主成分分析を用いた食事データ評価スコアの分類

氏名:福田 太一

1　はじめに

このレポートでは, 以前ご提供して頂いた食事データについて分類モデルを通して分析を行ったため, その際用いた手法や分析結果について詳細に述べていく. なお, 食事データの詳細については自明のものとして説明は割愛する.

2　データの前処理

データの分析を行う前に, 未入力の項目を持つレコードを 1 件削除した上で次の処理を行った.

文字列を含む変数のダミー化

文字列を含む変数 (表 1) といった数字ではない入力は分類モデルに投入する上で扱いが困難であることが多い. 従って, その変数に含まれるカテゴリごとに有無 (0:カテゴリを含まない, 1:カテゴリを含む) を示すダミー変数 (表 2) に変換する.

表 1: 文字列を含む変数

レコード番号	Type
1	breakfast
2	lanch
3	dinner

表 2: ダミー変数

レコード番号	Type_breakfast	Type_lanch	Type_dinner
1	1	0	0
2	0	1	0
3	0	0	1

標準化

数字を含む変数は各変数ごとに分布が異なるが, 次の処理を施すことによって平均値が 0, 分散が 1 となり, 各変数のスケールを統一することができる。

$$x' = \frac{x - \mu}{\sigma}$$

x : レコードの持つ変数, μ : 平均値, σ : 標準偏差

3　主成分分析

今回のデータ分析の手法として, 主に主成分分析を活用した. この方法は, データに含まれる多くの変数を低い次元の合成変数に縮約する方法である. これによって多くの変数を含むデータが有している情報を解釈しやすくすることができる. 変数 $x_1, x_2, ..., x_n$ をもとに合成変数 z_1 を作成する場合, 係数ベクトルを $a_1, a_2, ..., a_n$ とすると次の通りに定まる.

$$z_1 = a_1x_1 + a_2x_2 + ... + a_nx_n$$

ここで, z_1 の分散を最大にすることによって z_1 に含まれる情報量を多くできると考え, z_1 の分散を最大にし, かつ大きさが 1 のベクトル $a' = [a_1, a_2, ..., a_n]$ を導出する. このとき, z_1 を第 1 主成分とする. また, 主成分に変数の値を入力して得られる数値を主成分得点と呼ぶ.

合成変数 z_2 も同様に, 係数ベクトルを $b_1, b_2, ..., b_n$ とすると次の通りに定まる.

$$z_2 = b_1x_1 + b_2x_2 + ... + b_nx_n$$

z_2 の分散は z_1 の次に最大であるが, z_2 には z_1 に不足した情報を補足する役割があるため z_1 とは無相関になるように定める. z_2 の分散を z_1 の次に最大にする大きさ 1 のベクトルを $b' = [b_1, b_2, ..., b_n]$ とすると, z_1, z_2 が無相関になるためには a', b' が垂直であることが条件となっている.

$$\sum_{i=1}^{n} a_ib_i = 0$$

こうして導出される z_2 を第 2 主成分とする.

これらの方法を繰り返して第 n 主成分まで作成するが, 後に作成した主成分ほど情報量は小さくなっていくため, 情報量が十分に小さい主成分は切り捨てても問題ない. そのための指標として寄与率を用いる. 寄与率とは一つの主成分が全体に対して占める情報量の大きさを表している. 第 m 主成分に対応する固有値を λ_m, 主成分の分散を V とおくと第 m 寄与率 C_m は次の通りに定まる.

$$C_m = \frac{\lambda_m}{V = \sum_{i=1}^{n} V_i}$$

また, 第 m 主成分までの寄与率の合計を第 m 主成分までの累積寄与率と呼ぶ. 主成分を採用する際は基本的に累積寄与率が 70~ 80 パーセント程度が目安とされている.

