Math Everywhere

数理科学する明治大学

数理科学による商品のレコメンデーション

櫻井 義尚(明治大学　総合数理学部)

数理科学による商品のレコメンデーション

- 「数理科学」 → 数理モデリングに基づいた方法
- 商品のレコメンデーション（推薦）
  - それぞれの人の行動や興味・関心に基づいてその人に適した情報を提示・推薦する

例：Amazon.co.jp

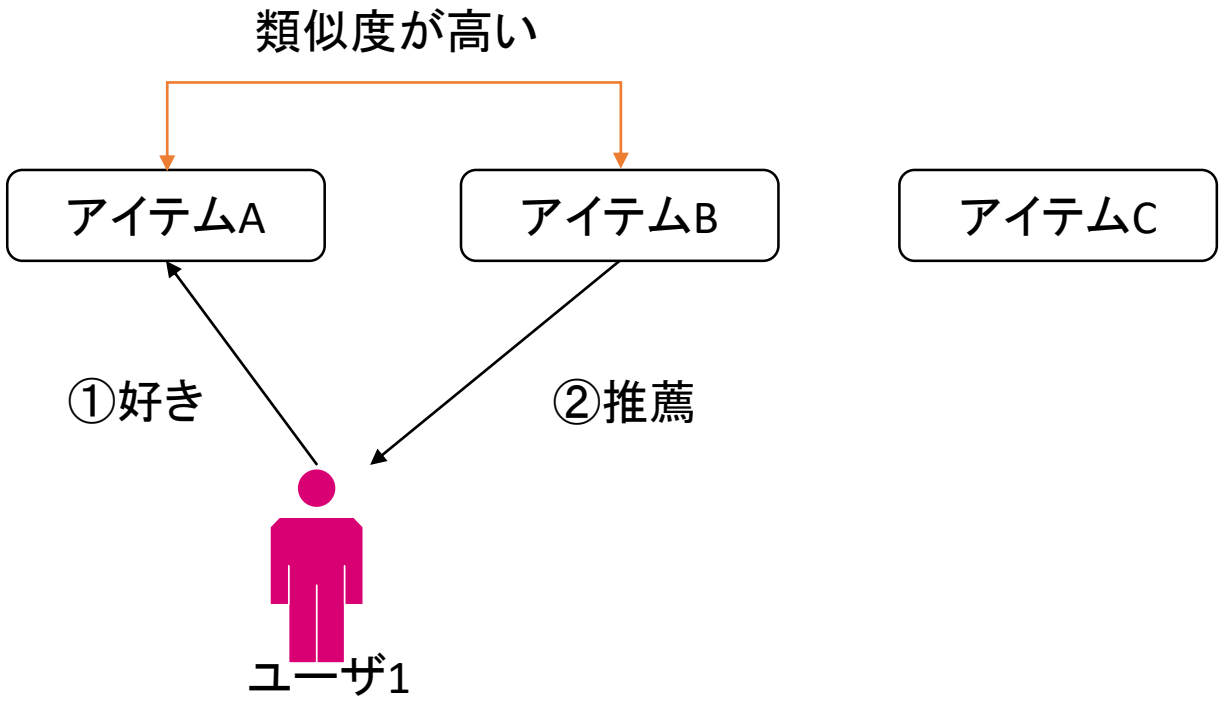
