

## 1. Въведение – Резюме

**Данни:** Записани факти.

**Информация:** Скрытые закономерности (patterns) в данните.

Фокусът е върху Machine Learning техники, които автоматично намират закономерности в данните.

### Представяне на закономерности / Patterns /

Намерените модели могат да бъдат представени по два начина:

- Structural descriptions (структурни описания)
- Black-box models (черна кутия – напр. невронни мрежи)

Machine Learning (Машинно обучение):

- Придобиване на знания чрез **учене, опит или преподаване**.
- Осъзнаване чрез **информация или наблюдение**.
- Запаметяване.
- Получаване на информация, проверка на факти, усвояване на инструкции.
- Трудно е да се измери.
- Тривиално за компютрите (те могат да запомнят и обработват огромни количества данни).

**Нещата „учат“, когато променят поведението си по начин, който ги прави по-ефективни в бъдеще. Примери: Does a slipper learn? (Научава ли се пантоф?) – Очевидно не.**

**Does learning imply intention? (Имплицира ли ученето намерение?) – Не е задължително, машините учат без намерение.**

## Data Mining

**-Цел: Намиране на закономерности в данните, които: Доставят полезна информация; Позволяват бързо и точно вземане на решения**

Основни проблеми при Data Mining:

- Повечето закономерности не са интересни.
- Закономерностите могат да бъдат неточни или фалшиви.
- Данните могат да бъдат замърсени или липсващи.

**Machine Learning** техники идентифицират модели в данните и предоставят инструменти за **Data Mining**. Основен интерес: Техники, които осигуряват **структурни описания** на данните.

Класификация срещу Асоциационни правила:

**Класификационно правило** -Предсказва стойността на даден атрибут (клас на примера). *Пример:* Ако пациент има симптоми X, Y, Z → Класифицираме го като „болен“ или „здрав“.

**Асоциационно правило**- Предсказва стойността на произволен атрибут (или комбинация от атрибути). *Пример:* Ако клиент купи „хляб“ и „масло“, вероятно ще купи и „сирене“.

**Machine Learning** помага на **Data Mining**, като открива закономерности в данните.

**Класификационните правила** прогнозират конкретни категории.

**Асоциационните правила** намират връзки между различни характеристики.

**Линейна регресия** - Това е математически модел, който предсказва стойности, използвайки права линия. Формула:  $y = a \cdot x + b$

$y$  – предсказаната стойност (напр. заплата);  $x$  – входната стойност (напр. часове работа);  $a$  – наклонът на правата (показва колко бързо се променя  $y$ );  $b$  – началната стойност (ако  $x = 0$ , каква е  $y$ )

**Дървета на решенията** - Това е модел, който взема решения, като разделя данните на стъпки (въпроси „да“ или „не“). Всяка стъпка води до ново разклонение, докато стигнем до крайния отговор.

Ключова разлика: Линеината регресия работи най-добре, когато има ясна зависимост между променливите. Дърветата на решенията са полезни, когато трябва да вземаме решения на база различни фактори.

### **Входът на Machine Learning:**

-Добавена експертна информация → по-сложни, но по-добри правила.

### **Научените правила превъзхождат ръчно създадените!**

Машинно обучение и статистика-Историческа разлика:

Statistics: Тестване на хипотези. Machine Learning: Намиране на правилната хипотеза. **Днес тези две дисциплини се припокриват.**

**Методи:** Decision Trees (C4.5 и CART); Nearest-neighbor methods.

-Повечето **Machine Learning** алгоритми използват **статистически техники**.

Generalization as search (Генерализацията като търсене)

-**Inductive learning (Индуктивно обучение):** Намиране на описание на концепция, което съответства на данните.

