

アドバンス・トップエスイー 最先端ソフトウェアゼミ成果報告 AI・データ分析グループ

アジェンダ

1. 全体ゼミでの文献調査方法の学習
2. 個別ゼミのチーム分け、目標設定
3. 実際にゼミでやったことの流れ
4. 自身の課題にゼミで得た知見をどう活かすか
5. 所感

全体ゼミにおける文献調査・発表

■ 目的

- ゼミの基本である文献調査方法・発表方法を学ぶ
 - 個別ゼミで技術調査・ツールの調査を行う上で必要なスキルを学ぶ
- 各メンバーが調査した技術を共有、内容について議論

■ 内容

- 各メンバーが興味を持っている技術の論文を調査・発表
 - AI、テストイング、アイトラッキング、形式手法、CPSなど
- 技術の知識を広げると共に、同じ分野に興味を持つメンバーを発見
- 個別ゼミで同じ興味を持つメンバーと、特定の技術を深く調査を進める
 - 本ゼミはAI・データ分析に興味を持ったメンバーが調査・試行を実施

個別ゼミでの目標・テーマ

■ 目標・テーマ

■ 機械学習を実践し問題の解き方を学ぶ

■ 学習手法として

- 何があるのか
- 何を使うのが最適なのか

■ 問題の解き方のアプローチ

■ 扱うテーマの選定

- メンバーが興味のあるトピックから選定(次ページ)
- 基本となる回帰・分類・クラスタリングの手法を調査・試行
 - 試行には複数のライブラリを選定、調査結果を共有することでライブラリの違いも確認
- 調査・試行をする中で分からなかったこと・足りないことを深く調査

個別ゼミでの目標・テーマ

実習の進め方
 (家にやるかゼミでやる?)
 前半パートの様子を見守り設定
 ツールをそれぞれ変える
 R, Python, C#

セミナーの時間と場所
 最も
 説明変数
 テキストマイニング

解きたい問題
 視線パターンの分類

AI(機械学習)に対する
 検証
 クラスタリング K-means
 ツーラーの使い方
 ツーラーの使い方

決定木
 遺伝的プログラミング
 画像処理
 データ分析
 データ増強

実装時の問題
 速度
 研究→論文→発表の
 具体的なサイクル

分類
 解きたい問題
 エラープログラムの原因による分類

個別ゼミ全体まとめ

- 実際に手を動かすことで分かったこと
 - 機械学習の使い方(ライブラリの使い方)
 - データの扱い方
- 使うだけではなく使いこなすために、理論についても調査
 - 数学の重要性(アルゴリズムの中身がわかっていないといけない)
 - チューニングするパラメーターの意味
 - 各アルゴリズムの特徴・弱点
- さらに複雑な問題に挑戦し学習前のデータ分析の重要性を認識
 - 可視化による扱うデータの選定
 - 相関関係によるデータの組み合わせなど
 - 欠損値の補完
- 個別ゼミでの取り組みについて
 - 週一回のゼミで上の内容が議論できてよかった
 - 時間割くのが大変。会社・部署のサポートが必要

実際にゼミでやったことの流れ

1. 簡単なサンプルデータでライブラリなど使って問題を解いた
 - 機械学習に不慣れなメンバーもここで機械学習に親しめるようになった
 - 同時に毎週のゼミでのスタイル(毎週4人で発表して議論する)が確立した
 - せっかくなので色々なライブラリを使ってみようと各人別々のものを使用
 - 使えたことは使えたものの、機械学習自体にぴんと来ていないというのもあって、2の調査を行うことにした
2. 機械学習手法の中身や特性をよくわかってないので分担してそれぞれ詳しく調べた
 - scikit-learnのチートシート等を参考に分類・回帰のアルゴリズム中心に調査
 - アルゴリズム自体の特徴等は学んだことがないメンバーが多く、実りの多い調査となった
 - ある程度行ったところで再び実践に戻って生かそうという事で、3に挑戦
3. 再びより難易度の高いデータの分析に挑んだ
 - 世界中で親しまれているデータ解析コンテストのkaggleにある問題を実施して、実戦でスコアがどの位まで上がるかチャレンジした

