scikit-learnとTensorFlowによる実践機械学習 1章

阿部 泰之

タイムスケジュール

17:30 ～　開場　受付開始
17:30 ～ 17:45　発表準備
17:45 ～ 18:45　1章　機械学習の現状
18:45 ～ 18:55　休憩
18:55 ～ 19:55　2章　エンドツーエンドの機械学習プロジェクト
20:00 ～ 20:30　今後のスケジュール、担当について
20:30 ～ 21:00　片づけ撤収

Wifiはありません。

自己紹介

・阿部 泰之　/ Hiroyuki Abe
・「scikit-learnとTensorFlowによる実践機械学習」の輪読会をやると決めた人で、場所の用意もしています。

・業務エンジニア
　（生命保険　主に保険金支払）

・twitter　　/ @taki_tflare
・https://tflare.com

グループ紹介

AI&機械学習しよう！（Do2dle）
・https://www.facebook.com/groups/do2dle/

AI&機械学習関連の勉強会を実施しています。
非公開グループですので、上記から参加をお願いします。

最近のニュース

個人的な補足

- 最近機械学習のフレームワークがどんどん使いやすくなっています。
- 例えばLSTMもtf.contrib.rnn.LSTMCellメソッドを呼び出せば良くなっています。

https://www.tensorflow.org/api_docs/python/tf/contrib/rnn/LSTMCell

最近のニュース

個人的な補足

- Google、AIツール構築サービス「Cloud AutoML」の画像認識版をβに、自然言語と翻訳もβで追加

http://www.itmedia.co.jp/news/articles/1807/25/news066.html

画像認識に最適化したサービス「Cloud AutoML Vision」

自然言語の「Cloud AutoML Natural Language」

翻訳の「Cloud AutoML Translation」

最近のニュース

個人的な補足

- Cloud AutoML Visionは、どう使えるサービスなのか

http://www.atmarkit.co.jp/ait/articles/1801/30/news020.html

最近のニュース

個人的な補足

- AutoKeras: The Killer of Google’s AutoML

https://towardsdatascience.com/autokeras-the-killer-of-googles-automl-9e84c552a319

To use Google’s AutoML for computer vision, it will cost you USD $20 per hour. That’s crazy!

