ソフトウェア演習Ⅲ〔課題 3:回帰分析クラス〕青野雅樹

この課題は 2Q の Java 言語（ベーシッククラス）で行ったものである。ターゲットの回帰の説明変数や未知データは 2Q とは違うので注意のこと。ここでは、回帰分析を行うクラス等を Python 言語で kadai3.py として作成してもらう。kadai3.py のプログラムと実行結果（kadai3.txt）をつけて ZIP にまとめ Moodle にアップロードせよ。締め切りは 10 月 27 日(火)までとする。

100ｇあたりのいろいろな食物のカロリーや炭水化物含有量などのデータ（https://www.kde.cs.tut.ac.jp/~aono/data/food.csv）(UTF-8)がある。引用元は https://www.eiyoukeisan.com/calorie/nut_list/index_nut.html である。

① このデータを読み込み、Food クラスを作成せよ。

② カロリー（calorie）を GI 値の単回帰（Regression クラス）で表現せよ。

③ カロリーを脂質(fat)の単回帰で表現せよ。

④ ２つの回帰モデルのどちらが「良い」回帰であるかを評価せよ。回帰の結果（よさ）は、寄与率で比較すること。結果は（以下で述べる a,b ならびに）寄与率 R2 を書き出して示すこと。回帰の良さ（寄与率の良さ）は、②③＋⑤を実行後、実行結果ファイル(kadai3.txt)に手で書き足すこと。

⑤ 寄与率がよかった説明変数で、未知データとして以下のカロリー値を予測せよ。負のカロリー値もありえるので注意。

[A] さくらんぼ（炭水化物=15.2, たんぱく質=1.0, GI=37, 脂質=0.2）

[B] バジル（炭水化物=4.0, たんぱく質=2.0, GI=5, 脂質=0.6）

[C] 豆乳（炭水化物=3.1, たんぱく質=3.6,GI=23, 脂質=2.0）

【説明変数は使うものだけ使用してよいし、すべてを保持してもよい】

【コメントとヒント】

多変量データに対する線形回帰（単回帰、重回帰）は、データマイニングの基礎技術のひとつであり、適応範囲が広く有名な技術です。単回帰モデルは、**目的変数**を y として、1 個の**説明変数** x を用いて n 個のサンプルから以下の式を推定することが目的です。

$$\mathbf{y} = a\mathbf{x} + b + \boldsymbol{\varepsilon}$$

ここで、$\boldsymbol{\varepsilon}$ は誤差を表し、a と b は係数（a を回帰係数、b を回帰切片と呼ぶ）を意味し、これらを推測することが単回帰の主たる問題となります。今回のデータは 49 個の「食べ物」データがあるので、$n = 49$ です。サンプルで式を書き直すと

$$y_i = a\,x_i + b + \varepsilon_i$$

となり、誤差の 2 乗和から、最小二乗法で a と b を推定します。最小二乗法の詳細は省略しますが、a と b の推定値（$\hat{a}$ と $\hat{b}$）は、以下の S_{xx}（x のサンプル平方和）、S_{yy}（y のサン

プル平方和)、S_{xy}（x と y のサンプル偏差積和）を用いて以下のように表現されます。

$$\hat{a} = S_{xy} / S_{xx} 、 \hat{b} = \bar{y} - \hat{a}\bar{x} 、 \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i 、 \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

$$S_{xx} = \sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2 、 S_{yy} = \sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2 、 S_{xy} = \sum_{i=1}^{n}\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)$$

変数の頭に hat(∧)がついているものは予測値です。また bar(￣)は平均値を表します。回帰の「良さ」は、いろいろな基準がありますが、以下の R²（寄与率が１つの基準として使われ、この値が 1.0 に近いほど、よい回帰であるとされます。なお、$\hat{a}$ と $\hat{b}$ は、それぞれサンプルデータから推定された回帰係数と切片です。

$$R^2 = \frac{\left(\sum_{i=1}^{n}\left(y_i - \bar{y}\right)\left(\widehat{y_i} - \bar{\widehat{y}}\right)\right)^2}{\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2 \sum_{i=1}^{n}\left(\widehat{y_i} - \bar{\widehat{y}}\right)^2}$$

