効果測定について

機械学習入門

実装機能の概要

「分類」フォルダーに２つ「回帰」フォルダーに３つのCSVファイルがあります。
「分類」フォルダーから1つ、「回帰」フォルダーから１つ、合計2つのCSVファイルを選択して、それぞれモデルを作成し、正解率：分類、決定係数：回帰がなるべく高くなるように、Colabにて実装してください。

◆ 提出ノートブックファイル名は
分類モデル：kouka_分類_学生番号_氏名.ipynb
回帰モデル：kouka_回帰_学生番号_氏名.ipynb

◆ 各ノートブックは各セルでどのような処理がなされているのかの説明（他の人が理解できるように）をテキストセルまたはPythonのコメントで記述してください。

実装機能要件

- 1．CSVファイルの読み込みと表示。
- 2．カテゴリ特徴量列があるときはダミー変数化
- 3．テストデータの分離がなされている
- 4．正しく前処理が実装されている
  - 4－1．欠損値処理の過程が実装されている
  - 4－2．以下を各手法を必要に応じて実装
    - 4－2－1．外れ値の処理が過程が理解できるように実装されている（外れ値があるかないかを判断する情報を提供すること）
    - 4－2－2．特徴量の絞り込みが理解できるように実装されている（特徴量の絞り込みの根拠を示すこと）
    - 4－2－3．標準化が実装されている
    - 4－2－4．多項式特徴量、交互作用特徴量が必要に応じて追加実装されている
    - 4－2－5．特徴量と正解の分割が実装されている
- 5．学習モデルの選択とそのモデルがなぜ選択されたかが理解できるように実装されている
- 6．選択された学習モデルで学習
- 7．正解率または決定係数の表示（未知のデータを使用して予測、判断させればさらに良い）
- 8．学習後のモデルを保存

回帰 **auto-mpg.csv**

列名	説明
mpg	燃料1ガロンあたりにつき何マイル走るかの燃費
cylinders	エンジンのシリンダー数
displacement	排気量
horsepower	馬力
weight	車重
acceleration	加速度
model year	発表年
origin	(origin == 1)はアメ車、(origin == 2)は欧州車、(origin == 3)は日本車

mpg を正解データとして、燃費を予測します。

回帰 diabetes.csv

列名	説明
age	年齢
sex	性別
bmi	BMI値
bp	平均血圧
s1	総血清コレステロール
s2	低密度リポタンパク質
s3	高密度リポタンパク質
s4	総コレステロール/ HDL
s5	血清トリグリセリドレベルの対数
s6	血糖値
target	進行状況

target を正解データとして、糖尿病患者の進行状況を予測します。

分類	digits.csv

列名	説明
pixel_0_0	１画素目のグレースケール（０：白、１：黒）以下同様
pixel_0_1	２画素目のグレースケール
pixel_0_2	３画素目のグレースケール
pixel_0_3	４画素目のグレースケール
pixel_0_4	５画素目のグレースケール
pixel_0_5	６画素目のグレースケール
pixel_0_6	７画素目のグレースケール
pixel_0_7	８画素目のグレースケース
：	：
pixel_7_6	６３画素目のグレースケール
pixel_7_7	６４画素目のグレースケール
target	正解データ（0, 1, 2, 3, 4, 5, 6, 7, 8, 9）

target を正解データとして、画像の数値を判断します。

分類 breast_cancer.csv

列名	説明	
mean radius	半径	平均
mean texture	テクスチャ―のグレースケールの標準偏差	平均
mean perimeter	外周長	平均
mean area	面積	平均
mean smoothness	中心から外周までの部分偏差	平均
mean compactness	コンパクト性（外周長 2÷ 面積－1.0）	平均
mean concavity	コンターの凹部強度	平均
mean concave points	コンターの凹点の数	平均
mean symmetry	対称性	平均
mean fractal dimension	フラクタル次元	平均

列名	説明	
radius error	半径	標準偏差
texture error	テクスチャ―のグレースケールの標準偏差	標準偏差
perimeter error	外周長	標準偏差
area error	面積	標準偏差
smoothness error	中心から外周までの部分偏差	標準偏差
compactness error	コンパクト性（外周長 2÷ 面積－1.0）	標準偏差
concavity error	コンターの凹部強度	標準偏差
concave points error	コンターの凹点の数	標準偏差
symmetry error	対称性	標準偏差
fractal dimension error	フラクタル次元	標準偏差

列名	説明	
worst radius	半径	最悪値
worst texture	テクスチャ―のグレースケールの標準偏差	最悪値
worst perimeter	外周長	最悪値
worst area	面積	最悪値
worst smoothness	中心から外周までの部分偏差	最悪値
worst compactness	コンパクト性（外周長 2÷ 面積－1.0）	最悪値
worst concavity	コンターの凹部強度	最悪値
worst concave points	コンターの凹点の数	最悪値
worst symmetry	対称性	最悪値
worst fractal dimension	フラクタル次元	最悪値

特徴量から、その人の腫瘍が良性か悪性を判断します。
target ＝ 0　　悪性
target ＝ 1　　良性

分類 mushrooms.csv

列名	説明
class	毒か食用か
cap-shape	カサの形 bell=b,conical=c,convex=x,flat=f, knobbed=k,sunken=s
cap-surface	カサの表面 fibrous=f,grooves=g,scaly=y,smooth=s
cap-color	カサの色 brown=n,buff=b,cinnamon=c,gray=g,green=r, pink=p,purple=u,red=e,white=w,yellow=y
bruises	bruises=t,no=f
odor	臭い almond=a,anise=l,creosote=c,fishy=y,foul=f, musty=m,none=n,pungent=p,spicy=s
gill-attachment	ひだがあるか？ attached=a,descending=d,free=f,notched=n
gill-spacing	ひだの間隔 close=c,crowded=w,distant=d
gill-size	ひだの大きさ broad=b,narrow=n
gill-color	ひだの色 black=k,brown=n,buff=b,chocolate=h,gray=g, green=r,orange=o,pink=p,purple=u,red=e, white=w,yellow=y
stalk-shape	茎の形 enlarging=e,tapering=t
stalk-root	茎の根 bulbous=b,club=c,cup=u,equal=e, rhizomorphs=z,rooted=r,missing=?

列名	説明
stalk-surface-above-ring	茎表面上方のリング fibrous=f,scaly=y,silky=k,smooth=s
stalk-surface-below-ring	茎下部下方のリング fibrous=f,scaly=y,silky=k,smooth=s
stalk-color-above-ring	茎の色（上方のリング） brown=n,buff=b,cinnamon=c,gray=g,orange=o, pink=p,red=e,white=w,yellow=y
stalk-color-below-ring	茎の色（下方のリング） brown=n,buff=b,cinnamon=c,gray=g,orange=o, pink=p,red=e,white=w,yellow=y
veil-type	partial=p,universal=u
veil-color	brown=n,orange=o,white=w,yellow=y
ring-number	none=n,one=o,two=t
ring-type	cobwebby=c,evanescent=e,flaring=f,large=l, none=n,pendant=p,sheathing=s,zone=z
spore-print-color	black=k,brown=n,buff=b,chocolate=h,green=r, orange=o,purple=u,white=w,yellow=y
population	abundant=a,clustered=c,numerous=n, scattered=s,several=v,solitary=y
habitat	grasses=g,leaves=l,meadows=m,paths=p, urban=u,waste=w,woods=d