Mechatronic Engineering program

Python for machine learning and data science:
2: Scientific method, exploratory data analysis

Ziemowit Dworakowski
AGH University of Krakow

1

Example no 1

1) A => B (A causes B)
2) B => A (B causes A)
3) C => B and C => A (something causes A and B)
4) A ! => B and B !=> A (coincidence)

2

Example no 1

1) A => B (A causes B)
2) B => A (B causes A)
3) C => B and C => A (something causes A and B)
4) **A ! => B and B !=> A (coincidence)**

Maybe gather more data
and observe whether the relations is maintained?

3

Example no 1

1) A => B (A causes B)
2) B => A (B causes A)
3) **C => B and C => A** **(something causes A and B)**

Maybe calculate more features and check for other correlated features?

Maybe try to artificially cause A or B and see if the other follows?

4

Example no 1

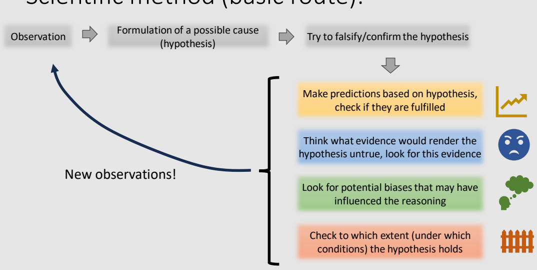
1) A => B (A causes B)
2) **B => A** **(B causes A)**

Maybe look for time-dependent relations? If something was observed first, it might have been the root cause

Maybe look for physical model to infer causality based on expert knowledge?

5

Scientific method (basic route):



6

This concept works in most of the typical scenarios:

We observe relations between features → We think one can be predicted based on the other → We build regression system and test it on **independent data subset** → We check if the results are statistically significant

We see that income in one group of occupations tends to be higher → We gather more data and check how likely it is that this relation was pure coincidence

We see which damage cases were recognized by a damage classifier and claim it can regognize up to 90% of cases → We gather more examples of different damage cases → We check if our hypothesis holds

7

We often go back-and-forth here...

Step 0: Initial research question or goal (optional)

- Transform variables
- Check experimental procedures
- Check and address data quality (outliers? Missing values?)

Step 1: Initial Data Analysis

- Build understanding of data
- Look for patterns and data relations (influences, correlations, causality)
- Draw hypotheses
- Model the data

Step 2: Exploratory Data Analysis

- Check if patterns hold for new data
- Check if models are still valid for new data
- Check if hypotheses are confirmed
- Double-check for biases

Step 3: Confirmatory Data Analysis

Note the gradient!

The more time we spend in the early steps, the less data we need for the later steps!

8

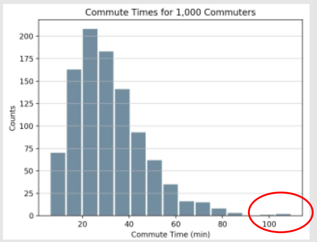
Different kinds of outliers and anomalies

Employment area	Earnings / mo	Age	Education	Opened term deposit?
Agriculture	3900 $	34	MSc	Yes
Agriculture	2400 $	45	College	Yes
Law	5400 $	41	**Illiterate**	Yes
Transportation	3000 $	30	College	No
Science & Ed	4200 $	36	MSc	No
Food industry	5100 $	32	BSc	Yes

0.7 % of people are marked as „Illiterate"
Some of them are definitely educated.
What to do with such data?

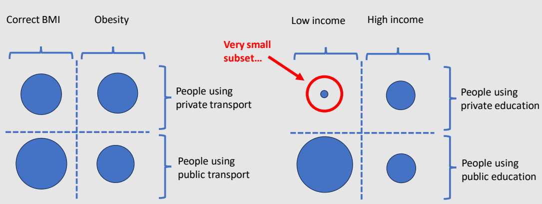
9

Different kinds of outliers and anomalies



10

Different kinds of outliers and anomalies



Outliers can be found also in features' relations to each other. Each feature value, when considered separately, can be perfectly typical, but their combination can form an anomaly

11

Should we remove them?

