＜ 修 士 論 文 ＞

識別モデルと生成モデルの比較について

滋 賀 大 学 大 学 院
デ ー タ サ イ エ ン ス 研 究 科
デ ー タ サ イ エ ン ス 専 攻

修了年度：２０２２年度

学籍番号：6021123
氏　　名：照井　健司
指導教員：椎名　洋

提出年月日：２０２３年１月９日

概要

ある個体に関して計測されたデータ x（連続値、離散値どちらの場合もあり）から、その個体が属するクラス（尺度変数、あるいはカテゴリー変数）y を予測する方法（以下、識別と呼ぶ）について、考える。この観点から、モデルを識別モデルと生成モデルに分けることができる。識別モデルは、識別を直接的な目的としたモデルであり、生成モデルは、データと似た分布を生成するモデルであり、データのモデルによる再現に目的がある。データのモデルによる再現ができれば、非常に幅広い目的に使用可能であり、もちろん識別にも使用可能である。本研究の目的は、識別モデルと生成モデルを様々な条件のもとで、識別に使用し、その結果を比較することにある。一般的に、前者の方が識別目的に特化したモデルなので、後者に比べて、「常に」識別性能が優れていると考えられているが、サンプル数が少ない間は、生成モデルの方がよい識別成績を示すことも多いことを Ng and Jordan（2001）が指摘した。本論文の目的は、サンプル数や説明変数の異なるいくつかのデータセットに、識別モデルと生成モデルを適用することで、この指摘を確認することにある。

論文の構成は以下のとおりである。第一章で、識別モデルと生成モデルについての解説、および今回使用した４つのデータセットの説明を行う。第ニ章では、ロジスティック回帰、SVM、深層学習の３つの識別モデルをこれらのデータに適用した結果を考察する。第三章では、汎用的な生成モデルの代表例として、ナイーブベイズ正規モデルを４つのデータセットに適用し、これに基づいた判別結果を紹介する。第四章では、識別モデルと生成モデルの比較に関するまとめを行う。なお、今回の研究では、すべてベイズモデルを採用している。すなわち、モデルのパラメーターはすべて確率変数であり、特に言及しない場合は、無情報事前分布を採用している。具体的には、cmdstan を R の上で動かしている。

第一章

1. 識別モデルと生成モデルについて

識別モデルは、各個体の資質を示す X（以下では説明変数と呼ぶ）というデータがあたえられたときに、これに基づいて Y を予測するモデルである。これに関しては、おおよそ二つの方式がある。一つは、X が与えられた時に、各クラスに識別される確率、つまり Y の X に関する条件付き確率 $P(Y|X)$を与えるモデルであり、この確率が最大になるクラスを選択することによって識別が行われる。この方式はソフトな識別方法と呼ばれ、ロジスティック回帰や深層学習（出力が確率のタイプ）が代表的なモデルである。一方、ハードな識別方法は、X を入力すると Y が出力として出てくるタイプのモデルであり、SVM、決定木、深層学習（出力がクラスラベルのタイプ）などがこれにあたる。

ソフトな識別モデルが X に関する Y の条件付き確率 $P(Y|X)$を直接モデル化することに対して、生成モデルは最初にクラス Y に関するデータ X の条件付き確率分布 $P(X|Y)$を求める。次に求めた $P(X|Y)$と $P(Y)$からベイズの定理を利用して X に関する Y の条件付き確率 $P(Y|X)$を求める。つまり識別モデルとは異なり、まずクラス Y の情報からデータ X の分布を推定するためのモデル $P(X|Y)$を構築し、これを $P(Y)$と組み合わせてベイズの定理で反転させて、Y の条件付き確率 $P(Y|X)$を構築する。X から Y を推定する識別モデルとは逆に、Y から X の分布を推定することから、生成モデルは逆推定とも呼ばれている。また、このモデルが生成モデルと呼ばれる理由は、$P(X|Y)$と $P(Y)$から、同時分布 $P(X,Y)$を導くことができる、すなわち X と Y をシミュレートできるようになるからである。（正確に言えば、識別モデルにおいても、ロジスティック回帰のように $P(Y|X)$をモデル化するものは、$P(X)$と組み合わせて同時分布 $P(X,Y)$ を作ることが可能であるが、あくまでも $P(Y|X)$を作ることが目的なので、識別モデルに分類されることが多い）。Y をダミー変数として説明変数に組み込んだ X に関する一般化線形モデルとなどが、これに属する。Y のクラスごとに、全く違った X に関するモデルをつくることも可能である。

2. 使用したデータについて

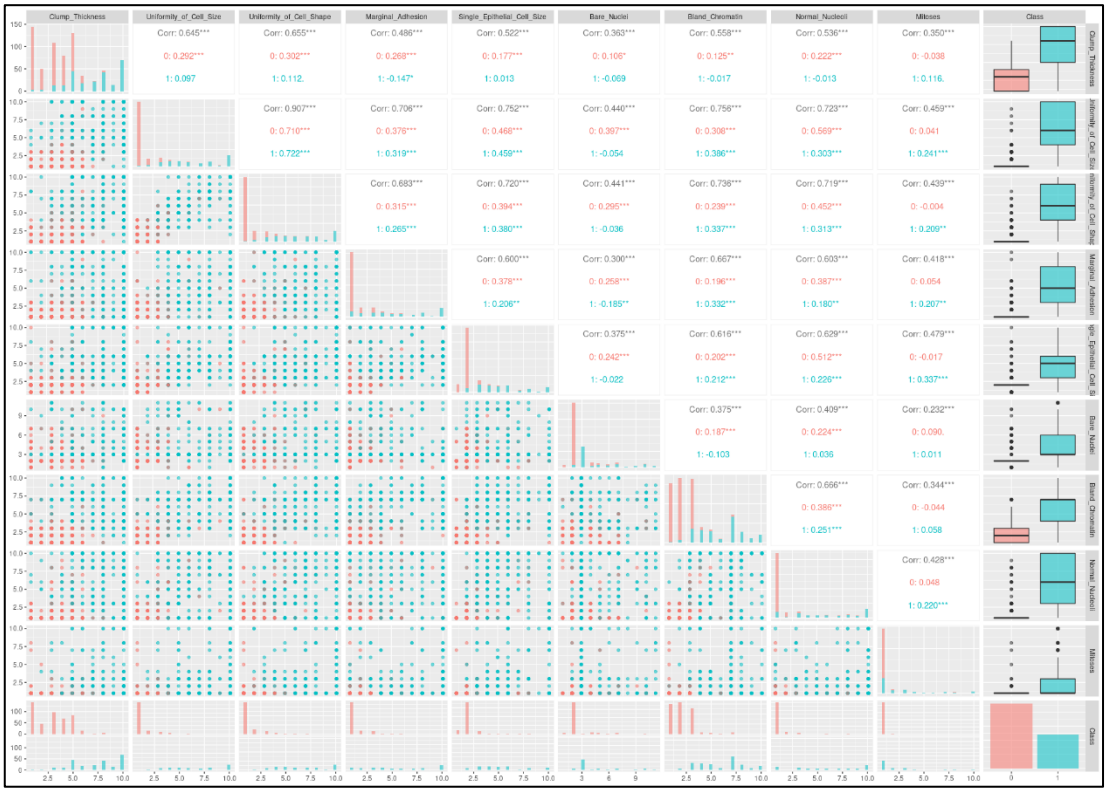
検証には The UCI Machine Learning Repository[1]からダウンロードした Breast Cancer(乳がん)、U.S voting(アメリカの選挙)、Sonar(ソナー)と、あるスポーツチームの顧客データ[2]を使用したデータの 4 種類を使用した。以下にこれらのデータセットの内容を記す。

Breast cancer ： The UCI Machine Learning Repository の乳癌患者のデータ
原題：Breast Cancer Wisconsin (Original) Data Set[3]
細針吸引による乳房の細胞診検査のデータセット。良性と悪性とで異なることが報告された 9 つの特徴について、それぞれ 1 から 10 に等級付けされている。値が大きいほど、よりサイズが大きい、より不均一、より量が多い、という等級になる。
患者のデータ数 N：683 人分
欠損しているデータがなかった為、元データと同じサンプル数を使用した。
Y：良性[0]、または悪性[1]の 2 値クラス
各クラス内に含まれるデータ数：[クラス 0: 444 個, クラス 1: 239 個]
説明変数の数：9 個
X_1：Clump Thickness：細胞塊の厚さ(癌細胞は多層構造で厚さがある)：1 から 10 の離散値
X_2：Uniformity of Cell Size：細胞の大きさの均一性(健康な細胞は大きさが均一であるが、癌細胞は大きさが様々である)：1 から 10 の離散値
X_3：Uniformity of Cell Shape：細胞の形状の均一性(健康な細胞は形状が均一であるが、癌細胞は形状が様々である)：1 から 10 の離散値
X_4：Marginal Adhesion：細胞と周囲との境界の癒着(正常な細胞は互いに癒着している傾向があるが、癌細胞は癒着していない傾向がある)：1 から 10 の離散値
X_5：Single Epithelial Cell Size：単一上皮細胞の大きさ(上皮細胞：臓器などの表面を覆う細胞であり、癌細胞はこの細胞が肥大していることがある) ：1 から 10 の離散値
X_6：Bare Nuclei：裸核の量(裸核：乳酸によって細胞質が溶解し、細胞の核だけが残ったものであり良性の腫瘍にみられることが多い) ：1 から 10 の離散値
X_7：Bland Chromatin：無菌性のクロマチンの量(クロマチン：DNA とタンパク質の複合体であり、癌細胞ではこれが塊となっていることがある)：1 から 10 の離散値
X_8：Normal Nucleoli：正常な核小体の大きさ(核小体：細胞核の中にある分子密度の高い領域であり、正常な細胞では小さいものだが癌細胞では大きいことがある) ：1 から 10 の離散値
X_9：Mitoses：有糸分裂(有糸分裂期にある細胞の量。癌は本質的には有糸分裂の疾患である。) ：1 から 10 の離散値

---

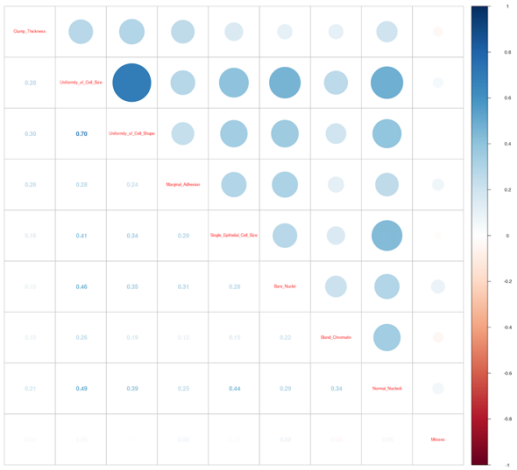
[1] UCI Machine Learning Repository より
https://archive.ics.uci.edu/ml/index.php
(2022 年 12 月 25 日確認)
[2] 筆者も参加した滋賀大学と企業との共同研究で使用したデータを、修論にも使うことを許可していただいた
[3] UCI Machine Learning Repository より
https://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+%28original%29
(2022 年 12 月 30 日確認)

Breast Cancer データセット：ヒストグラムと相関

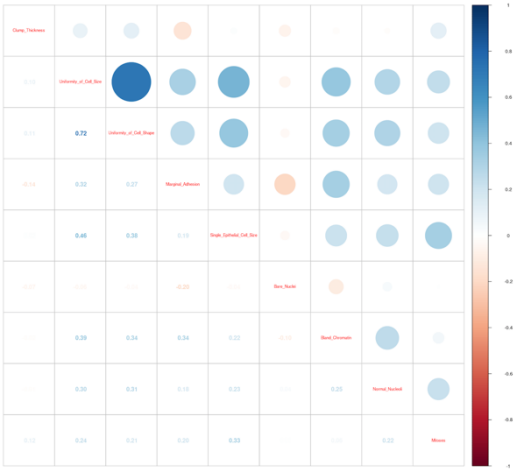


Breast Cancer データセット：説明変数（X）のクラス別の相関行列のヒートマップ

Class0（良性）



Class1（悪性）



U.S. Voting ： The UCI Machine Learning Repository の米国下院議員の投票データ

原題：Congressional Voting Records Data Set[4]