簡単なサンプルデータでライブラリなど使って問題を解いた

■ Irisの種別分類・住宅価格予測

- 初学者も多いので、機械学習で入門的に用いられる上記の有名なデータセットを利用して発表した。
 - 分類: アヤメ(植物)の花のデータを学習させて、データからアヤメの中でどの花(ヒオウギアヤメ、バージニアアイリス...)であるかを推定させる。
 - 回帰: 住宅地の周辺情報など(犯罪率とか税率、川の近くかどうか...)を用いて、住宅価格を推定させる。
- 分類・回帰がどういう概念か、どういう風に行うかという事は各人習得できた
 - とはいえ、アルゴリズムやパラメータを選ぶ基準はほぼ「なんとなく」とか「デフォルト」止まりである事が多く、かつその意味が理解できていないケースも多かった。

簡単なサンプルデータでライブラリなど使って問題を解いた

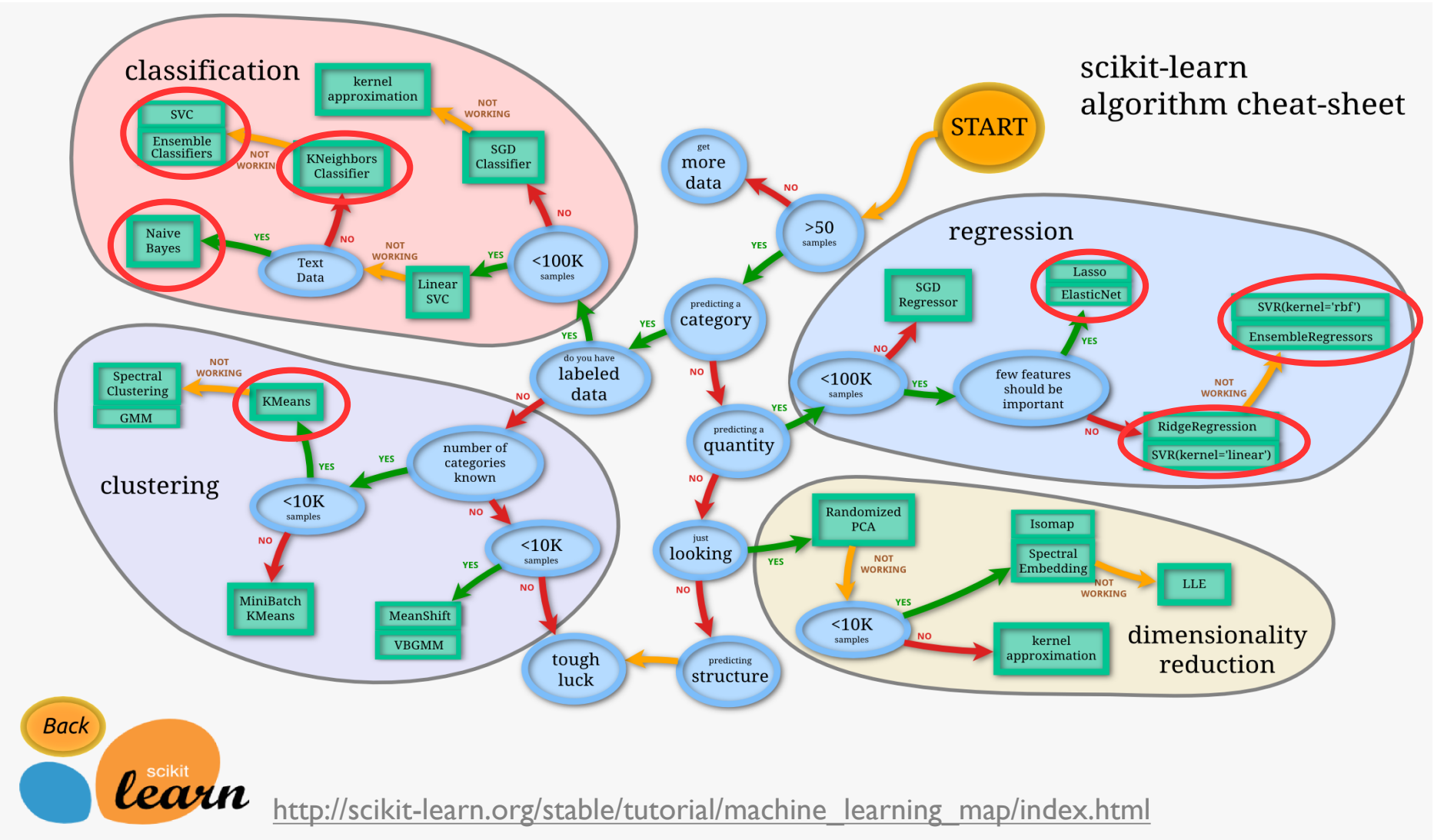
■ 使ったライブラリの話

- scikit-learn(Python)、Tensorflow(Python)、Weka(Java)、R(R言語)、accord.net(C#)を使用した。
- 基本的な環境構築や実施方法を学べた。
- GPGPU環境構築等、若干手間のかかるものもあったが、基本的には知識があればどのライブラリも数時間程度で構築・実施出来る事を確認。
- ライブラリ毎にデータの整形の仕方や整形するレベル感等が異なることが発表から確認できた。
- ライブラリ毎に何が使える、使えないか、どのように拡張できるかなどをざっくりと確認できた。
- より理解を深めるために、アルゴリズムそのものを学ぼうという事で、次のステップに移行。

機械学習手法の中身や特性がよくわかってないので分担してそれぞれ詳しく調べた

- 前のステップまでの流れで、アルゴリズム自体の差異や特徴、どういうアルゴリズムがあるかをあまり把握していないことが発覚した。
- そのため、より正しい機械学習でのデータ分析などを実施するために色々なアルゴリズムについて学んで見ることになった。
- どのようなアルゴリズムを学んで見るべきかという事で、機械学習のライブラリとして最もよく使用されているものの一つであるscikit-learn、そのサイトで開設されているcheatsheetに出てくるアルゴリズムであれば、使用頻度も高く実用的な意味のあるアルゴリズムが多いだろうと判断し、その中から選定した。