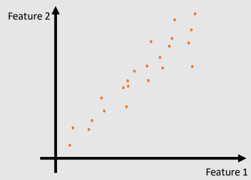
It depends:

- In general outliers caused by measurement errors should probably be deleted from training dataset. If we can provide a well-described methodology for deleting them, we can also do so for testing dataset. Otherwise it is better to leave them for testing – to actually measure expected final accuracy

- Anomalies that are just rare but correctly gathered examples of data are not actually outliers (they are important for us!). Still, they will decrease efficiency of training because they will result in poorly-sampled regions of feature space. We can either **accept** them or **include an anomaly detector** to mark them in the future.

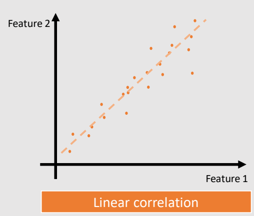
12

Features' correlations



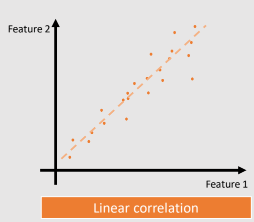
13

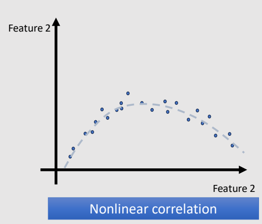
Features' correlations



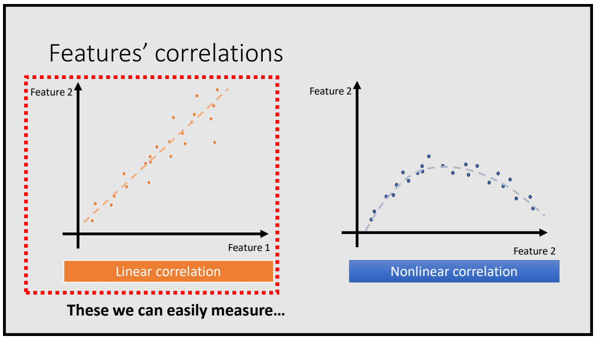
14

Features' correlations

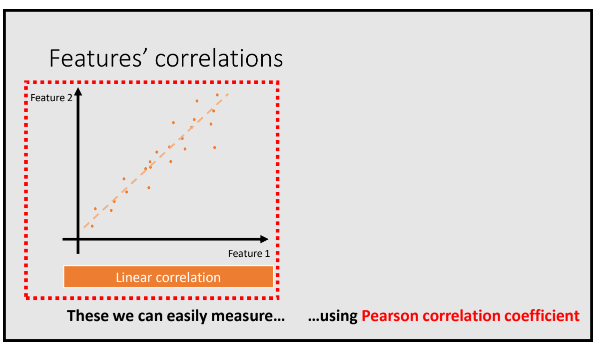




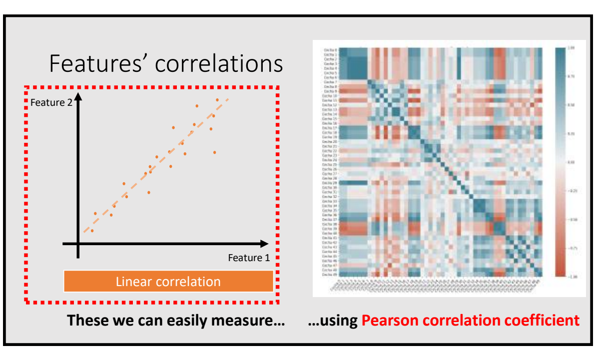
15


Features' correlations
Feature 2
Feature 1
Linear correlation
Feature 2
Feature 2
Nonlinear correlation
These we can easily measure…

16


Features' correlations
Feature 2
Feature 1
Linear correlation
These we can easily measure…
…using Pearson correlation coefficient

17


Features' correlations
Feature 2
Feature 1
Linear correlation
These we can easily measure…
…using Pearson correlation coefficient

18

Features' correlations

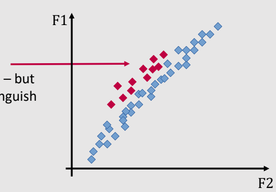
We want the features to **be highly correlated** with target value (in regression) but we also want the features **to not be correlated** with each other *(to reduce redundancy in our dataset).* And there is a risk in that:

19

Features' correlations

We want the features to **be highly correlated** with target value (in regression) but we also want the features **to not be correlated** with each other *(to reduce redundancy in our dataset).* And there is a risk in that:

Example:
F1 is strongly correlated with F2 – but they both are necessary to distinguish classes from each other



20

Principal Component Analysis (PCA)

21

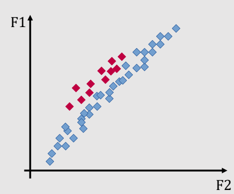
Principal Component Analysis (PCA)

If we want to quickly reduce dimensionality of the problem and transform features to form uncorrelated ones, PCA is a standard tool to do so.