民主党、共和党のそれぞれに所属する米国下院議員がどういった法案に賛成および反対票を投じたのかが記録されたデータセット

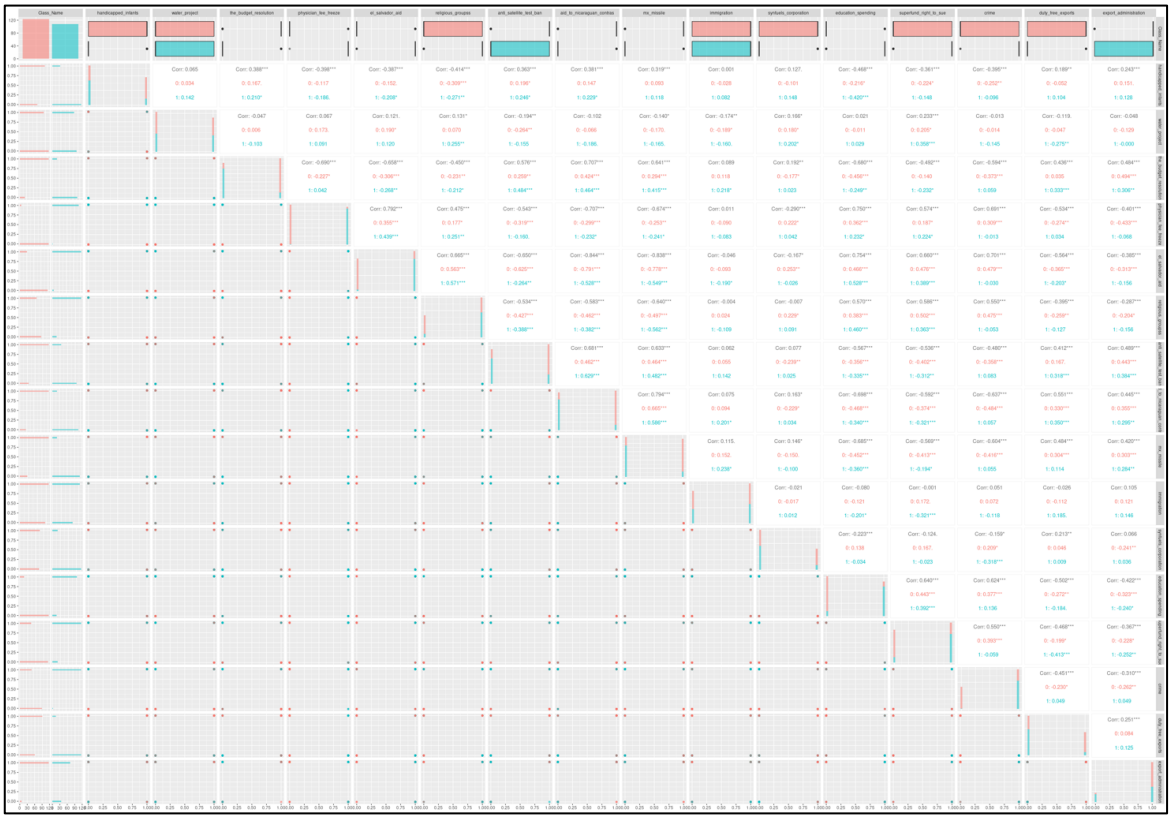
議員のデータ数 N：232 人分

The UCI Machine Learning Repository にある元データのサンプル数は 435 個であるが、203 個のサンプルにおいて説明変数 X のデータが欠損していた為、全てのデータが揃っている 232 個のサンプルを検証に使用した。

[4] UCI Machine Learning Repository より
https://archive.ics.uci.edu/ml/datasets/congressional+voting+records
（2022 年 12 月 30 日確認）

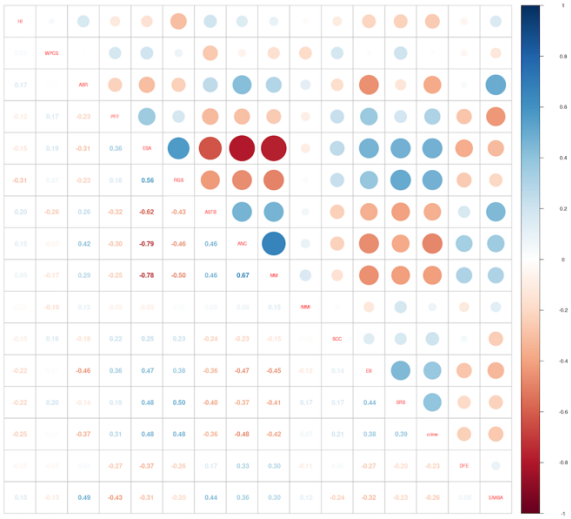
Y: 民主党[0]、または共和党[1]の２値クラス
各クラス内に含まれるデータ数：[クラス 0: 124 個, クラス 1: 108 個]
以下の説明変数は賛成[1]、または反対[0]の２値データ
説明変数の数：16 個
X_1：handicapped-infants：障害児に関する投票
X_2：water-project-cost-sharing：水の供給に関する投票
X_3：adoption-of-the-budget-resolution：予算案の採択に関する投票
X_4：physician-fee-freeze：診療報酬に関する投票
X_5：el-salvador-aid：エルサルバドル援助に関する投票
X_6：religious-groups-in-schools：学校内における宗教グループに関する投票
X_7：anti-satellite-test-ban：衛星に対する実験禁止に関しての投票
X_8：aid-to-nicaraguan-contras：ニカラグアの反政府民兵への支援に関する投票
X_9：mx-missile：mx ミサイルに関しての投票
X_{10}：immigration：移民に関する投票
X_{11}：synfuels-corporation-cutback：米国政府資金提供の会社 Synfuels に関する投票
X_{12}：education-spending：教育に関する投票
X_{13}：superfund-right-to-sue：superfund-right に関する投票
X_{14}：crime：犯罪に関する法案への投票
X_{15}：duty-free-exports：免税に関する投票
X_{16}：export-administration-act-south-africa：南アフリカに対する輸出管理法に関しての投票

U.S.Voting データセット：ヒストグラムと相関

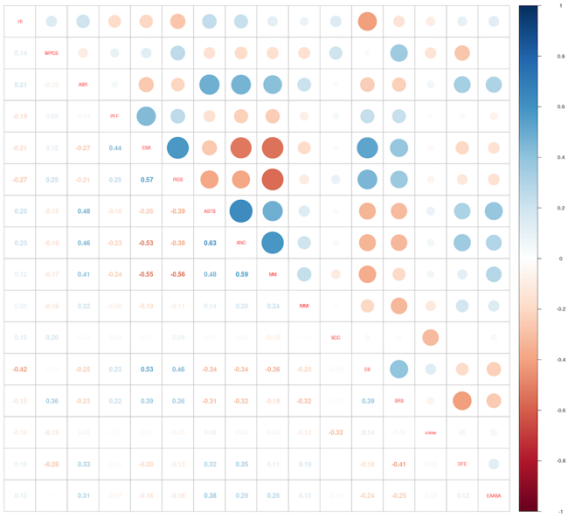


U.S.Voting データセット：説明変数（X）のクラス別の相関行列のヒートマップ

Class0(反対)



Class1(賛成)



Sonar ：ソナーで岩か地雷（金属シリンダー）かを判定したデータ

原題：Connectionist Bench (Sonar, Mines vs. Rocks) Data Set[5]

ソナー信号を海底にある円柱の形状をした岩および地雷(金属シリンダー)にあて、跳ね返ってきた信号パターンを計測したデータ。跳ね返ってきた信号を特定の周波数帯域内ごとに、そのエネルギーを一定期間にわたって計測し、積分した値が 0 から 1 の数値として 60 個の説明変数として記録されている。ソナー信号は岩および地雷に様々な角度や条件で当てており、その条件と角度の組み合わせのパターンがそれぞれ、岩が 97 パターン、地雷が 111 パターンあり、合計 208 パターンがデータセットとなっている。

データ数 N：208

欠損しているデータがなかった為、元データと同じサンプル数を検証に使用した。

Y：岩[0]または地雷[1]（金属シリンダー）の 2 値クラス

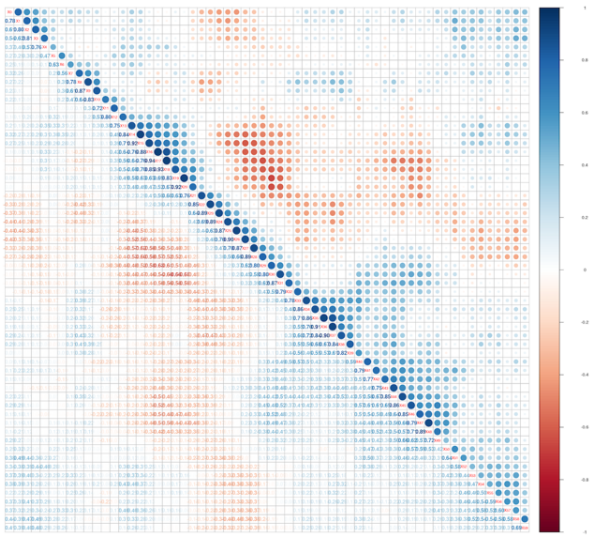
各クラス内に含まれるデータ数：[クラス 0: 97 個, クラス 1: 111 個]

説明変数の数：60 個

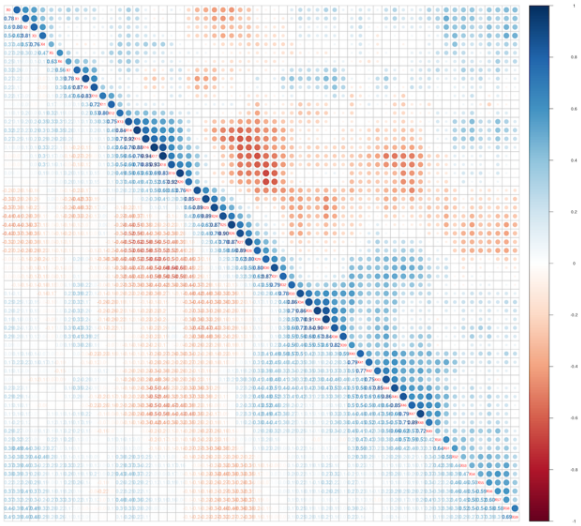
X_1~X_{60}：60 個の説明変数： 全て連続値のデータ

Sonar データセット：説明変数（X）のクラス別の相関行列のヒートマップ

Class0(岩)



Class1(地雷)



[5] UCI Machine Learning Repository より
http://archive.ics.uci.edu/ml/datasets/connectionist+bench+(sonar,+mines+vs.+rocks)
(2022 年 12 月 30 日確認)

スポーツチームの顧客データ ： ある企業から滋賀大学に提供頂いたスポーツチームの顧客データ
この顧客データは性別、年齢、住所等の顧客情報と顧客に対してとられたアンケートの回答結果の情報を含んでいる。アンケートの内容は回答者が当該スポーツチームへ積極的に関与する意欲があるかを測る 6 択の選択形式の質問である。回答者は最も自身に当てはまると感じた回答を選ぶ。6 択の回答はそれぞれ、チームへ関与する意欲の強さが最も弱いことを示す回答を１、最も強い意欲を示す回答を６としている。なお、最もチームへの意欲が弱いことを示す回答である回答 1 と回答 2 は、答えの内容に大きな差異がないと思われるので、これら２つの回答をしたクラスを１つにまとめ、5 つのクラスとしたデータを識別モデルの検証で使用している。チームへの関与する意欲が最も弱いクラスを 1、最も意欲が強いクラスを 5 とする。また、識別モデルと生成モデルを比較した検証においてはアンケート回答結果のクラスを 6 つから、さらに小さく 3 つにまとめたデータで検証を行った。チームへの関与する意欲が最も弱いクラスを 1、最も意欲が強いクラスを 3 とする。

アンケート回答の選択肢
1 ファンではない
2 ファンになりたて、これから応援するのが楽しみ
3 これからもずっと継続して応援したい
4 これまで以上に、夢中になって応援したい
5 もっと役に立ちたい、一緒に関わって盛り上げていきたい
6 もはや人生の一部、これからも一緒に生きていきたい

