データの加工と可視化

データの加工で扱う内容(1)

データに関する知識
- ベクトル、行列、テンソル
- Machine readableなデータ
- 量的データ
- 質的データ
- 比例尺度、間隔尺度、順序尺度、名義尺度
- 離散データ、連続データ

数値データの前処理
- 欠損値、異常値、外れ値
- それぞれの対処

文字データの前処理
- 重複、誤記、表記ゆれ
- それぞれの対処（名寄せ）

データの加工で扱う内容(2)

適切なグラフの使い方

- 棒、折れ線、円、帯、ヒストグラム、積み上げ棒、散布図
- データの読み込み
- グラフの作成

レコードデータの加工

- データ構造の加工
  - 抽出、集約、結合、分割、生成、展開
- データ内容の加工
  - 数値型、カテゴリ型、日時型、文字型

データの加工で扱う内容(3)

マルチメディアデータの加工

- 画像データ
  - 画像読み込みと表示
  - 画像データの構造
  - 輝度値の加工
  - 輝度値標準化・正規化
  - 輝度値の可視化（輝度値ヒストグラム）

見出データの加工で扱う内容(4)

実データによる演習

- 可視化によるデータの理解
- 教師無し学習のためのデータ加工
- 教師あり学習のためのデータ加工

スカラー、ベクトル、行列、テンソル

※数学的には厳密な定義があるが、あくまでプログラミングにおけるデータ構造として考える

スカラー：1つの数値

ベクトル：1次元配列として表現されるもの

行列：2次元配列として表現されるもの

テンソル：スカラー、ベクトル、行列の一般的な概念

- 配列の次元数はテンソルの階数と対応する
- 階数0：スカラー、階数1：ベクトル、階数2：行列
- 3次元配列として表されるものは階数3のテンソル

Machine readableなデータ

総務省のオープンデータの定義

（地方公共団体のオープンデータの推進、https://www.soumu.go.jp/menu_seisaku/ictseisaku/ictriyou/opendata/）

- 営利目的、非営利目的を問わず二次利用可能なルールが適用されたもの
- 機械判読に適したもの
- 無償で利用できるもの

goo国語辞典の記述

- machine readable（機械可読）：データやコンテンツなどがデジタル化されており、機械やコンピューターで直接読み取って利用できる形式であること。具体的には、一般的なアプリケーションソフトで利用可能なファイル形式や、タグなどのマークアップによって構造化されたデータのこと、また画像・音声・動画がデジタル化されていることなどを指す。機械可読。

Machine readableとHuman readable

Machine readable

- レコードデータ
  - csvやエクセルの表
- デジタル画像
  - 行列として扱うことが可能
- その他、フォーマットが厳密に定義されているデジタルデータ

Human readable

- 紙に書かれた数値やテキストなど
- アナログデータ
- pdf
- テキストが書かれた画像
- 構造化されていない数値やテキスト群
  - 奥村晴彦, "「ネ申Excel」問題",情報教育シンポジウム2013論文集, p.93-98
- デジタルデータでもMachine readableでないものもある