検索・閲覧

推薦

レコメンデーションの手法

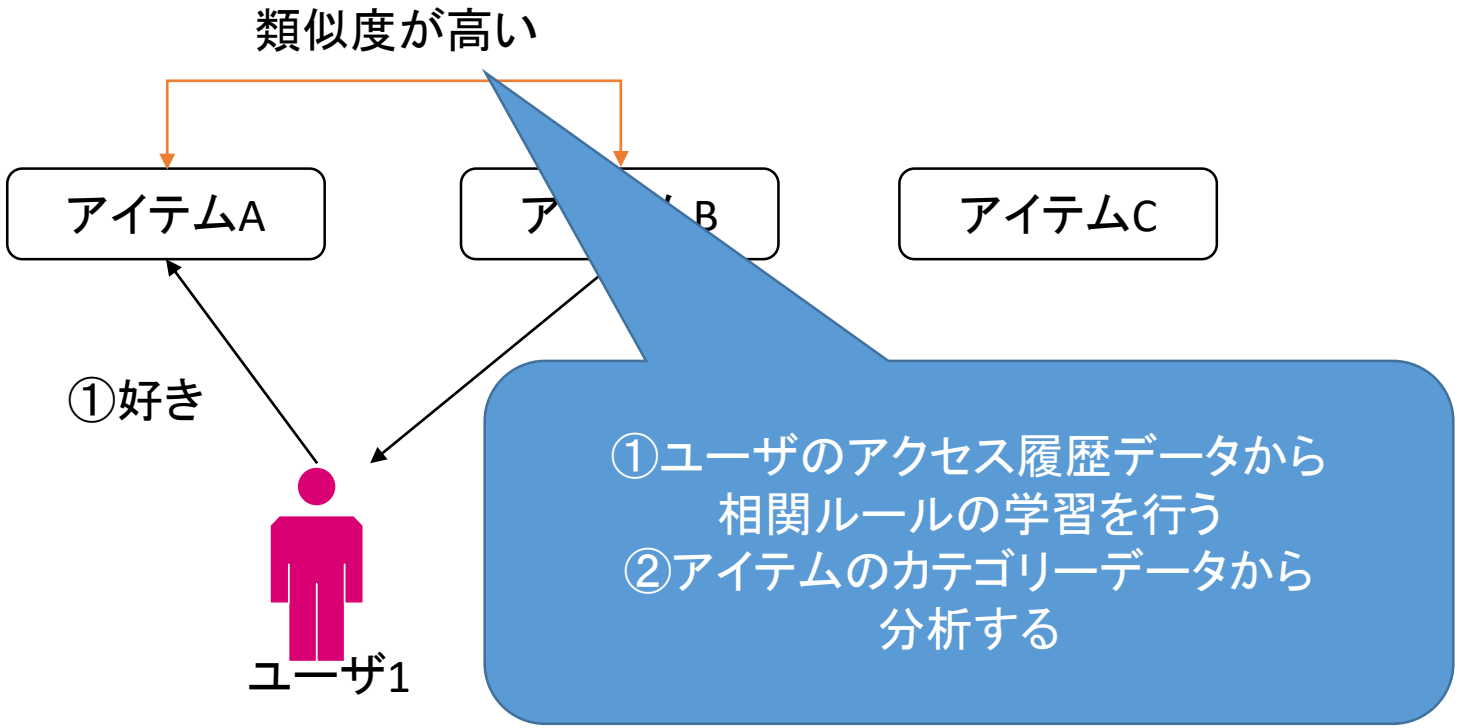
- 類似度マッチング
  - 似ている商品・ユーザの情報から推薦
  - 対象をベクトル化し、対象間の距離を求める
  - 様々な方法が存在
- 頻出パターンマイニング
  - 一緒に買われている組合せから推薦
  - 教師なし機械学習
- 推薦ルールのモデリング
  - 過去の事例に基づいて推薦
  - 教師あり機械学習