最近のニュース

個人的な補足

- 機械学習の一般化が進んでいるように見えます。
- これからは、機械学習＋αが求められていくでしょう。

アジェンダ

下記の1章について輪読を行います。

はじめに

随時コメント下さい。

本の内容に沿っていない補足の箇所は、その部分の右上に下記をつけています。（すべて補足の場合はページ右上につけます。）

個人的な補足

ついていない場合は、本の内容のままです。

1. 機械学習の現状

個人的な補足

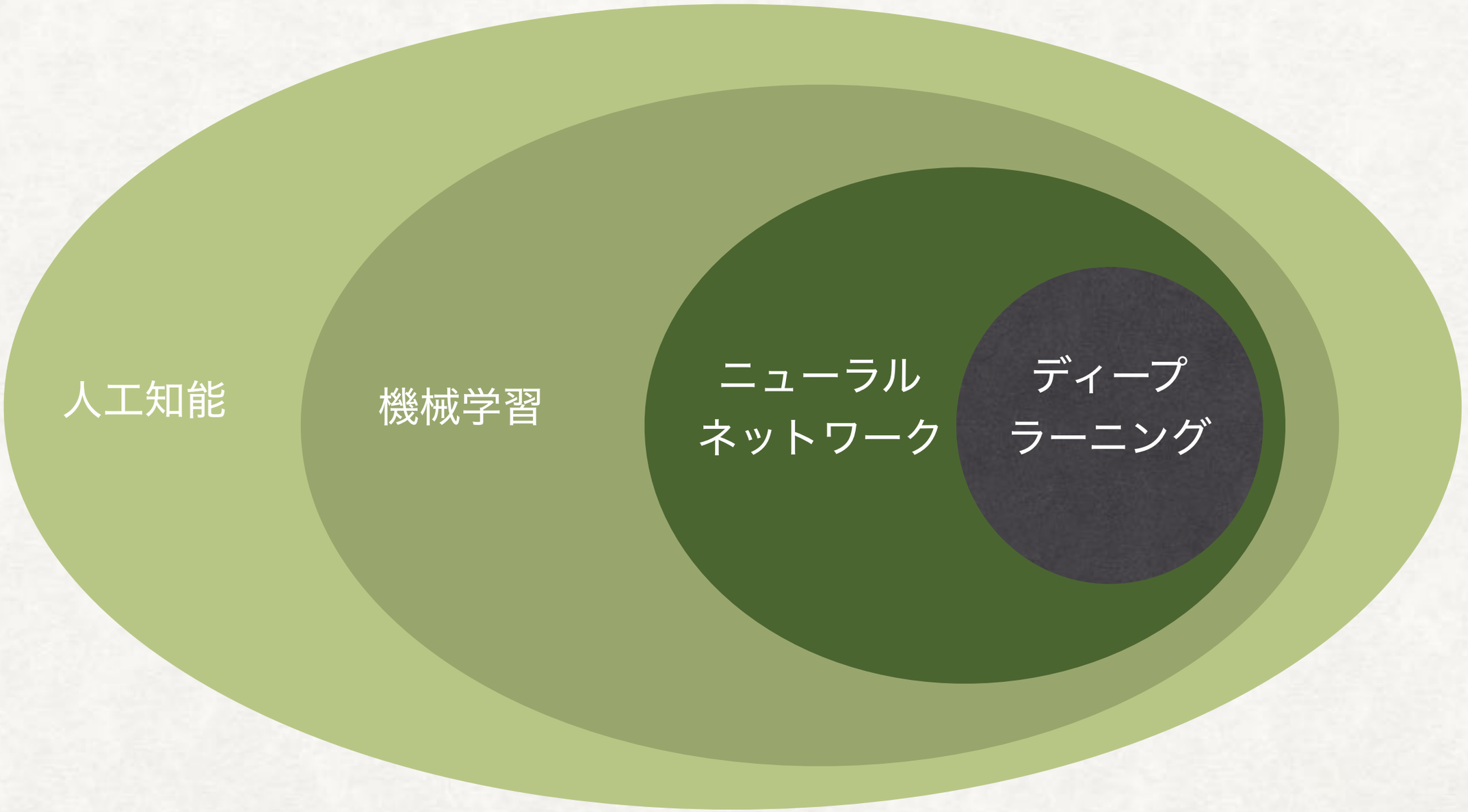
- 「機械学習」（Machine Learning：ML）という言葉を聞いてほとんどの人がイメージするのはロボットだろう。

- とあるが、日本では「機械学習」を勉強している人以外には通じません。

- 人によって「人工知能」と言いわける必要がありますね。

用語についてのまとめ

個人的な補足



1. 機械学習の現状

- OCR（何十年も前）
- スパムフィルタ（1990年代～）
- その後数百の製品やサービスで日常的に使われるようになっている。
- 例えば
- 　おすすめの商品の提案
- 　音声検索

1章では

1. 機械学習の現状

- 機械学習とは何なのか、なぜ機械学習を使いたいと思うときがあるのかというところから話を始めていく。

- 教師あり学習
- 教師なし学習

- オンライン学習
- バッチ学習

- インスタンスベース学習
- モデルベース学習

- 典型的なMLプロジェクトのワークフローをながめ、直面する課題がどのようなものかを検討し、機械学習システムを評価、調整するための方法を説明する。

1. 機械学習の現状

- この章では、すべてのデータサイエンティストが頭のなかに叩き込んでおかなければならない基本概念（および専門用語）をたくさん紹介する。この章は概要を俯瞰的に説明するものになる（ちなみに、コードがあまりない唯一の章である）。書かれているのはすべて比較的単純なことだが、本書の続きの部分を読むためには、ここで説明することをしっかりと頭に叩き込まなければならない。

1.1　機械学習とは何か

- 機械学習とは、コンピュータがデータから学習できるようにするためのコンピュータプログラミングについての科学である。

1.2　なぜ機械学習を使うのか

機械学習は変化への対応が可能なため（スパムフィルタなど）

scikit-learnとTensorFlowによる実践機械学習 1.1参照ください。

機械学習は次のような問題を得意とする。

- 既存のソリューションでは、手作業による大量のチューニングや、ルールの長いリストが必要な問題。ひとつの機械学習アルゴリズムがあれば、コードが単純化され、性能が上がることが多い。
- 伝統的な方法ではよいソリューションが作れないような複雑な問題。最良の機械学習テクニックならソリューションを見つけられる。
- 変動する環境。機械学習システムなら新しいデータに対応できる。
- 複雑なシステムや大量のデータについての知見の獲得。