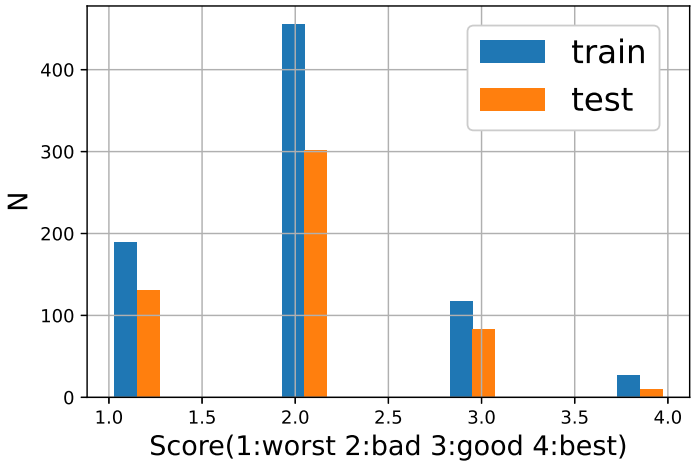

図 1: 訓練データとテストデータに含まれるスコアごとのレコード数

4　食事データの分類実験の説明

Python の scikit-learn ライブラリに含まれるランダムフォレストモデル (RandomForestClassifier) を用いて学習を行い, 食事データの目的変数となる食事に対する評価スコア (1:worst, 2:bad, 3:good, 4:best) を分類し, その精度を測定する. モデルの評価指標は正解率 (正解数÷テストデータのレコード数) と F 値 (各スコアをどれだけ正しく分類できているか) とする.

モデルへの入力

食事データに含まれるモデルへ入力する説明変数 (目的変数以外の変数) に対して主成分分析による次元削減を行い, 得られた主成分得点をモデルへの入力とし学習を行う.

訓練データとテストデータ

今回, レコード数 1314 の食事データのうち訓練データとテストデータの比率が 3:2 になるように分割した (図 1) . しかし, 訓練データとテストデータの両方においてスコア 3, 4 のレコード数がスコア 1, 2 のレコード数と比較して不均衡であることが判明した. 従って, モデルが学習を行う際にスコア 3, 4 を他のスコアと誤って分類したときのペナルティを重くする必要があり, これによりレコード数が少ないスコアの正解率を上げることができる. 今回はランダムフォレストモデルのパラメータ class_weight を”balanced”に設定することでこの仕様を実装した.

5　分類実験の結果

まず最初に第 1~ 第 n 主成分得点 (以下「PCn」とする) までモデルに投入してそれぞれの正解率 (図 2), F 値 (図 3) を算出した. それぞれにおいて主成分の数に比例して精度が単調増加するのではなく, 所々精度が増加または減少している箇所が見られた. 即ち, 評価スコアの分類の精度に良い影響をもたらす主成分とノイズとなる主成分がそれぞれ存在していると考えられる.