類似度マッチング（アイテムベース）



アイテム間の類似度に基づいて推薦

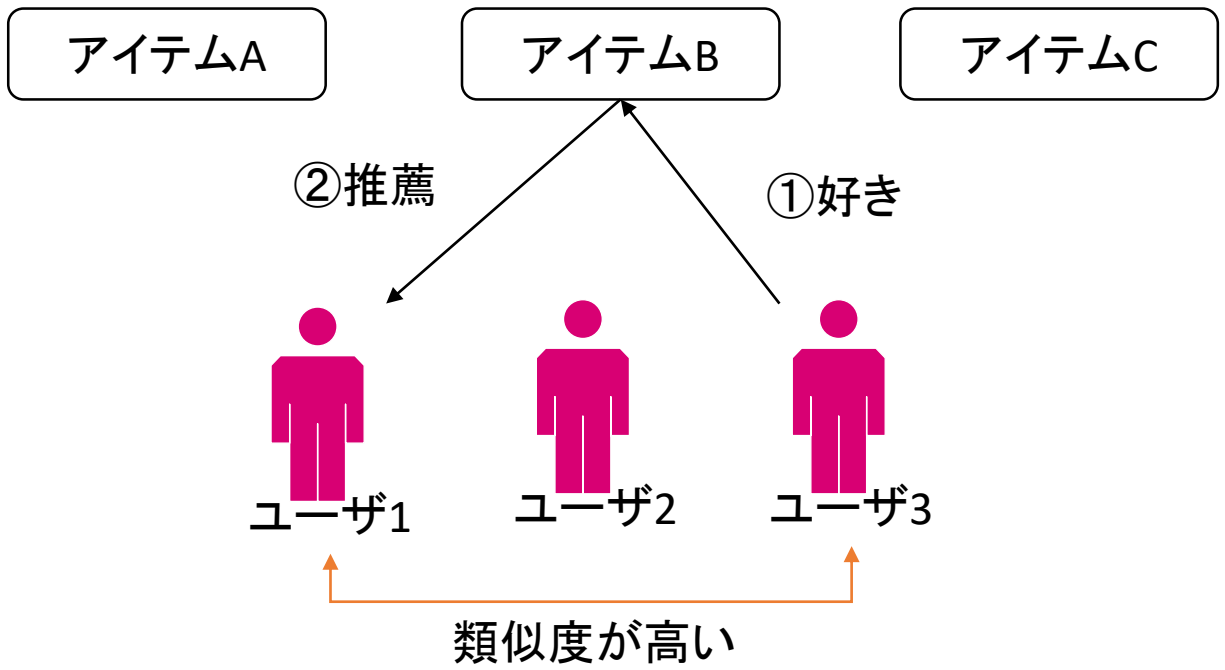
類似度マッチング（アイテムベース）



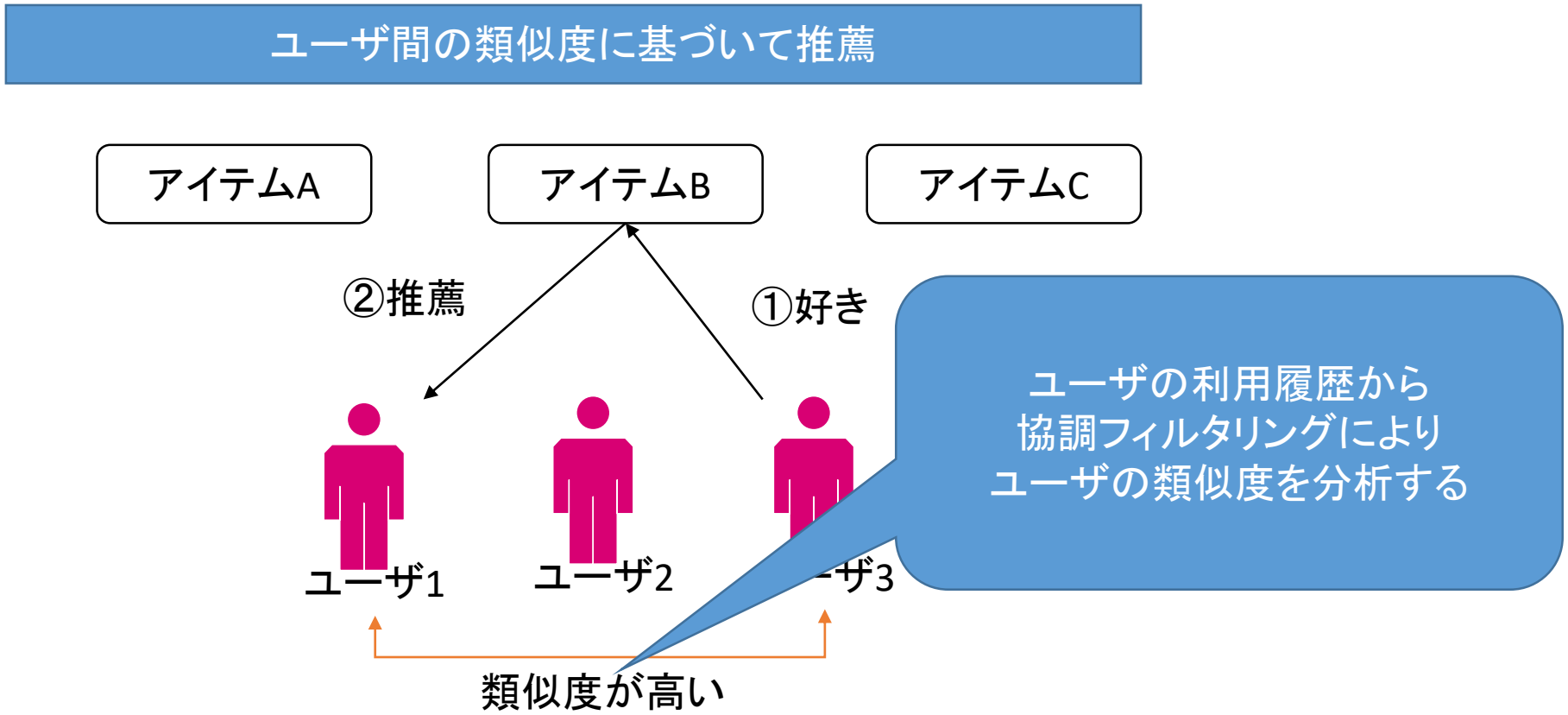
アイテム間の類似度に基づいて推薦

類似度マッチング（ユーザベース）

ユーザ間の類似度に基づいて推薦



類似度マッチング（ユーザベース）



どうやって「類似」を求める？

- アイテムの類似度
  - ユーザ行動履歴
    - 「この商品を買った人はこんな商品も買っています」
  - カテゴリー
    - どの分野、分類の商品か
  - 関連コンテンツ
    - 説明テキストや画像の相関。スペック内容
- ユーザの類似度
  - ユーザ行動履歴
    - 閲覧や購入、WEBサイト内の行動（選択、評価）
  - プロファイル
    - 年代、性別、住まい、興味のある分野などの登録・アンケート情報