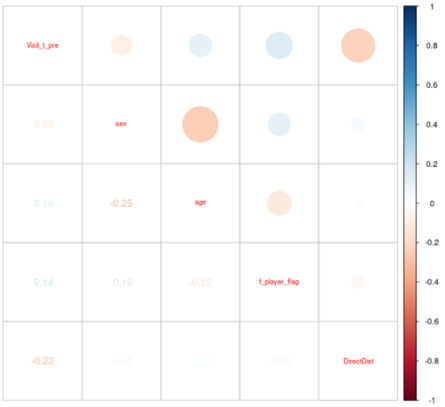
回答 1 の「ファンではない」と回答 2 の「ファンになりたて」は一見すると全く異なる内容の回答である。しかし、このアンケートはスポーツチームのファンがチケットやグッズ等を購入するために無料および有料のファンクラブ会員登録を行い、さらにメールマガジン配信希望をしている顧客にのみ実施されたものであることから、チームにある程度のレベルで興味がある人が回答している。チームへ関与する意欲の度合いという観点で、「ファンではない=消極的なファン」と「ファンになりたて=意欲がこれから高まる」と解釈し、大きな差のない回答としている。

データ数 N : 14, 620
元のアンケート回答データと顧客情報データ両方が揃った顧客データは 14, 938 件であったが、いくつかの顧客データが欠損していた 318 件データのデータを除外した 14, 620 件を検証に使用した。
説明変数の数：5
Y： 識別モデルの検証では 5 クラス、識別モデルと生成モデルの比較では 3 クラス
各クラス内に含まれるデータ数：5 クラスの場合：
[クラス 1:336 個，クラス 2:6, 397 個，クラス 3:1, 826 個，クラス 4:1, 262 個，クラス 5:4, 799 個]
各クラス内に含まれるデータ数：3 クラスの場合：
5 クラスから、クラス 1, 2 をクラス 1 に、クラス 3, 4 をクラス 2 に、クラス 5 をクラス 3 としている
[クラス 1： 6, 733 個，クラス 2： 3, 088 個，クラス 3： 4, 799 個]
X_1：スポーツ会場への来場回数：離散値
X_2：性別：男性[1]、女性[2]の２値データ
X_3：年齢：離散値
X_4：お気に入り選手の有無：有[1]、無[0]の２値データ
X_5：顧客の住所とスポーツ会場との直線距離（会場と住居との遠さの情報）、連続値

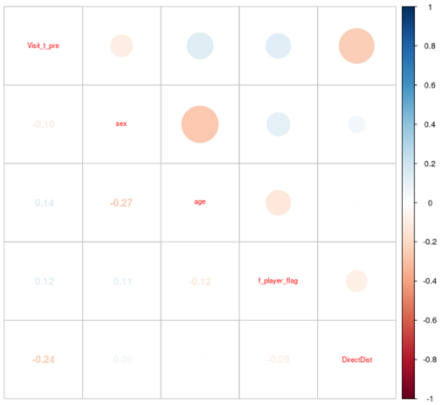
スポーツチームの顧客データセット：ヒストグラムと相関

スポーツチームの顧客データセット：説明変数（X）のクラス別の相関行列のヒートマップ

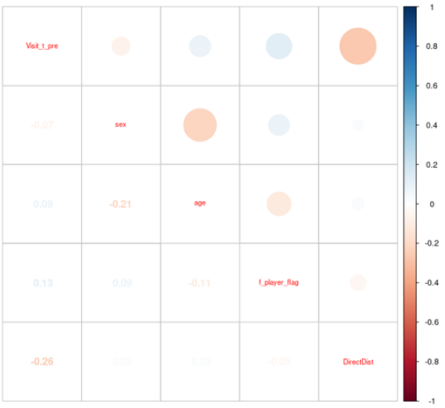
Class0



Class1



Class3



モデルの識別精度の基準
本研究ではモデルの識別精度を比較する為の基準としてエラー率と F1 スコアを使用している。

	クラス1の正解データ (Positive)	クラス0の正解データ (Negative)
クラス1と予測 (Positive)	TP	FP
クラス0と予測 (Negative)	FN	TN

エラー率：予測値と正解データの一致している割合である正答率を 1 から引いたもの、予測値の中で誤った予測をしているデータの割合を計算したもの

正答率(Accuracy)

$$\frac{TP+TN}{TP+TN+FP+FN}$$

エラー率(Error rate)

$$\frac{FN+FP}{TP+TN+FP+FN}$$

F1 スコア：適合率と再現率を調和平均した値であり、2 値のクラスを識別するモデルの評価指標

再現率(Recall)

$$\frac{TP}{TP+FN}$$

適合率(Precision)

$$\frac{TP}{TP+FP}$$

F1 スコア

$$\frac{2\times\ \textit{再現率}\ \times\ \textit{適合率}}{\textit{再現率}+\textit{適合率}}$$

F1 macro 平均：多クラスの場合の F1 スコアのひとつ
多クラスを特定の 1 つのクラスとその他全てのクラスの 2 値とすることで、各クラスの F1 スコアを計算することができる。マクロ平均 F1 スコアは各クラスの F1 スコアを計算し、その平均をとったもので多クラス識別の評価指標として利用される。

$(F1Score)_k$をクラス k とそれ以外のクラスで計算した F1 スコア、K を目的変数 Y のクラスの数とすると、

$$macroF1\ =\ \frac{1}{K}\sum_{k=1}^{K}(F1Score)_k$$

学習用データと検証用データの選び方

本研究では、学習に使用するサンプル数 n の数を変化させたときに、識別性能がどうのように変化するかを見る。どのように、n を変化させるかについては、以下の通りである。

例：元データのサンプル数が N のデータセットの場合

1. サンプル N 個全てを使用し、クロスバリデーションを 10 分割(9 学習:1 検証)で検証。結果的に $n=N\times0.9$。
2. データセットのうち 8 割のサンプルをランダムサンプリングで抽出し、クロスバリデーションを 8 分割（学習：検証＝７：１）で検証。$n=N\times0.7$。
3. データセットのうち 6 割をランダムサンプリングで抽出、クロスバリデーションを 6 分割（学習：検証＝５：１）で検証。$n=N\times0.5$。
4. データセットのうち 4 割をランダムサンプリングで抽出、クロスバリデーションを 4 分割（学習：検証＝３：１）で検証。$n=N\times0.3$。
5. データセットのうち 2 割をランダムサンプリングで抽出、クロスバリデーションを 2 分割（学習：検証＝１：１）で検証。$n=N\times0.1$。

検証データは、いずれの場合も $N\times0.1$ 個になる。

１．から５．のひとつひとつで、三回ずつ反復を行い、すべての結果（エラー率、F1 値）は三回の平均値をとっている。

第二章

識別モデル(Discriminative model)

本研究では識別モデルに分類される代表的な手法のうち、ロジスティック回帰、サポートベクターマシン、ディープニューラルネットワークの３つの手法を検証に使用した。なお、識別モデルでは、クラスごとにデータの偏りが多いと、数の多いクラスに識別してしまう欠点が明らかになるが、この点は、オーバーサンプリングを実行することである程度解消できることを確認している（各モデルの最後にオーバーサンプリングによる検証結果（但し、スポーツチームの顧客データのみ）を付け加えている）。

今回は、すべてベイズモデリングによって、実証を行っている。パラメーターは事後分布（学習データから得られる）にしたがってその都度ランダムに発生させ、そのパラメーターと検証データから、さらにモデルに従ってランダムに Y を発生させている。この Y から、エラー率や F1 値を計算している。

第一節

ロジスティック回帰（Logistic Regression）

説明変数 X と目的変数 Y の関係 $p(X) = Pr\,(Y = 1|X)$ を表す最もシンプルなモデルに線形回帰モデルがある。目的変数 Y が 0 と 1 の２値である場合、$p(X) = \beta_0 + \beta_1 x_1 + \cdots + \beta_M x_M$ という形である線形回帰モデルでは、X の値によっては $p(X)$ が 0 以下となる場合や 1 以上となる場合がある。この $p(X)$ を 0 と 1 の間に収めることができるように、線形結合をさらに変換するための代表的な関数の１つがロジスティック関数である。

多クラスの識別への拡張：多項ロジスティック回帰（Multinomial Logistic Regression）

ロジスティック回帰は 2 クラスの識別に使用されるが、これを多クラスに拡張したものが多項ロジスティック回帰である。多項ロジスティック回帰を活用した多クラスの識別問題には予測値を決定するための２つの異なる方法がある。One vs All と Softmax 関数を使用した方法である。One vs All(One vs Rest とも呼ばれる)方法は、多クラスのデータをあるクラスとその他のクラスの２つに分けて２値識別を行う。この２値識別をクラスの数だけ行い、もっとも確率の高かったクラスに識別することで多クラスの識別を可能にする。一方、Softmax 関数を使用した方法は、Softmax 関数を使用することで対象のデータが各クラスに含まれる確率を計算することができる。Softmax 関数の出力が最も確率の高かったクラスが予測値となる。今回の検証ではこの２つの異なる方法である One vs All と Softmax 関数を使用する方法をそれぞれ使用し識別を行っている。

ロジスティック回帰の式
M は説明変数 X の数、K は目的変数 Y のクラスの数を表す。
以下の式から最も確率が高いクラス Y が予測値となる。

2 クラス識別の場合（$K=2$）

$$Pr\ (Y=0|X) = \frac{1}{1+\ e^{\beta_0+\ \beta_1\ x_1\ +\ \cdots+\beta_M\ x_M}}$$

$$Pr(Y=1|X) = \frac{e^{\beta_0+\ \beta_1\ x_1\ +\ \cdots+\beta_M\ x_M}}{1+\ e^{\beta_0+\ \beta_1\ x_1\ +\ \cdots+\beta_M\ x_M}}$$

Softmax 関数の式（K が 3 以上の場合に使用する）
M は説明変数 X の数、K は目的変数 Y のクラスの数を表す。

$$P(Y=i)\ =\ \frac{\frac{1}{1+\ e^{-(\beta_{i0}+\ \beta_{i1}\ x_1\ +\ \cdots+\beta_{iM}\ x_M)}}}{\sum_{k=1}^{K}\frac{1}{1+\ e^{-(\beta_{k0}+\ \beta_{k1}\ x_1\ +\ \cdots+\beta_{kM}\ x_M)}}} \qquad (i=1,2,\cdots,K)$$

実質的なパラメータの数 P は、以下のようになる。

$$P\ =(M+1)\ \times(K-1)$$

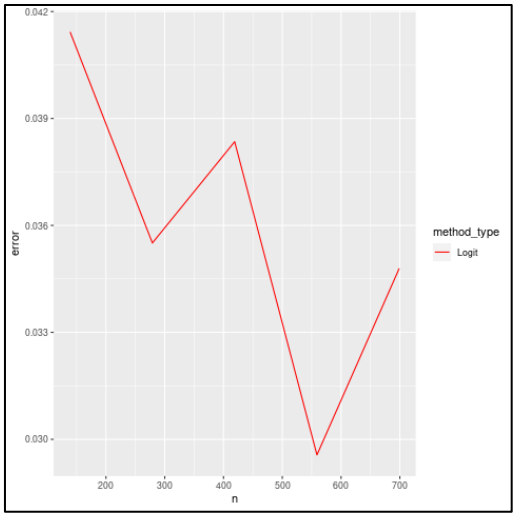
データセット別のパラメータ数一覧：ロジスティック回帰
パラメータ数をサンプル数で割り、パラメータ数とサンプル数の比を出している。

データセット名 および サンプル数 N	説明変数の数 M	クラス Y の数 K	ロジスティック回帰： パラメータ数	ロジスティック回帰： パラメータ数と サンプル数の比
Breast Cancer：683	9	2	10	0.015
U.S.Voting：232	16	2	17	0.073
Sonar：208	60	2	61	0.293
スポーツチームの 顧客データ：14,620	5	5	24	0.0016