機械学習の入力となるデータ

機械学習では、同一の形のテンソルとなるデータを入力とする

- ベクトルであれば、使うデータは全て同じ要素数のベクトル
- 行列であれば、全て同じ行数と列数の行列

例えば↓のような形のデータは使えない（要素数の異なるベクトル）

データ1：2,3,1,5,4

データ2：3,5,4

データ3：2,5,3,2

量的データと質的データ

量的データ

- 個数や長さ、重さのような数量を表すもの
- 5個、10cm、2.5kg、15000円、100km/sなど

質的データ

- 性別や生物の種など、数量でない分類項目を表すもの
- 男/女、犬/猫/ネズミ、A/B/O/AB、大/中/小、など

4つの尺度

質的データ

名義尺度

- 単に分類するためにつけたラベル

順序尺度

- 順序には意味があるが、その間隔には意味がない数値を割り当てたもの

量的データ

間隔尺度

- メモリが等間隔であるもの、または等間隔と仮定されているもの

比例尺度

- 原点（0）の決め方が決まっており、間隔と比率両方に意味があるもの

情報量としては、名義＜順序＜間隔＜比例

- 大は小を兼ねることができる
- 順序尺度で意味があるものは間隔尺度以上でも意味が有る

名義尺度

プログラミングでは名義尺度のラベルとして整数を使うことが多い

- 2種類ならブーリアン型を使うこともある
- 0：猫、1：犬、2：ネズミ、3：コアラ、・・・など

区別するためだけに用いるため、等しいか等しくないかだけに意味がある

意味が有る統計量

- 度数、最頻値

意味が無い統計量

- 平均、分散、標準偏差など、ラベルの値を使った計算全般

順序尺度

こちらもプログラミングではラベルとして数値を使うことが多い

- 1：1位、2：2位、3：3位、・・・など

大小の比較は可能だが、間隔や比率などの「どのくらい」には意味が無い

- 1位と0.5ゲーム差の2位と、1位と10ゲーム差の3位でも、データとしてわかることは1位、2位、3位という順序だけ（ゲーム差は比例尺度）

意味が有る統計量

- 度数、最頻値、中央値

便宜的に順位などの値を
比例尺度のように扱うこともある

意味が無い統計量

- 平均、分散、標準偏差など、ラベルの値を使った計算全般

間隔尺度

知能指数や摂氏温度、アンケート調査など

プログラミングでは整数型、実数型の数値データとして扱われる

意味が有る統計量

- 順序尺度で意味が有るもの、値の和と差、平均、分散など

意味が無い統計量

- 比率
- 数値データなので比率を計算できるように見えるが、その値に意味は無い
- 摂氏30℃は摂氏10℃より20℃暑いが、3倍暑いとは言えない
- アンケート調査
  - 回答と値の割り当てによって比率は一定ではない、差は一定

便宜的に値を
比例尺度のように
扱うこともある

比例尺度

身長、体重、金額、絶対温度など

和差積商の計算がすべて意味のある値となる

全ての統計量が意味のあるものとして計算可能

数値データの前処理

欠損値

- 値が無い状態
- コンピュータプログラムが実行時エラーを起こす

異常値

- 主に計測ミスや記録ミスによって起こるありえない値
  - 正の値しかとらないのに負の値（データ範囲の異常）
  - 部屋の温度の記録で3000℃（現実的でない値）