協調フィルタリングとコンテンツベース

- 類似度を求めて推薦を自動化する2つの方法

協調フィルタリング

- アイテム利用者の行動履歴（他人の意見）を元にレコメンドする
- 例：Amazonの『この商品を買った人は、こんな商品も』機能

コンテンツベースフィルタリング

- アイテムの特徴（カテゴリーやキーワード）を元にレコメンドする
- アイテムの特徴ベクトルで類似度ソートしてレコメンド

協調フィルタリング

嗜好が似ている人が好きな（購入している）ものを推薦

	アイテム1	アイテム2	アイテム3	アイテム4	類似
ユーザX	購入	購入			
ユーザA	購入	購入	購入		似ている
ユーザB				購入	似ていない
ユーザC	購入		購入		中間

ユーザXに推薦したい→他の利用者と嗜好データを比較
ユーザXに似ている人のおすすめを尊重

協調フィルタリングの長所と短所

分類	協調フィルタリング	コンテンツベース
多様性	○	×
ドメイン知識	○不要	×必要
スタートアップ問題	×	△
利用者数	×多数必要	○少数可
被覆率	×	○
類似アイテム	×	○
少数派の利用者	×	○

神嶌敏弘著『推薦システムのアルゴリズム』より引用

ハイブリッドフィルタリング

- 実際には複数の手法を組み合わせて使う事が多い

ハイブリッドフィルタリング

- 2種類の方法を組み合わせる
- 協調フィルタリングとコンテンツベースフィルタリングは補完関係にあるので組み合わせて使う

知識ベースフィルタリング

- 具体的な嗜好の提示
- 推薦ルールを複数登録する

レコメンデーションの手法

- 類似度マッチング
  - 似ている商品・ユーザの情報から推薦
  - 対象をベクトル化し、対象間の距離を求める
  - 様々な方法が存在
- 頻出パターンマイニング
  - 一緒に買われている組合せから推薦
  - 教師なし機械学習
- 推薦ルールのモデリング
  - 過去の事例に基づいて推薦
  - 教師あり機械学習

靴の最適サイズ推薦

- ビジネス課題
  - 実店舗：フィッティング時間を短くしたい
    - フィッティング時間10分～1時間/人（平均15分程度）
    - 混雑時の機会損失
  - ECサイト：サイズ違いによる返品をなくしたい
    - 返品によるコスト増大は大きい
  - 靴によってその人の最適サイズは異なる

→靴の最適サイズを推定する事で解消

靴サイズ推定 ―概要―

入力

顧客ID	122404	靴タイプ	ラウンド
右足	(mm)	左足	(mm)
足囲	234.6	足囲	232.2
足長	224.2	足長	224.1
足高	48.1	足高	47.8
足幅	94.3	足幅	94.1
踵幅	61.2	踵幅	60.2
外踏まず長	145.6	外踏まず長	146.3
内踏まず長	165.1	内踏まず長	163.9

計測値＆靴タイプ

機械学習モデル
(プログラム)

出力

顧客ID	122404	靴タイプ	ビーナス
右足	(cm)	左足	(cm)
推奨サイズ	23.0	推奨サイズ	23.5

推薦靴サイズ

手法の候補

- K近傍法
- ロジスティック回帰
- サポートベクターマシン(SVM)
- ランダムフォレスト
- Gradient Boosting Decision Tree(GBDT)

評価指標

		予測値	
		Positive	Negative
真値	Positive	TP	FN
	Negative	FP	TN

正解率(Accuracy)

- 「購入する/しない」の予測が何%当たっていたか

再現率(Recall)

- 実際に「購入した」データのうち、「購入する」と予測されたデータの割合(%)

$$Recall = \frac{TP}{TP+FN}$$

適合率(Precision)

- 「購入する」と予測したデータのうち、実際に「購入した」ものの割合(%)

$$Precision = \frac{TP}{TP+FP}$$

F-measure(F1-socre)

- 適合率と再現率の調和平均。

$$F-measure = \frac{2Recall*Precision}{Recall+Precision}$$

多クラス分類の評価

- マクロ平均（macro average）

全クラスの結果の平均

- マイクロ平均(micro average)

