Глава 2: Входни данни, концепции, екземпляри и атрибути

Компоненти на входните данни за обучение

Концепция -***модел*** или ***правило***, което алгоритъмът се опитва да научи от данните. Това може да бъде:

- **Classification (Класификация)** – предсказване на дискретна категория (имейлът е спам или не).
- **Association (Асоцииране)** – намиране на връзки между различни елементи (анализ на потребителско поведение – „Клиентите, които купуват хляб, често купуват и масло“).
- **Clustering (Клъстериране)** – групиране на сходни елементи без предварително зададени категории.
- **Numeric Prediction (Числова прогноза)** – предсказване на числова стойност (прогнозиране на цената на жилище).

Примерът(example) - *индивидуална наблюдавана стойност* (наблюдение) в набора от данни:

- **Relations (Релации)** – свързани обекти в база от данни.
- **Flat Files (**Плоски файлове**)**– таблици с редове и колони (CSV).
- **Recursion (Рекурсия)** –примерите са свързани помежду си в йерархични или рекурсивни структури.

Атрибутът измерва различни характеристики на един пример. Разделяме ги в няколко вида:

- **Nominal (Номинален)** – категории без подредба (цвят на кола: червен, син, зелен).

- **Ordinal (Ординален)** – категории с подредба, но без равномерна разлика между тях (оценка: нисък, среден, висок).
- **Interval (Интервален)** – числови стойности със смислена разлика между тях, но без истинска нулева стойност (темпер.).
- **Ratio (Отношение)** – числови стойности с истинска нулева точка (тегло, височина).

Подготовка на входните данни- критичен етап в процеса на машинно обучение. Ако данните не са правилно обработени, моделът може да даде неточни резултати. Включва:

- **Sparse Data (Редки данни)** – случаи, при които повечето стойности са нули или липсващи.
- **Attributes (Атрибути)** – Изборът на правилните характеристики за модела е от решаващо значение. Често се налага трансформиране на данните, за да станат по-информативни.
- **Missing and Inaccurate Values (Липсващи и неточни стойности)** – обработка на пропуснати данни и шум в данните. Те могат да бъдат коригирани чрез: запълване със средни стойности, премахване на грешни записи или използване на модели за предсказване на липсващите данни.
- **Unbalanced Data (Небалансирани)** – някои класове са представени с много повече примери от други (95% "не-спам", а само 5% "спам"). Моделът може да игнорира редките случаи. За справяне с това се използват техники като балансиране на класовете, претегляне на примерите или генериране на нови данни чрез методи като SMOTE.

- **Запознаване с данните**– анализиране и визуализиране на входните данни преди обучение.

За да работи машинното обучение, то използва входни данни, които се състоят от няколко основни елемента – **Компоненти на входните данни**:

- **Concepts (Концепции):** знания, които можем да извлечем от данните. **Цел:** Да разберем и опишем ясно какво означава дадена концепция. Ако анализираме данни за времето, концепция може да бъде „дали ще вали или не“.
- **Instances (Екземпляри):** отделните примери в набора от данни. Всеки екземпляр представлява един запис или случай. **Важно**: Понякога има връзки между примерите. При анализ на пациентски данни, здравословното състояние на един човек може да бъде свързано с неговото минало лечение.
- **Attributes (Атрибути):** Характеристики, които измерват различни аспекти на даден екземпляр. **Видове атрибути**:

- Категориални (номинални) – „цвят кола“ (червен, син, зелен).
- Числови – „височина в сантиметри“, „температура в градуси“.

Concept (Концепция) -знанието, което трябва да бъде научено от алгоритъма. **Concept description (Описание на концепция)**: резултатът (изходът) от процеса на обучение.

Стилове на обучение

Classification Learning (Класификационно обучение) -
Предсказване на дискретен клас- **конкретни категории (етикети), към които се отнасят обектите**

- данни за времето, контактни лещи, ириси

- **supervised (контролирано) обучение**
- Алгоритъмът получава **реални изходни стойности** (етикети).
- Изходната стойност се нарича **class (клас) на примера**.
- Успехът се измерва върху **test data**, където класовете са известни.
- На практика успехът често се измерва **субективно**.

Асоциативно обучение -Прилага се, когато **няма предварително зададен клас**, а всяка структура се счита за „интересна“. Разлика спрямо класификационното обучение:

-предсказва **стойността на всеки атрибут, а не само класа**.

-предсказва **повече от един атрибут едновременно**.

- много повече асоциативни правила, отколкото класификационни.

- необходими ограничения, като минимално покритие и минимална точност.

- търговията - асоциативно обучение, за да се открият зависимости между продукти, които често се купуват заедно. Ако клиент закупи хляб, има голяма вероятност да закупи и мляко. Асоциациите се представят като правила:"Ако клиент купи хляб, ще купи и мляко".

