

# Машинно обучение - въведение

Доц. д-р Ивайло Пенев

Кат. „Компютърни науки и технологии“

# Примери за машинно обучение

- Извличане на данни (database mining) – натрупване на данни от web, напр. действия на потребители (web click data)
- Приложения (алгоритми), които не могат да бъдат програмирани „ръчно“ – напр. автономни превозни средства, разпознаване на ръкописен текст (handwriting recognition), обработване на човешки език (Natural Language Processing - NLP), компютърни зрение (Computer Vision)
- Препоръчващи приложения (self-customizing programs) – напр. препоръчване на продукти от Amazon, Netflix, Facebook
- Наподобяване на човешко обучение (human learning) – имитиране на човешки разум, опити за истински изкуствен интелект (Artificial Intelligence - AI)

# Какво е машинно обучение (machine learning)?

- Две дефиниции за машинно обучение

- Arthur Samuel (1959) - по-стара (неформална) дефиниция

Наука, която изучава способността на компютрите да се обучават за изпълнение на задача без да бъдат програмирани за конкретната задача.

- Tom Mitchell (1998) – по-съвременна (формална) дефиниция

Казваме, че дадена компютърна програма се обучава чрез опит  $E$  за изпълнение на клас от задачи  $T$  с мярка за ефективност  $P$  тогава, когато ефективността на програмата при изпълнение на задача от класа  $T$ , измерена чрез мярката  $P$ , се подобрява с увеличаване на опита  $E$ .

# Пример за игра на шах

- Игра
  - $E$  – опит, натрупан в резултат от изиграване на голям брой игри
  - $T$  – задачата на играта (победа над противник)
  - $P$  – вероятността обучената компютърна програма да спечели играта срещу нов (непознат) съперник

# Пример

Имаме програма за email, която следи потребителя кои писма отбелязва като spam и така се обучава да разпознава spam в email. Какво представлява задачата T при тази постановка?

- Класифицира (определя) email като spam или не spam.
- Следи потребителите кои писма отбелязват като spam.
- Брой (или съотношение) на писмата, които правилно са класифицирани като spam (съответно не spam).
- Никое от изброените, защото това не е алгоритъм за машинно обучение.

# Видове машинно обучение

- Надзиравано (контролирано) машинно обучение (Supervised machine learning) или машинно обучение с надзирател (учител)
- Ненадзиравано (неконтролирано) машинно обучение (Unsupervised machine learning)

# Прост пример за обучение

- Родители учат децата как да постъпват в дадени ситуации – обучение с надзирател (supervised learning) Децата се учат как да постъпват в дадени ситуации само на базата на свой предишен опит – обучение без надзирател (unsupervised learning)

# Обучение с надзирател

- Разполагаме с дадено множество от входни данни
- **Знаем какви са коректните изходни стойности (результати) за конкретни входни данни**
- Търсим връзка (зависимост) между входните данни и резултатите
- Задачите, решавани чрез надзиравано машинно обучение се разделят в две групи:
  - Регресионни задачи (Regression problems) – изходните стойности (результати) са **непрекъсната функция** на входните данни
  - Класификационни задачи (Classification problems) – изходните стойности (результати) са **дискретни стойности (обикновено малък брой)**, т.е. търси се съответствие на дадени входни данни с точно една дискретна стойност



# Обучение с надзирател - примери

- Регресионна задача
  - Разполагаме с данни за цени на жилища с дадена площ. Да се предскаже цената на жилище с известна площ.
  - Цената е непрекъснатата функция на площта, следователно задачата е регресионна
- Класификационна задача
  - Имаме зададена цена за продаване на жилище. За каква цена ще бъде продадено жилището – по-ниска или по-висока от зададената?
  - В тази задача дадени входни данни (т.е. жилище с известна площ) се отнасят (класифицират) към една от две категории, следователно задачата е класификационна

# Обучение с надзирател – други примери

- Регресионна задача
  - По дадена снимка на мъж/жена да се предскаже възрастта на човека
- Класификационни задачи
  - По дадена снимка на мъж/жена да се предскаже дали човекът е ученик, студент или завършил
  - По дадена кредитна история на клиент банката решава дали да отпусне заем на клиента

# Контролен въпрос

- Разработват алгоритъм с машинно обучение за решаване на следните две задачи:
  - Задача 1. Имате голям склад с множество стоки. Опитват се да предскажете колко от тези стоки ще бъдат продадени през следващите 6 месеца.
  - Задача 2. Разработват софтуер, който проверява индивидуални профили на клиенти и за всеки профил се опитва да определи дали е пробит (т.е. хакнат).
- Как ще решите тези задачи – като регресионни или класификационни?
  - а) Двете задачи са класификационни
  - б) Задача 1 е класификационна, а задача 2 е регресионна
  - в) Задача 1 е регресионна, а задача 2 е класификационна
  - г) Двете задачи са регресионни