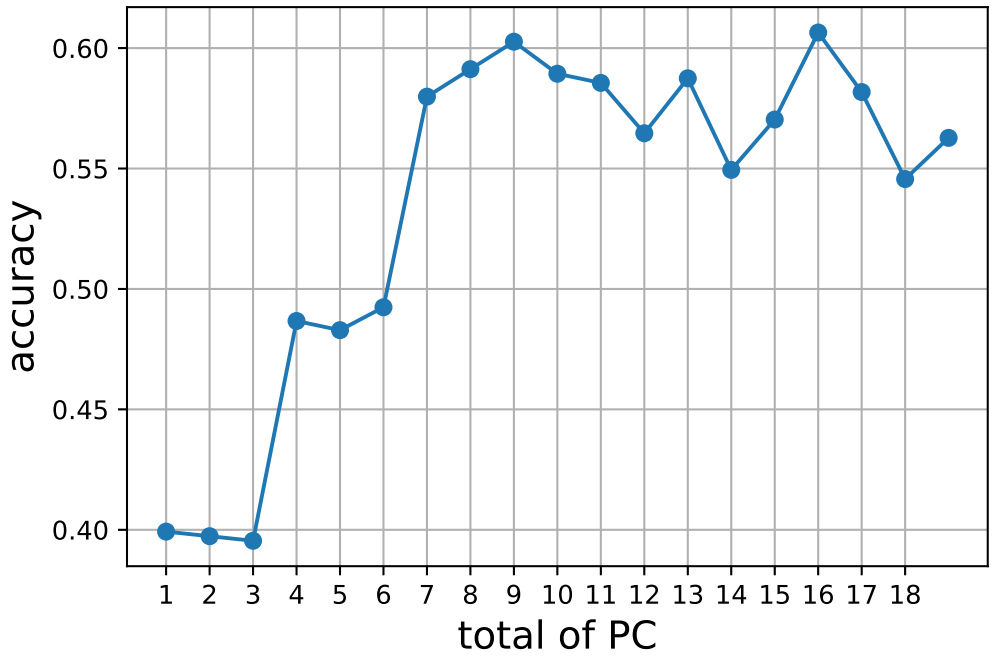

図 2: 累積寄与率の大きい順に主成分を投入したモデルの正解率

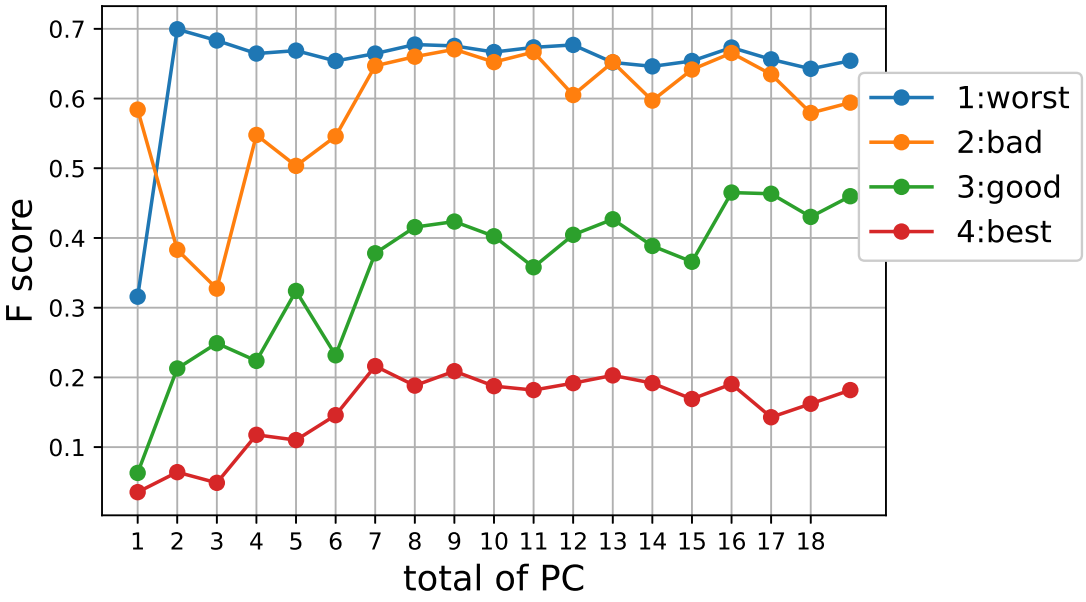

図 3: 累積寄与率の大きい順に主成分を投入したモデルの F 値

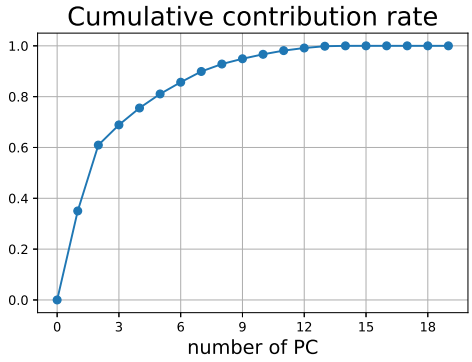

図 4: 累積寄与率

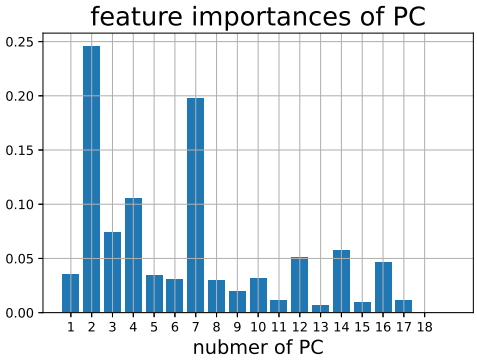

図 5: モデルの特徴量に対する重み

主成分の累積寄与率とモデルの特徴量に対する重み

今回得られた主成分の累積率と前主成分投入後のモデルの特徴量 (ここでは主成分) に対する重みをそれぞれ図 4, 図 5 に示す. PC_6 までで累積寄与率は 80 パーセントを超えており, それ以降の主成分の持つ寄与率は十分に小さいことが言える. ただ, それにも関わらず PC_7 や PC_{14} といった重みが 2 番目, 5 番目に大きい特徴量が存在する. このことから, 情報量の小さい主成分であっても学習後のモデルが重要視するような質の高い情報を持つことが示唆される. よって, 自分は「寄与率の大小に関わらず重みの大きい特徴量が分類の精度の改善をもたらす」 という仮説を立てた.

