

(YouTube ピクアカ)

【はじめてのデータサイエンス PART3】

機械学習の特徴と活用例を知ろう

教師あり学習の代表的な用途

1. スパムメール判定
2. 購買履歴の分析
3. 画像認識
4. レコメンド機能

汎用的なモデルは存在しない

NFLT ノーフリーランチ定理

no-free-lunch theorem

あらゆる問題に対応できる

汎用的モデルは存在しない

決定木モデル

構造化データを扱う
(変数の意味が明確)

ビジネスで広く使われる

学習結果の解釈が容易、施策に繋げやすい

購買履歴データの分析

決定木系のモデルを活用

- ① 基本の決定木
- ② 決定木のアンサンブル学習器
 - ランダムフォレスト
 - 勾配ブースティング回帰木

分析コンペティションでは

勾配ブースティング回帰木 が人気

決定木系モデルには

多種多様な特徴量を扱う **柔軟性**がある

決定木の概要

特徴量ごとにデータをソート

条件分岐でグループに分割

閾値：分割後なるべく同じ属性で構成されるように
分割を**再帰的**に行う

決定木のメリット

学習過程が把握しやすい **決定木の可視化**

学習結果の説明が容易 **特徴量の重要度を可視化**



ビジネスに向く

決定木の弱点

過学習 **overfitting**

学習データに合わせすぎて

未知データへの汎用性が失われている状態

モデルが複雑なほど過学習しやすい

アンサンブル学習器

単純なモデルを複数組み合わせる
過学習を抑制、精度も改善

ランダムフォレスト

決定木を複数構築

多数決で結果を求める



データと特徴量の両方をサンプリング

レコメンドエンジンには

k 近傍法 k Nearest Neighbor (KNN)

予測対象データ点と学習データ点の間の
距離に基づいて予測

スパムメール判定

ナイーブベイズを多用

身近なカテゴリ分類問題

特に**文章分類**に適する

単語の出現頻度に注目

単語間の関連性は無いという**単純仮定**

高速リアルタイム処理に強いが**精度は高くない**

ナイーブベイズ

単語の出現頻度に注目

出現頻度 = 特徴量

単語同士の**関連性**は深く考えない

特徴量が無相関と仮定

ナイーブベイズで扱うデータ

学習データ

ある程度の量の非スパム、スパムメール

学習済みモデル

特定の単語がスパムに含まれる確率
(確率データベース)

メールの分類

メール本文の**前処理**が必要

単語単位、n文字単位 で分割

自然言語処理の場合の特徴量

文章数 × 単語数 → **単語頻度行列**

次元数は 10,000以上

<ここまで>