-Разликата спрямо класификационното обучение: няма предварително зададен клас, а се откриват асоциации между атрибути и може да се предсказва повече от един атрибут едноврем.

Clustering (Клъстериране)-Откриване на групи от подобни обекти. Клъстерирането е **unsupervised (неконтролирано)**. **Класът на даден пример не е известен предварително.**Успехът често се измерва **субективно**.

- В маркетинга може да се използва клъстериране, за да се открият групи от потребители, които имат подобни навици за

пазаруване. Клъстерирането ще групира потребителите в различни сегменти, като "млади хора, които често купуват спортни стоки" или "пенсионери, които предпочитат книги".
- Клъстерирането е неконтролирано обучение (unsupervised)- не знаем предварително какви групи ще открием. Успехът на алгоритъма зависи от това как ще се интерпретират тези групи.

Numeric Prediction (Числова прогноза)- Вариант на класификационното обучение, при който **“class” (клас) е числова стойност**. Нарича се още **regression (регресия)**. Обучението е **supervised (контролирано)**. Алгоритъмът получава **целеви стойности (target value)**.Успехът се измерва върху **тестови данни**.

- В икономиката може да се използва числова прогноза за предсказване на стойността на акциите на базата на исторически данни. Вместо да се класифицират акциите в категории ("расте" или "пада"), числовата прогноза дава конкретна стойност /прогноза за стойността на акциите/.
- вид регресия (regression), където целевата стойност е числова, а успехът на модела се измерва чрез точността на предсказаните стойности върху тестови данни.

Instance (Екземпляр): Конкретен пример за концепцията, която трябва да бъде класифицирана, асоциирана или клъстерирана. **Независим отделен обект**, който има **предварително дефиниран набор от атрибути**. Входът за алгоритъма за обучение е **набор от екземпляри (dataset, набор от данни)**. Представя се като **единична релация/плосък файл (flat file)**. **Ограничения:** Липса на **връзки между обектите**; **Най-често срещаната форма в практическото извличане на данни.**

Generating a Flat File (Генериране на плосък файл) -Процесът на **“flattening” (изравняване)** се нарича **denormalization (денормализация)**. **Обединяват се няколко релации в една.** Може да се направи с всяко крайно множество от крайни релации.

- **Проблем:** Връзки без предварително зададен брой обекти.
- **Денормализацията може да доведе до фалшиви зависимости**, отразяващи структурата на базата данни.

Recursion (Рекурсия) - функция или процес извиква сам себе си, докато не достигне някакво крайно условие. **Неограничените релации изискват рекурсия**, защото имаме данни, които са свързани в **непредвиден брой нива**.

- Подходящите техники са известни като **Inductive Logic Programming (ILP, Индуктивно логическо програмиране)**.
- Пример за ILP метод: **Quinlan's FOIL rule learner (FOIL – алгоритъм за индукция на правила)**.
- **Основни проблеми:**

(a)Шум в данните (noise) - Неточни или противоречиви данни могат да доведат до грешни изводи.

(b) Изчислителна сложност (computational complexity) - При големи данни рекурсията може да стане бавна и ресурсоемка.

Multi-Instance Concepts (Концепции с множество екземпляри) - Всеки индивидуален пример съдържа **„чувал" (bag)** от екземпляри. Всички екземпляри се описват чрез **същите атрибути**. Един или повече екземпляри могат да определят класификацията на примера.

- Цел – създаване на описание на концепцията.
- **Важни реални приложения:** Прогнозиране на действието на лекарства; Класификация на изображения: Лекарството се разглежда като „чувал" от различни геометрични конфигурации на молекулите; Изображението се представя като „чувал" от компоненти на изображението.

Attribute (атрибут)- Всеки екземпляр се описва чрез предварително зададен **набор от характеристики (features), наречени атрибути. Но:** Броят на атрибутите може да варира в

практиката. **Възможно решение:** Използване на **флаг „irrelevant value" (нерелевантна стойност)**. **Свързан проблем:** Съществуването на даден атрибут може да зависи от стойността на друг атрибут.

Възможни типове атрибути (Levels of Measurement – Нива на измерване):

- **Nominal (Номинален):** Категории без подредба (цвят на кола: червен, син, зелен).
- **Ordinal (Ординален):** Категории с подредба, но без равномерна разлика между тях (оценка: нисък, среден, висок).
- **Interval (Интервален):** Числови стойности със смислена разлика, но без абсолютна нулева точка (температура).
- **Ratio (Отношение):** Числови стойности с абсолютна нулева точка (тегло, височина, време).

Нива на измерване (Levels of Measurement):

Nominal Levels of Measurement (Номинални нива на измерване): Стойностите са просто символи – служат **само като етикети или имена**. **Nominal** идва от латинската дума за **„име"**.

