate	No
Mary	₽8,156.20	Nov 1, 2020	HK	N/A	27	Graduate	Yes
Max	€7475,11	Nov 8, 2020	UK	GER	32	Primary	No
Peter	₽500.00	Nov 9, 2020	Hong Kong	RUS	0	Postgraduate	No
Anson	₽7,475.11	Nov 9, 2020	Hong Kong	RUS	20	Postgraduate	Yes

Observations

Feature

Target





Assignment Project Exam Help
https://powcoder.com
Add WeChat powcoder

Copyright (c) by Daniel K.C. Chan. All Rights Reserved.

Feature Learning

- Creates brand-new features from existing features making no assumption on the shape of the data
  - Feature learning algorithms are not parametric
- Relies on stochastic learning
  - Instead of applying the same equation to the data every time, algorithms will discover the best features by looking at the data over and over again (in epochs) and converge onto a solution (potentially different ones at runtime)
- Can learn fewer or more features than in the original dataset and the exact number of features to learn depends on the problem and can be grid-searched

Assignment Project Exam Help
https://powcoder.com
Add WeChat powcoder

Copyright (c) by Daniel K.C. Chan. All Rights Reserved.

Feature Transformation vs Feature Learning

	Feature Transformation Algorithms	Feature Learning Algorithms
Parametric	Yes	No
Simple to use	Yes	No
New feature set	Yes	Yes
Deep learning	No	Yes
Algorithms	PCA, LDA	Deep learning

Assignment Project Exam Help
https://powcoder.com
Add WeChat powcoder

- A model being non-parametric does not mean that no assumptions are made at all by the model during training
- Feature learning algorithms forgo the assumption on the shape of the data but they still may make assumptions on other aspects of the data (e.g., the variable values)

Copyright (c) by Daniel K.C. Chan. All Rights Reserved.

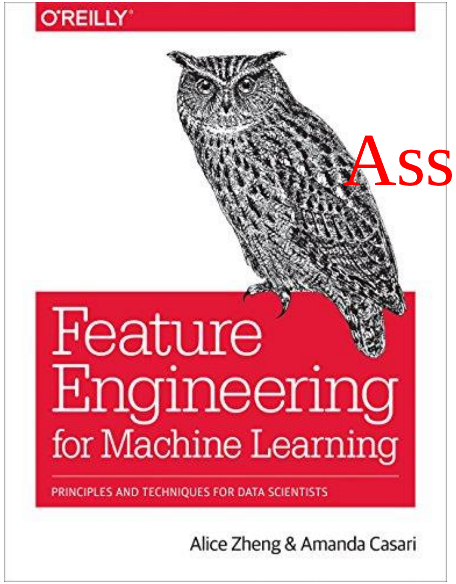
References

Assignment Project Exam Help

https://powcoder.com

Add WeChat powcoder

References



"Feature Engineering for Machine Learning: Principles and Techniques for Data Scientists", Alice Zhang & Amanda Casari, O'Reilly Media, April 2018, ISBN-13: 978-1-491-95324-2

"Python Feature Engineering Cookbook", Soledad Galli, Packt Publishing, January 2020, ISBN-13: 978-1-789-80631-1

"Feature Engineering Made Simple", Susan Ozdemir & Divya Susarla, Packt Publishing, January 2018, ISBN-13: 978-1-787-28760-0

Assignment Project Exam Help
https://powcoder.com
Add WeChat powcoder

Feature Engineering

- "How to Choose a Feature Selection Method for Machine Learning", Jason Browniee, November 2019 (https://machinelearningmastery.com/feature-selection-with-real-and-categorical-data/)
- "Correlation Coefficient Calculator" (https://byjus.com/correlation-coefficient-calculator/)
- "How to Perform Feature Selection with Categorical Data", Jason Browniee, November 2019 (https://machinelearningmastery.com/feature-selection-with-categorical-data/)
- "A Gentle Introduction to the Chi-Squared Test for Machine Learning", Jason Browniee, June 2018 (https://machinelearningmastery.com/chi-squared-test-for-machine-learning/)
- "Chi-Squared Test for Feature Selection in Machine Learning", Sampath Kumar Gajawada, October 2019 (https://towardsdatascience.com/chi-square-test-for-feature-selection-in-machine-learning-206b1f0b8223)
- "Chi-Squared Test Calculator" (https://www.socscistatistics.com/tests/chisquare2/default2.aspx)
- "A Tutorial on Principal Components Analysis", Lindsay Smith, February 2002 (http://www.cs.otago.ac.nz/cosc453/student_tutorials/principal_components.pdf)

Assignment Project Exam Help
https://powcoder.com
Add WeChat powcoder

Copyright (c) by Daniel K.C. Chan. All Rights Reserved.

Assignment Project Exam Help

https://powcoder.com

THANK YOU

Add WeChat powcoder