Data Mining and Machine Learning
Introduction

Gergely Horváth

September 15, 2021

Outline

You are expected to ...

You are expected to ...

- Pass the test covering the material of the semester

You are expected to ...

- Pass the test covering the material of the semester
- Pass the short quizzes, the passing score is given by the average of their results (At the beginning of every lab session, 50% passing score)

You are expected to ...

- Pass the test covering the material of the semester
- Pass the short quizzes, the passing score is given by the average of their results (At the beginning of every lab session, 50% passing score)
- Complete the semester project – (the task will be posted later)

You are expected to ...

- Pass the test covering the material of the semester
- Pass the short quizzes, the passing score is given by the average of their results (At the beginning of every lab session, 50% passing score)
- Complete the semester project – (the task will be posted later)
- You can miss a maximum of 3 lab sessions

Today, you will learn about ...

Today, you will learn about ...

- Basics of the IDE of your choice

Today, you will learn about ...

- Basics of the IDE of your choice
  - Google Colab

Today, you will learn about ...

- Basics of the IDE of your choice
    - Google Colab
    - PyCharm (professional is suggested because of Jupyter support)

Today, you will learn about ...

- Basics of the IDE of your choice
  - Google Colab
  - PyCharm (professional is suggested because of Jupyter support)
  - VSCode (easy to use)

Today, you will learn about ...

- Basics of the IDE of your choice
  - Google Colab
  - PyCharm (professional is suggested because of Jupyter support)
  - VSCode (easy to use)
  - Other platforms (it might happen that I will not be able to assist with those)

Today, you will learn about ...

- Basics of the IDE of your choice
  - Google Colab
  - PyCharm (professional is suggested because of Jupyter support)
  - VSCode (easy to use)
  - Other platforms (it might happen that I will not be able to assist with those)
- Basics of Python required for the course

Today, you will learn about ...

- Basics of the IDE of your choice
    - Google Colab
    - PyCharm (professional is suggested because of Jupyter support)
    - VSCode (easy to use)
    - Other platforms (it might happen that I will not be able to assist with those)
- Basics of Python required for the course
    - Pandas

Today, you will learn about ...

- Basics of the IDE of your choice
  - Google Colab
  - PyCharm (professional is suggested because of Jupyter support)
  - VSCode (easy to use)
  - Other platforms (it might happen that I will not be able to assist with those)
- Basics of Python required for the course
  - Pandas
- The concept of EDA

Today, you will learn about ...

- Basics of the IDE of your choice
  - Google Colab
  - PyCharm (professional is suggested because of Jupyter support)
  - VSCode (easy to use)
  - Other platforms (it might happen that I will not be able to assist with those)
- Basics of Python required for the course
  - Pandas
- The concept of EDA
- Bayesian networks

IDE

IDE

- Google Colab → easiest to set-up, straightforward to use (advised to have a Gmail account)

IDE

- Google Colab $\rightarrow$ easiest to set-up, straightforward to use (advised to have a Gmail account)

- PyCharm $\rightarrow$ widely used professional development environment, professional licence can be obtained via your student e-mail address (community version does not have Jupyter support)

IDE

- Google Colab → easiest to set-up, straightforward to use (advised to have a Gmail account)
- PyCharm → widely used professional development environment, professional licence can be obtained via your student e-mail address (community version does not have Jupyter support)
- VSCode → easy to set-up, customize, use and rich of features compared to other freeware options (will be the IDE of my choice)

IDE

- Google Colab → easiest to set-up, straightforward to use (advised to have a Gmail account)
- PyCharm → widely used professional development environment, professional licence can be obtained via your student e-mail address (community version does not have Jupyter support)
- VSCode → easy to set-up, customize, use and rich of features compared to other freeware options (will be the IDE of my choice)
- Any other environment of your choice

Python

Python

- The language of prototyping and machine learning

Python

- The language of prototyping and machine learning
- Packages that will be used the most frequently: Pandas, Scikit-learn

Python

- The language of prototyping and machine learning
- Packages that will be used the most frequently: Pandas, Scikit-learn
- Anyone would like some help or assistance in Python?