外れ値

- 明らかな異常とは判定できないが、他の値から大きく外れた値
- 実用上は、外れ値と異常値を区別できない場合もあるので注意が必要

欠損値の対応

値の除去

- リストワイズ除去
- ペアワイズ除去

何らかの値で埋める

- 決められた値（平均値、中央値など）
  - 欠損値を意味する何らかの値を使うこともある（0など）
- 前の行の値など、近傍の値
- k-近傍法での予測値

値の補間

- 線形補間
- スプライン補間
- その他の曲線による補間

値の除去

A	B	C
1	0.3	3
3	0.1	4
4		8
5	0.4	3

欠損値のあるデータ

A	B	C
1	0.3	3
3	0.1	4
5	0.4	3

リストワイズ除去
欠損値のある行そのものを除去

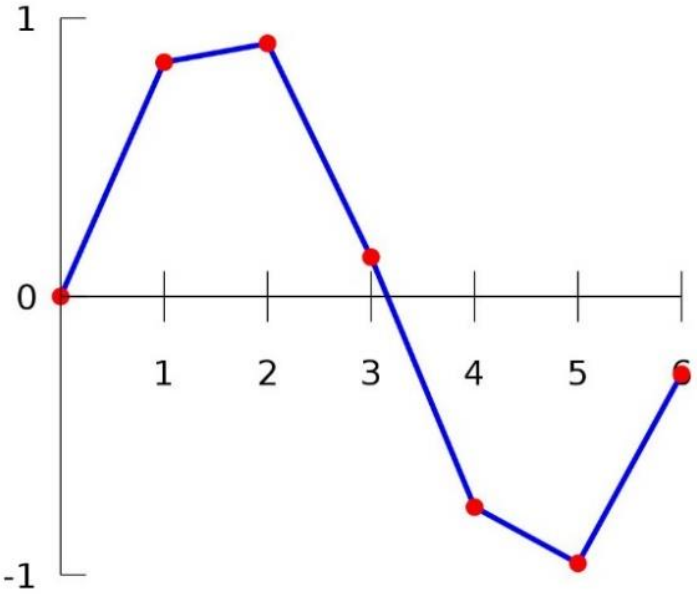
A	C
1	3
3	4
4	8
5	3

ペアワイズ除去
目的に必要な列だけ抜き出した後、リストワイズ除去

値の補間

線形補間

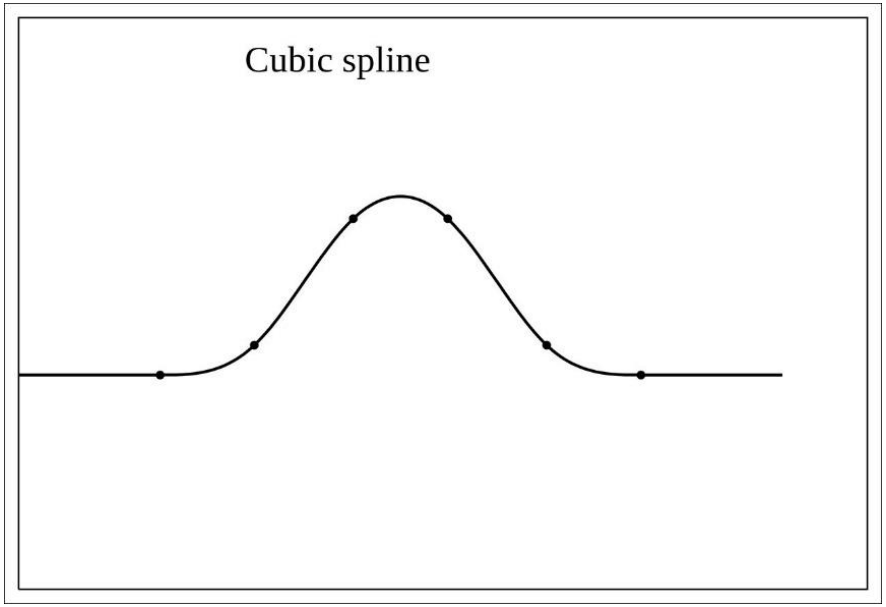
- 線形（1次）多項式を使った補間



https://ja.wikipedia.org/wiki/線形補間

スプライン補間

- スプライン曲線を使った補間



https://ja.wikipedia.org/wiki/線形補間
Stamcose - 投稿者自身による作品, CC 表示-継承 3.0,
https://commons.wikimedia.org/w/index.php?curid=16560057による

異常値の対応

値の修正が可能かどうかを検討する

- 数値の読み間違えなど

除去が可能かどうかを検討する

- 除去が可能であれば、対応は欠損値の対応に準ずる

異常値の対応には、対象と計測方法の知識が必要

外れ値の対応

外れ値があると、正規化や予測モデルの構築に悪影響を与える

しかし、外れ値を除去することは、極端な値の状況を考慮しないことになる

慎重に除去が可能かどうかを検討する必要がある

外れ値の見極め

- データを可視化して、目視で見つけた外れ値を恣意的に除去する
- 正規分布を前提にして、平均値と標準偏差を使って見つける
  - 平均値±定数×標準偏差
  - 定数は3以上を使うことが多い
  - 正規分布の場合、平均±3×標準偏差の範囲に約99.73%の値が収まる

文字データの前処理

重複、誤記、表記ゆれデータはまとめる

表記ゆれ

- 異字体（斎藤、斉藤、齋藤、齊藤など）
- コンピュータとコンピューターなど

その他

- アルファベットや数字の全角文字と半角文字
- 空白文字（スペースとタブ）、区切り文字（, と ; など）

データの不整合性に対する対処をデータクレンジングと呼ぶ

人手で行う必要がある作業も多い

pythonにおけるデータの読み込み

pandasによるcsvファイルの読み込み

- csv形式のレコードデータ
- 「09_1_データの読み込みと可視化.ipynb」を参照

読み込んだデータは、データフレーム型となる

- 形としてはいわゆる2次元配列
- 列ごとに異なるデータ型が可能
- numpyのarray形式は全て同じ型

グラフの種類

「09_1_データの読み込みと可視化.ipynb」を参照

棒グラフ： pyplot.bar

折れ線グラフ：pyplot.plot

ヒストグラム：pyplot.hist

散布図：pyplot.scatter

棒グラフ

pyplot.bar

主な引数

left (必須)	各棒の X 軸の数値
height (必須)	各棒の高さ
width	棒の太さ (デフォルト値: 0.8)
bottom	各棒の下側の余白。(積み上げ棒グラフを作るときに使用)
color	棒の色。
edgecolor	棒の枠線の色
linewidth	棒の枠線の太さ。
tick_label	X 軸のラベル