- 各人それぞれが毎週1つをテーマにして研究、発表をして議論するというスタイルで知識を得ていった。

機械学習手法の中身や特性がよくわかってないので分
担してそれぞれ詳しく調べた(赤線部が研究対象)



機械学習手法の中身や特性がよくわかってないので分担してそれぞれ詳しく調べた

- どのようなアルゴリズムなのかと言う数学的な内容をメインで語られた。
- 広く分類・回帰のアルゴリズムを中心に学ぶことができたため、これまで「なんとなくこれ」という感じで1つのアルゴリズムを中心に見ていることが多かったが、複数のアルゴリズムを選択肢に持てるようになった。
- アルゴリズムを学んだことによって、ライブラリなどで設定できるパラメータの意味が理解できるようになり、より精度の高い分析が可能になった。
- 本当に精度の高い分析が可能になったかどうかを試すために、再びデータ分析の実践を行ってみようという事になった。

難易度の高いデータの分析に挑戦

■ KaggleのTitanic: Machine Learning from Disasterに挑戦

- Kaggleとは:
企業や研究者がデータを投稿し、世界中の統計家やデータ分析家がその最適モデルを競い合うプラットフォーム

<https://www.kaggle.com/>








■ Titanic: Machine Learning from Disaster

- Titanicの乗客データから生存・死亡を予測するモデルを作成しスコアを競う
- データは右表の形式
- 学習データ(Survivalの値あり)からモデルを作成し、テストデータ(Survivalの値なし)のSurvivalの値を予測
- スコアは正解数/全テストデータ数のAccuracyで表現される

変数	詳細
Survival	死亡:0, 生存:1
Pclass	チケットの階級(1 st , 2 nd , 3 rd)
Name	名前
Sex	性別
Age	年齢
Sibsp	共に乗船した兄弟・配偶者の人数
Parch	共に乗船した親・子供の人数
Ticket	チケット番号
Fare	旅客運賃
Cabin	キャビン番号
Embarked	Titanicへの乗船港

学んだ機械学習手法の活用、しかし...