Python

- The language of prototyping and machine learning
- Packages that will be used the most frequently: Pandas, Scikit-learn
- Anyone would like some help or assistance in Python?
- Our favorite data structure on the course: Dataframe

Python

- The language of prototyping and machine learning
- Packages that will be used the most frequently: Pandas, Scikit-learn
- Anyone would like some help or assistance in Python?
- Our favorite data structure on the course: Dataframe
- Knowing and understanding OOP will be a benefit in understanding the codes

Python

- The language of prototyping and machine learning
- Packages that will be used the most frequently: Pandas, Scikit-learn
- Anyone would like some help or assistance in Python?
- Our favorite data structure on the course: Dataframe
- Knowing and understanding OOP will be a benefit in understanding the codes
- We will be using Jupyter most of the time, although you are welcomed to do scripting

Python

- The language of prototyping and machine learning
- Packages that will be used the most frequently: Pandas, Scikit-learn
- Anyone would like some help or assistance in Python?
- Our favorite data structure on the course: Dataframe
- Knowing and understanding OOP will be a benefit in understanding the codes
- We will be using Jupyter most of the time, although you are welcomed to do scripting
- What is Jupyter? → "Project Jupyter exists to develop open-source software, open-standards, and services for interactive computing across dozens of programming languages."

EDA

EDA

- Machine learning is essentially data-driven

EDA

- Machine learning is essentially data-driven
- Widely applicable mathematical frameworks, where the quality inherently depends on the quality of the data

EDA

- Machine learning is essentially data-driven
- Widely applicable mathematical frameworks, where the quality inherently depends on the quality of the data
- Therefore you have to have insights, how your data looks like

EDA

- Machine learning is essentially data-driven
- Widely applicable mathematical frameworks, where the quality inherently depends on the quality of the data
- Therefore you have to have insights, how your data looks like
- Descriptive statistics

EDA

- Machine learning is essentially data-driven
- Widely applicable mathematical frameworks, where the quality inherently depends on the quality of the data
- Therefore you have to have insights, how your data looks like
- Descriptive statistics
- Exploratory Data Analysis (EDA) is often time-consuming

EDA

- Machine learning is essentially data-driven
- Widely applicable mathematical frameworks, where the quality inherently depends on the quality of the data
- Therefore you have to have insights, how your data looks like
- Descriptive statistics
- Exploratory Data Analysis (EDA) is often time-consuming
- EDA can be done with: Pandas, Numpy, Matplotlib, Seaborn, etc.

EDA

- Machine learning is essentially data-driven
- Widely applicable mathematical frameworks, where the quality inherently depends on the quality of the data
- Therefore you have to have insights, how your data looks like
- Descriptive statistics
- Exploratory Data Analysis (EDA) is often time-consuming
- EDA can be done with: Pandas, Numpy, Matplotlib, Seaborn, etc.
- An automated EDA tool: Sweetviz

Bayesian networks – I.

Bayes' theorem:

$$P(A \mid B) = \frac{P(A \wedge B)}{P(B)}$$

$$P(A \wedge B) = P(B \mid A) \cdot P(A)$$

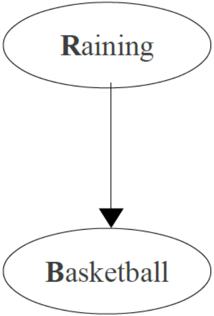
$$P(A \mid B) = \frac{P(B \mid A) \cdot P(A))}{P(B)}$$

First set of exercises:

$$P(\boldsymbol{R} = T) = \quad ?$$

$$P(\boldsymbol{B} = T) = \quad ?$$

$$P(\boldsymbol{B} = T \wedge \boldsymbol{R} = T) = \quad ?$$



P(**R**=T)
0.1

r	P(**B**=T | **R**=r)
T	0.2
F	0.7

Bayesian networks – I.

Bayes' theorem:

$$P(A \mid B) = \frac{P(A \wedge B)}{P(B)}$$

$$P(A \wedge B) = P(B \mid A) \cdot P(A)$$

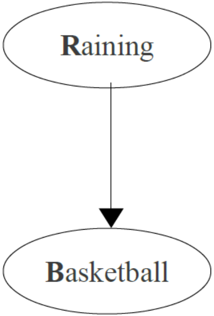
$$P(A \mid B) = \frac{P(B \mid A) \cdot P(A))}{P(B)}$$

First set of exercises:

$$P(\boldsymbol{R} = T) = \quad ?$$

$$P(\boldsymbol{B} = T) = \quad ?$$

$$P(\boldsymbol{B} = T \wedge \boldsymbol{R} = T) = \quad ?$$

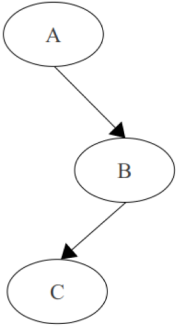


P(**R**=T)
0.1

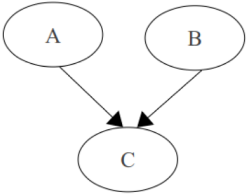
r	P(**B**=T | **R**=r)
T	0.2
F	0.7

Solutions: 0.1, 0.65, 0.02.