重みの大きい主成分得点から投入して精度を測定

最初の実験 (以下「実験 1」とする) では寄与率の大きい主成分得点から順に投入して精度を測定したが, 次に行った実験 (以下「実験 2」とする) では先程立てた仮説に基づいて重みの大きい主成分得点から順にモデルに投入した. 実験 2 の正解率, F 値をそれぞれ図 6, 図 7 に示す. 5 番目に重みの大きい主成分まで投入したモデルの正解率は 0.6103, スコア 2, 4 の F 値はそれぞれ 0.6859, 0.2456 と実験 2 の結果の中では最大値を記録した. また, この場合のモデルは正解率と各 F 値の合計が最も優秀であったため, 食事データの分類には 5 番目に重みの大きい主成分得点まで, 即ち $PC_2, PC_7, PC_4, PC_3, PC_{14}$ が有効であることが言える.

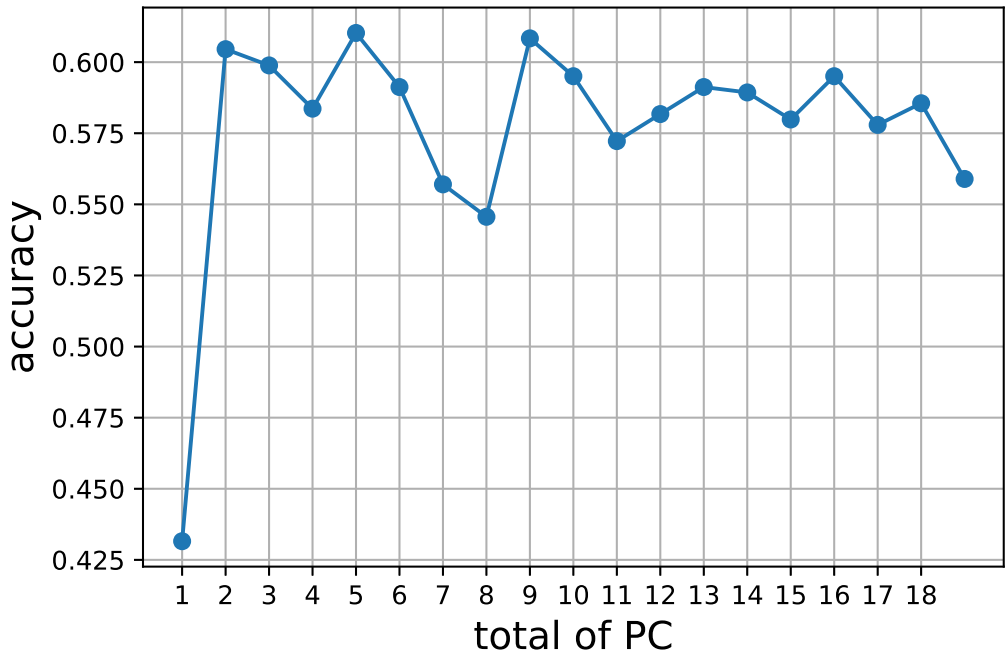

図 6: 重みの大きい順に主成分得点を投入したモデルの正解率

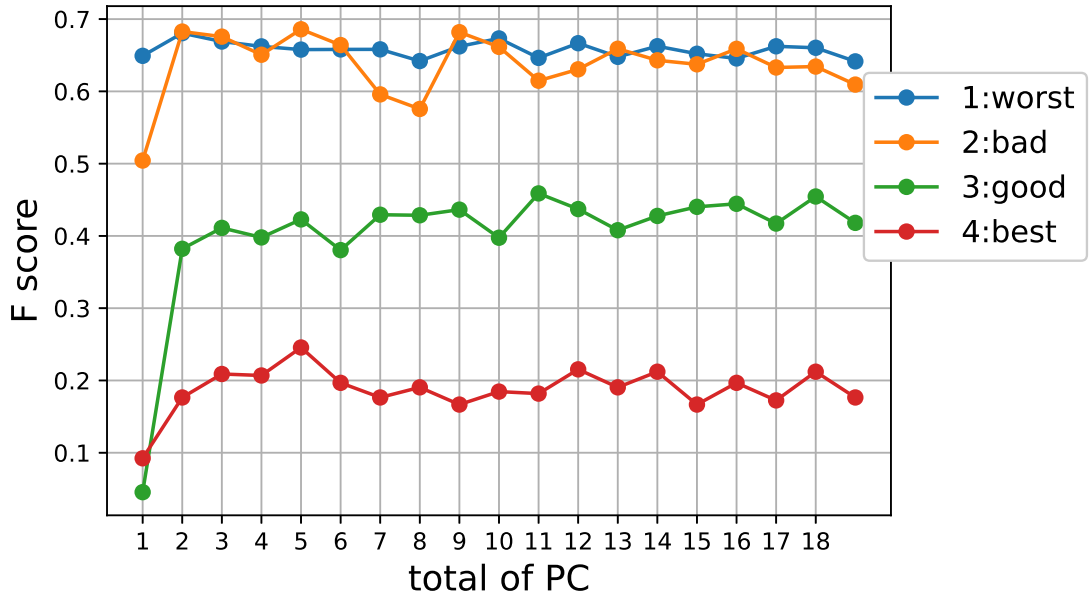

図 7: 重みの大きい順に主成分得点を投入したモデルの F 値

表 3: 二つの実験の最良モデルの精度比較

	正解率	F 値				合計
		スコア 1	スコア 2	スコア 3	スコア 4	
実験 1	0.6065	0.6731	0.6654	0.4651	0.1905	2.606
実験 2	0.6103	0.6578	0.6859	0.4229	0.2456	2.622

