

1-4 データ・AI利活用のための技術

東京大学 数理・情報教育研究センター

2020年4月28日

2024年6月17日 改訂

概要

- データサイエンスやAI利活用の現場ではどういう技術が用いられるのでしょうか？
- ここでは基本的なものを見ていくことで、データ・AIを活用するために使われている技術の概要を知ることが目標とします

本教材の目次

1. データの1次分析と可視化	4
2. データ利活用のための技術	15
3. ビッグデータとAI	21
4. 参考文献	27

1-4-1 データの1次分析と可視化

データの1次分析

- 基礎統計量（平均値、最小値、最大値、分散など）、欠損値、偏りがいないかなどを調べます
- データ可視化も1次分析に含まれます

データの確認

```
# show first 5 records -> seems ok
df.head()
```

	id	type	region	code	prefecture	city	town	nearest_station	nearest_station_min	price	...	future_usage	orientation	road_type	road_area
0	1	宅地 (土地)	住宅地	1101	北海道	札幌市中央区	旭ヶ丘	円山公園	28	86000000	...	その他	南	市道	10.2
1	2	宅地 (土地)	住宅地	1101	北海道	札幌市中央区	旭ヶ丘	円山公園	26	5000000	...	NaN	西	私道	4.0

基礎統計量

```
In [68]: # summary statistics
df[\"price\"].describe()
```

```
Out[68]: count    3.664002e+06
          mean     2.761324e+07
          std      1.432475e+08
          min      1.000000e+02
          25%      5.000000e+06
          50%      1.400000e+07
          75%      2.900000e+07
          max      6.100000e+10
          Name: price, dtype: float64
```

欠損値の確認

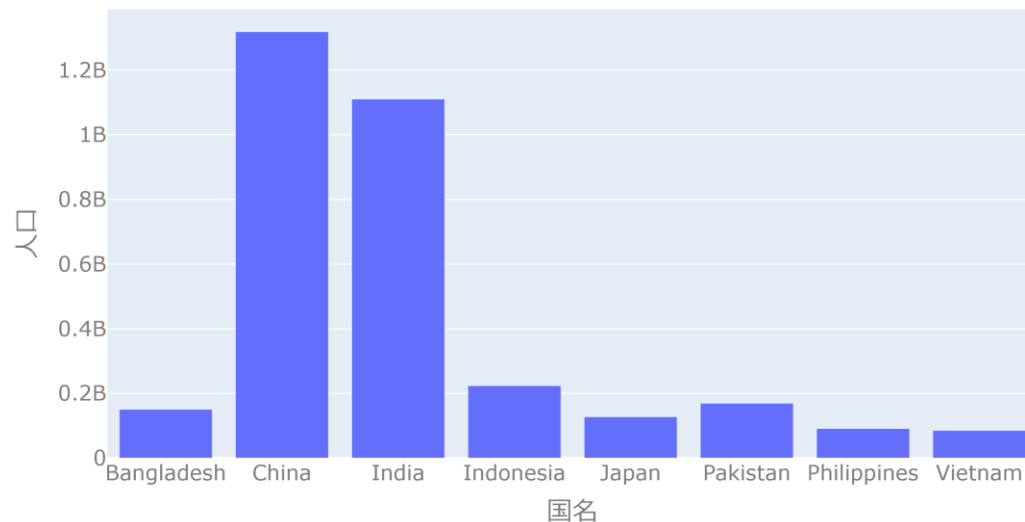
```
In [5]: # ratio of null records
df.isnull().sum() / len(df)
```

```
Out[5]: id                0.000000
         type              0.000000
         region            0.285000
         code              0.000000
         prefecture        0.000000
         city              0.000000
         town              0.002065
         nearest_station    0.143826
         nearest_station_min 0.150569
         price              0.000000
         unit_price         0.644265
         floor_plan         0.846775
         square_meters      0.000000
         unit_price_square  0.644265
         shape              0.287836
         frontage           0.349677
         total_floor        0.663158
         year_built         0.521600
         structure          0.509749
         purpose            0.514253
         future_usage       0.722026
         orientation        0.288267
         road_type          0.304437
         road_area          0.311708
         city_plan          0.134972
         coverage           0.182164
         volume             0.182164
         transaction_date   0.000000
         renovation         0.855279
         circumstances      0.939687
         dtvpe: float64
```