混同行列全体でTP, FP, FNを集計し、それに基いて二値分類と同様に評価指標を算出

TP= TA+TB+TC

- 重み付き平均(weighted average)

各クラスのデータ数の偏りを考慮し、データ数の比率で重み付き平均

		予測値		
		A	B	C
真値	A	TA	FB(A)	FC(A)
	B	FA(B)	TB	FC(B)
	C	FA(C)	FB(C)	TC

$$P_A = \frac{TA}{TA+FA(B)+FA(C)}$$

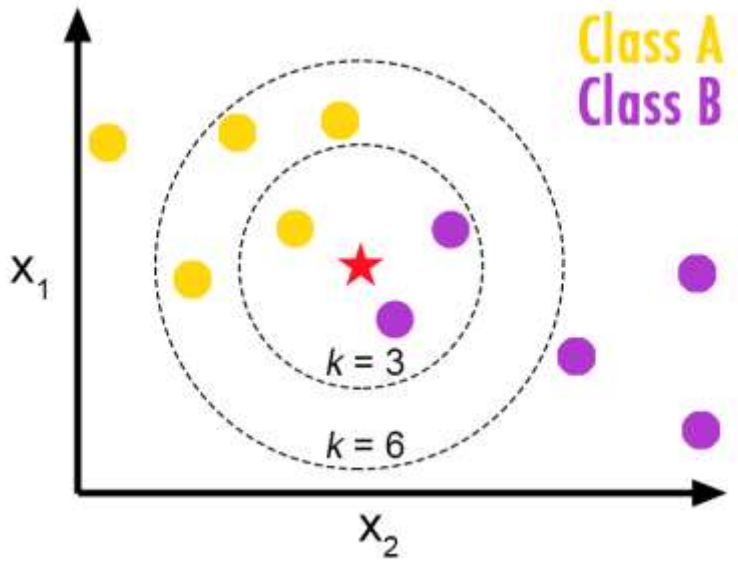
$$R_A = \frac{TA}{TA+FB(A)+FC(A)}$$

$$F_A = H(P_A, R_A)$$

P:Precision,R:Recall,F:F1−measure,H:harmonicmean

k近傍法

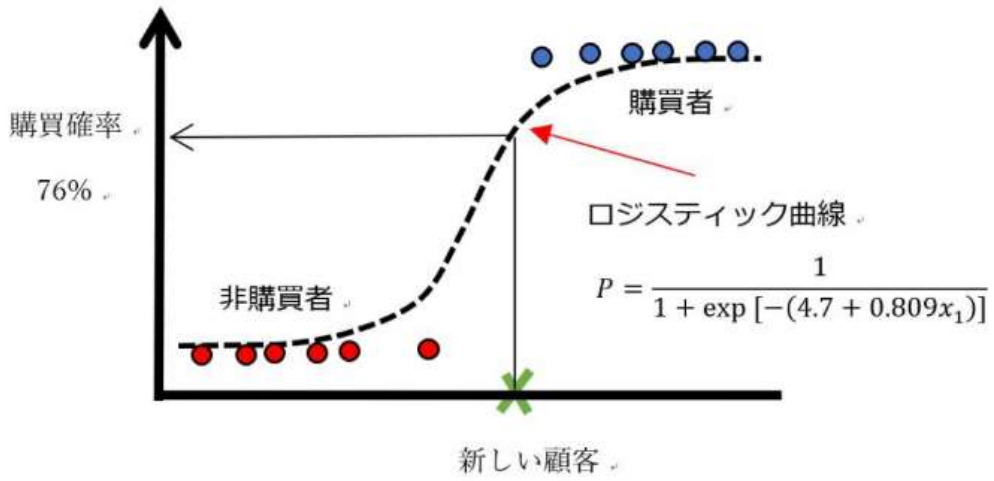
- **アルゴリズム**
  - あるデータ点と近いk(任意)個のデータ点の中で、最も多いクラスを出力とする。