ロジスティック回帰 での検証結果：エラー率

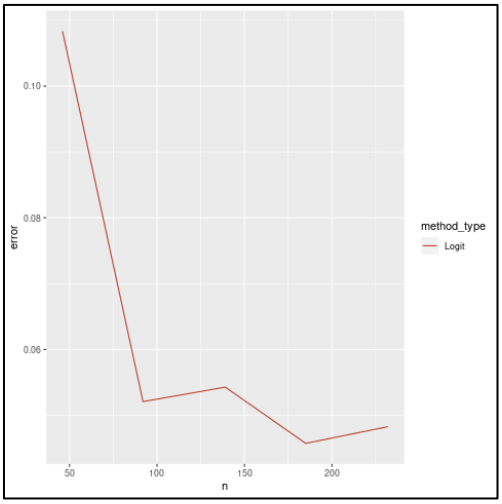
学習に使用するデータ n を増やしながら、エラー率の推移をグラフにした。多クラスの識別であるスポーツチームの顧客データに関しては One vs All と Softmax 関数を使用したそれぞれの方法をグラフにしている。縦軸がエラー率、横軸が分析に使用するデータ数である。

Breast cancer



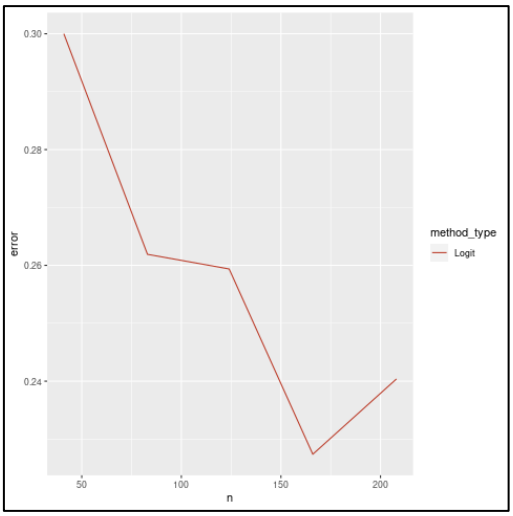
パラメータとサンプル数との比：0.015

U.S. Voting



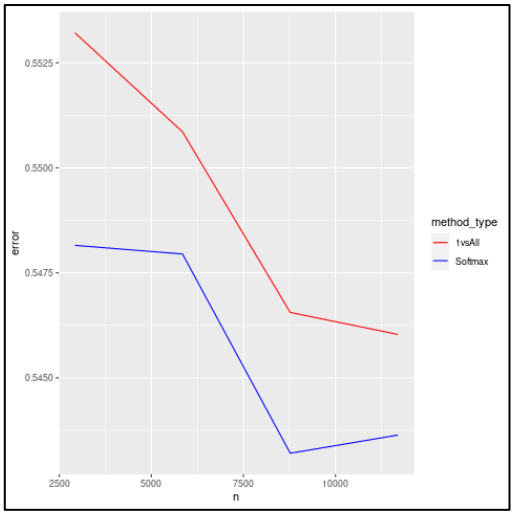
パラメータとサンプル数との比：0.073

Sonar



パラメータとサンプル数との比：0.293

スポーツチームの顧客データ



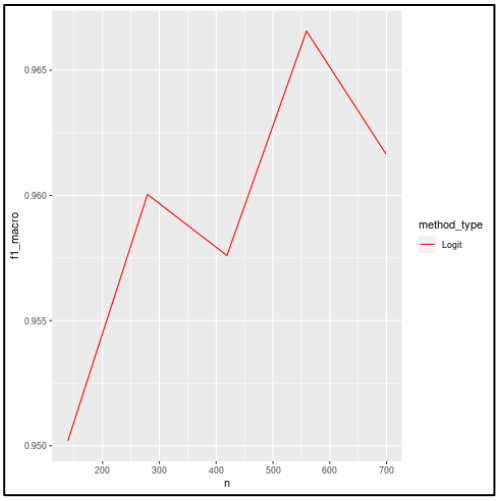
パラメータとサンプル数との比：0.0016

パラメータ数とサンプル数の比が比較的小さい Breast Cancer、U.S Voting データセットではエラー率が低いものの、パラメータ数とサンプル数の比が比較的大きい Sonar ではエラー率が高い。スポーツチームの顧客データはパラメータのサンプルの比が最も小さいものの、エラー率は他のデータセットと比べて高い。これは他のデータセットと比べてスポーツチームの顧客データはクラスが 5 つと多いこと、また、Y のラベルがアンケートによる回答であり、その区別が他のデータに比べて曖昧な点であることが、識別を困難にしている原因と考えられる。One vs All と Softmax 関数を使用した 2 つの方法に関しては大きな差は見られなかった。

ロジスティック回帰 での検証結果：F1 スコア

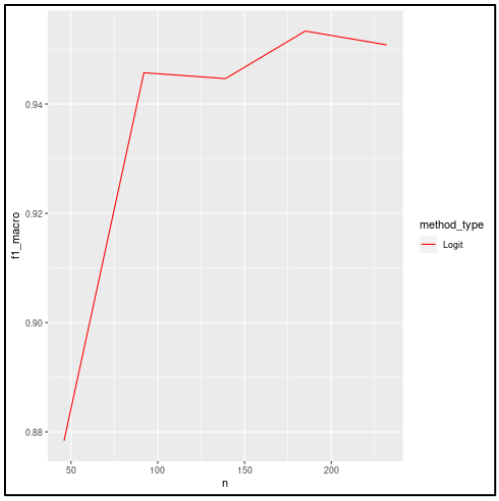
分析に使用するデータ n を増やしながら、One vs All と Softmax 関数を使用したそれぞれの方法について F1 スコアの推移をグラフにした。なお、Breast Cancer、U.S. Voting、Sonar は目的変数のクラスが 2 値であるが、スポーツチームの顧客データは目的変数のクラスが 5 つである為、グラフに示している F1 スコアは F1 macro 平均の値を記載している。

Brest cancer



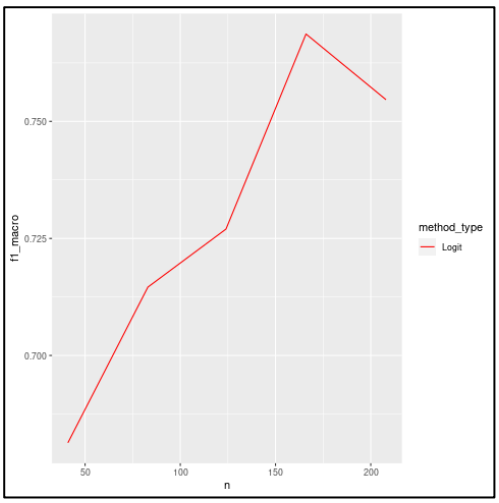
パラメータとサンプル数との比：0.015

U.S. Voting



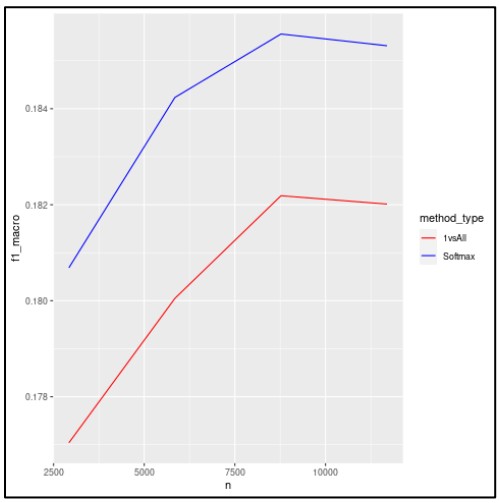
パラメータとサンプル数との比：0.073

Sonar



パラメータとサンプル数との比：0.293

スポーツチームの顧客データ



パラメータとサンプル数との比：0.0016

パラメータ数とサンプル数の比が比較的小さい Breast Cancer、U.S Voting データセットは F1 スコアが高く、パラメータ数とサンプル数の比が比較的大きい Sonar では F1 スコアが Breast Cancer、U.S Voting と比べて高い。スポーツチームの顧客データはパラメータのサンプルの比が最も小さいものの、エラー率について述べたような理由で、他のデータセットと比べて高い。One vs All と Softmax 関数を使用した２つの方法に関しては大きな差は見られなかった。

＜付録：オーバーサンプリングによる検証＞

ロジスティック回帰での検証結果：偏りのあるデータに対しての分析

スポーツチームの顧客データでは、各クラス内に含まれるデータ数は、以下の通りである。
クラス１：336、クラス２：6,397、クラス３：1,826、クラス４：1,262、クラス５：4,799。
クラス２とクラス５にデータ数が偏っているのが分かる。

全てのデータをもとに識別をおこなった際の混同行列は以下のとおりである。act_x は実際のクラス、pred_x は予測したクラスである。

One vs All：混同行列

	pred_1	pred_2	pred_3	pred_4	pred_5
act_1	0	41	0	0	14
act_2	0	804	0	0	158
act_3	0	202	0	0	62
act_4	0	149	0	0	44
act_5	0	534	0	0	185

Softmax：混同行列

	pred_1	pred_2	pred_3	pred_4	pred_5
act_1	0	44	0	0	4
act_2	0	777	0	0	183
act_3	0	214	0	0	76
act_4	0	118	0	0	69
act_5	0	487	0	0	221

予測値はクラス２とクラス５に集中しており、クラス１、クラス３、クラス４と予測されたデータはゼロであった。クラス内に含まれるデータの偏りが予測値に影響を及ぼしていることが分かる。

オーバーサンプリングの手法：SMOTE

SMOTE（Synthetic Minority Oversampling Technique）はオーバーサンプリング手法の１つであり、少数のクラスを以下の手順で生成し増やす手法である。

1. 少数のクラスからある特定のデータを１つ選ぶ。
2. K 個の近傍データが選択される。（K は任意のパラメータ）
3. 選ばれた K 個のデータからランダムに１つのデータが選択される。
4. 1 で選ばれたデータと 3 で選ばれたデータの中間に新しいデータを生成する。
5. 1 から 4 を繰り返すことで少数のクラスのデータを増やす。

ロジスティック回帰での検証結果：オーバーサンプリングの効果について

偏りがあるデータに対して、今回の検証では SMOTE でのオーバーサンプリングを行いクラス内に含まれるデータ数を均一にして分析を行った。すべてのデータを使った場合の混同行列が以下の表である。予測値は特定のクラスのみに集中することはなく、全体的にバラついているのが分かる。

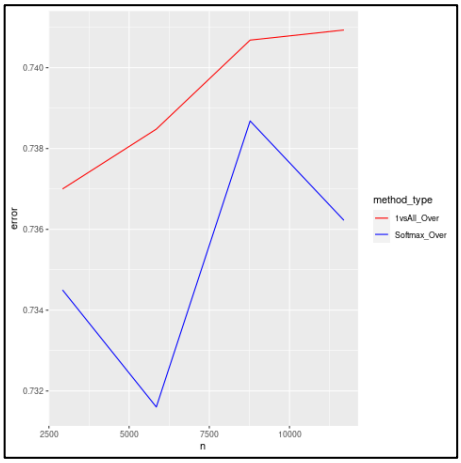
One vs All：混同行列

	pred_1	pred_2	pred_3	pred_4	pred_5
act_1	225	290	49	214	167
act_2	167	450	38	206	140
act_3	152	285	56	300	134
act_4	180	305	50	326	103
act_5	153	343	45	224	196

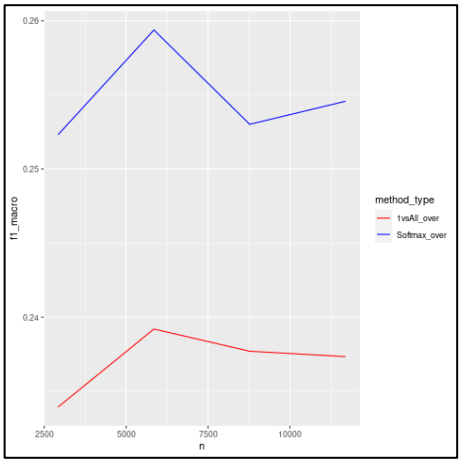
Softmax：混同行列

	pred_1	pred_2	pred_3	pred_4	pred_5
act_1	141	304	197	147	169
act_2	128	415	182	104	165
act_3	77	308	265	199	125
act_4	127	251	196	224	165
act_5	110	289	155	145	210