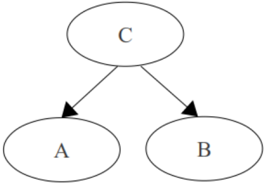
Bayesian networks – II.



1.
If event B is known:
A and C are independent.
$$P(C|A \wedge B) = P(C|B)$$
$$P(C \wedge A|B) = P(C|B) \cdot P(A|B)$$



2.
If event C is <u>not</u> known:
A and B are independent.
$$P(A \wedge B) = P(A) \cdot P(B)$$



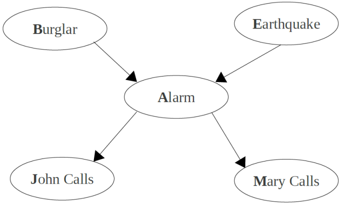
3.
If event C is known:
A and B are independent.
$$P(A|C \wedge B) = P(A|C)$$
$$P(A \wedge B|C) = P(A|C) \cdot P(B|C)$$

Second set of exercises:

$$P(J \wedge M \wedge A \wedge \neg B \wedge \neg E) = \quad ?\,, \quad P(B \mid J) = \quad ?\,, \quad P(J) = \quad ?$$

P(**B**=T)
0.001

P(**E**=T)
0.002



b	e	P(**A**=T | **B**=b ^ **E**=e)
T	T	0.950
T	F	0.940
F	T	0.290
F	F	0.001

a	P(**J**=T | **A**=a)
T	0.900
F	0.050

a	P(**M**=T | **A**=a)
T	0.700
F	0.010

Bayesian networks – III.

$$P(J \wedge M \wedge A \wedge \neg B \wedge \neg E) = \quad ?$$

Bayesian networks – III.

$$P(J \wedge M \wedge A \wedge \neg B \wedge \neg E) = \quad ?$$

$$P(J \wedge M \wedge A \wedge \neg B \wedge \neg E) = P(J \wedge M \mid A \wedge \neg B \wedge \neg E) \cdot P(A \wedge \neg B \wedge \neg E) =$$

Bayesian networks – III.

$$P(J \wedge M \wedge A \wedge \neg B \wedge \neg E) = \quad ?$$

$$\begin{aligned} P(J \wedge M \wedge A \wedge \neg B \wedge \neg E) &= P(J \wedge M \mid A \wedge \neg B \wedge \neg E) \cdot P(A \wedge \neg B \wedge \neg E) = \\ &= P(J \wedge M \mid A) \cdot P(A \mid \neg B \wedge \neg E) \cdot P(\neg B \wedge \neg E) = \end{aligned}$$

Bayesian networks – III.

$$
\begin{aligned}
P(J \wedge M \wedge A \wedge \neg B \wedge \neg E) &= \quad ? \\
P(J \wedge M \wedge A \wedge \neg B \wedge \neg E) &= P(J \wedge M \mid A \wedge \neg B \wedge \neg E) \cdot P(A \wedge \neg B \wedge \neg E) = \\
&= P(J \wedge M \mid A) \cdot P(A \mid \neg B \wedge \neg E) \cdot P(\neg B \wedge \neg E) = \\
&= P(J \mid A) \cdot P(M \mid A) \cdot P(A \mid \neg B \wedge \neg E) \cdot P(\neg B) \cdot P(\neg E) =
\end{aligned}
$$

Bayesian networks – III.

$$
\begin{aligned}
P(J \wedge M \wedge A \wedge \neg B \wedge \neg E) &= \quad ? \\
P(J \wedge M \wedge A \wedge \neg B \wedge \neg E) &= P(J \wedge M \mid A \wedge \neg B \wedge \neg E) \cdot P(A \wedge \neg B \wedge \neg E) = \\
&= P(J \wedge M \mid A) \cdot P(A \mid \neg B \wedge \neg E) \cdot P(\neg B \wedge \neg E) = \\
&= P(J \mid A) \cdot P(M \mid A) \cdot P(A \mid \neg B \wedge \neg E) \cdot P(\neg B) \cdot P(\neg E) = \\
&= 0.9
\end{aligned}
$$

Bayesian networks – III.

$$
\begin{aligned}
P(J \wedge M \wedge A \wedge \neg B \wedge \neg E) &= \quad ? \\
P(J \wedge M \wedge A \wedge \neg B \wedge \neg E) &= P(J \wedge M \mid A \wedge \neg B \wedge \neg E) \cdot P(A \wedge \neg B \wedge \neg E) = \\
&= P(J \wedge M \mid A) \cdot P(A \mid \neg B \wedge \neg E) \cdot P(\neg B \wedge \neg E) = \\
&= P(J \mid A) \cdot P(M \mid A) \cdot P(A \mid \neg B \wedge \neg E) \cdot P(\neg B) \cdot P(\neg E) = \\
&= 0.9 \cdot 0.7
\end{aligned}
$$