- **考察**
  - 他手法と比較して精度が低め
  - データ数の少ないクラスの分類について多めに行われた

ロジスティック回帰

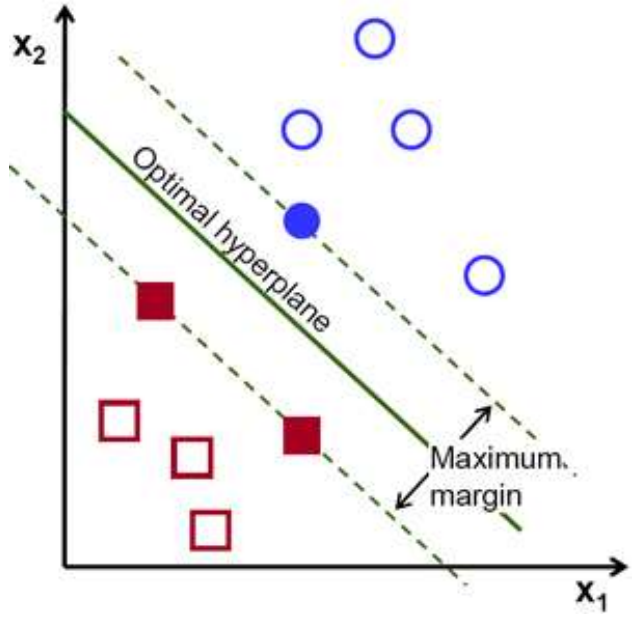
- **アルゴリズム**
  - 入力の各数値に「重み」を設定し、シグモイド関数を用いて出力を0~1の範囲に変換する(0:NO, 1:YES)



- **考察**
  - データ数の少ないクラスについて予測がされなかった
  - 精度についてはやや高め

SVM

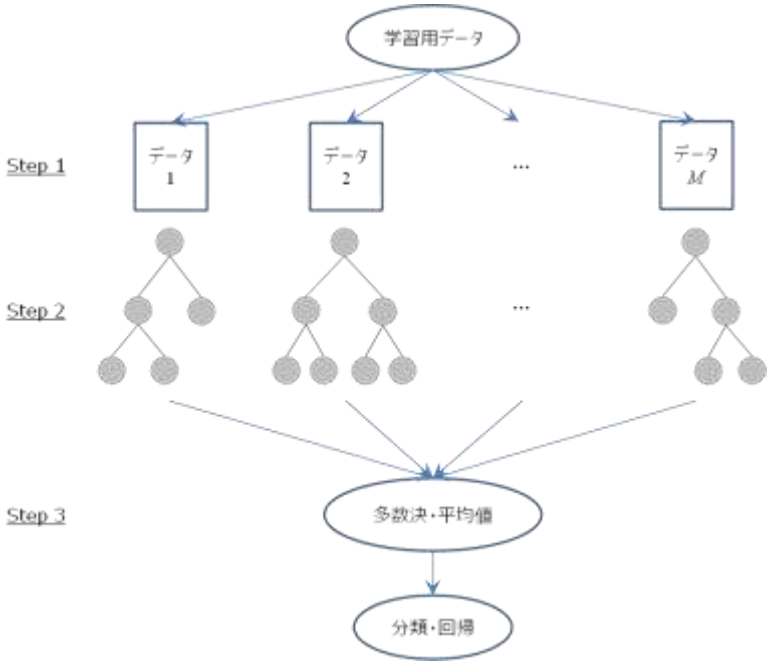
- **アルゴリズム**
  - クラスを分類する境界線を作成する。このとき、クラス間のマージンを最大化する。



- **考察**
  - 学習に非常に時間がかかった
  - しかし時間当たりの精度は高くない

ランダムフォレスト

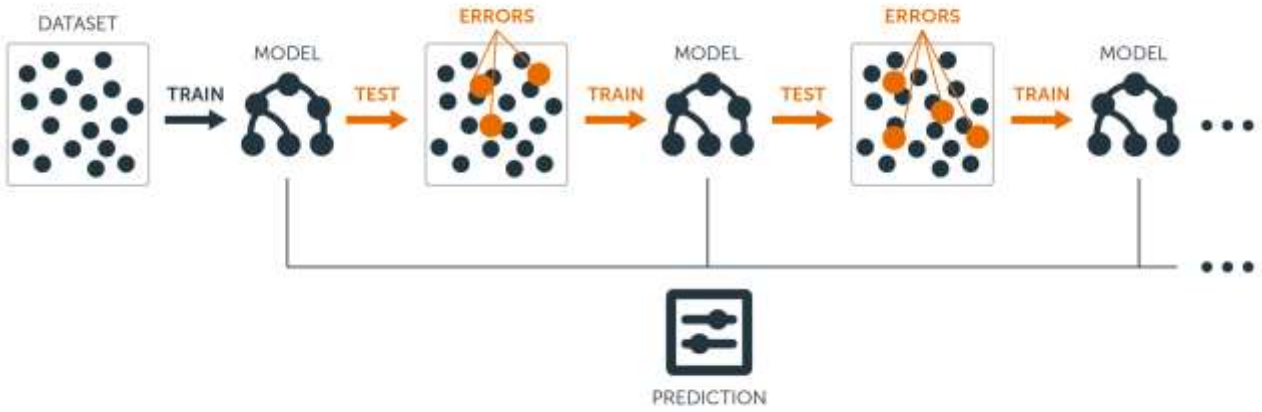
- **アルゴリズム**
  - 決定木をランダムに複数作成し、多数決を行う。



- **考察**
  - 他モデルより正解率と再現率がやや低い

GBDT(Gradient Boosting Decision Tree)