エラー率



F1macro スコア



第二節

サポートベクターマシン（Support Vector Machine）：略称　SVM

SVM は、対象のデータを正しいクラスに分ける境界を引く、ということが基本的な考え方になる。すなわち、データが境界線のどちら側にあるかで識別を行う。SVM では、このデータを分ける境界を分離超平面と呼び、データを正しいクラスに分ける最適な分離超平面を求めることが SVM の目的となる。分離超平面から最も近い位置に存在するデータをサポートベクターと呼び、このサポートベクターから分離超平面までの距離をマージンと呼ぶ。このマージンが最大の距離になるような分離超平面を決めるようにデータの学習を行う為、このアルゴリズムはサポートベクターマシンと呼ばれている。

SVM：カーネル関数について

データが線形の境界を引くことができない状態の場合、データを線形の境界で分けることができる形に変換する必要がある。データを線形の境界で分けることができる高次元の形に変換した上で境界を引く方法の一つが、カーネル法である。カーネル関数によって、各データをより高次元（場合によっては、無限次元の）関数空間に写すが、実際の計算は、カーネルトリックによって計算量低下を実現させている。SVM ではカーネル関数として、多項式カーネル(polynomial)、動径基底関数(RBF)、シグモイドカーネル関数(Sigmoid)がよく使用される。

SVM での識別方法と今回の検証に使用したカーネル関数について

k 個の多クラス識別を行う SVM ではデータが特定の 1 つのクラスに含まれるのか、その他のクラスに含まれるのかを識別する識別機を k クラス分作成する One vs Rest 識別と、クラスの異なるペアごとに個別の識別器を $k(k-1)/2$ 個作り、その多数決で識別されるクラスが決定する One vs One 識別の二種類が使われることが多い。今回の検証では RBF カーネル関数を使用し、One vs Rest と One vs One の両方の識別方法を比較した。またハイパーパラメータの調整は accuracy が最大になるものを選択する方法と、f1_score が最大になるものを選択する方法の 2 種類があり、その両方をそれぞれグリッドサーチによって行った。

カーネル関数の式
多項式カーネル(polynomial)：dは次数、pは説明変数の数を表す。

$$K(x_i, x_{i'}) = (1 + \sum_{j=1}^{p} x_{ij} x_{i'j})^d$$

動径基底関数(RBF)：γは正の定数であり、小さいほど単純な境界になり、大きいほど複雑な境界となる。

$$K(x_i, x_{i'}) = exp(-\gamma \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2)$$

シグモイドカーネル関数(Sigmoid)：αは重み、cは正の定数を表す。

$$K(x_i, x_{i'}) = tanh(\alpha x_i^T x_{i'} + c)$$

今回は、カーネル関数に動径基底関数を使用している為、モデル自体の次元は、無限次元となるが、実質的にはサンプル数の数だけのパラメーターでモデルが決まる（カーネルトリック）、すなわち、$P = N$となる。ハイパーパラメータは境界の複雑さを決める　γ　と SVM でどの程度誤識別を許容するのかを決めるハイパーパラメータ Cの２つとなる。Cが小さい場合、 SVM のマージンが広がり誤識別とされるデータは増える。逆に Cが大きい場合、SVM のマージンは狭くなり誤識別とされるデータは少なくなる。

データセット別のパラメータ数一覧：SVM

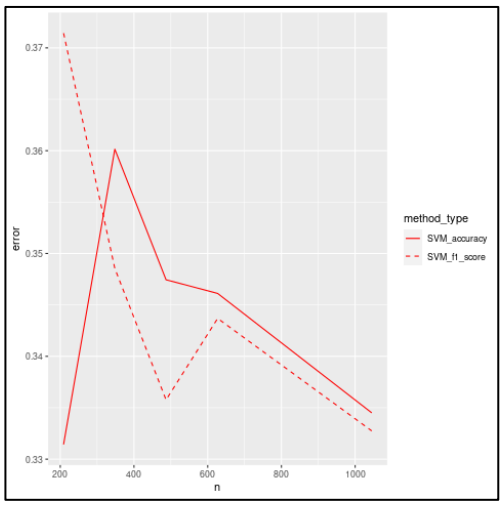
データセット名 および サンプル数 N	説明変数の数 M	クラス Y の数 K	SVM： パラメータ数	SVM： パラメータ数と サンプル数の比
Breast Cancer：683	9	2	683	1
U.S.Voting：232	16	2	232	1
Sonar：208	60	2	208	1
スポーツチームの 顧客データ：14,620	5	5	14,620	1

ハイパーパラメータの候補について：グリッドサーチ
ハイパーパラメータ Cと　γ　はグリッドサーチにて最適化を行った。以下はパラメータの候補である。
Cの候補：[0.001, 0.01, 0.1, 1, 10, 100]
γ の候補：[0.001, 0.01, 0.1, 1, 10, 100]

SVM での検証結果：エラー率

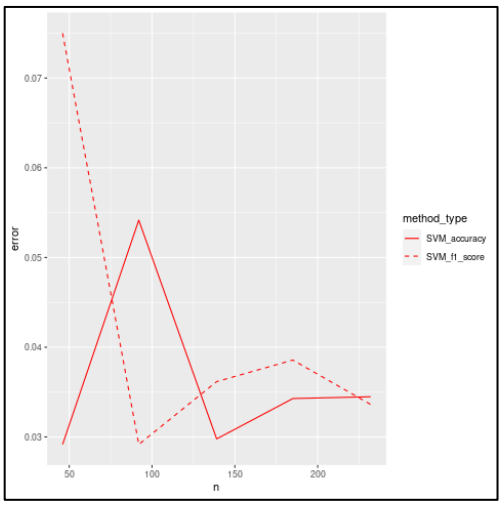
学習に使用するデータ n を増やしながらエラー率の推移をグラフにした。ハイパーパラメータの C および γ をグリッドサーチにて調整する際の基準として accuracy(正答率)を最大にする方法と F1 スコアを最大にする方法のそれぞれで検証を行っている。またスポーツチームの顧客データは 5 クラスである為、One vs Rest と One vs One のそれぞれの方法について accuracy(正答率)を最大にするハイパーパラメータ調整方法を使用して検証を行った。グラフの縦軸がエラー率、横軸が分析に使用するデータ数である。

Breast cancer



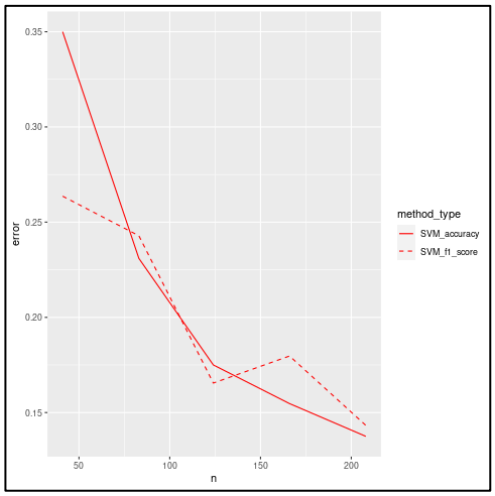
パラメータとサンプル数との比：1

U.S. Voting



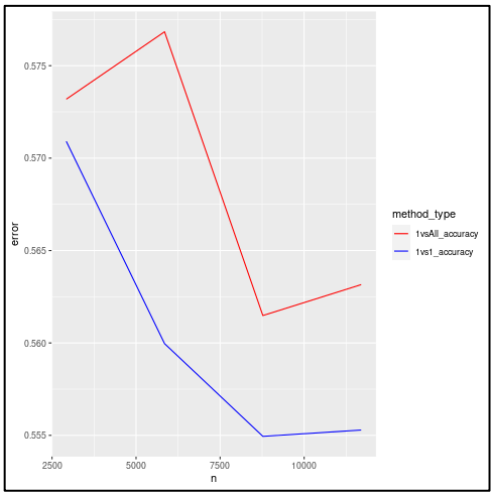
パラメータとサンプル数との比：1

Sonar



パラメータとサンプル数との比：1

スポーツチームの顧客データ



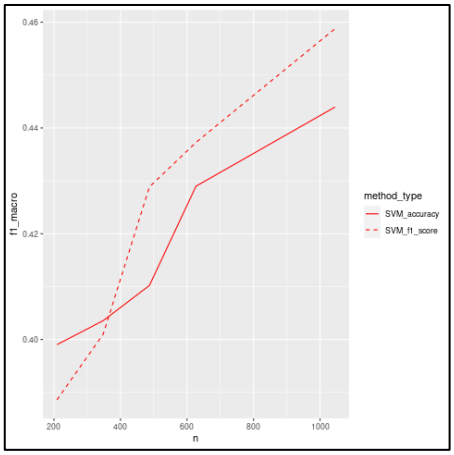
パラメータとサンプル数との比：1

Breast Cancer、U.S.Voting、Sonar の 3 つのデータセットに対して、ハイパーパラメータをグリッドサーチにて accuracy(正答率)を最大にする場合と F1 スコアを最大にする場合では大きな差はなかった。各データセットのパラメータとサンプル数との比に差はないが、各データセットのエラー率は異なっている。U.S.Voting のエラー率が最も低く、続いて Sonar、Breast Cancer、スポーツチームの顧客データの順にエラー率が低い。

SVM での検証結果：F1 スコア

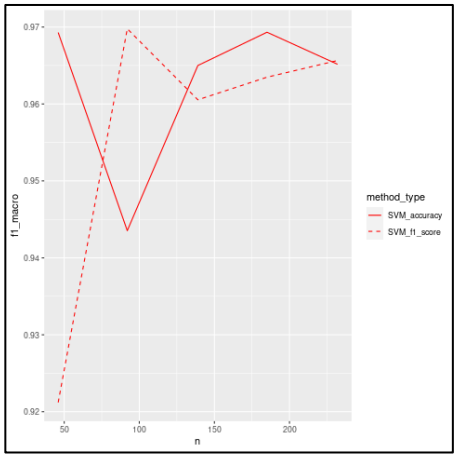
分析に使用するデータ n を増やしながら F1 スコアの推移をグラフにした。ハイパーパラメータの C およびγをグリッドサーチにて調整する際の基準として accuracy(正答率)を最大にする方法と F1 スコアを最大にする方法のそれぞれで検証を行っている。またスポーツチームの顧客データは 5 クラスである為、One vs Rest と One vs One のそれぞれの方法について検証を行っており、グラフに示している F1 スコアは F1 macro 平均の値を記載している。なおハイパーパラメータの調整は accuracy(正答率)を最大にする方法を使用した。グラフの縦軸が F 1 スコア、横軸が分析に使用するデータ数である。

Breast cancer



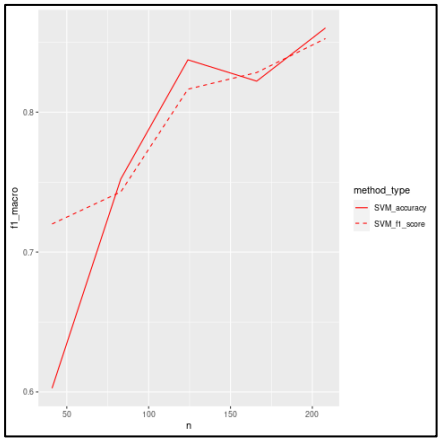
パラメータとサンプル数との比：1

U.S. Voting



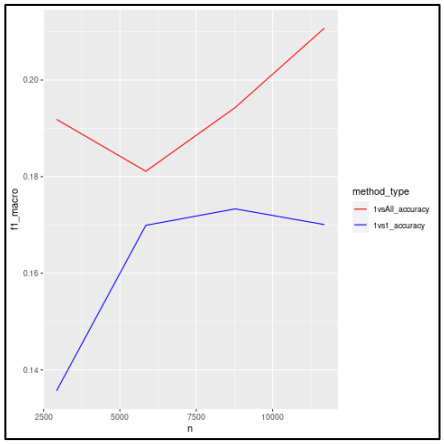
パラメータとサンプル数との比：1

Sonar



パラメータとサンプル数との比：1

スポーツチームの顧客データ



パラメータとサンプル数との比：1

各データセットのパラメータとサンプル数との比に差はないが、各データセットの F1 スコアは異なっている。U.S.Voting の F 1 スコアが最も高く、続いて Sonar、Breast Cancer、スポーツチームの顧客データの順に F1 スコアが高い。Breast Cancer、U.S.Voting、Sonar の 3 つのデータセットに対して、ハイパーパラメータグリッドサーチにて探す場合、accuracy(正答率)を最大にする場合と F1 スコアを最大にする場合では大きな差はなかった。

