G検定対策講座　第 5 回

機械学習

データから自分で法則を見つける

今日学ぶこと

1. **機械学習とは**
   データから自動でパターンを学ぶ
2. **ルールベースとの違い**
   昔ながらのやり方と比べる
3. **特徴量エンジニアリング**
   機械学習の腕の見せどころ
4. **ビッグデータと次元の呪い**
   注目された背景と、高次元に潜む落とし穴
5. **代表的な応用例**
   迷惑メール・レコメンド・統計的自然言語処理

そもそも
AIが「学ぶ」とは？

人間がルールを書くのではなく、データを見せれば自分で法則を見つけていく

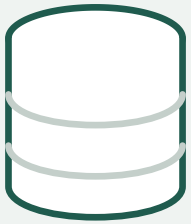
第三次AIブーム

機械学習とは

大量のデータから**自動でパターンを学習する**技術

人間が用意するのは
学習のしくみだけ
ルールはAIが見つける

具体例 学ぶ流れ



大量のデータ

学習

法則（ルール）を見つける

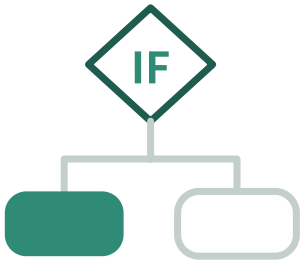
データの中から

新しいデータを
予測・分類する

POINT **人間が知識を書かなくてよい**という発想で、第二次AIブームの壁を突破した

ルールベース と 機械学習

同じ「AI」でも、ルールを誰が用意するかが違う



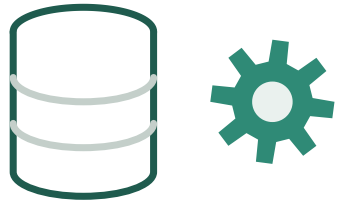
ルールベース

人間がルールを書く

◯ 中身が明快で説明しやすい

× 知識を集めるのが大変・複雑な問題に弱い

≠



機械学習

データから自動で学ぶ

◯ 大量のデータに強く、複雑なパターンも扱える

× 中身が見えにくい・データの品質しだい

POINT 試験では「**機械学習はデータが命／ルールベースは説明のしやすさで上**」と覚える

腕の見せどころ

特徴量エンジニアリング

データの中で**どこに注目するか**を、
人間が設計すること

学習そのものはAIに任せるが
注目点は人間が決める

具体例 **「猫」を見分けるなら**

- 耳の形
- 目の色
- ヒゲの本数

→

3 機械学習

特徴量は

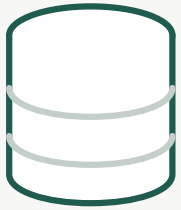
人間が設計する

4 深層学習

特徴量も

AIが自分で見つける

POINT **特徴量を誰が決めるか**——ここがレベル3とレベル4の分かれ目



機械学習の燃料

ビッグデータ

ウェブ・センサー・SNSから集まる**膨大なデータ**

ダグ・レイニーが整理した
3つのVという考え方

具体例 ビッグデータの3つのV

Volume
量が多い

Variety
種類が多様

Velocity
発生が速い

→

膨大なデータ

それを支える計算力

アルゴリズムの進歩

POINT この**3つがそろったこと**が、第三次AIブームの大きな転換点になった

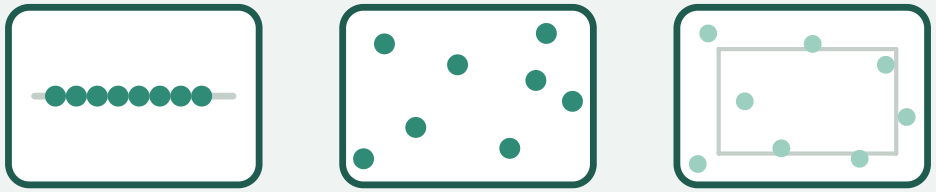
ベルマン

次元の呪い

特徴量を増やすほど、
必要なデータが爆発的に増える

数学者ベルマンが
動的計画法の研究で名づけた

具体例 同じ8個の点でも、次元が上がるとスカスカになる



→

点どうしの距離が
意味をなさなくなる

対策：次元削減・特徴量選択
正則化

POINT 「**次元を増やすとデータがまばらになる**」と「名づけたのはベルマン」をセットで

応用その1

迷惑メールのフィルタ

過去の迷惑メールと正常なメールから
見分け方を学ぶ

古くから使われてきたのが
ナイーブベイズという手法

具体例 単語の出やすさで判定する

過去のメール

→

特徴 「無料」が出やすい

特徴 「当選」が出やすい

特徴 送信元が初めて …

→

迷惑メールらしさを判定

POINT 人間がルールを書ききるのは無理でも、**データからなら学べる**。身近な成功例

応用その2

レコメンデーション

利用者の**行動の履歴**から好みを学ぶ

Amazonのおすすめ商品
Netflixの動画・Spotifyの音楽

具体例 主な2つのやり方

A 協調フィルタリング

似た好みの人から推薦する

「この商品を買った人はこちらも」

B コンテンツベース

商品そのものの特徴から推薦する

「同じジャンル・同じ監督の作品」

POINT データが多いほど精度が上がる（規模の経済がはたらく）。**巨大プラットフォームの強さの源**

応用その3

統計的自然言語処理

ルールではなく**統計の力**で言葉を扱う方法

いまはニューラルネット方式へ進化している

具体例 大量の文章から学ぶこと

大量の文章 → 単語の出やすさ / 言葉どうしの関わり → 文書分類 / 機械翻訳 / 情報抽出 / 感情分析

POINT 第2回に出た**統計的機械翻訳**もこの仲間。言葉を数学的に扱う対象へ変えた転換点

機械学習

機械学習
データから自動でルールを学ぶ

特徴量
注目点は人間が決める

ビッグデータ
3つのV が土台

次元の呪い


増やすとスカスカに

代表的な応用
迷惑メール・推薦・言語処理