- **アルゴリズム**
  - 基本的にはランダムフォレストと同じだが、他の決定木の弱点を補う木を順に作成することでより分類の精度を高める。



- **考察**
  - あらゆる評価指標が他のモデルよりも高かった。
  - 学習時間も他モデルに比べてとくに大きくなかった。

購入が予測される靴タイプ分類

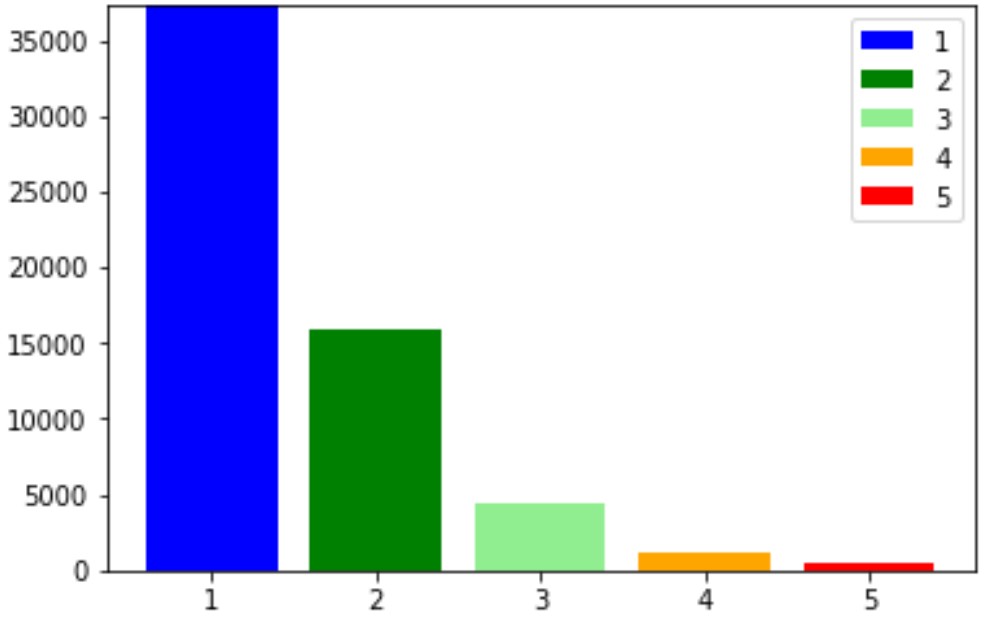
精度比較

手法	正解率	適合率	再現率	f1-score
k近傍法	0.41	0.37	0.42	0.38
ロジスティック回帰	0.44	0.37	0.44	0.40
SVM	0.44	0.37	0.44	0.40
ランダムフォレスト	0.43	0.37	0.43	0.39
GBDT	**0.45**	**0.40**	**0.45**	**0.41**

チューニング済みモデルの推定結果

- GBDTをハイパーパラメータチューニングした結果

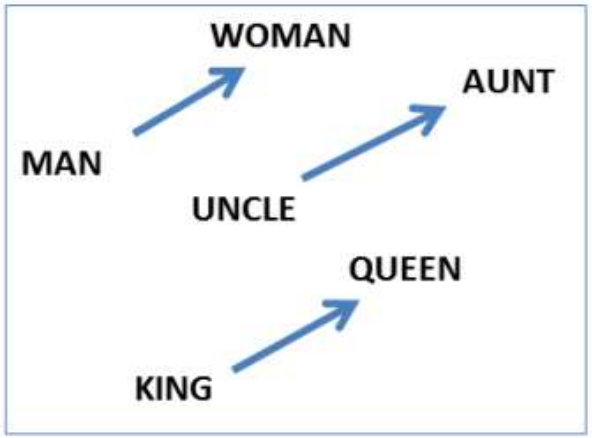
- 正解率　0.62
  - 一足目での精度
    62.31%
  - 二足目までの精度
    88.89%