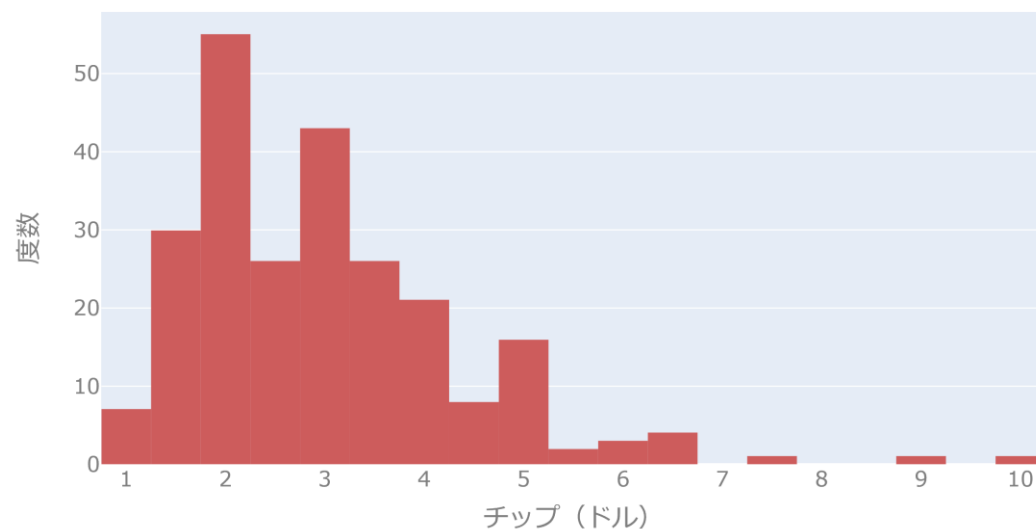
データ可視化

- 可視化の例を見ていきましょう

棒グラフ
(各国の人口)



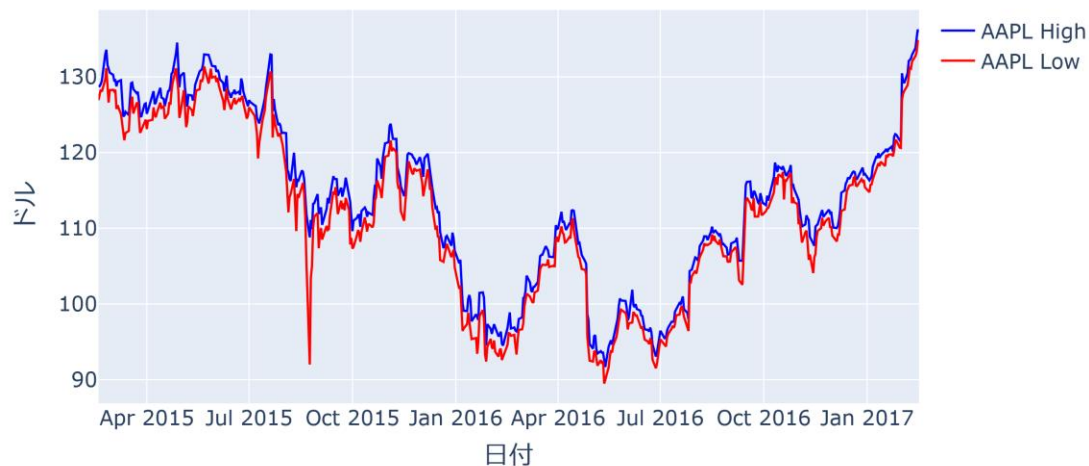
ヒストグラム
(米国のチップ)