- 数値データを学習器に入力し、モデル作成
 - 死亡:0 or 生存:1 の分類問題のため、SVM・ランダムフォレスト等を各自選択
- しかし、作成した予測モデルの精度はあまり高くない

6790	▼ 120	Srikanth Srungarapu		0.71770
6791	▼ 120	Shanger Lin		0.71770
6792	new			0.71770
Your Best Entry ↑				
Your submission scored 0.71770  Tweet this!				
6793	▼ 121	AnuragKumar 2		0.71292
6794	▼ 121	Navneet Pal		0.71292

- スコアは約0.72
 - 7700人中6800位
- 学習に用いたデータの内容を確認し、変数を選択

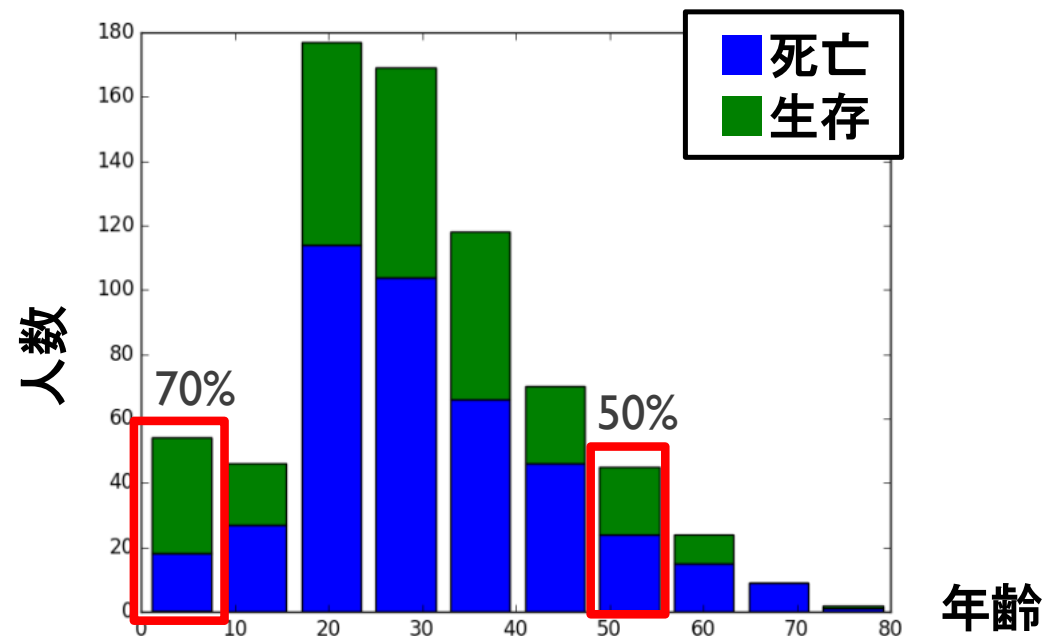
データの可視化

- 相関係数を表で確認

変数	Survivalとの相関係数
Pclass	-0.338
Sex(male)	-0.543
Age	-0.077
Sibsp	-0.035
Parch	0.082
Fare	0.257