＜付録：オーバーサンプリングによる検証＞

スポーツチームの顧客データはクラス内に含まれるデータ数に偏りがあることから、One vs Rest と One vs One のそれぞれの方法でクラスごとの F1 を確認した。なお、ハイパーパラメータの選択は accuracy(正答率)を最大にする方法を用いている。

One vs All：混同行列

	pred_1	pred_2	pred_3	pred_4	pred_5
act_1	2	53	2	1	16
act_2	10	942	33	33	231
act_3	6	257	21	14	89
act_4	2	179	13	11	52
act_5	13	579	40	31	294

One vs One：混同行列

	pred_1	pred_2	pred_3	pred_4	pred_5
act_1	0	51	0	0	10
act_2	0	1143	0	0	127
act_3	0	305	0	0	46
act_4	0	233	0	0	40
act_5	0	807	0	0	162

予測値はクラス 2 とクラス 5 に集中している。One vs All ではわずかにクラス 3 とクラス 4 も予測がされていが、One sv One ではクラス 2 とクラス 5 のみを予測している。この結果からクラス内に含まれるデータの偏りが予測値に影響を及ぼしていることが分かる。

SVM での検証結果：オーバーサンプリングの効果について

SMOTE でのオーバーサンプリングを行いクラス内に含まれるデータ数を均一にして分析を行った。予測値は特定のクラスのみに集中することはなく、全体的にバラついている。偏りのあるデータに関して、一定の効果があることがわかった。

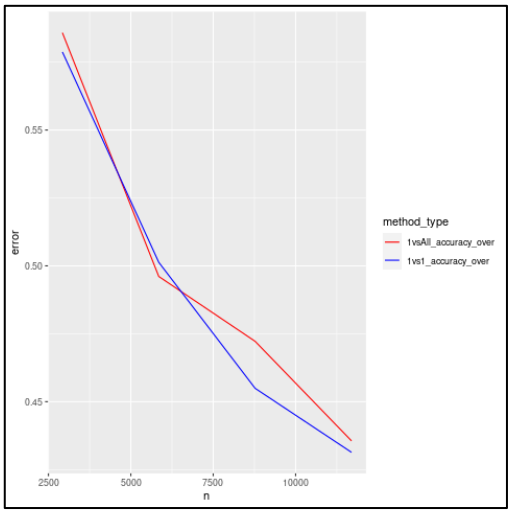
One vs All：混同行列
(accuracy を基準にパラメータを調整)

	pred_1	pred_2	pred_3	pred_4	pred_5
act_1	995	166	52	37	46
act_2	121	549	194	148	276
act_3	68	193	822	93	82
act_4	79	201	98	859	92
act_5	98	339	155	136	498

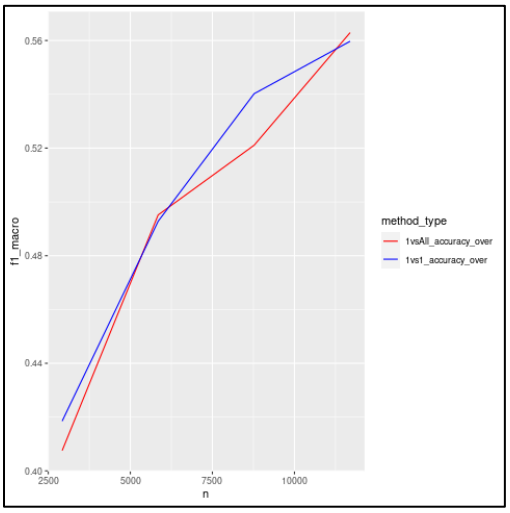
One vs One：混同行列
(accuracy を基準にパラメータを調整)

	pred_1	pred_2	pred_3	pred_4	pred_5
act_1	1044	104	44	48	58
act_2	140	444	226	181	332
act_3	84	167	791	104	110
act_4	87	112	102	847	106
act_5	122	331	168	163	482

エラー率



F1macro スコア



SVM での検証結果：グリッドサーチで選択されたハイパーパラメータについて

ランダムサンプリングごとに最適な C と　γ を探索した。サンプル数 n を変えながら、一つの n ごとにランダムサンプリング 3 回づつ行っている。以下の表はスポーツチームの顧客データでの、選ばれたハイパーパラメータの値である。値が大きいほど境界の複雑さを決める　γ（gamma）と、値が大きいほど SVM のマージンが狭まり誤識別とされるデータが減るパラメータ C があり、一方が大きくなると、もう一方が小さくなる関係であることが分かる。

One vs All スコア

n : 11696	C: 10	gamma: 1
	C: 10	gamma: 1
	C: 10	gamma: 1
n : 8772	C: 100	gamma: 0.1
	C: 10	gamma: 1
	C: 100	gamma: 0.1
n : 5848	C: 100	gamma: 0.1
	C: 100	gamma: 0.1
	C: 100	gamma: 0.1
n : 2924	C: 10	gamma: 0.1
	C: 0.1	gamma: 1
	C: 100	gamma: 0.1

One vs One

n : 11696	C: 100	gamma: 0.1
	C: 100	gamma: 0.1
	C: 100	gamma: 0.1
n : 8772	C: 100	gamma: 0.1
	C: 100	gamma: 0.1
	C: 100	gamma: 0.1
n : 5848	C: 100	gamma: 0.1
	C: 100	gamma: 0.1
	C: 100	gamma: 0.1
n : 2924	C: 100	gamma: 0.1
	C: 100	gamma: 0.1
	C: 100	gamma: 0.1

第三節

ディープニューラルネットワーク(Deep Neural Network)：略称　DNN

DNN は、入力層、隠れ層、出力層の３種類の層で構成される。今回は、出力が確率であるタイプを用いている。例えば、説明変数のデータ X の数が５つでクラス Y 数が３つの場合、入力層はこの５つの情報が与えられ、出力層では３つのクラスのうちそれぞれに属する確率が出力される。

最適化アルゴリズムについて

予測値と正解の値の差である損失を最小にするパラメータを効率的に求める最適化アルゴリズムについてふれる。

１）最急降下法

最適化アルゴリズムとして基本的なものに最急降下法があり、これは損失関数が最小になるように勾配を利用してパラメータを更新していく。勾配の谷の底に近づくにつれ勾配が小さくなり、底に到達すると傾きがゼロとなる為、勾配が小さくなるとパラメータの更新が完了したとみなされる。この最急降下法には欠点があり、最も深い位置にある谷の底が目指すべき最適な解であるものの、最適な解ではなくかつ勾配がゼロになる極小値と呼ばれる場所を目指してパラメータが更新されていく場合がある。極小値を目指してパラメータの更新が行われると、最適な解にたどり着く前に更新が終了してしまう。最急降下法の式は、以下の様になる。

w_t：t 回目の更新パラメータ、α　：学習率（変化の幅）、∇_w　：パラメータでの微分、$\mathcal{L}(w)$　：損失関数

$$w_{t+1} = w_t - \alpha \nabla_w \mathcal{L}(w)$$

２）確率的勾配降下法（Stochastic Gradient Descent:略称 SGD）
極小値を目指してパラメータが更新される問題を解消したアルゴリズムが確率的勾配降下法（Stochastic Gradient Descent:略称 SGD）である。最急降下法は一回のパラメータ更新を全てのデータを使って行う為、一度極小値に向かってパラメータが更新されるとそこから抜け出せなくなる。それに対して SGD はランダムに選ばれた１つのデータを使ってパラメータの更新を一回行い、次のパラメータ更新には別のランダムで選ばれたデータを使用する。あるデータで極小値を目指してパラメータの更新が行われても、別のデータでパラメータの更新が行われる為、たとえ極小値でパラメータの更新が終了しても別のデータのパラメータ更新は最適解を目指して継続される。SGD は逐次的にパラメータ更新が行われるが、複数のランダムに選ばれたデータを使用して並列的にパラメータを更新できるように改良されたアルゴリズムにミニバッチ SGD がある。

３）adam
SGD でのパラメータ更新を効率化させたアルゴリズムが Adam である。更新の幅が小さすぎると、一度の更新でパラメータは僅かにしか変化しない。また更新の幅が大きすぎると、一回の更新で勾配の谷の底を通り越すことを繰り返す振動を起こしてしまい、最適なパラメータになかなかたどり着かない。SGD に対して振動を抑える仕組みであるモーメンタムと RMSProp を組み込んだアルゴリズムを Adam と呼ぶ。モーメンタムは大きな振動に対して移動平均をとることで、振動を緩やかにするものである。また、RMSProp は勾配の大きさに応じで学習率を調整するものである。今回の検証ではこの Adam を最適化アルゴリズムとして使用している。

活性化関数：Relu 関数
負の値が入力されると 0 を、0 か正の値が入力されると入力された値をそのまま出力として返す関数。DNN の中間層の活性化関数として利用される。

損失関数：多クラス交差エントロピー(categorical crossentropy)
損失関数とはニューラルネットワークの識別がどのくらい間違っているかを表す指標であり、そのひとつに交差エントロピー誤差　(cross entropy error)がある。これは多クラス識別問題の損失関数に用いられることから多クラス交差エントロピー(categorical crossentropy)とも呼ばれる。

$$E = -\sum_{k=1}^{K} t_k \, log_e \, y_k$$

y_kはニューラルネットワークの出力層の活性化関数に softmax 関数を用いた際の出力、t_kは正解ラベル、K は出力層の出力数(識別するクラスの数)を表す。softmax 関数の出力y_kは 0 から 1 の値であり、$k=1$ から $k=K$ の総和は 1 となる為、各出力y_k*は*入力されたデータが各クラス k のうちどれにあたるかを表す確率として考えることができる。t_k*は*正解ラベルとなるインデックスのみが 1、それ以外が 0 の One-hot 表現になっており、正解ラベルが 1 の時に対応するニューラルネットワークの出力y_kの自然対数が計算される。正解ラベルt_kと softmax 関数の出力y_k*の両方の値が*1 *のとき*、つまり正解ラベルが 1 であり対応する出力y_kの値が 1(クラスkの確率が 100%)*のときに*最小値の 0 となる。

DNN での検証に使用したニューラルネットワークの構成
使用する主な python ライブラリ：Keras
ニューラルネットワーク構成：入力層
活性化関数：Relu 関数
中間層：1 層　ユニット数：8 つ
出力層
出力層の活性化関数：softmax 関数
損失関数：categorical crossentropy
最適化アルゴリズム：Adam

実質的なパラメータの数は、$P = (M+1) \times 8 + 9 \times (K-1)$ となる。

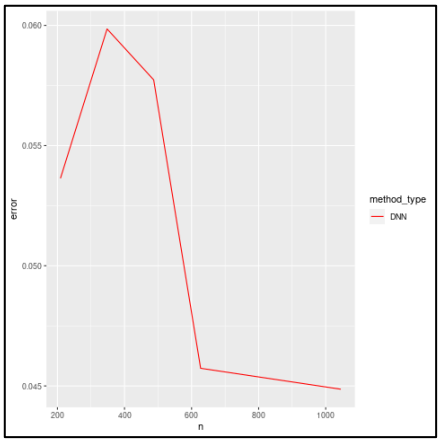
データセット別のパラメータ数一覧：DNN
パラメータ数をサンプル数で割り、パラメータ数とサンプル数の比を出している。

データセット名 および サンプル数 N	説明変数の数 M	クラス Y の数 K	DNN： パラメータ数	DNN： パラメータ数と サンプル数の比
Breast Cancer：683	9	2	89	0.130
U.S.Voting：232	16	2	145	0.625
Sonar：208	60	2	489	2.350
スポーツチームの 顧客データ：14,620	5	5	84	0.006