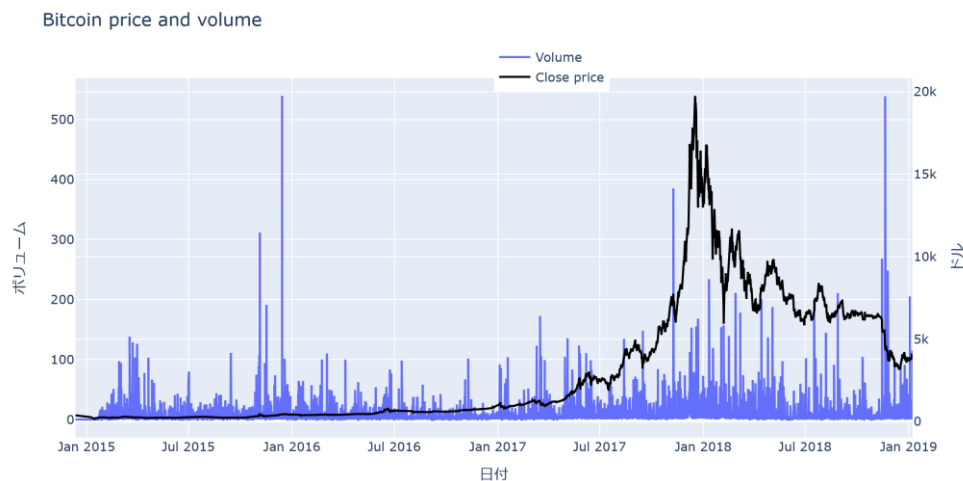
時系列

- 時間発展していくデータでは横軸を時間軸として扱います

アップル株の時系列

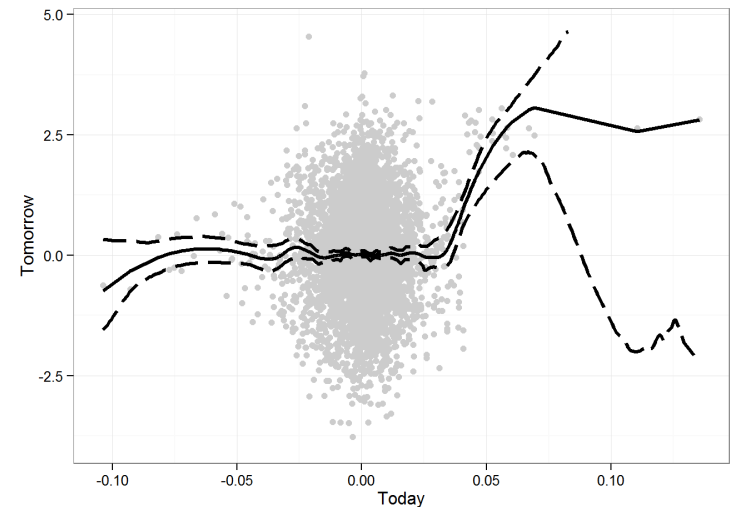
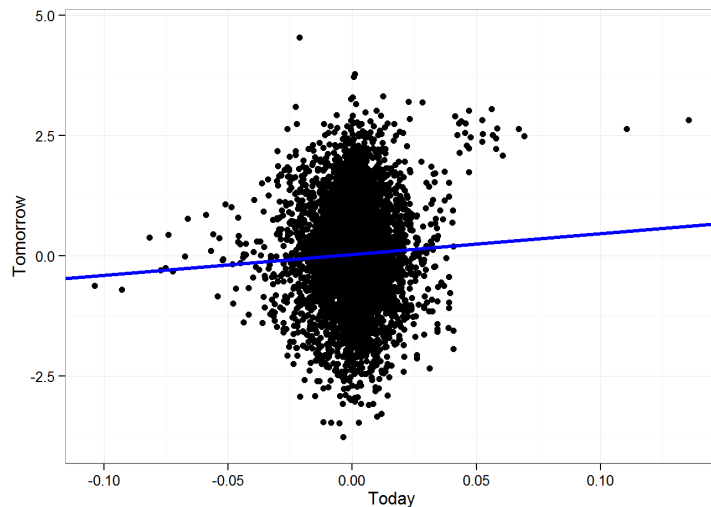


ビットコイン価格の時系列
➔右図のように複合チャートとして描画することもあります



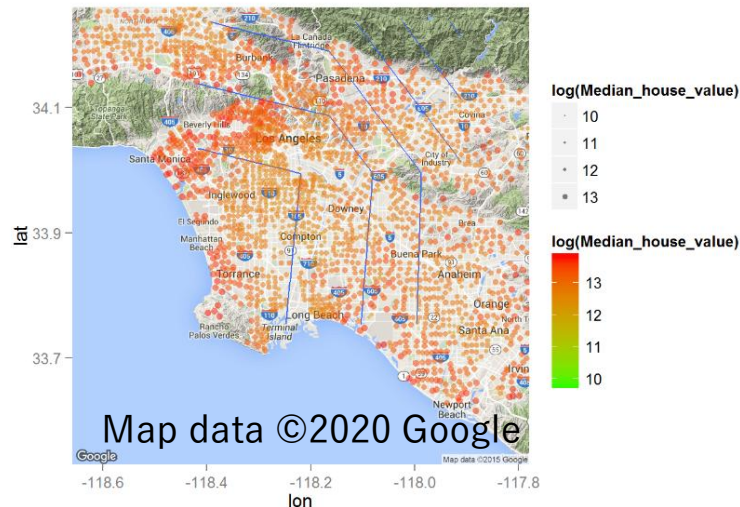
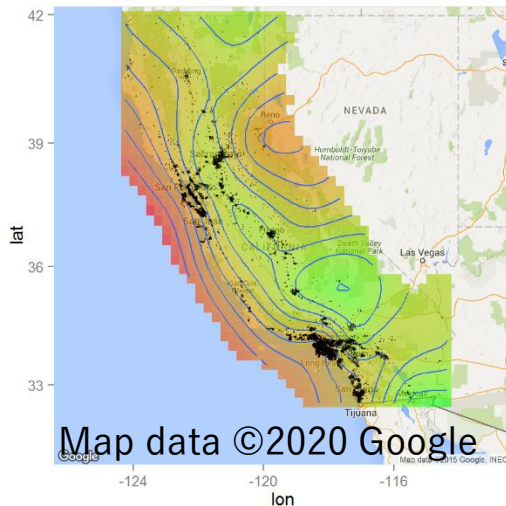
散布図と回帰