22

Principal Component Analysis (PCA)

If we want to quickly reduce dimensionality of the problem and transform features to form uncorrelated ones, PCA is a standard tool to do so.

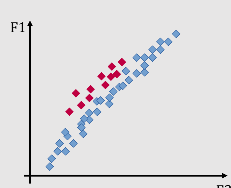


23

Principal Component Analysis (PCA)

If we want to quickly reduce dimensionality of the problem and transform features to form uncorrelated ones, PCA is a standard tool to do so.

PCA maps data into new space where each dimension **(principal component)** is orthogonal to all others and maximizes remaining variance

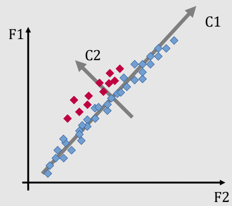


24

Principal Component Analysis (PCA)

If we want to quickly reduce dimensionality of the problem and transform features to form uncorrelated ones, PCA is a standard tool to do so.

PCA maps data into new space where each dimension **(principal component)** is orthogonal to all others and maximizes remaining variance



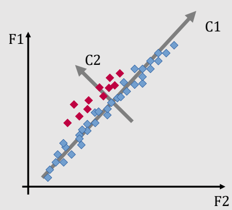
25

Principal Component Analysis (PCA)

If we want to quickly reduce dimensionality of the problem and transform features to form uncorrelated ones, PCA is a standard tool to do so.

PCA maps data into new space where each dimension **(principal component)** is orthogonal to all others and maximizes remaining variance

And then we just take the first few principal components and use them for classification or regression.



26

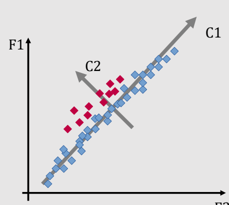
Principal Component Analysis (PCA)

If we want to quickly reduce dimensionality of the problem and transform features to form uncorrelated ones, PCA is a standard tool to do so.

PCA maps data into new space where each dimension **(principal component)** is orthogonal to all others and maximizes remaining variance

And then we just take the first few principal components and use them for classification or regression.

Note! This still poses a risk of getting rid of important information!

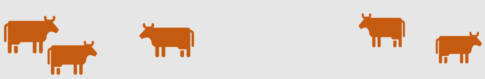


27

Hypotheses testing

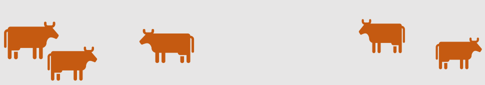
28

Hypotheses testing



29

Hypotheses testing



How many cows do we need to see to confirm
hypothesis that they are usually brown?

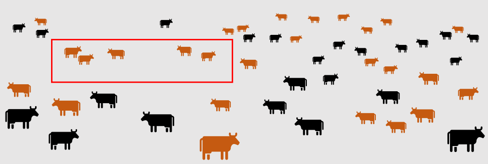
30

Hypotheses testing

How many cows do we need to see to confirm hypothesis that they are usually brown?

31

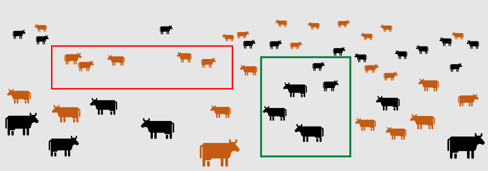
Hypotheses testing



How many cows do we need to see to confirm hypothesis that they are usually brown?

32

Hypotheses testing



How many cows do we need to see to confirm hypothesis that they are usually brown?