- チケット階級、性別、運賃と高い相関

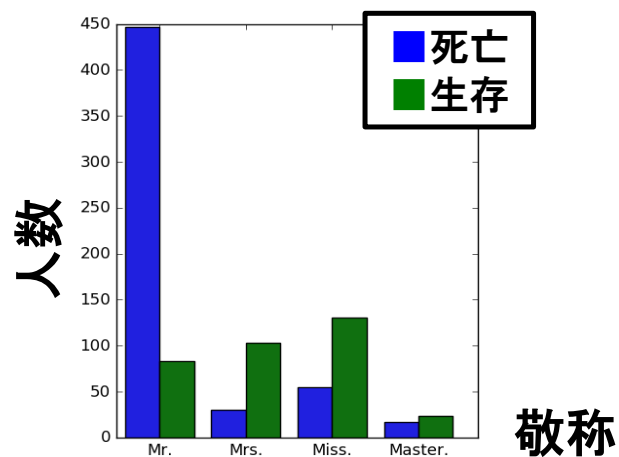
- 変数ごとの生存率をヒストグラムで可視化



- 10歳以下の生存率が高い
50～60歳の生存率も高い

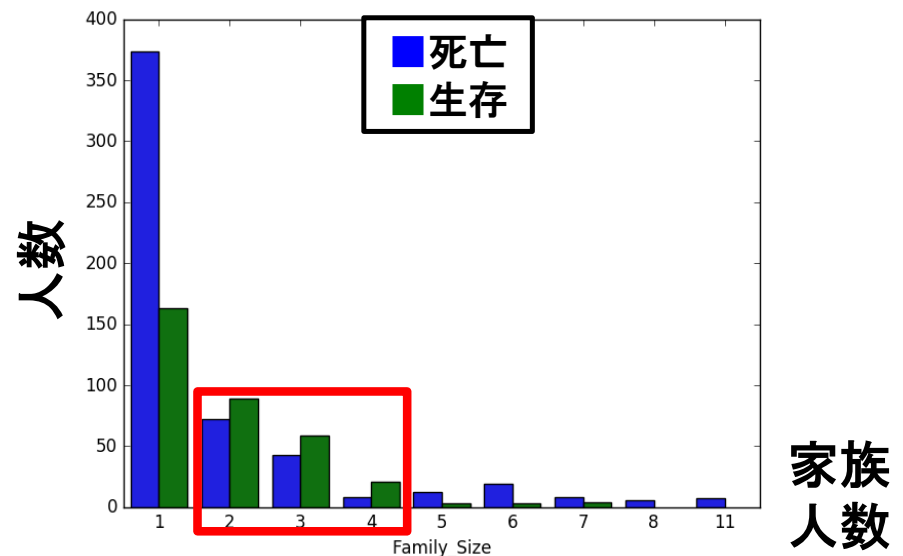
可視化して新たな変数を追加

- 名前は文字列のためそのままでは扱いにくいので、Title(敬称)ごとの生存率をヒストグラムで可視化して分析
例) Braund, **Mr.** Owen Harris
Heikkinen, **Miss.** Laina