- 散布図は2つの変数の関係を確認するためによく使われます
- 散布図に沿うように直線や曲線を入れることもあります
➔ 左は線形回帰（直線で二変数の関係を表現）で右はスプライン回帰（滑らかな曲線で二変数の関係を表現）の例



地図上の可視化・ダイナミック可視化

- 地図上に表示したり動的に可視化したりすることで初めて見えるものも多いです
- 左下の図によるとカリフォルニア州では海外沿いになるほど住宅価格が高いことがわかります
- もっと手の込んだ図やダイナミック可視化については右下のリンクを参照してください

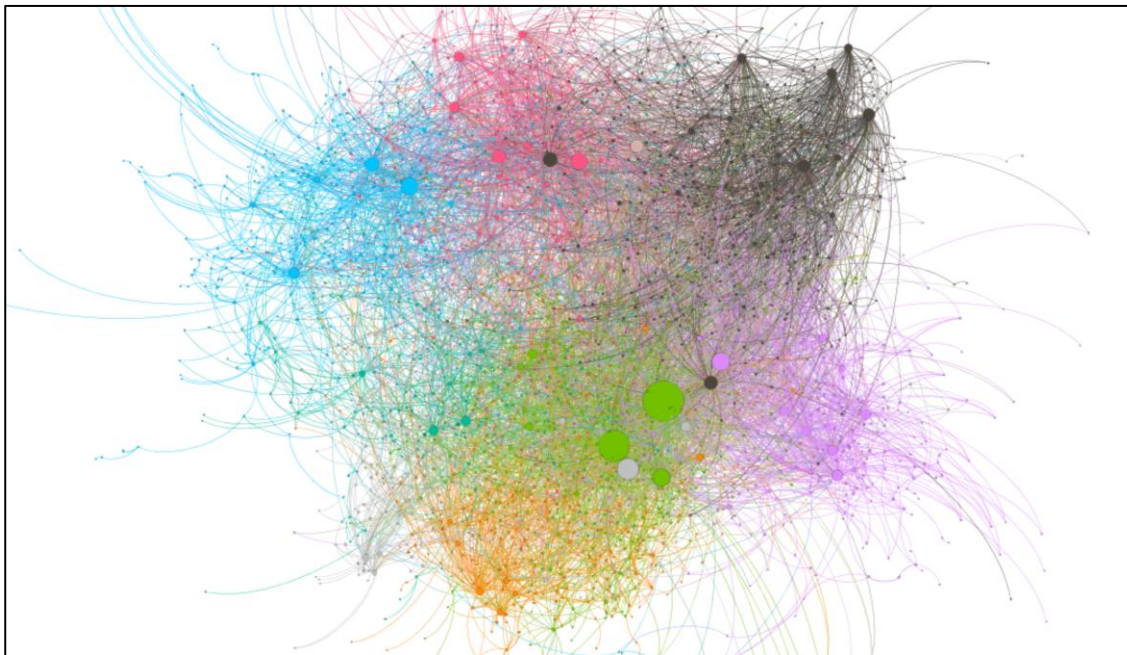


挙動・軌跡の可視化、
ダイナミックな可視化、
リアルタイム可視化の手の込んだ例
については次を参照
してください

<https://eng.uber.com/keplergl/>

関係性の可視化

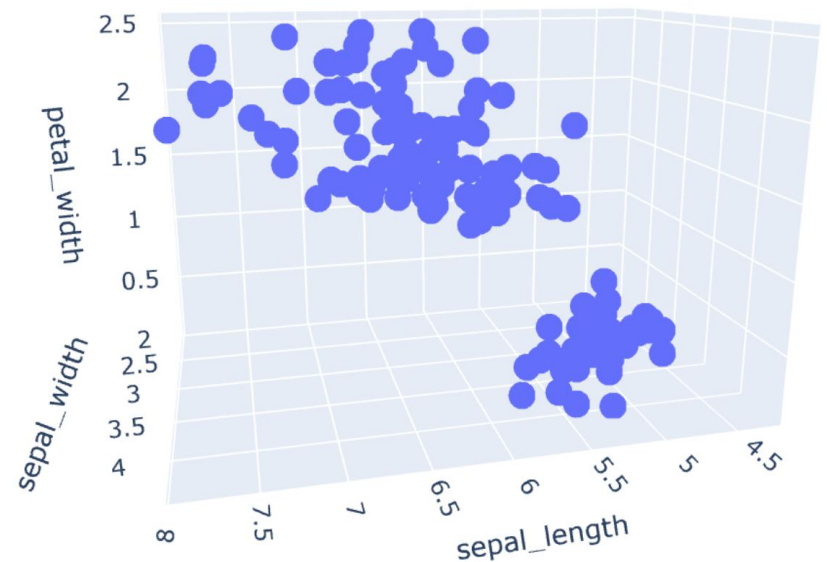
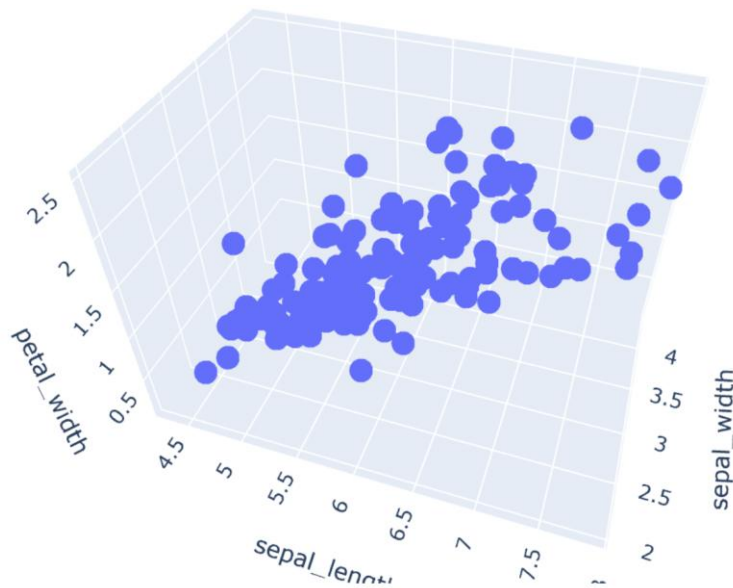
- 関係性の可視化にはネットワークや行列が使用されます
 - 行列は4-1-5で詳しく説明します
- 下記はネットワーク図として上場企業同士の関係を描画したものです



上場企業同士の関係

多次元の可視化

- 3次元の散布図を描くこともあります
 - 下記はアヤメの品種のデータでがく片の長さ、がく片の幅、花びらの幅の関係を可視化しています



言語処理

- テキストの分析では文書を単語の出現頻度行列としてまとめたり単語ごとにベクトルを与えたりすることで分析しやすい形に変換します
 - 文書を単語の出現頻度行列としてまとめたものをバッグオブワーズと呼びます

	the	is	of	finance	bank	tributuary	river
文書1	34	12	23	7	3	0	0
文書2	8	4	12	0	2	3	5
文書3	2	3	3	3	9	0	0
...							
文書N	12	43	12	0	5	0	5

バッグオブワーズの例

- 単語をベクトルに変換する手法もあります (word2vec[Milkoć et al.2013]など)

画像処理と認識技術

- 画像は画素（ピクセル）ごとにベクトルが記録されています
- そのままだと分析しづらいのでイメージを分割しタグ付けするなど前処理を加えることがあります
 - オブジェクト認識と呼ばれる認識技術の一種です
- 車載カメラで使用するオブジェクト認識の具体例はCityscape(<https://www.cityscapes-dataset.com>)で見ることができます