折れ線グラフ

pyplot.plot

主な引数

xdata (必須)	X 軸方向の数値
ydata (必須)	Y 軸方向の数値
linewidth	線の太さ
linestyle	線のスタイル。 solid（デフォルト）, dashed, dashdot, dotted
color	線の色
marker	マーカーの種類。 参考 https://matplotlib.org/api/markers_api.html#module-matplotlib.markers
markersize	マーカーの大きさ
markeredgewidth	マーカーの枠線の太さ
markeredgecolor	マーカーの枠線の色
markerfacecolor	マーカーの塗りつぶしの色

散布図

pyplot.scatter

主な引数

x, y	グラフに出力するデータ
s	サイズ (デフォルト: 20)
c	色、または、連続した色の値
marker	マーカーの形 (デフォルト: ‘o’= 円)
cmap	カラーマップ。c が float 型の場合のみ利用可能です。

ヒストグラム

pyplot.hist

主な引数

x (必須)	度数分布ではない生データの配列。
bins	階級数。(デフォルト: 10)
range	ビンの最小値と最大値を指定。(デフォルト: (x.min(), x.max()))
align	各棒の中心を X 軸目盛上のどの横位置で出力するか。 left, mid（デフォルト）, right
orientation	棒の方向。 horizontal, vertical(デフォルト)
rwidth	各棒の幅を数値または、配列で指定。
color	ヒストグラムの色。
label	凡例

pandasによる欠損値処理

「09_2_レコードデータの加工.ipynb」を参照

欠損値の発見

- isnull()

欠損値の除去

- dropna()

欠損値の置換

- fillna()
  - 直近の前の値
  - 直近の後の値
  - 特定の値

Pandasによる異常値、外れ値処理

「09_2_レコードデータの加工.ipynb」を参照

異常値、外れ値の発見

- ソートしてグラフ描画し、目視で確認
- 平均値と標準偏差による切り分け

異常値、外れ値の除去

- 欠損値処理に準ずる

レコードデータの加工

データ構造の加工

- 行や列に対して、抽出、集約、結合、分割を行うことで、入力データとして加工する
  - 機械学習の入力データは、テンソルの形
- 取得できたデータをそのまま入力データにできないこともある
  - 余分なデータが含まれる
  - 複数ファイルに収められている

データ内容の加工

- 型変換
- 比例尺度から名義尺度・順序尺度へ（数値データからカテゴリデータへ）
- カテゴリデータの集約（細かいカテゴリから大まかなカテゴリへ）

マルチメディアデータの加工

画像データ

- 画像データは、色チャンネル数を含めた3次元配列（3階テンソル）として扱われる
  - グレースケール：1チャンネル
  - カラー：3チャンネル
  - 透過画像：4チャンネル（カラー＋アルファ）

