

k-近傍法

正田 備也

masada@rikkyo.ac.jp

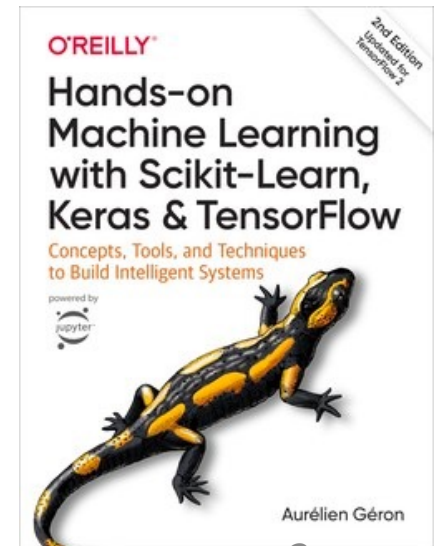
参考書

- 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。
 - 日本語訳あります。

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/>





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「通常」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、スパムか否かを判定したい。

素朴な手法 (p.17)

- 新しく来たメールと同じメールが、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

- 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（5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。

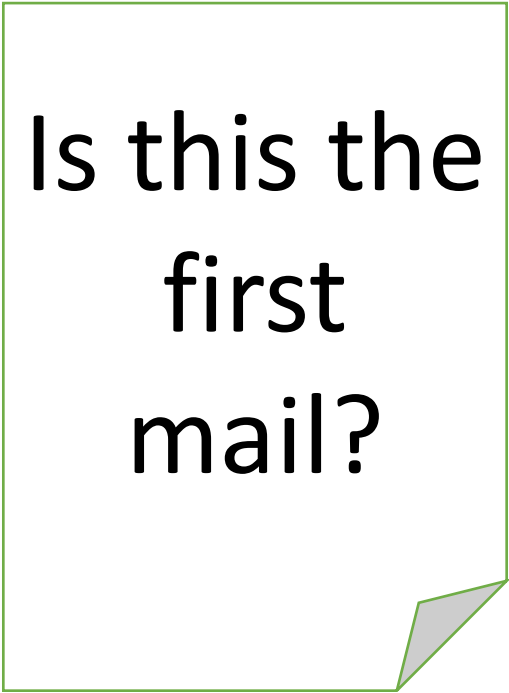


類似度の例

- 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.

k-近傍法

k-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-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-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-近傍法のk

- k-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: 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>
 - <https://worldhappiness.report/ed/2021/> (別バージョン)
 - <https://ourworldindata.org/grapher/gdp-vs-happiness>

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

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