# Обучение без надзирател (Unsupervised machine learning)

- Разполагаме с дадено множество от входни данни
- Разполагаме с **МАЛКО (ИЛИ С НИКАКВА)** информация за **коректните изходни стойности (результати)** за конкретни входни данни
- В процеса на обучение не разполагаме с „учител“ за корекция на предсказания резултат
- Задачите, решавани чрез ненадзиравано машинно обучение, се разделят в две групи:
  - Задачи за клъстеризация (Clustering problems)
  - Задачи за не-клъстеризация (Non-clustering problems)

# News.google.com – обучение без надзирателя

The screenshot shows the Google News homepage in a browser window. The address bar displays the URL `https://news.google.com/?hl=en-US&gl=US&ceid=US:en`. The page features a search bar at the top with the text "Search for topics, locations & sources" and a "Sign in" button. On the left, there is a navigation menu with categories like "Top stories", "For you", "Favorites", "Saved searches", and various topics such as "U.S.", "World", "Business", "Technology", "Entertainment", "Sports", "Science", and "Health".

The main content area is titled "Headlines" and contains several news items. Two items are circled in blue:

- Headline 1:** "Federal prosecutors subpoena Trump's inaugural committee". It includes sub-headlines from AOL (5 hours ago), CNN (4 hours ago), The Washington Post (5 hours ago), and The New York Times (today). A small image of Donald Trump is visible to the right.
- Headline 2:** "Joshua Trump's name got him bullied. Now the sixth-grader is going to the State of the Union." It includes a sub-headline from The Washington Post (4 hours ago) and Fox News (today).

Other visible headlines include "VA lawmaker who's known Northam for a decade: He 'cannot effectively govern'" and "Senator Lindsey Graham Warns Of Possible GOP 'War' Over Trump's Wall".

On the right side, there is a weather widget for Varna showing "Partly cloudy" and "46°F", a "Fact check" section with several items, and a "Spotlight" section.

The Windows taskbar at the bottom shows the system tray with the date "10:06 5.2.2019" and language "ENG".

Week1.pdf - Adobe Acrobat Reader DC  
File Edit View Window Help

Home Tools Week1.pdf x

125%

Share

The slide is divided into four quadrants, each illustrating a different application of machine learning. The top-left quadrant shows two server racks with the caption "Organize computing clusters". The top-right quadrant shows a network graph with nodes of different colors and connecting arrows, captioned "Social network analysis". The bottom-left quadrant shows a 3D pie chart with three segments (green, blue, red) and human icons on each, captioned "Market segmentation". The bottom-right quadrant shows a colorful nebula in space, captioned "Astronomical data analysis".

Organize computing clusters

Social network analysis

Market segmentation

Astronomical data analysis

Andrew Ng

Источник: coursera.org, Andrew Ng, Machine Learning

# Обучение без надзирател - примери

- Задача за клъстеризация
  - Дадени са 1000 есета за икономиката на САЩ. Да се разпределят статиите на групи (клъстери) по определени критерии – напр. честота на срещане на дадени думи, дължина на изречение, брой страници и др.
- Задача за не-клъстеризация
  - “Cocktail Party Algorithm” – открива структури в множество от несвързани данни (напр. разпознаване на отделни гласове от звуците в тълпа от хора)

# Контролен въпрос

- Кои от следните задачи се решават чрез обучение **без надзорател**? (Отбележете всички подходящи)
  - а) При дадени писма, означени като спам или не спам, да се разработи филтър за спам
  - б) В даден списък от новини в web да се групират статиите по еднакви теми
  - в) В база от данни за потребители да се открият автоматично пазарните сегменти и да се групират потребителите по сегменти
  - г) По данни за пациенти, диагностицирани с диабет или не, да се разработи класификатор за нови пациенти, който определя дали са болни от диабет



# Методи и алгоритми

- Регресия (регресионен анализ)
- Алгоритъм на най-близките съседни (k-Nearest Neighbors)
- Алгоритъм k-Means
- Машини с поддържащи вектори - Support Vector Machines (SVM)
- Невронни мрежи
- Класификация чрез Бейс (Bayes)
- Принципен компонентен анализ (Principal component analysis)

# Програмни средства и технологии за машинно обучение

- Python
  - Python-базирани библиотеки и frameworks за задачи от машинно обучение
    - NumPy
    - Scikit-learn
    - Theano
    - PyTorch/Torch
    - TensorFlow
    - Keras
- Microsoft Cognitive Toolkit
- R
- Matlab
- Готови библиотеки в популярните езици за програмиране, напр.:
  - C# - Accord.NET
  - Java - WEKA