DNN での検証結果：エラー率

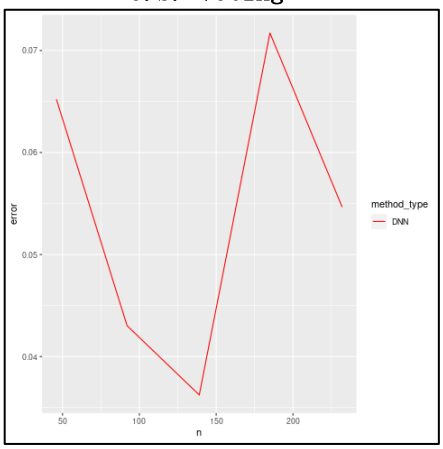
学習に使用するデータ n を増やしながら、DNN を使用した識別のエラー率の推移をグラフにした。縦軸がエラー率、横軸が分析に使用するデータ数である。

Breast cancer



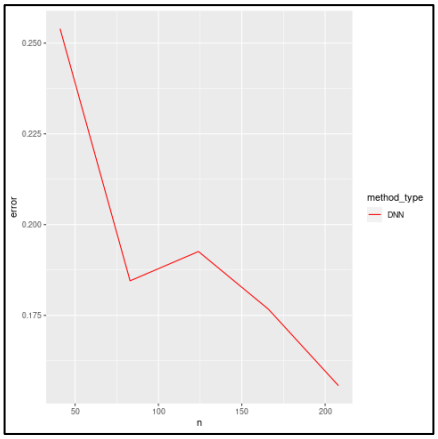
パラメータとサンプル数との比： 0.130

U.S. Voting



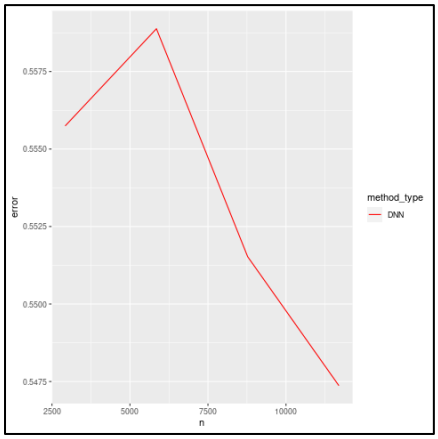
パラメータとサンプル数との比：0.625

Sonar



パラメータとサンプル数との比：2.350

スポーツチームの顧客データ



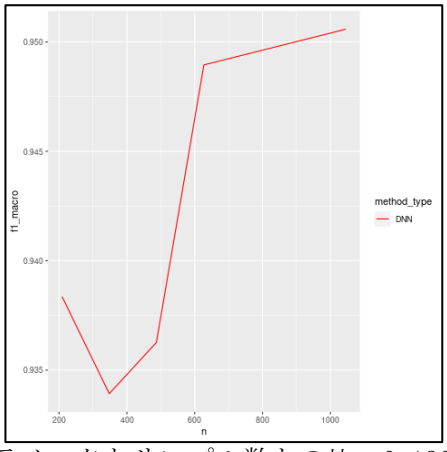
パラメータとサンプル数との比：0.006

パラメータとサンプル数の比が小さい Breast Cancer、U.S.Voting はエラー率も低く、パラメータとサンプル数の比が大きい Sonar は Breast Cancer と U.S.Voting と比べてエラー率が高い。スポーツチームの顧客データはパラメータとサンプル数の比が最も小さいもののエラー率は最も高い。これはクラス数が５つであるために誤識別の数が多くなっていることが理由として考えられる。

DNN での検証結果：F1 スコア

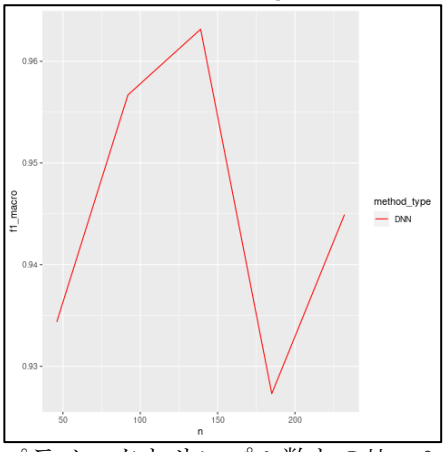
分析に使用するデータ n を増やしながら、F1 スコアの推移をグラフにした。なお、Breast Cancer、U.S. Voting、Sonar は目的変数のクラスが 2 値であるが、スポーツチームの顧客データは目的変数のクラスが 5 つである為、グラフに示している F1 スコアは F1 macro 平均の値を記載している。

Brest cancer



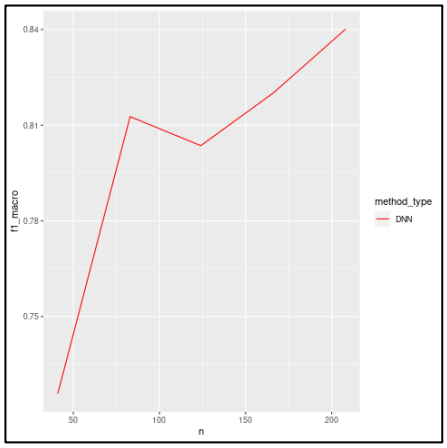
パラメータとサンプル数との比：0.130

U.S. Voting



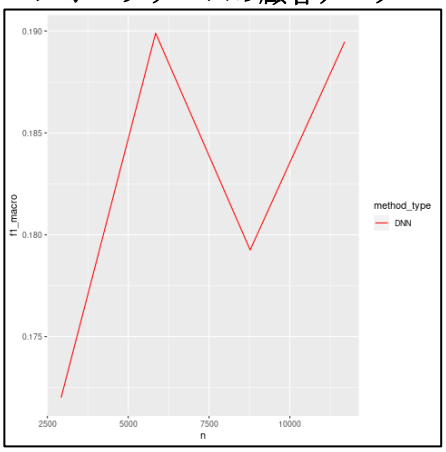
パラメータとサンプル数との比：0.625

Sonar



パラメータとサンプル数との比：2.350

スポーツチームの顧客データ



パラメータとサンプル数との比：0.006

パラメータ数とサンプル数の比が比較的小さい Breast Cancer、U.S Voting データセットは F1 スコアが高く、パラメータ数とサンプル数の比が比較的大きい Sonar では F1 スコアが Breast Cancer、U.S Voting と比べて低い。スポーツチームの顧客データはパラメータのサンプルの比が最も小さいものの、他のデータセットと比べて低い。これはスポーツチームの顧客データはクラス数が 5 つと多い為に他のデータセットと比べて識別が困難であることが理由として考えられる。

＜付録：オーバーサンプリングによる検証＞

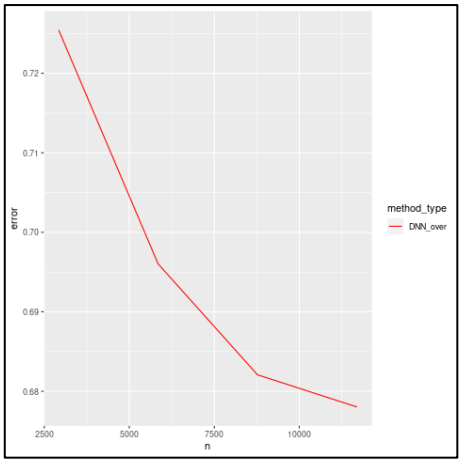
予測値はクラス２、クラス５に集中した。またわずかにクラス１に予測がされているがクラス 3、クラス４には予測がされなかった。

confusion matrix	pred_1	pred_2	pred_3	pred_4	pred_5
act_1	0	32	0	0	3
act_2	1	546	0	0	93
act_3	0	141	0	0	39
act_4	0	94	0	0	31
act_5	0	360	0	0	122

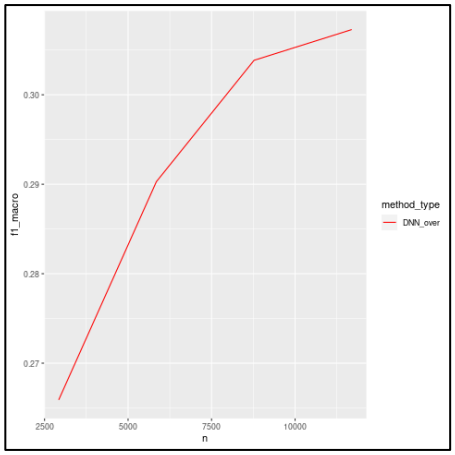
偏りがあるデータに対して、今回の検証では SMOTE でのオーバーサンプリングを行いクラス内に含まれるデータ数を均一にして分析を行った。予測値は特定のクラスのみに集中することはなく、全体的にバラついている。偏りのあるデータに関して、一定の効果があることがわかった。

confusion matrix	pred_1	pred_2	pred_3	pred_4	pred_5
act_1	277	31	157	126	44
act_2	99	112	229	131	77
act_3	89	58	231	217	45
act_4	105	50	165	253	64
act_5	110	82	196	154	96

エラー率



F1macro スコア



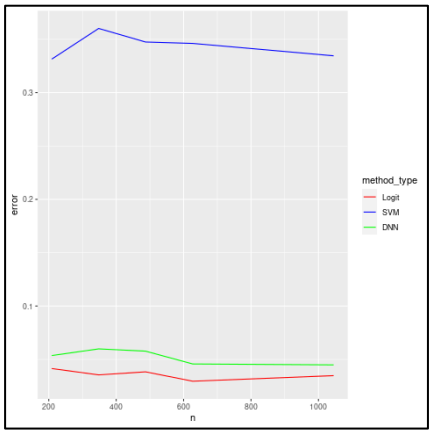
第四節

識別モデルの比較

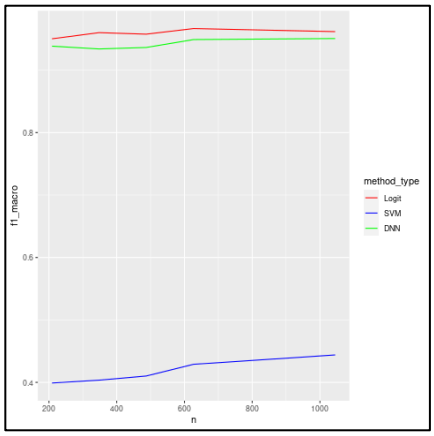
識別モデルであるロジスティック回帰、SVM、DNN の分析結果をエラー率と F1 スコアの２つの基準で比較を行った。ロジスティック関数は 1vsAll にて予測値を決める方法、SVM はハイパーパラメータを C と γ (gamma)を正答率(accuracy)を最大にするものを選択し 1vsAll にて予測値を決める方法でそれぞれ検証をした結果をグラフに記載している。