**Пример:** Правилата като описателен език; Огромно, но крайно пространство за търсене.

- 1. Решение: Изброяване на възможните концепции;  
Елиминирание на несъответстващите описания;  
Оставащите описания съдържат целевата концепция.**

Enumerating the concept space (Изброяване на концептуалното пространство)

**Пример: Проблем с времето (weather problem);  $4 \times 4 \times 3 \times 3 \times 2 = 288$  възможни комбинации; С 14 правила  $\rightarrow 2.7 \times 10^{34}$  възможни правила**

**Други практически проблеми:**

- Може да оцелеят повече от едно описание.
- Възможно е нито едно описание да не е правилно.
- Езикът може да не позволява описанието на целевата концепция.
- Данните може да съдържат шум.

**Друг поглед върху генерализацията като търсене:**

- **Hill-climbing в описателното пространство** - Представи си, че се катериш по планина и винаги избираш пътя нагоре. В машинното обучение това означава, че постепенно подобряваме даден модел, докато стигнем най-доброто възможно решение според определен критерий. Проблем: Може да попаднем на локален максимум (не най-доброто решение, а само "височина", от която няма по-добър път нагоре).

-**Хевристични алгоритми** - Това са алгоритми, които не търсят перфектното решение, а просто добро решение. Те използват правила и догадки, за да намерят резултат по-бързо. Недостатък: Не гарантират, че ще намерят най-доброто възможно решение. Пример: Ако търсиш най-бързия маршрут до вкъщи, но нямаш карта: Хевристика: Винаги избираш най-широката улица или тази с най-малко коли. Може да не е оптималният път, но ще стигнеш достатъчно бързо.

**Bias (Пристрастия в обучението)- предпочитания или изкривявания,** които влияят на начина, по който машината взема решения или научава нещо. Например, ако обучаваме система да разпознава животни, но използваме само снимки на кучета и котки, моделът може да не научи как изглеждат други животни.

**Основни решения в обучаващите системи: Concept description language (Език на концепцията); Order of search (Ред на търсене); Overfitting-avoidance bias (Пристрастие за избягване на презапасване).**

**Language bias (Пристрастие в езика) -** Универсален ли е езикът или ограничава какво може да се научи? **Универсален език** може да изразява произволни подмножества от примери. Ако езикът включва логическо "или" (disjunction), той е универсален.

**Пример: Rule sets (Набори от правила).** Домейн знанията изключват някои концепции предварително.

Search bias (Пристрастие в търсенето)

**Хевристики за търсене:**

**-Greedy search:** Извършва най-доброто възможно действие на момента.

**-Beam search:** Поддържа няколко алтернативи.

**Посока на търсенето:**

**1 General-to-specific (От общо към специфично) -** Уточняване на правило чрез добавяне на условия.

**2 Specific-to-general (От специфично към общо) -** Обобщаване на конкретен пример в правило

Overfitting-avoidance bias (Избягване на презапасване)

**Модифициран критерий за оценка** (балансиране между сложност и грешки).

**Модифицирана стратегия за търсене: Pruning (Окастрияне)** → Оптимизиране на описанията; **Pre-pruning: Спира търсенето рано**, преди описанието да стане твърде сложно; **Post-pruning: Генерира сложни описания, след което ги опростява.**

Data Mining and Ethics I (Етика в Data Mining – Част I)

**Етични въпроси в практическите приложения-**

**Анонимизирането на данни е трудно. 85% от американците могат да бъдат идентифицирани само по пощенски код, дата на раждане и пол.**

**-Data mining често се използва за дискриминация Пример: При разглеждане на кредитни заявки неетично е да се използват данни като пол, религия, раса.**

**-Етичният проблем зависи от приложението. В медицината същата информация може да е приемлива.**

**-Някои атрибути съдържат скрита проблемна информация. Пример: Кодът на областта може да корелира с расата.**

**Важни въпроси: Кой има достъп до данните? С каква цел са събрани данните? Какви изводи могат да се направят? Задължително е да има предупреждения към резултатите.**

**Machine Learning използва регресии, дървета на решенията и хевристично търсене. Пристрастията в обучението определят ефективността на алгоритмите**