- Атрибут „outlook" (прогноза) в данните за времето: стойности: **„sunny", „overcast" (облачно), „rainy"**.
- **Няма връзка или подредба между стойностите** (няма ред, няма мерки за разстояние).
- Единствената възможна операция е **проверка за равенство**.

Ordinal Levels of Measurement (Ординални нива на измерване): Добавя ред към стойностите, но **няма дефинирано разстояние между тях**. Атрибут „temperature" при данни за време:стойности: **„hot" > „mild" (умерено) > „cool" (хладно)**. **Събиране и изваждане на стойности не е смислено.** Ако **temperature < hot → play = yes (играем = да)**.

- **Разликата между номинални и ординални стойности не винаги е ясна** (например атрибутът „outlook“).

Interval Quantities (Интервални величини): Редът е дефиниран, но **стойностите са измерени в равномерни интервали**. Темпер.

- **Разликата между две стойности има смисъл**. **Но:** Нулата **не е дефинирана като абсолютна стойност**. **Сумиране и умножение не са смислени операции**.

Ratio Quantities (Отношения величини):Дефинирана е абсолютна нулева точка. distance между един обект и себе си е **0**.

- **Отношения величини се третират като реални числа**. **Всички математически операции са позволени**.
- **Но:** Винаги ли има „естествена“ нулева точка? **Отговорът зависи от научните знания. Фаренхайт първоначално не е знаел за най-ниската възможна температура**

Типове атрибути, използвани на практика: Много алгоритми за Data Mining работят само с два типа измервания: номинални и ординални. Други алгоритми работят само с ratio (отношения).

- **Номиналните атрибути - “categorical” (категориални), “enumerated” (изброими) или “discrete” (дискретни)**. Но: **“enumerated” и “discrete” предполагат ред**, което не винаги е вярно. **Специален случай:** Dichotomy (дихотомия) → **Boolean (булев) атрибут** („да“/„не“).
- **Ординалните атрибути** понякога се кодират като **“numeric” (числови) или “continuous” (непрекъснати)**. Но: **“continuous” предполага математическа непрекъснатост**, което не винаги е вярно.

Metadata (Метаданни)- информация за данните- данни за данните. Може да се използват за **ограничаване на пространството за търсене** на алгоритъма.

- **Примери: Дименсионални ограничения** (математически изрази трябва да са дименсионално правилни); **Циклични подредби** (градуси в компаса); **Частични подредби** (релации между обобщения и специализации).

Подготовка на входните данни

- **Денормализацията не е единственият проблем** при подготовката на данни за обучение.
- **Проблем:** Различни източници на данни (отдел продажби, отдел фактуриране на клиенти и др.).
- **Разлики:** Различни стилове на съхранение на записи; Кодиране на данни; Различни времеви периоди; Агеграция на данни; Различни първични ключове; Различни видове грешки.
- **Данните трябва да бъдат събрани, интегрирани и почистени.**
- **“Data warehouse” (Хранилище на данни):** Единна и последователна точка за достъп до данните.
- **Понякога са необходими външни данни („overlay data“).**
- **Критично:** Видът и нивото на агрегиране на данните

Допълнителни типове атрибути: Форматът ARFF поддържа:
1/**“string” (низови) атрибути**: Подобни на **номинални**, но **списъкът с възможни стойности не е предварително зададен**;
2/**“date” атрибути** - формат **ISO-8601: yyyy-MM-dd-THH:mm:ss**.

Relational Attributes (Релационни атрибути) - представяне на **multi-instance (многоекземплярни) проблеми** във формат ARFF. Всяка стойност на релационен атрибут представлява **отделен „чувал“ от екземпляри**. Всеки „чувал“ съдържа **едни и същи атрибути**.

Sparse Data (Рехави/Разредени данни) - повечето стойности на атрибутите са нули.

- **Решение:** Използване на **компактен начин за съхранение**.

- **Пример:** Броене на думи при категоризация на текстове.
- Някои алгоритми работят **много по-ефективно** със sparse data.

Missing Values (Липсващи стойности) -Често се маркират с **извън обхвата стойности** за даден атрибут. **Видове липсващи стойности:“Unknown” (неизвестна); “Unrecorded” (незаписана); “Irrelevant” (нерелевантна)**

- **Причини:** Проблеми с оборудването; Промени в експерименталния дизайн; Комбиниране на различни набори от данни; Невъзможност
- **Липсваща** за измерване **стойност може да носи собствено значение!**/Липсващ тест в медицин. преглед може е значещ/.

Unbalanced Data (Небалансирани данни) -единият клас е много по-често срещан от останалите.

- **Пример:** Откриване на рядко заболяване.
- **Решение:** Използване на техники, които **отчитат неравните разходи за грешна класификация**.