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）の課題で行ったものと基本的に同じで、それを Python で実装し、違うデータで試してみよう、というものである。ここでは、回帰分析（単回帰）を行うクラスとデータを表現する RealEstate クラス（不動産クラス）を含む kadai3.py のプログラムを作成し、実行結果（kadai3.txt）をあわせ ZIP にまとめ Moodle にアップロードせよ。締め切りは 11 月 1 日(火)までとする。

築年数、最寄り駅からの距離、ならびに単位面積あたりの価格を含む物件データ（訓練データ(RS_train.csv)とテストデータ(RS_test.csv)をサーバ上においてある。https://www.kde.cs.tut.ac.jp/~aono/data/RS_train.csv が訓練データ、https://www.kde.cs.tut.ac.jp/~aono/data/RS_test.csv がテストデータで、いずれも同じフォーマットで（訓練が 400 件、テストは 10 件）用意している。引用元は https://archive.ics.uci.edu/ml/datasets/Real+estate+valuation+data+set である。いずれも、UTF-８符号にしてある。実数のプリント時は、小数点以下２桁程度とする。

① これらのデータを読み込め。その際、個々のデータを保持する RealEstate クラス（不動産クラス）を作成せよ。

② 単位面積価格（value）を目的変数とし、築年数(howOld)を説明変数とする単回帰を行え。その際、あとで詳細を述べる Regression クラス（単回帰クラス）を作成せよ。回帰実行後、寄与率をプリントせよ。

③ 目的変数を変えないで、説明変数を最寄駅距離(howFar)を説明変数とする単回帰を行え。回帰実行後、寄与率をプリントせよ。

④ テストデータ(10 件)から、物件番号の末尾と学籍番号の末尾が一致する１つのデータに対して、単位面積価格を予測せよ。この予測値と真値を比較し、誤差の絶対値をプリントせよ。④は、②と③のそれぞれの実行内の末尾で行うものとする。

【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 400 件の「不動産」データがあるので、$n = 400$ 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、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。

不動産クラス（クラス名=RealEstate）

メンバー変数:

メンバー変数名	型	概要
id	整数値	物件番号
howOld	実数値	築年数
howFar	実数値	最寄り駅からの距離
value	実数値	単位面積あたりの価格

コンストラクタ（イニシャライザ）:

引数の数	引数の型	概要
4	(id,howOld,howFar,value)	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

xm	実数値	説明変数の平均値（計算用）
ym	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
xlist	リスト	説明変数データ
ylist	リスト	目的変数データ

コンストラクタ(イニシャライザ)：

引数の数	引数	概要
２	(xlist, ylist)	２つの引数をアトリビュート（メンバ変数）に代入。同時に samples をセット。ほかのアトリビュートの初期化

メソッド（関数）：

メソッド名	引数	戻り値型	概要
compMean	なし	なし	xlist と ylist から xm と ym 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、テストデータにある未知な説明変数データ(x)を与えて目的変数の値(y)を予測し返す

実行時に$ python kadai3.py RS_train.csv RS_test.csv S >> kadai3.txt のように実行。ただし、S は一文字で O なら築年数(howOld)、F なら最寄駅距離(howFar)を説明変数とする。学籍番号の末尾とテストデータ内の物件番号の末尾を比較する際、自分の学生番号あるいは、その末尾はプログラムに埋め込んでよい。

【実行例】

以下は、411番～414番の物件（注：実際のUCI archiveには全体で414件のデータがあります）に対して、２種類の単回帰を実行した例です。（回帰の際のa,b,R2などの値は伏せてあります）。

```
$ python kadai3.py RS_train.csv RS_test2.csv O
************************************************************
課題3: RS_train.csvから単回帰クラスを作成し学習。
その後、RS_test.csvからテストデータを読み、単位面積価格を予測
青野雅樹, 01162069
日付: 2022-10-22 22:01:34.344975
************************************************************
```

築年数で予測します
a（回帰係数） = xxxx
b (回帰切片) = xxxx
R2 (寄与率) = xxxx

物件番号411の単位面積価格の予測は41.11です
物件番号411の単位面積価格の真値は50.00です
予測値と真値の絶対誤差は8.89です

物件番号412の単位面積価格の予測は37.76です
物件番号412の単位面積価格の真値は40.60です
予測値と真値の絶対誤差は2.84です

物件番号413の単位面積価格の予測は40.47です
物件番号413の単位面積価格の真値は52.50です
予測値と真値の絶対誤差は12.03です

物件番号414の単位面積価格の予測は40.88です
物件番号414の単位面積価格の真値は63.90です
予測値と真値の絶対誤差は23.02です

```
$ python kadai3.py RS_train.csv RS_test2.csv F
************************************************************
課題3: RS_train.csvから単回帰クラスを作成し学習。
その後、RS_test.csvからテストデータを読み、単位面積価格を予測
青野雅樹, 01162069
日付: 2022-10-22 22:10:43.602464
************************************************************
```

最寄駅距離で単回帰します
a（回帰係数） = xxxx
b (回帰切片) = xxxx
R2 (寄与率) = xxxx

テストデータでの予測と絶対誤差
物件番号411の単位面積価格の予測は45.24です
物件番号411の単位面積価格の真値は50.00です
予測値と真値の絶対誤差は4.76です

物件番号412の単位面積価格の予測は43.06です
物件番号412の単位面積価格の真値は40.60です
予測値と真値の絶対誤差は2.46です

物件番号413の単位面積価格の予測は45.13です
物件番号413の単位面積価格の真値は52.50です
予測値と真値の絶対誤差は7.37です

物件番号414の単位面積価格の予測は45.24です
物件番号414の単位面積価格の真値は63.90です
予測値と真値の絶対誤差は18.66です