

# k-最近傍法

正田 備也

[masada@rikkyo.ac.jp](mailto:masada@rikkyo.ac.jp)

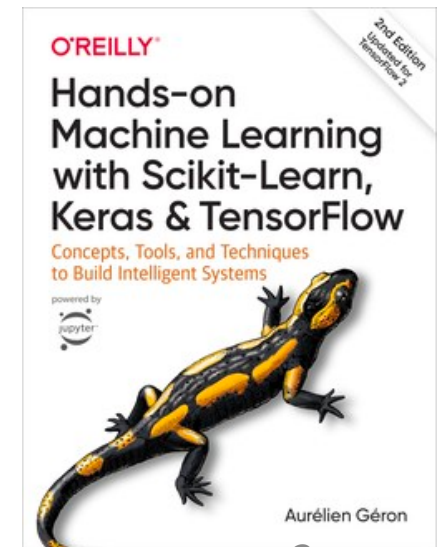
# 参考書

- 最初の回でも述べたように、この演習では独自の説は述べません
  - ですので、授業内容については安心してください。
- 機械学習関連の事項については、下記の本を参考書にして授業します
  - 買う必要はないです。

Aurélien Géron.

Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow, 2nd Edition.

<https://www.oreilly.com/library/view/hands-on-machine-learning/9781492032632/>



Spam



## 例題: スパムフィルタ (p.8)

- メールがスパムか否かを判定するシステムを作りたい
- 設定：
  - ある程度の数のメールを、すでに持っている。
  - 全てのメールに、スパムか否かのラベルが付けられている。
  - このラベルをうまく使って、新しく来たメールについて、スパムかどうかを判定したい。

## 素朴な手法 (p.17)

- 新しく来たメールと全く同じメールが、すでに持っているメールの中にないか、探す
- もしあれば、そのメールと同じラベルを答えとして出力する

## 演習4-1

- 前スライドの手法の問題点は何か

# instance-based vs model-based learning

- 機械学習にはinstance-basedな手法とmodel-basedな手法がある
- 先ほどの手法はinstance-basedな手法
  - すでに持っているメール = 実例(instance)をそのまま使うから。
- とはいえ、演習4-1で考えたとおり、問題がある

# 類似性に基づく instance-based method

- 同じメールが見つからなくても、似ているメールはあるだろう
- 新しく来たメールと似ているメールがあるなら、それと同じラベルを答えとして出力すれば良いのでは？
- 問：メールが似ているとは、どういうことか？

## 演習4-2

- 2通のメールが似ているか似ていないかを調べる手法を、考えてください（10分待ちます）
- 計算機に実行させることができる手法でないと、ダメです。
- スпамか否かの判定に役立つ手法でないと、ダメです。



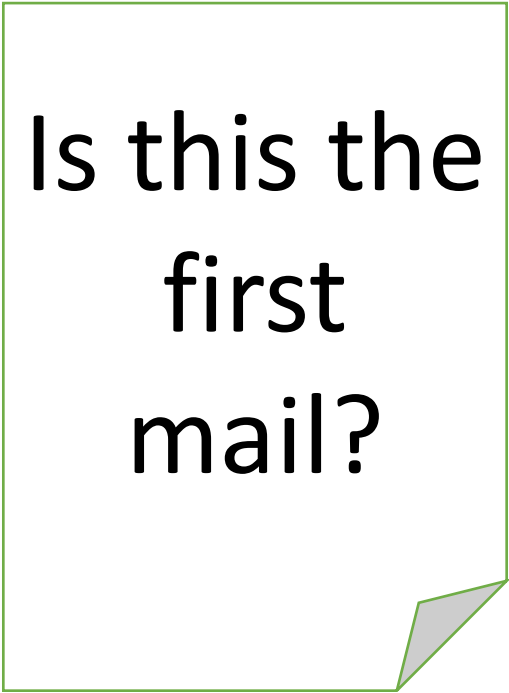


# measure of similarity

- 例えば、2つのメールに共通して出現する単語の数を数えて、それが多いほど似ている、とする  
(p.18)
- 他にどんな類似度の尺度が考えられるか？



This is  
the first  
mail.



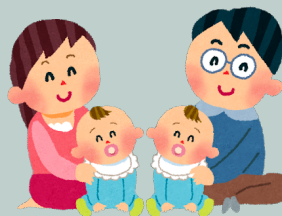
Is this the  
first  
mail?



This is  
the  
second  
mail.

データの上でなら近いペア

実物が近いペア



実物の類似度をうまく求めるには？

# k-最近傍法

## k-最近傍法 (k-Nearest Neighbors) (p.22)

- 新しく得られたinstanceについて、
- すでに正解が分かっているinstancesの中から、それと最も類似しているものをk個選び、
- それらk個の正解を利用して予測を実現する手法

# 予測問題の二種類

1. クラスの予測

2. 数値の予測

# クラスの予測 = 分類(classification) (p.8)

- 分類 (classification)

- 未知のinstanceを、複数のクラスのいずれかへグループ分け
- その際、グループ分けがすでに済んでいるデータを利用する
  - グループ分けが済んでいる = 正解が分かっている
  - 正解がすでに分かっているデータを「訓練データ(training data)」と呼ぶ
- 例：スパムフィルタ、手書き数字認識、など

# 数値の予測 = 回帰(regression) (p.8)

- 回帰(regression)
  - 未知のinstanceについて、関心がある数値(target value)を予測
  - その際、target valueがすでに分かっているinstancesを利用する
    - target valueが分かっている = 正解が分かっている
    - 正解がすでに分かっているデータを「訓練データ(training data)」と呼ぶ
  - 例：住宅価格の予測、CTRの予測、など