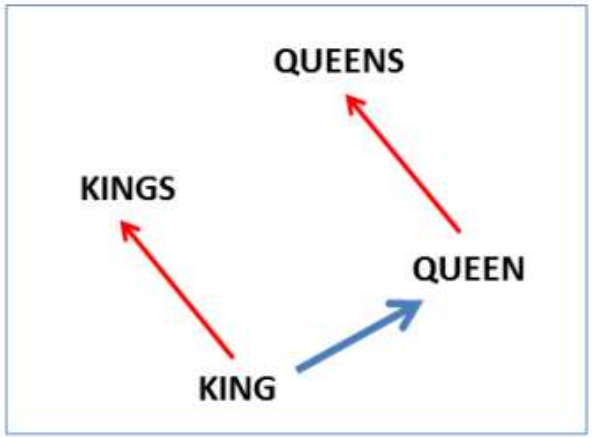


GBDT分類器におけるn番目での正解者数

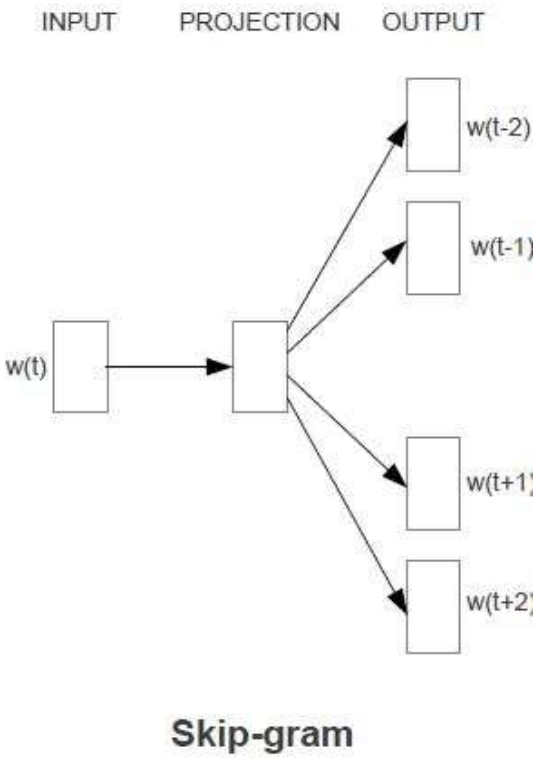
Word2Vec　単語の分散表現

- 単語のベクトル化
  - 単語の特徴どころかその意味構造までもがベクトル化されている
  - 単語同士の類似度や、単語間での加算・減算などができる！
  - 例："king" + "woman" - "man" = "queen"
  - 隠れ層と出力層の 2 層からなる単純なニューラルネットワーク
    - 深層学習に使われる自己符号化器 (オートエンコーダ) のようなもの





(Mikolov et al., NAACL HLT, 2013)



Skip-gram

item2vec アイテムのベクトル化

- **word2vecをアイテム集合に適用したもの**
- word2vec：文書　　item2vec：購買履歴POSデータなど
- 単語　→　アイテムID
- 文章(単語列)　→　ユーザごとに購入したアイテムIDリストの文字列
- NNモデル　Skip-gram
  - 周辺語予測　→　購入商品群予測

まとめ

- レコメンデーションでは商品とユーザの対応関係を如何にモデリングするのかが重要
  - 商品、ユーザ特徴の意味を含んだベクトル化、データ間距離が有用
  - 自動で集積させるデータから作り出す

平成28年度 私立大学研究ブランディング事業選定 Math Everywhere

ご清聴ありがとうございました

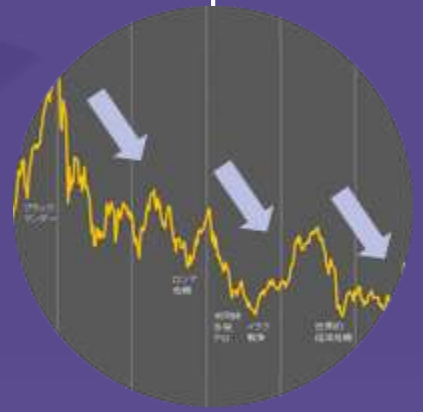
数理科学する明治大学

自己組織化

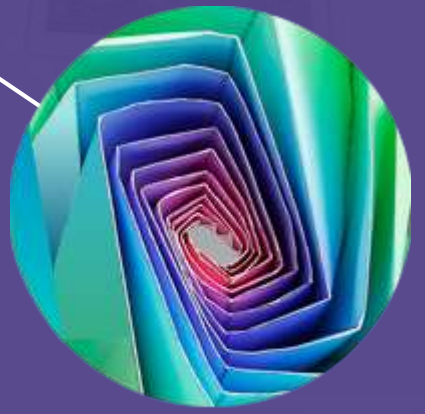
自己組織化とは

生物・社会システムの
形成と破たん

錯覚現象の解明と利用



金融危機の解明と予測



折紙工学の産業化

快適な介護空間の構築