具体的にどのようなものに活用されようとしているか

個人的な補足

- 機械学習による異常検知を活用したクレジットカード不正利用の検知

- 侵入検知システム（IDS：Intrusion. Detection System）や侵入防止システム（IPS：. Intrusion Prevention System）の不正パケットの検知

具体的にどのようなものに活用されようとしているか

個人的な補足

具体的にどのようなものに活用されようとしているか（個人的な補足）

個人的な補足

- OpenAIの人工知能「OpenAI Five」がDota 2の5対5バトルで人間チームに勝利
- トレーニングに256 P100 GPUs and 128,000 CPU cores on GCPを使用している
- https://gigazine.net/news/20180626-openai-five-dota-2-defeating/
- https://blog.openai.com/openai-five/
- https://www.youtube.com/watch?v=UZHTNBMAfAA

DOTA 2とは

個人的な補足

- 5人の2チームに分かれ敵の本拠点を破壊するゲーム

https://www.gamespark.jp/article/2014/03/14/47039.html

- プレイ時間は30〜90分程度で、平均45分

- Eスポーツ上最大の賞金額のゲーム

- Dota2 -The International 2016

  賞金総額　$ 20,770,640　(約23億)

  優勝賞金　$ 9,139,002　(約10億)

1.3　機械学習システムのタイプ

- scikit-learnとTensorFlowによる実践機械学習 1.3参照ください。

1.3　機械学習システムのタイプ

- 1.3.1　教師あり／教師なし学習
- 1.3.1.1　教師あり学習
- 1.3.1.2　教師なし学習
- 1.3.1.3　半教師あり学習
- 1.3.1.4　強化学習

- 1.3.2　バッチ学習とオンライン学習
- 1.3.2.1　バッチ学習
- 1.3.2.2　オンライン学習

- 1.3.3　インスタンスベース学習とモデルベース学習
- 1.3.3.1　インスタンスベース学習
- 1.3.3.2　モデルベース学習

1.4　機械学習が抱える難問

- 1.4.1 データとアルゴリズムのどちらが有効か

十分なデータを与えれば、非常に異なる機械学習アルゴリズム（ごく単純なものも含む）が、複雑な自然言語の曖昧性解消問題に対してほぼ同じ程度の性能を示すことを明らかにした（図1-20)

しかし、中小規模のデータセットは今でもごく一般的であり、訓練データの入手がいつも簡単でコストがかからないわけではないので、アルゴリズムを捨ててしまうにはまだ早いということに注意していただきたい。

1.4　機械学習が抱える難問

- 1.4.2 現実を代表しているとは言えない訓練データ

汎化の性能を上げるためには、訓練データが汎化の対象となる新データをよく代表するものになっていることがきわめて重要である。これはインスタンスベース学習でもモデルベース学習でも変わらない。

- サンプリングバイアスの例

scikit-learnとTensorFlowによる実践機械学習 1.4.2参照ください。

1.4　機械学習が抱える難問

- 1.4.3 品質の低いデータ

当然ながら、訓練データに誤り、外れ値、ノイズがたくさん含まれている場合（たとえば、測定品質の低さのために）、システムが背後に隠れているパターンを見つけるのは難しくなり、システムの性能が高くなる可能性は下がる。訓練データをクリーンアップするために時間を割くとよい場合が多い。実際、ほとんどのデータサイエンティストは、作業時間のかなりの部分をデータのクリーンアップのために使っている。