音声データ

- 音声データは、波形データ（信号の強さの時系列データ）として扱われる

画像データの加工

「09_3_画像データの加工.ipynb」を参照

OpenCVによる画像データ加工

- 画像データはnumpyのarray形式のデータとして扱われる
- numpyによる数値操作が可能
- OpenCV独自の処理も可能

Jupyter notebook上でmatplotlibによる画像表示をする場合は、色チャンネルの順番が逆になるので注意が必要

輝度値の加工

画素値は0～255の整数

OpenCVで画像を読み込んだ時点では、uint8（符号なし8ビット整数）型

輝度値を操作して実数にする場合は、型変換しないと桁落ちする

numpyの値操作

- 配列全体を対象にした演算
- 要素の直接演算
- スライス記法による対象範囲の指定
- 平均や標準偏差の計算

輝度値の正規化、標準化

画像の正規化（値を0～1の範囲にする）

- 画素値は0～255の範囲であるため、単純にそれぞれの画素値を255で割る

画像の標準化

- 画像全体の画素値の平均と標準偏差を計算し、それぞれの画素値から平均値を引いて標準偏差で割る

輝度値の分布の可視化

輝度値ヒストグラム

- 0～255の画素値の数（画素値の度数分布）をヒストグラムにする
- 正規化や標準化した画像の場合でも、最小値～最大値の範囲でbin（階級）数を256としたヒストグラムを作ると分布は変わらない
- 隣接画素の情報は持っていないが、画素値の分布は画像特徴量の一つ
- カラー画像の場合は、3つのチャンネル全体でヒストグラム化する場合と、3つのヒストグラムを作成する場合がある。
- matplotlibのhistでヒストグラムを作成する場合、1つのベクトル（1次元配列）が1つのデータ群であるため、2次元配列の場合は平坦化（1次元配列化）する必要がある。

実データによる演習

Kaggleのデータを使った演習

可視化によるデータの理解

- 必要に応じてMachine readableに加工し、グラフ化してデータの傾向を観察する

教師無し学習のためのデータ加工

- Machine readableなデータに対して、欠損値処理、異常値処理を行い、機械学習の入力となるテンソルの形に加工する

教師あり学習のためのデータ加工**（今回はこちらを演習）**

- テンソルの形のデータと教師データを作成し、学習用データと評価用データに分割する

Kaggleのデータを使った例（1）

Craft Beers Dataset
2K+ craft canned beers from the US and 500+ breweries in the United States.

https://www.kaggle.com/nickhould/craft-cans

「09_4_データ加工の例_beers.ipynb」を参照

styleの分類がこのデータから可能かを調べるための予備分析のようなことを行う

データの内容と型

Unnamed: 0	int64	上から順のID
abv	float64	アルコール度数 (Alcohol By Volume)
ibu	float64	苦みの強さ (International Bitterness Unit
id	int64	ID（nameに対応）
name	object	銘柄
style	object	ビールの種類（ラガー、エールなど）
brewery_id	int64	醸造所ID
ounces	float64	一缶の量（オンス）

Craft Beers Datasetの加工

- データ種類のチェック
  - 数値データとして意味を持つものは3つ
  - abv, ibu, ounces
- 欠損値のチェック
  - 欠損値があるので欠損値処理を行う
- 外れ値のチェック
  - 今回は外れ値無しとみなす
- 分類目的のstyleの様子を観察
  - styleの数、style毎の銘柄数
  - ごく少数の銘柄しかないstyleもある
  - 今回は仮に10以上の銘柄があるstyleを対象とする
- 相関を見るために散布図を描く
- 主成分分析で次元圧縮し、主成分で散布図を描く
  - 特徴量としての主成分得点

- 教師なし学習＝入力と同じデータを教師とする学習
  - 教師あり学習データの入力のみ

Kaggleのデータを使った演習（2）

Solar Radiation Prediction. Task from NASA Hackathon.

- https://www.kaggle.com/dronio/SolarEnergy

「09_5_データ加工の例_solar.ipynb」を参照

データの内容と型

UNIXTime	int64	Unix時間（1970年1月1日午前0時0分0秒からの経過時間）
Data	object	日付（Dateのスペルミス？）
Time	object	時刻
Radiation	float64	日射量
Temperature	int64	気温（華氏）
Pressure	float64	気圧（Hg）
Humidity	int64	湿度（%）
WindDirection(Degrees)	float64	風向（角度）
Speed	float64	風速（miles / h）
TimeSunRise	object	日の出時刻
TimeSunSet	object	日没時刻

Radiationの値を
他のデータから予測する

Solar Radiation Predictionの加工

- 欠損値の確認
- 折れ線グラフでのデータの確認
- 散布図でのデータの確認
- データ加工（派生データ作成、昼間の時間）
  - 日付データへの変換
  - 時間データへの変換
  - 数値型への変換
- 相関行列のヒートマップ
  - radiationの予測に影響しそうなデータを目視で確認

教師あり学習のためのデータ作成

「09_6_教師あり学習データ作成.ipynb」を参照

- 不使用のデータ列を削除
- 学習用データと評価用データに分割

「09_7_教師あり学習演習.ipynb」を参照

- DNNで教師あり学習を行い、回帰モデルを作成する
  - Radiationを予測する
  - どのようなデータを入力とするか、標準化または正規化の有無、DNNの構造、バッチサイズ、その他さまざまな要素を考慮して精度を上げてみてください。