33

Hypotheses testing

P-value: Probability of obtaining observed results by coincidence (while the assumed hypothesis is not true)

34

Hypotheses testing

P-value: Probability of obtaining observed results by coincidence (while the assumed hypothesis is not true)

In a (gross) simplification:

1) Start from a formulated hypothesis (e.g. **more than 90%** of cows are brown)

35

Hypotheses testing

P-value: Probability of obtaining observed results by coincidence (while the assumed hypothesis is not true)

In a (gross) simplification:

1) Start from a formulated hypothesis (e.g. **more than 90%** of cows are brown)
2) Assume a threshold for „hypothesis acceptance" -> **α = 0.05** This means that we accept the fact, that 5% of our studies will end in a false positive conclusion (in strict tests **α** probably should be higher)

36

Hypotheses testing

P-value: Probability of obtaining observed results by coincidence (while the assumed hypothesis is not true)

In a (gross) simplification:

1) Start from a formulated hypothesis (e.g. **more than 90%** of cows are brown)
2) Assume a threshold for „hypothesis acceptance" -> **α = 0.05** This means that we accept the fact, that 5% of our studies will end in a false positive conclusion (in strict tests **α** probably should be higher)
3) Consider experimental data (e.g.: we observed 5 brown cows)

37

Hypotheses testing

P-value: Probability of obtaining observed results by coincidence (while the assumed hypothesis is not true)

In a (gross) simplification:

1) Start from a formulated hypothesis (e.g. **more than 90%** of cows are brown)
2) Assume a threshold for „hypothesis acceptance" -> **α = 0.05** This means that we accept the fact, that 5% of our studies will end in a false positive conclusion (in strict tests **α** probably should be higher)
3) Consider experimental data (e.g.: we observed 5 brown cows)
4) Calculate how likely it is, that a test result returned 5 brown cows if the underlying probability of a brown cow is **at most** 90%
(That we obtained such results despite hypothesis being not true)

38

Hypotheses testing

P-value: Probability of obtaining observed results by coincidence (while the assumed hypothesis is not true)

In a (gross) simplification:

1) Start from a formulated hypothesis (e.g. **more than 90%** of cows are brown)
2) Assume a threshold for „hypothesis acceptance" -> **α = 0.05** This means that we accept the fact, that 5% of our studies will end in a false positive conclusion (in strict tests **α** probably should be higher)
3) Consider experimental data (e.g.: we observed 5 brown cows)
4) Calculate how likely it is, that a test result returned 5 brown cows if the underlying probability of a brown cow is **at most** 90%
(That we obtained such results despite hypothesis being not true)

$$p = 0.9^5 = 0{,}59$$

39

Hypotheses testing

P-value: Probability of obtaining observed results by coincidence (while the assumed hypothesis is not true)

In a (gross) simplification:

1) Start from a formulated hypothesis (e.g. **more than 90%** of cows are brown)
2) Assume a threshold for „hypothesis acceptance" -> **α = 0.05** This means that we accept the fact, that 5% of our studies will end in a false positive conclusion (in strict tests **α** probably should be higher)
3) Consider experimental data (e.g.: we observed 5 brown cows)
4) Calculate how likely it is, that a test result returned 5 brown cows if the underlying probability of a brown cow is **at most** 90%
(That we obtained such results despite hypothesis being not true)
5) If this probability (**p-value**) is lower than **0.05**, we say that the test was statistically significant (It is unlikely to get this data given false hypothesis)

$p = 0.9^5 = 0{,}59$

40

Hypotheses testing

P-value: Probability of obtaining observed results by coincidence (while the assumed hypothesis is not true)

Brown cows	Black cows	Assumed hypothesis	Null hypothesis (alternative to assumed?)	p-value
5	0	More than 90% brown	90% or less are brown	0,590

$p = 0.9^5 = 0{,}590$

41

Hypotheses testing

P-value: Probability of obtaining observed results by coincidence (while the assumed hypothesis is not true)

Brown cows	Black cows	Assumed hypothesis	Null hypothesis (alternative to assumed?)	p-value
5	0	More than 90% brown	90% or less are brown	0,590
50	0	More than 90% brown	90% or less are brown	0,005

$p = 0.9^5 = 0{,}590$
$p = 0.9^{50} = 0{,}005$

42

Hypotheses testing

P-value: Probability of obtaining observed results by coincidence (while the assumed hypothesis is not true)

Brown cows	Black cows	Assumed hypothesis	Null hypothesis (alternative to assumed?)	p-value
5	0	More than 90% brown	90% or less are brown	0,590
50	0	More than 90% brown	90% or less are brown	0,005
5	0	More than 50% brown	50% or less are brown	0,031