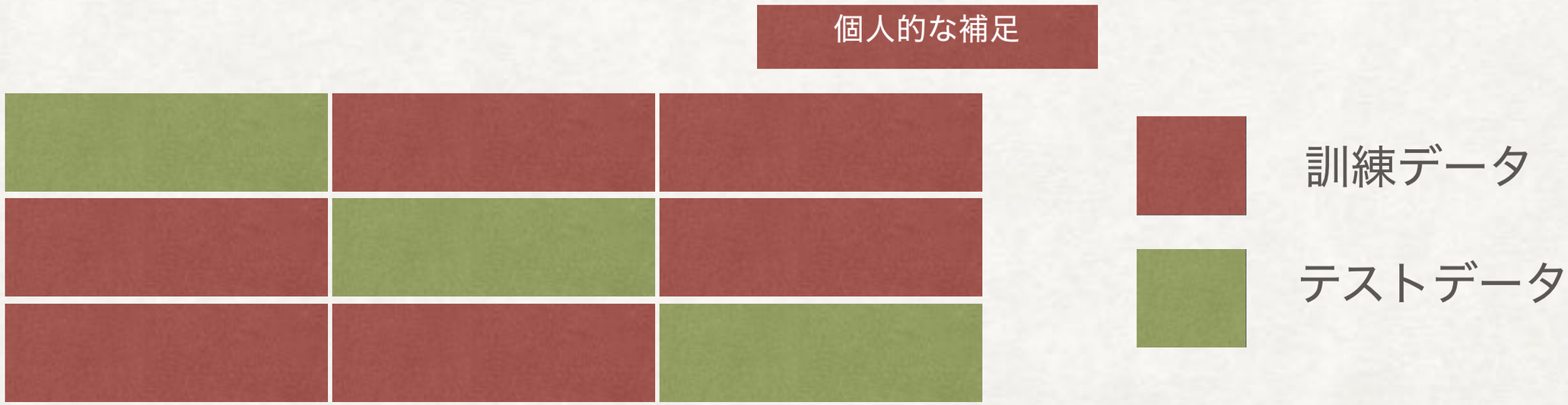
1.4　機械学習が抱える難問

- 1.4.4 無関係な特徴量
- 1.4.5 訓練データへの過学習
- 1.4.6 訓練データへの過小適合

1.5 テストと検証

- 交差検証（cross-validation）

検証セットのために訓練データを「無駄」にし過ぎないためによく使われているのは、交差検証（cross-validation）というテクニックである。訓練セットを複数のサブセットにきれいに分割し、サブセットの別々の組み合わせを使って各モデルを訓練し、残ったサブセットで検証する。モデルのタイプとハイパーパラメータを選択したら、訓練セット全体を対象とし、選択したハイパーパラメータを使って最終的なモデルを訓練する。そして、テストセットを使って汎化誤差を測定する。



1.5 テストと検証

個人的な補足

3分割して、結果を平均し、評価を出します。

<scikit-learnでのコード例>

```
results = []
names = []
for name,model in models:
    result = cross_val_score(model, train_data[predictors], train_data["Survived"],  cv=3)
    names.append(name)
    results.append(result)

for i in range(len(names)):
    print(names[i],results[i].mean())
```

1.5 テストと検証

個人的な補足

実行結果

LogisticRegression 0.785634118967
SVC 0.687991021324
LinearSVC 0.58810325477
KNeighbors 0.701459034792
DecisionTree 0.766554433221
RandomForest 0.796857463524
MLPClassifier 0.785634118967

ノーフリーランチ（タダ飯なし）定理

- デビッド・ウォルパートは、1996年の有名な論文（http://goo.gl/3zaHIZ）†11で、データに対して前提条件を何も設けなければ、あるモデルを別のモデルよりもよいと評価する理由はないことを実証した。これをノーフリーランチ（nofreelunch:NFL）定理と呼ぶ。あるデータセットで最良のモデルは線形モデルであり、別のデータセットではニューラルネットワークになる。アプリオリによりよい性能が得られるモデルはない（定理の名前はここから来ている）。どのモデルがもっともよいかをはっきりと知るためには、それらすべてを評価してみるしかない。しかし、そのようなことは不可能なので、現実には、データに対して何らかの合理的な前提条件を設け、合理的ないくつかのモデルだけを評価する。たとえば、単純なタスクではさまざまなレベルの正則化を加えた線形モデルを評価し、複雑な問題ではさまざまなニューラルネットワークを評価するのである。

コード

個人的な補足

- Google Colab上にコードを置いています。

下記共有しているので、セルごとに実行可能です。

https://colab.research.google.com/drive/1IWTBxhi2bUSbu0Z_aHeO8afbQqHwzMy0

すべてのセルをまとめて実行したい場合、コピーしてから実行して下さい。