認識技術の
一例です



出典: Cityscape <https://www.cityscapes-dataset.com/examples/#fine-annotations>

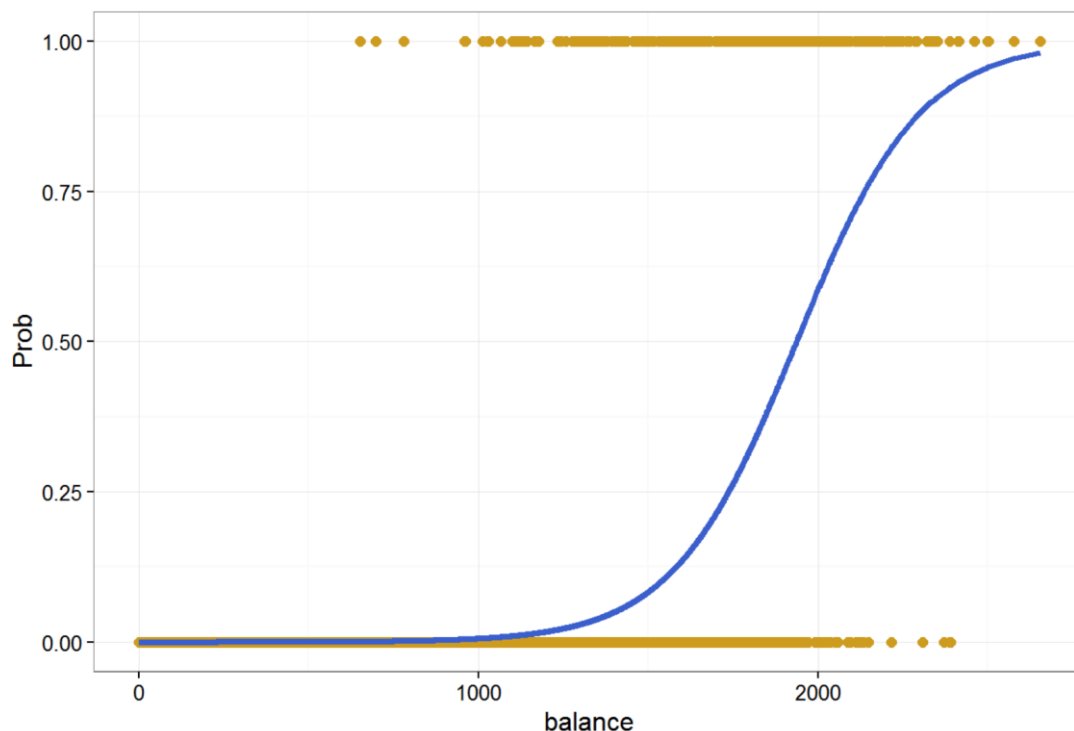
音声処理

- オーディオ信号は分析しやすい形式に変換し分析することが多いです
- 変換の例としては短時間フーリエ変換、メル周波数係数、定数Q変換、クロマグラムなど様々な方法があります
 - 詳細は[Choit et al.2018]
(<https://arxiv.org/pdf/1709.04396.pdf>)を参照してください

1-4-2 データ利活用のための技術

予測

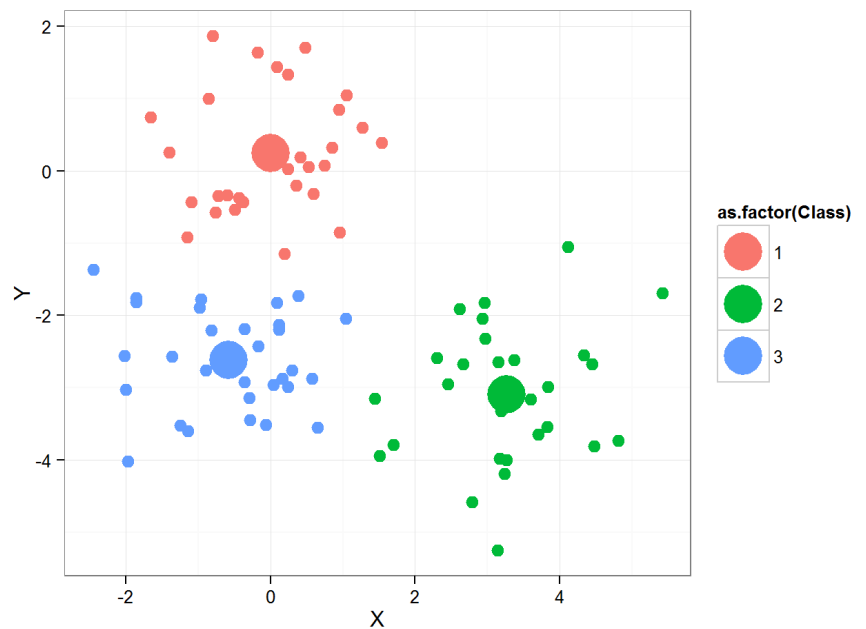
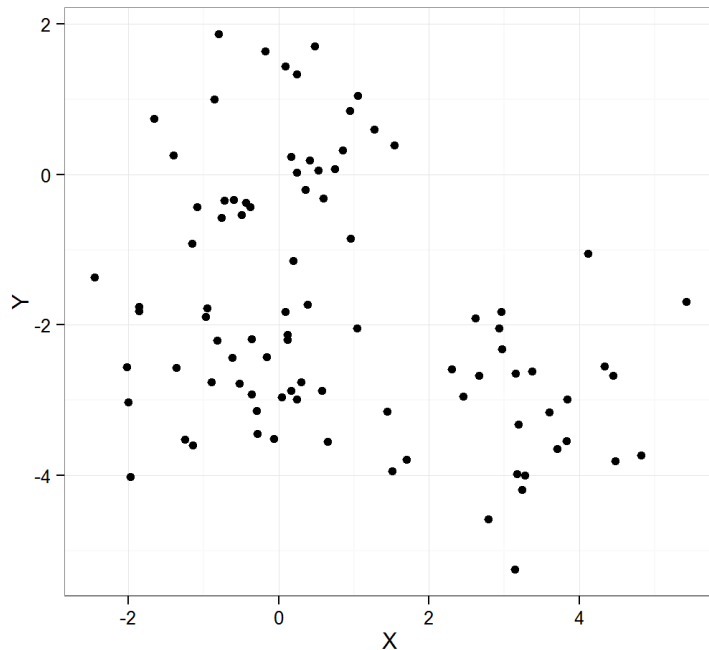
- 訓練データからパターンを学習し、同様のパターンをもっていると考えられる未知のデータの分類や予測に役立てます
 - 下記以外にも時系列予測やネットワークのリンク予測など応用は豊富です