Bayesian networks – III.

$$
\begin{aligned}
P(J \wedge M \wedge A \wedge \neg B \wedge \neg E) &= \quad ? \\
P(J \wedge M \wedge A \wedge \neg B \wedge \neg E) &= P(J \wedge M \mid A \wedge \neg B \wedge \neg E) \cdot P(A \wedge \neg B \wedge \neg E) = \\
&= P(J \wedge M \mid A) \cdot P(A \mid \neg B \wedge \neg E) \cdot P(\neg B \wedge \neg E) = \\
&= P(J \mid A) \cdot P(M \mid A) \cdot P(A \mid \neg B \wedge \neg E) \cdot P(\neg B) \cdot P(\neg E) = \\
&= 0.9 \cdot 0.7 \cdot 0.001
\end{aligned}
$$

Bayesian networks – III.

$$
\begin{aligned}
P(J \wedge M \wedge A \wedge \neg B \wedge \neg E) &= \quad ? \\
P(J \wedge M \wedge A \wedge \neg B \wedge \neg E) &= P(J \wedge M \mid A \wedge \neg B \wedge \neg E) \cdot P(A \wedge \neg B \wedge \neg E) = \\
&= P(J \wedge M \mid A) \cdot P(A \mid \neg B \wedge \neg E) \cdot P(\neg B \wedge \neg E) = \\
&= P(J \mid A) \cdot P(M \mid A) \cdot P(A \mid \neg B \wedge \neg E) \cdot P(\neg B) \cdot P(\neg E) = \\
&= 0.9 \cdot 0.7 \cdot 0.001 \cdot 0.999
\end{aligned}
$$

Bayesian networks – III.

$$P(J \wedge M \wedge A \wedge \neg B \wedge \neg E) = \quad ?$$

$$\begin{aligned}
P(J \wedge M \wedge A \wedge \neg B \wedge \neg E) &= P(J \wedge M \mid A \wedge \neg B \wedge \neg E) \cdot P(A \wedge \neg B \wedge \neg E) = \\
&= P(J \wedge M \mid A) \cdot P(A \mid \neg B \wedge \neg E) \cdot P(\neg B \wedge \neg E) = \\
&= P(J \mid A) \cdot P(M \mid A) \cdot P(A \mid \neg B \wedge \neg E) \cdot P(\neg B) \cdot P(\neg E) = \\
&= 0.9 \cdot 0.7 \cdot 0.001 \cdot 0.999 \cdot 0.998
\end{aligned}$$

Bayesian networks – III.

$$P(J \wedge M \wedge A \wedge \neg B \wedge \neg E) = \quad ?$$

$$\begin{aligned} P(J \wedge M \wedge A \wedge \neg B \wedge \neg E) &= P(J \wedge M \mid A \wedge \neg B \wedge \neg E) \cdot P(A \wedge \neg B \wedge \neg E) = \\ &= P(J \wedge M \mid A) \cdot P(A \mid \neg B \wedge \neg E) \cdot P(\neg B \wedge \neg E) = \\ &= P(J \mid A) \cdot P(M \mid A) \cdot P(A \mid \neg B \wedge \neg E) \cdot P(\neg B) \cdot P(\neg E) = \\ &= 0.9 \cdot 0.7 \cdot 0.001 \cdot 0.999 \cdot 0.998 = 0.00063 \end{aligned}$$

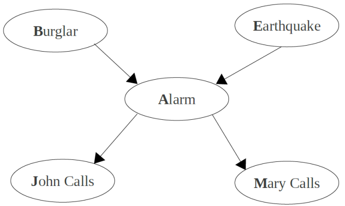
Bayesian networks – III.

Second set of exercises:

$$P(J \wedge M \wedge A \wedge \neg B \wedge \neg E) = \quad ?\,, \qquad P(B\,|\,J) = \quad ?\,, \qquad P(J) = \quad ?$$

P(**B**=T)
0.001

P(**E**=T)
0.002



b	e	P(**A**=T | **B**=b ^ **E**=e)
T	T	0.950
T	F	0.940
F	T	0.290
F	F	0.001

a	P(**J**=T | **A**=a)
T	0.900
F	0.050

a	P(**M**=T | **A**=a)
T	0.700
F	0.010

Solutions: 0.00063, 0.00085, 0.016.