表 4: 二つの実験の最良モデルの構成

	モデルの構築方法	投入した主成分得点
実験 1	寄与率の高い順に主成分得点を投入	PC_1から PC_{16}
実験 2	特徴量の重みの大きい順に主成分得点を投入	$PC_2, PC_7, PC_4, PC_3, PC_{14}$

実験 1 と実験 2 の結果の比較

二つの実験で最良であったモデルの精度比較を表 3, それぞれの構成を表 4 に示す. 実験 2 の最良モデルは実験 1 のモデルと比較して総合的な精度で上回っていることや, 投入するデータの次元数が 1/3 程小さいことを考えると実験 1 のモデルよりも優秀であることが言える. これにより自分が立てた「寄与率の大小に関わらず重みの大きい特徴量が分類の精度の改善をもたらす」という仮説は立証された.

6　得られた主成分に関する考察

ここでは実験 2 の最良モデルに投入された $PC_2, PC_7, PC_4, PC_3, PC_{14}$ についてこれらが食事データの評価スコアの分類にもたらす影響を考える.

主成分の持つ意味について

得られた主成分の係数ベクトルを表 5 に示す. 主成分の持つこの値の大きさと正負から主成分の意味について考える. 例えば PC_2 は P[g], Salt[g] の係数が大きいうえに 0 よりも大きいため, 「タンパク質と塩分・摂取量の大きさ」という意味付けができる. このようにして行った各主成分の意味付けを表 6 に示す. いずれも食事に対する評価スコアに影響するものと思われるが, その影響力については主成分得点を可視化して考える.