ただし、$\widehat{y_i} = \hat{b} + \hat{a}x_i$ です。$\bar{\widehat{y}}$ は $\widehat{y_i}$ の平均値です。

Regression(回帰)クラスでは、少なくとも以下の値をクラスに保持してください。これら以外のメンバー変数やメンバー関数は自由です。

食べ物クラス（クラス名=Food）

メンバー変数: 以下を少なくとも含むこと（CSV にあるすべての値を保持してもよい）

メンバー変数名	型	概要
name	文字列	食べ物の名前
GI	実数値	GI 値(Glycemic Index)
fat	実数値	脂質（含有量）
calorie	実数値	カロリー

コンストラクタ：(CSV にあるすべての値を含むものでもよい)

引数の数	引数の型	概要
4	(name,GI,fat,calorie)	すべてのメンバー変数をセット

回帰クラス（クラス名=Regression）

メンバー変数:

変数名	型	概要
a	実数値	係数
b	実数値	係数
R2	実数値	寄与率

xmean	実数値	説明変数の平均値（計算用）
ymean	実数値	目的変数の平均値（計算用）
samples	整数値	データのサンプル数
data	リスト	観測データ（説明変数）
labels	リスト	目的変数（本課題ではカロリー）
predicted	リスト	目的変数の予測値(サンプル数個)

コンストラクタ：

引数の数	引数	概要
2	(data, labels)	２つの引数をメンバー変数に代入。同時に samples をセット。他の実数値のメンバー変数は 0.0 で初期化。

メソッド（関数）：

メソッド名	引数	戻り値型	概要
compMean	なし	なし	data と labels から xmean と ymean を計算
doRegression	なし	なし	単回帰を計算し predicted, a, b, R2 をセットする
predict	x	実数	doRegression のあとに呼び出す関数で、未知な説明変数データ(x)を与えて目的変数の値（ここではカロリー値）を予測し返す

[1] 実行時に$ python kadai3.py food.csv X で実行。ただし、X は一文字で G なら GI 値、F なら脂質(fat)を説明変数とする。未知データの値（さくらんぼ、バジル、豆乳）は、プログラムに埋め込んでよい（doRegression を呼び出した後で、predict を呼び出し、結果をプリントすること）。

[2] 実行結果は、たとえば kadai3-1.txt (G の場合)と kadai3-2.txt(F の場合)、のように実行後、cat コマンド等で２つをまとめ、最後に手で、寄与率はどちらがよかったかを書き添え、２つまとめて、kadai3.txt というファイル名とすること。なお、どちらが GI 値での実行か、脂質の実行かもわかるように、すること。

CSVファイルの読込みは、第3回の資料参照

【実行例】

以下は、炭水化物で単回帰を実行した例です。実際の課題では、説明変数はGI値と脂質で2回実行し、寄与率をプリントし、(末尾に)どちらがよい回帰モデルであるか書き足してください。未知データの予測は、いずれの場合も必要です。負の予測値もありえますので驚かないように。

```
$ python kadai3.py food.csv C
******************************************
課題3：青野雅樹, 01162069
日付:2020年9月3日15時10分13秒
内容：カロリーを炭水化物で単回帰した場合
未知データは、「さくらんぼ」「バジル」「豆乳」
******************************************
炭水化物で単回帰します
a（回帰係数） = 4.33551
b（回帰切片） = 61.41730
R2（寄与率） = 0.40979

未知データの予測
さくらんぼ のカロリー予測は 127.31709です
バジル のカロリー予測は 78.75935です
豆乳 のカロリー予測は 74.85739です
```