ロジスティック回帰による
デフォルト確率の予想
(この例についての詳細は
[James et al.2013]を
参照してください)

グルーピング

- グルーピング（クラスタリング）とはデータをいくつかのまとまりに分割することです
- 下は二次元で表現されたデータを3つのクラスタに分割した例です
 - （参考）下記はK-meansの例です



パターン発見とルールベース

- データ内のパターンを発見する手法も多くあります

マーケットバスケット分析はデータの中からよく出現するパターンを見つける手法のことです。例えば[Hastie et al.2009]のpg494では米国のアンケートのデータを用い

「結婚しており、家を所有している」なら
「住居はアパートではない」など

データからルールを見つけることに成功しています。こうしたものをルールベースによるアプローチと言います。

トピックモデルは文書からトピックと呼ばれる潜在構造を推論する手法です

<<Gozo Stadium>>

...The main mission of the new board of the GFA is to improve the facilities of the Gozo Stadium in order to reach UEFA and FIFA standards so in the future the stadium could host international matches of the Maltese National Team.

...The main mission of the new board of the GFA is to improve the facilities of the Gozo Stadium in order to reach UEFA and FIFA standards so in the future the stadium could host international matches of the Maltese National Team.

<<bet365 Stadium>>

The bet365 Stadium is an all-seater football stadium in Stoke-on-Trent England and the home of Premier League club Stoke City.

The bet365 Stadium is an all-seater football stadium in Stoke-on-Trent England and the home of Premier League club Stoke City.

<<University of Colorado Law School>>

The University of Colorado Law School is one of the professional graduate schools within the University of Colorado System....United States Supreme Court Justice Wiley Blount Rutledge graduated from the University of Colorado Law School in 1922.

The University of Colorado Law School is one of the professional graduate schools within the University of Colorado System....United States Supreme Court Justice Wiley Blount Rutledge graduated from the University of Colorado Law School in 1922.

トピック1 トピック2 トピック3 トピック4

stadium	he	bank	state
world	football	war	university
cup	club	financial	mathematical
de	first	banks	school
international	league	banking	theory

最適化

- 大きくは連続最適化、離散（組み合わせ）最適化に分かれます
 - 連続最適化：変数が連続値（実数値）をとる最適化問題が対象
 - 離散最適化：組み合わせなど微分不可能な関数が対象
- 数理・計算機の発達と共に従来では「解く」ことが難しかった問題も近似的に解けるようになりました
 - 深層学習の発展とも関連します

モデル化・シミュレーション・データ同化

- 実機の使用が困難な時や想定シナリオを作る際には、モデル化を経て、シミュレーションが用いられます
- シミュレーションをより現実的なものにするために実データから得られた観測値を統合することをデータ同化と呼びます

アンリアルエンジンを用いた自動運転シミュレーター

<https://unrealengine.com/ja/spotlights/carla-democratizes-autonomous-vehicle-r-d-with-free-open-source-simulator>