- Mrs. Miss.(女性)および Master(若い男性)の生存率が高い

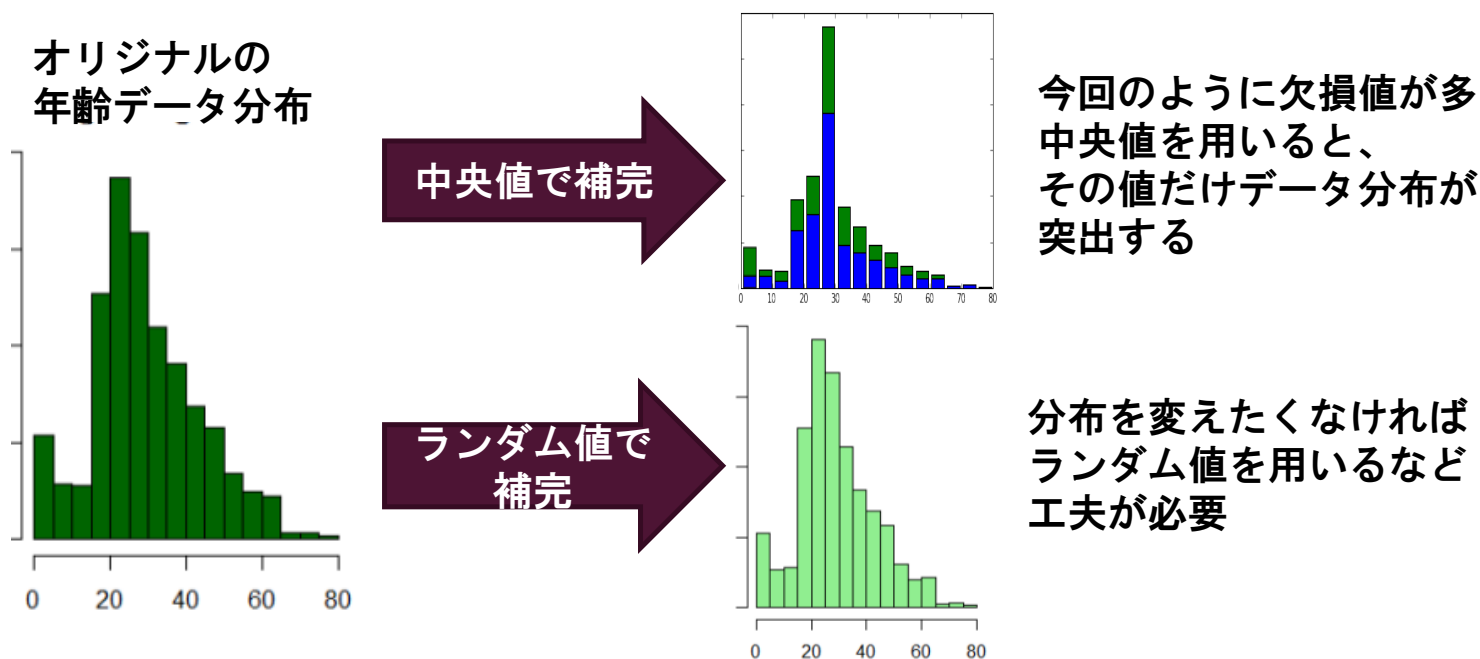
- Sibsp(兄弟/配偶者)と Parch(親/子供)の人数を足して Family_Size という変数を作成・可視化



- 家族が2~4人の生存率が高い

欠損値の補完



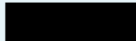




- Titanic乗客の実データのため、欠損しているデータも多い
891データ中、年齢:177件, キャビン:687件, 乗船港:2件が欠損
- しかし、年齢データは可視化の結果、生存率と相関がありそうなので
利用したい→欠損値を補完する必要がある
 - 平均値・中央値やランダム値、欠損地以外の他の変数から予測などで補完



TITANIC生存者予測への取り組み 最終結果とまとめ

■ 結果

- 可視化によるデータ分析で
変数選択・新たな変数追加を試み
最終的には約0.80までスコアが上
(7700人中6800位→1500位)

1492	▲1056	Allia Bordes		0.79904	22
1493	new	TrentRogers		0.79904	11
1494	▲1669			0.79904	20
Your Best Entry ↑					
Your submission scored 0.79904				 Tweet this!	
1495	▼43	Kevin White		0.79426	6
1496	▼43	TaekYoung Kim		0.79426	17

■ 取り組みで得た知見のまとめ

- 機械学習活用に必要な知識は学習機そのものの知識だけではない
以下の技術も必要と実感
- 可視化してデータ分析し、必要な変数を選択する技術
- 文字列データの数値化、欠損値の補完の技術

個別ゼミ全体まとめ

- 実際にツールを使って手を動かして分かったこと
 - 機械学習の動かし方(ライブラリの使い方)
 - データの扱い方
- 数学の重要性(中身わかっていないといけない)
- 各アルゴリズムの特徴・弱点わかった
- 可視化とকাশないとデータ何使えばいいとか
- 週一回のゼミで上の内容が議論できてよかった
- 時間割くのが大変。会社・部署のサポートが必要