## k-最近傍法 (k-Nearest Neighbors) (p.22)

- 新しく得られたinstanceについて、
- すでに正解が分かっているinstancesの中から、それと最も類似しているものをk個選び、
- それらk個の正解を利用して予測を実現する手法

# 類似度をどう決めるか

- 演習4-2の問題を言い換えると・・・

「メールとメールの間に、  
どのような類似度を定義すれば、  
スパムフィルタのシステムに有用か？」

# 近傍 = 似ているもの

- **k-最近傍法**においては、**instance**間の類似度を、うまく決める必要がある
  - スпамフィルタ：二つのメールが似ている、とは？
    - 似ているメールは、クラスが同じになるように。
  - 住宅価格予測：二つの住宅が似ている、とは？
    - 似ている住宅は、価格が近くなるように。

# k-最近傍法のk

- k-最近傍法では、最も似ているk個を選ぶ
  - 分類：k個で多数決をとる
  - 回帰：k個のtarget valueの平均をとる
- 個数kは、手動で調整する必要あり
- 予測性能ができるだけ良くなるようにkを選ぶ

实践

## 例題: 一人当たりのGDPから生活満足度を予測する

- 参考書のp.19にある例
  - <https://github.com/ageron/handson-ml2/tree/master/datasets/lifesat>
- 一人当たりのGDPの出典
  - <http://goo.gl/j1MSKe>
- 生活満足度の出典
  - <http://stats.oecd.org/index.aspx?DataSetCode=BLI>

# 授業用のデータファイル

- `lifesat.csv`

- Blackboardの「教材/課題/テスト」→「プランナークラス」→「data」フォルダの下にある

4回目の授業の教材欄にも置いてある

- Pythonプログラムから読んで使う（pandasを使用）

1	Country	GDP per capi	Life satisfaction
2	Russia	9054.914	6
3	Turkey	9437.372	5.6
4	Hungary	12239.894	4.9
5	Poland	12495.334	5.8
6	Slovak Repub	15991.736	6.1
7	Estonia	17288.083	5.6
8	Greece	18064.288	4.8
9	Portugal	19121.592	5.1
10	Slovenia	20732.482	5.7
11	Spain	25864.721	6.5
12	Korea	27195.197	5.8
13	Italy	29866.581	6
14	Japan	32485.545	5.9
15	Israel	35343.336	7.4
16	New Zealand	37044.891	7.3
17	France	37675.006	6.5
18	Belgium	40106.632	6.9
19	Germany	40996.511	7
20	Finland	41973.988	7.4
21	Canada	43331.961	7.3
22	Netherlands	43603.115	7.3
23	Austria	43724.031	6.9
24	United Kingd	43770.688	6.8
25	Sweden	49866.266	7.2
26	Iceland	50854.583	7.5
27	Australia	50961.865	7.3
28	Ireland	51350.744	7
29	Denmark	52114.165	7.5
30	United State	55805.204	7.2



## 課題4

- 一人当たりの**GDP**から生活満足度を予測してみよう (p.22)
- 日本について、生活満足度を予測しよう
  - 他の国の生活満足度は、すべて分かっていると考えてよい。
- 予測の良し悪しは、実際の値との差の絶対値で評価しよう

# k-最近傍法による解き方

- 日本の生活満足度は未知だと仮定する
  - 真の値は分かっているが、いまはあくまで実験上の予測なので。
- 日本の一人当たり**GDP**と他の国々の一人当たりの**GDP**との距離を求める
  - 距離は差の絶対値でよい。
- 距離が近い順に、他の国々を並び替える
- 距離が最も近いkカ国の生活満足度の、平均を計算する
- その平均と、真の値とのズレを求める（これが小さいほど良い）

# 皆さんにさせていただくこと

- $k$ をどうやって決めればいいのかを考える
  - 日本の生活満足度は未知だと仮定しているので、 $k$ の値を選ぶときに使ってはいけません。
- **そのとき、 $k$ をいくらにすると最も良い予測が得られるかを、日本以外のデータで明らかにしてください。**
- 日本以外の一つの国について同じ実験をして、同じ $k$ の値が得られるか、調べてください（これは必須ではない）

次のステップ：  
多次元のデータを扱う

## 演習4-3

- 生活満足度を予測したいときに、今回の課題4のような設定を使うことに、どのような問題があるか

# 特徴量(features)

- それを使って予測を行うところの、各種のデータ
  - 例：住宅価格の予測をするときに使う、住宅の位置（経度・緯度）や部屋数、近隣地域の世帯年収の中央値、 etc
- 属性(attributes)と呼ばれることもあるが、正確には・・・
  - 身長 = 属性(attribute) ...身長はいろいろな値をとりうる
  - 170cm = 値(value) ...身長以外にも170cmとなる属性はいろいろある
  - 170cmの身長 = 特徴量(feature)

# 次のステップ：多次元データの分析へ

- たった一つの特徴量で、一人の人間、一つの企業、一つの国、等々を、表現すれば足る、なんてことはない
- 複数の特徴量の組によって、一人の人間、一つの企業、一つの国、などなどを表すのが普通
- だから、ベクトル（ $\equiv$ 順序のついた複数の実数値の組）を使うし、
- だから、線形代数（ $\equiv$ ベクトルとその変換に関する学問）を使う