上のリンクの例はまるでゲームのようですが、リアルなシミュレーションを構成することで実機を使用せず自動運転のプログラムをテストできるため注目を集めています

NASAによるハリケーンのシミュレーションの例は

<https://www.youtube.com/watch?v=p-3aB9hJ8Hc>

で見ることができます

1-4-3 ビッグデータとAI

AIとビッグデータ

- 「1-1 社会で起きている変化」で多くの例を紹介しましたが、本節で紹介した分析技術を適用できるデータは数多くあります
 - 物理学、生物学、化学、政治学、法律文書、文学、音楽、テキスト、動画、経済学、ファイナンス、ビジネス、アプリ開発、農業、機械制御など
- こうしたデータを活用することでAIはパターンを学習し社会における様々な分野で役立てられます

特化型AIと汎用型AI

- 「1-1 社会で起きている変化」の生成AIの項でも紹介しましたが、特化型AIとはあるタスクの処理に特化したAIのことです。それに対して汎用AIとは様々なタスクの処理に対応できるAIのことです。
- 本節で紹介した技術を突き詰めることで多くの特化型AIは作成されています
- 複数の技術を組み合わせることで汎用AIに近いものもできるかもしれませんが、現代ではまだ模索段階です
 - データの処理性能などでAIは人間をすでに上回っていますが、理解を与えたり、未知の状況に対応したりするなど様々な面でまだ課題を抱えています
 - 大規模自然言語や基盤モデルは、学習時に訓練されたタスク以外にもそれなりの精度でパフォーマンスを出せるという意味で、汎用AIに一步近づいたとも言えますが、依然として課題を抱えています
 - 今のAIでできないことは研究開発のフロンティアになっています

生成AIの活用（プロンプトエンジニアリング）

- ChatGPTを少し触るとわかると思いますが、ちょっとした指示の出し方で答えが変わることがよくあります。自分の望んだような振る舞いをさせるために入力を工夫することは自然な発想です。
- プロンプトエンジニアリングとは人工知能への指示の出し方を工夫し最適化する枠組みのことです
 - Chain of thought[Wei et al.2022]とは答えにたどり着くまでの思考過程を明記した問いと答えのペアをプロンプトに含めた上で、本当に聞きたい質問を聞く工夫です
(<https://www.promptingguide.ai/jp/techniques/cot>)
 - ペルソナ設定は特定のキャラクターや性格を持つようにする工夫です（下図）

以下のペルソナを設定してください。

年齢：20歳

職業：大学生（東京大学工学部）

性別：男性

ペルソナを加味しておすすめの海外旅行先を教えてください



1. シリコンバレー, アメリカ

- 理由：工学部の学生にとって、シリコンバレーはテクノロジーの中心地として非常に魅力的です。多くのスタートアップや大手IT企業が集まっており、最先端の技術やイノベーションを体験できます。
- アクティビティ：GoogleやAppleのキャンパスツアー、スタートアップの見学、テックイベントやミートアップへの参加。

- プロンプトエンジニアリングを自動化しようとする試みもあります
 - <https://www.tensorops.ai/post/top-tools-for-prompt-engineering>

自動機械学習

- 「1-1 社会で起きている変化」でも紹介しましたが、本節で見た機械学習モデルの開発ステップを自動化する試みもあります
 - GoogleのAutoML(<https://cloud.google.com/automl>)
 - MicrosoftAutoML(<https://docs.microsoft.com/ja-jp/azure/machine-learning/concept-automated-ml>)
- そうした自動化技術の進化した先にもしかしたら未来のAIがあるのかもしれない

自然言語処理 × 画像生成（マルチモーダル）

- 「1-1 社会で起きている変化」の生成AIの項でも説明しましたが、インターネットスケールで手に入るデータはテキスト情報だけではありません。テキスト情報と画像情報など二つ以上のドメインのデータを同時に活用できるように深層学習モデルを創ることをマルチモーダル学習といいます。
 - 下記の例では機械とのインターフェースに自然言語を利用しながら高度な画像生成タスクを処理することができるようになっていきます

GLIDE

<https://gpt3demo.com/apps/openai-glide>

DALL-E2

<https://openai.com/index/dall-e-2>

1-4-4 参考文献

参考文献

[Choi et al.2018] Keunwoo Choi et al., "A Tutorial on Deep Learning for Music Information Retrieval", <https://arxiv.org/pdf/1709.04396.pdf>

[Hastie et al.2009] Trevor Hastie, Robert Tibshirani, Jerome Friedman, "The Elements of Statistical Learning: Data Mining, Inference, and Prediction", Second Edition, February 2009

[James et al. 2013] James, G., Witten, D., Hastie, T., Tibshirani, R., "An Introduction to Statistical Learning with Applications in R", Springer, 2013.

[Mikolov et al.2013] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S. Corrado, Jeff Dean, "Distributed Representations of Words and Phrases and their Compositionality", Part of: Advances in Neural Information Processing Systems 26 (NIPS 2013)