$p = 0.9^5 = 0{,}590$
$p = 0.9^{50} = 0{,}005$
$p = 0.5^5 = 0{,}031$

43

Hypotheses testing

P-value: Probability of obtaining observed results by coincidence (while the assumed hypothesis is not true)

Brown cows	Black cows	Assumed hypothesis	Null hypothesis (alternative to assumed?)	p-value
5	0	More than 90% brown	90% or less are brown	0,590
50	0	More than 90% brown	90% or less are brown	0,005
5	0	More than 50% brown	50% or less are brown	0,031
4	1	More than 90% brown	90% or less are brown	0.328

$p = 0.9^5 = 0{,}590$
$p = 0.9^{50} = 0{,}005$
$p = 0.5^5 = 0{,}031$
$p = 5 \cdot (0.9^4 \cdot 0.1) = 0.328$

44

Hypotheses testing

P-value: Probability of obtaining observed results by coincidence (while the assumed hypothesis is not true)

Brown cows	Black cows	Assumed hypothesis	Null hypothesis (alternative to assumed?)	p-value
5	0	More than 90% brown	90% or less are brown	0,590
50	0	More than 90% brown	90% or less are brown	0,005
5	0	More than 50% brown	50% or less are brown	0,031
4	1	More than 90% brown	90% or less are brown	0.328
			better: 80% are brown	0.409

$p = 0.9^5 = 0{,}590$
$p = 0.9^{50} = 0{,}005$
$p = 0.5^5 = 0{,}031$
$p = 5 \cdot (0.9^4 \cdot 0.1) = 0.328$
$p = 5 \cdot (0.8^4 \cdot 0.2) = 0.409$

45

Hypotheses testing

P-value: Probability of obtaining observed results by coincidence (while the assumed hypothesis is not true)

Brown cows	Black cows	Assumed hypothesis	Null hypothesis (alternative to assumed?)	p-value
5	0	More than 90% brown	90% or less are brown	0,590
50	0	More than 90% brown	90% or less are brown	0,005
5	0	More than 50% brown	50% or less are brown	0,031
4	1	More than 90% brown	90% or less are brown	0.328
			better: 80% are brown	0.409
0	5	More than 90% brown	90% or less are brown	0.00001?
			better: 0% are brown	1

$p = 0.9^5 = 0{,}590$

$p = 0.9^{50} = 0{,}005$

$p = 0.5^5 = 0{,}031$

$p = 5 \cdot (0.9^4 \cdot 0.1) = 0.328$

$p = 5 \cdot (0.8^4 \cdot 0.2) = 0.409$

$\cancel{p < 0.1^5 = 0.00001}$

$p < 1^5 = 1$

46

Hypotheses testing

Low enough p-value tells us only that data support hypothesis in a statistically significant way. It is **not** a final confirmation of the hypothesis.

47

Hypotheses testing

Low enough p-value tells us only that data support hypothesis in a statistically significant way. It is **not** a final confirmation of the hypothesis.

It works under the assumption that the tests are independent to each other (*often not true! We can just happen to observe one pasture with cows of the same color and it will tell us nothing regarding the whole cow population*)

48

Hypotheses testing

Low enough p-value tells us only that data support hypothesis in a statistically significant way. It is **not** a final confirmation of the hypothesis.

It works under the assumption that the tests are independent to each other (*often not true! We can just happen to observe one pasture with cows of the same color and it will tell us nothing regarding the whole cow population*)

High p-value **does not** tell us that the hypothesis is false.

49

Hypotheses testing

Low enough p-value tells us only that data support hypothesis in a statistically significant way. It is **not** a final confirmation of the hypothesis.

It works under the assumption that the tests are independent to each other (*often not true! We can just happen to observe one pasture with cows of the same color and it will tell us nothing regarding the whole cow population*)

High p-value **does not** tell us that the hypothesis is false.

p-value **should not** be used for early stopping of the experiment or to select a subset of data to confirm a hypothesis <- this is *p-hacking*

50

Things to remember:

1. Explain an overview on scientific method
2. Explain steps of data analysis (IDA, EDA, CDA)
3. Explain sources for anomalies, explain how they affect model preparation
4. What does it mean that data are correlated? How do we measure correlation?
5. How does PCA work?
6. What are the risks associated with looking blindly into correlation information or using PCA?
7. How do we test hypotheses?
8. Explain what a p-value is and how is it used. Note what p-value does **not** allow.

51