ソフトウェア演習Ⅱ〔課題5:クラスの継承（応用編)〕青野雅樹

この課題では、クラスの継承を、伝統的な機械学習の一種である回帰分析とあわせて学習することを主旨とする。以下の問題に対する Java 言語でのプログラム（Kadai5.java, Regression.java, Food.java 等）を作成し，プログラムと実行結果（Kadai5.txt）を ZIP にまとめ Moodle にアップロードせよ。**〆切は 7 月 27 日(火)**までとする。

100 g 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）がある。

① このデータを読み込み、Food クラスを作成せよ。クラスのメンバー変数としては、以下の回帰で用いるカロリー、脂質、タンパク質は少なくとも持たせること。

② 後述する変数や関数を有する抽象クラス(Regression クラス)を作成せよ。

③ Regression クラスを継承するクラスとして FoodRegression クラスを作成せよ。（詳細は後述参照）

④ Kadai5.java クラスを作成せよ。ここには全体の main 関数を保持し、カロリー（calorie）を目的変数とし、脂質(fat)を説明変数とする単回帰、ならびにカロリーを目的変数とし、タンパク質(protein)の説明変数とする単回帰の 2 種類（の係数 a,b）の FoodRegression クラスを作成せよ。

⑤ ④で述べた回帰モデルの評価を行え。回帰の評価は、寄与率で比較すること。すなわち、回帰の実行末尾に、寄与率 R2（注：R2 は R の 2 乗の意味）を書き出せ。

⑥ 未知データとして以下のカロリー値を予測せよ。

[A] 落花生（炭水化物=19.6, タンパク質=26.5, GI=28, 脂質=49.4）

[B] 絹豆腐（炭水化物=2, タンパク質=4, GI=42, 脂質=3）

[C] しいたけ（炭水化物=4.9, タンパク質=3,GI=28, 脂質=0.4）

【{A},{B},[C]で与えている変数のうち、回帰で使うものだけ使用してよい】

【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、データサイエンス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目的です。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。aとbの推定値（$\hat{a}$と$\hat{b}$）が求まると、目的変数の推定値（$\widehat{y_i}$ ）は

$$\widehat{y_i} = \hat{a}x_i + \hat{b}$$

で近似できます。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、$\bar{\widehat{y}} = \frac{1}{n}\sum_{i=1}^{n}\widehat{y_i}$ である。Java の 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(たとえば Food の GI 値など)やメンバー関数は自由です。

① 食べ物クラス（クラス名=Food, ファイル名=Food.java）

メンバー変数: アクセス修飾子（制御子）は **private** (注 CSV の出現順とは異なる)

メンバー変数名	型	概要
name	String	食べ物の名前
fat	double	脂質（含有量）
protein	double	タンパク質（含有量）
calorie	double	カロリー

コンストラクタ：以下の２つを用意すること。アクセス修飾子は **public**

引数の数	引数の型	概要
なし	—	何もしない
4	(String,double,double,double)	すべてのメンバー変数をセット

メソッド（関数）：アクセス修飾子はすべて **public**

メソッド名	引数型	戻り値型	概要
getName	なし	String	name を返す

getCarbon	なし	double	carbon を返す
getProtein	なし	double	protein を返す
getCalorie	なし	double	calorie を返す

② 抽象回帰クラス（クラス名=Regression, ファイル名=Regression.java）を以下の条件で作成せよ。なお、Regression クラスは抽象クラスとする。

メンバー変数: アクセス修飾子（制御子）は **protected**

変数名	型	概要
a	double	係数
b	double	係数
R2	double	寄与率
xmean	double	説明変数の平均値（計算用）
ymean	double	目的変数の平均値（計算用）
samples	int	データのサンプル数
variables	自由	説明変数（計算対象説明変数のみ）
values	自由	目的変数（本課題ではカロリー）
predicted	自由	目的変数の予測値(サンプル数個)

コンストラクタ：以下を用意すること。アクセス修飾子は **public**

引数の数	引数	概要
2	(variables, values)	型は上の変数の定義参照。2 つの引数をメンバー変数に代入。同時に samples をセット。他は 0.0 で初期化。

メソッド（関数）：アクセス修飾子:get メソッドが **public** その他は **public abstract**

メソッド名	引数	戻り値型	概要
compMean	なし	自由	variables と values から xmean と ymean を計算
doRegression	なし	自由	単回帰を計算し predicted, a, b, R2 をセットする
computeR2	なし	double	R2 を返す
getA	なし	double	a を返す
getB	なし	double	b を返す

③ Regression クラスの継承クラス（クラス名=FoodRegression, ファイル名=FoodRegression.java）

メンバー変数：なし

コンストラクタ：

アクセス修飾子は **public**

引数の数	引数	概要
2	(variables, values)	基底クラスのコンストラクタを呼び出す。

メソッド（関数）：ここで Food のデータ構造に応じた、以下の２つの関数を実装する。アクセス修飾子:**public**

メソッド名	引数	戻り値型	概要
compMean	なし	自由	variables と values から xmean と ymean を計算
doRegression	なし	自由	単回帰を計算し predicted, a, b, R2 をセットする
computeR2	なし	double	R2 を返す

④ 全体の制御（クラス名=Kadai5, ファイル名=Kadai5.java）ここの main 関数で全体の制御を行う

[1] 実行時は $ java(u) Kadai5 food.csv X >> Kadai5.txt のようなコマンドラインで実行することとする。ただし、X は一文字で F なら fat（脂質)、P ならタンパク質(protein)を説明変数とする。

[2] 未知データ（落花生、絹豆腐、しいたけ）は、プログラム内に埋め込んでよい。

[3] 実行結果は、2 回の実行（F の場合)と(P の場合)をまとめて、Kadai5.txt というファイル名として、どちらで実行したかがわかるように出力させること。

[4] 2 種類の異なる回帰モデルの実行の末尾（Kadai5.txt の末尾）に、適当なエディタで（手動でいいので)、どちらの寄与率がより、1.0 に近かったかを 1 行程度、書き込み比較結果を述べてください。

【実行例（単独の実行例)】

以下は、GI値で単回帰を実行した例です。実際の課題では、説明変数は脂質とタンパク質で、それぞれ実行したものを、結合してください。（グラフのプロットは不要ですが、余裕がある人は、オプションで添付しても結構です。）

```
$ javau Kadai5 food.csv G
**************************************
課題5：青野雅樹, 01162069
日付:2021年7月15日18時10分25秒
内容：カロリーをGI値で単回帰した場合
**************************************
a（回帰係数） = 3.1167
b（回帰切片） = -18.3106
R2（寄与率） = 0.344
「落花生」のカロリー予測 = 68.9573
「絹豆腐」のカロリー予測 = 112.5912
「しいたけ」のカロリー予測 = 68.9573
```