表 5: 主成分の係数ベクトル

	PC_2	PC_7	PC_4	PC_3	PC_{14}
Type_breakfast	-0.116	0.03	0.071	-0.007	0.04
Type_dinner	0.078	-0.031	0.046	0.015	-0.065
Type_lunch	0.038	0.001	-0.117	-0.008	0.025
gender_female	0.063	-0.045	0.072	0.198	0.012
gender_male	-0.063	0.045	-0.072	-0.198	-0.012
age	0.031	-0.029	0.51	-0.679	0.001
height	-0.105	-0.004	-0.258	-0.383	-0.004
weight	-0.121	-0.021	0.007	-0.381	0.03
EER[kcal]	-0.164	-0.003	0.123	0.172	-0.003
P target(15%)[g]	-0.164	-0.001	0.123	0.18	0.004
F target(25%)[g]	-0.164	-0.005	0.125	0.18	-0.011
C target(60%)[g]	-0.164	-0.003	0.123	0.171	-0.004
number of dishes	0.245	0.281	0.618	0.148	-0.013
E[kcal]	0.42	0.165	-0.116	-0.054	0.783
P[g]	0.4	0.214	-0.092	0.032	-0.151
F[g]	0.36	0.444	0.019	-0.036	-0.44
C[g]	0.291	-0.282	-0.249	-0.102	-0.401
Salt[g]	0.371	-0.211	-0.144	0.051	0.046
Vegetables[g]	0.299	-0.721	0.308	0.069	0.004

表 6: 各主成分の意味付け

	意味
PC_2	タンパク質・塩分の摂取量の大きさ
PC_7	脂肪の摂取量の大きさと野菜の不足分
PC_4	食事の回数と年齢の高さ
PC_3	年齢の低さと身長・体重の大きさ
PC_{14}	一日に得られるカロリーの大きさと脂肪・ビタミン C の不足分

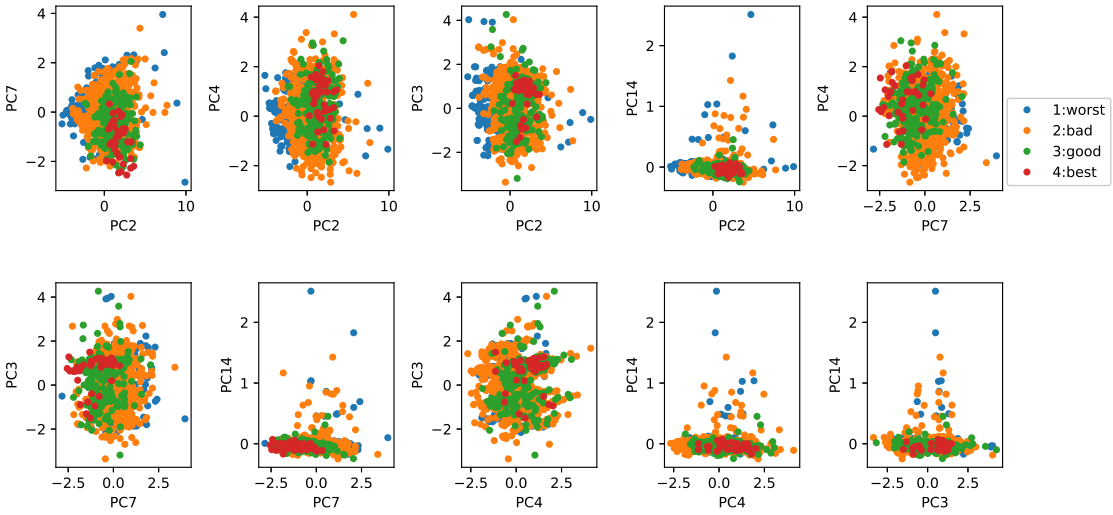

図 8: 食事の評価スコアごとの主成分得点のプロット

食事の評価スコアごとの主成分得点

食事の評価スコアごとの主成分得点を図 8 に示す. PC_2, PC_4, PC_3 に関しては平均に近いほど, PC_7 は値が小さいほど, PC_{14} は値が 0 に近いほどスコアが高くなる傾向が見られた. それぞれの分布から, 食事の評価スコアが高くなる人物の特徴として次のものが考えられる.

- PC_2, PC_4, PC_3 から考えられること
  タンパク質・塩分を過不足なく摂取し, 年齢・身長・体重・食事回数は平均的である. (特に食事回数は 2, 3 回であると考えられる.)

- PC_7 から考えられること
  脂肪を多く摂らず, 野菜を沢山食べる.

- PC_{14} から考えられること
  必要最低限のカロリーとビタミン C を摂取している.