識別モデルの結果

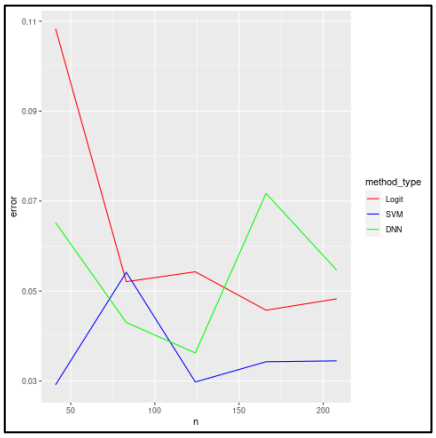
Breast cancer：エラー率



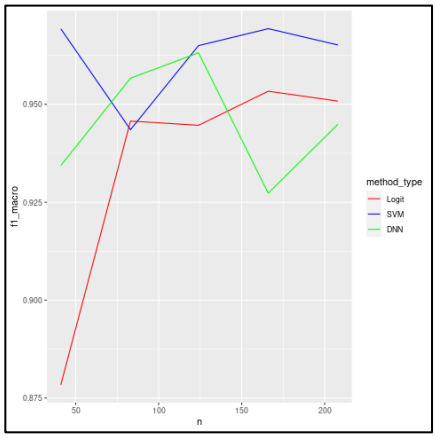
Breast cancer：F1 スコア



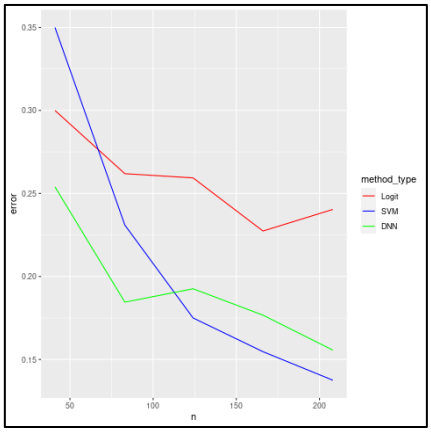
U.S. Voting：エラー率



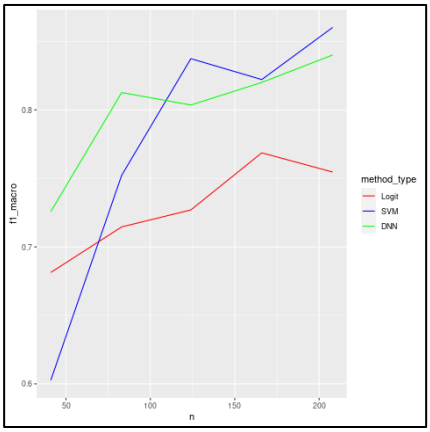
U.S. Voting：F1 スコア



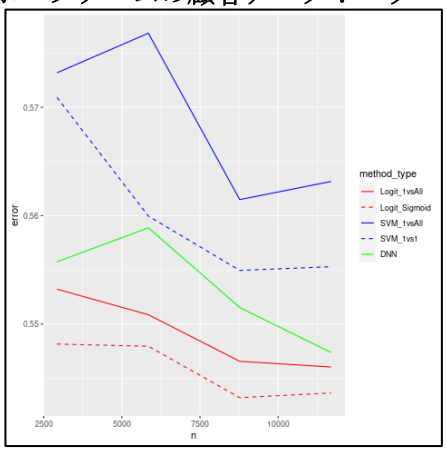
Sonar：エラー率



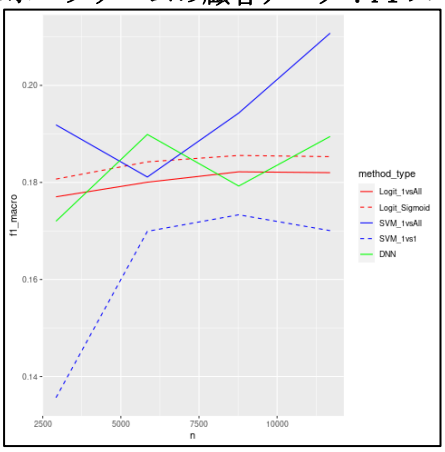
Sonar：F1 スコア



スポーツチームの顧客データ：エラー率



スポーツチームの顧客データ：F1 スコア



データセット名 および サンプル数 N	ロジスティック回帰： パラメータ数と サンプル数の比	SVM： パラメータ数と サンプル数の比	DNN： パラメータ数と サンプル数の比
Breast Cancer：683	0.015	1	0.015
U.S.Vote：232	0.073	1	0.625
Sonar：208	0.293	1	2.350
スポーツチームの 顧客データ：14,620	0.0016	1	0.006

最もエラー率が低いモデル、および最も F1 スコアが高いモデルはデータセットごとに異なっている。クラス内のデータ数に偏りがないデータセットである Breast Cancer、U.S.Voting、Sonar データはエラー率が低いモデルは、同様に F1 スコアが高いモデルである。しかし、クラス内のデータ数に偏りがあるスポーツチームの顧客データでは、SVM と 1vsAll の組み合わせのモデルはエラー率が他のモデルと比較して高いものの、F1 スコアは最も高い。全てのモデルにおいて特定のデータ数が偏ったクラスに予測値が集中することでエラー率が低下しているが、SVM と 1vsAll の組み合わせのモデルはその中では予測値が若干バラついていることから、エラー率は高くなっているものの F1 スコアのマクロ平均が高くなっている。

第三章

生成モデル（Generative model）

今回の検証には生成モデルとして代表的なナイーブベイズ正規モデルを使用した。サンプルサイズが少ない場合、複雑なモデルにすると推定すべきパラメータの数が増える為、十分な学習を行うにはサンプルの数が足りなくなる場合がある。ナイーブベイズはシンプルでありパラメータの数が少ないため、サンプルサイズが少ない場合でも使用することができる。

生成モデルに関する説明の流れ

ナイーブベイズ を用いてのクラス Y の推定方法について説明する。その後、生成モデルについて識別モデルとの分析結果の比較を通してその特徴を考察していく。

生成モデル（逆推定）：ナイーブベイズ を用いてのクラス Y の推定方法

Y を条件としたときの X の確率モデルにおいて、X の各要素が独立であるという仮定をおくのが、ナイーブベイズモデルである。

$$P(X_1, X_2, \dots, X_M|Y) = P(X_1|Y)P(X_2|Y)\cdots P(X_M|Y)$$

次に $X_1, X_2, \dots, X_M$ に関する Y の条件付き分布をベイズの定理を用いて構築する。

$$P(Y|X_1, X_2, \dots, X_9) = \frac{P(X_1|Y)\cdots P(X_M|Y)P(Y)}{P(X_1, X_2, \dots, X_M)}$$

この Y の確率を比べ、最も高い確率の Y を予測値とする。比較の際には、上式右辺の分母の計算は不要である点が重要である。

Y を与えたときに X の各要素が独立になるという設定はかなり厳しい仮定であり、現実のデータではこれが満たされることはほとんどないにも関わらず、ナイーブベイズ生成モデルが優れた「識別」パフォーマンスを示すことが多い。Zhang(2005)が指摘しているように、本来存在しているはずの X の要素間の相関が、Y のクラス間であまり違わないような状況では、これを無視しても識別に大きな影響を与えないことが一つの理由として考えられる。一方、ナイーブベイズが、X の共分散構造の次元を低く抑えていることで（いわゆる「次元の呪い」を受けない）、サンプル数が少数でも良いパラメーター推定につながっていることも、また一つの原因と思われる。

生成モデル（逆推定）： ナイーブベイズ正規モデル

説明変数データ X の、Y が与えられた時の条件付き分布は、ナイーブベイズの仮定から全て独立である。今回は正規モデル正規分布を仮定しており、個々の説明変数は、

$$X \sim N(\alpha + \beta Y, \sigma^2)$$

という形になる。つまり、各説明変数Xの分布は、与えられた Y によって平均が変化する（分散はクラス間で共通）ことになる。推定をするパラメータは、X の各要素につき、切片 α、係数 β、分散 σ の３つである。事前分布としては、MCMC が収束しやすいように、以下のようなコーシー分布を利用している。

$$\alpha \sim cauchy(0,2.5)$$
$$\beta \sim cauchy(0,2.5)$$
$$\sigma \sim cauchy(0,2.5)$$

モデルのパラメーターの数は、$p = 3 \times M + K - 1$ となる。*K-1* の部分は、Y のクラスが K 個あり、その多項分布としてのパラメーターの数が、*K-1* であるからである。

データセット別のパラメータ数一覧：ナイーブベイズ
パラメータ数をサンプル数で割り、パラメータ数とサンプル数の比を出している。

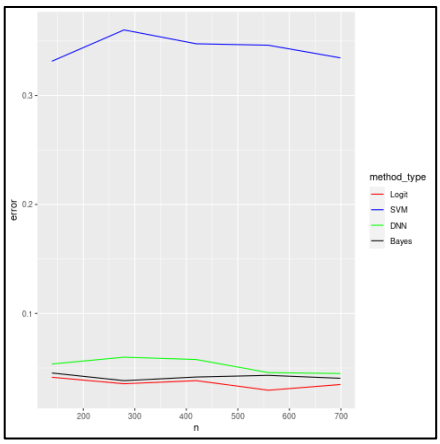
データセット名 および サンプル数 N	説明変数の数 M	クラス Y の数 K	ナイーブベイズ ： パラメータ数	ナイーブベイズ ： パラメータ数と サンプル数の比
Breast Cancer : 683	9	2	28	0.041
U.S.Voting : 232	16	2	49	0.211
Sonar : 208	60	2	181	0.870
スポーツチームの 顧客データ : 14,620	5	3	17	0.001

第四章

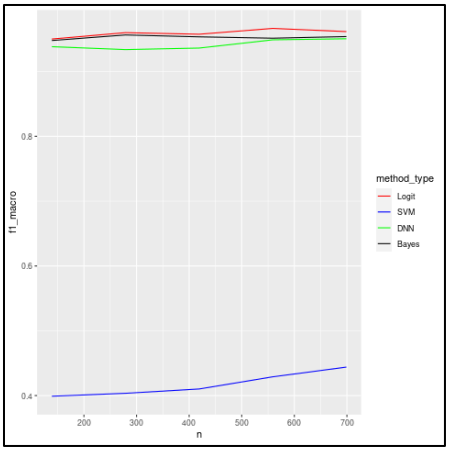
生成モデルと識別モデルの比較

識別モデルであるロジスティック回帰、SVM、DNN の分析結果および生成モデルであるナイーブベイズでの分析結果をエラー率と F1 スコアの２つの基準で比較を行った。ロジスティック関数は 1vsAll にて予測値を決める方法、SVM はハイパーパラメータを C と γ (gamma)を正答率(accuracy)を最大にするものを選択し 1vsAll にて予測値を決める方法でそれぞれ検証をした結果をグラフに記載している。

Breast cancer：エラー率

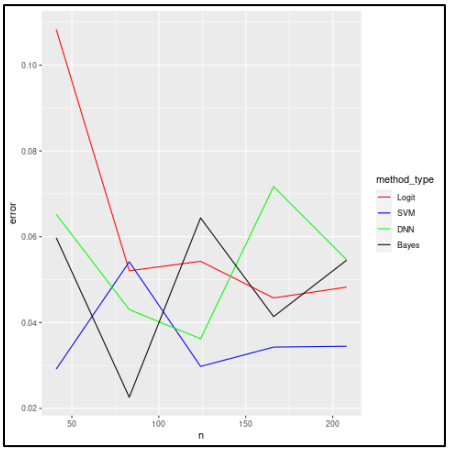


Breast cancer：F1 スコア

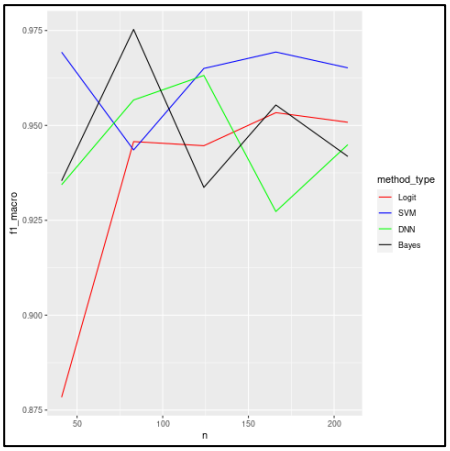


パラメータとサンプル数との比
ナイーブベイズ　：0.041

U.S. Voting：エラー率

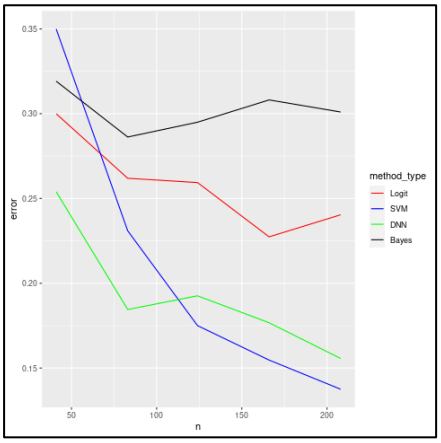


U.S. Voting：F1 スコア



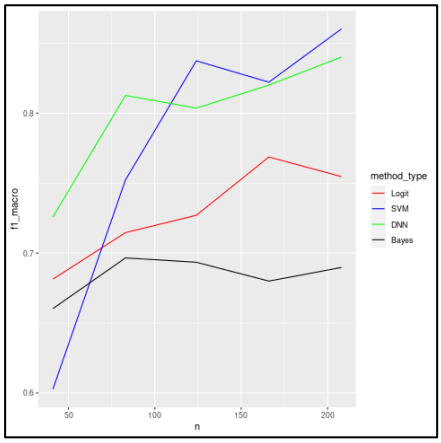
パラメータとサンプル数との比
ナイーブベイズ　：0.211

Sonar：エラー率

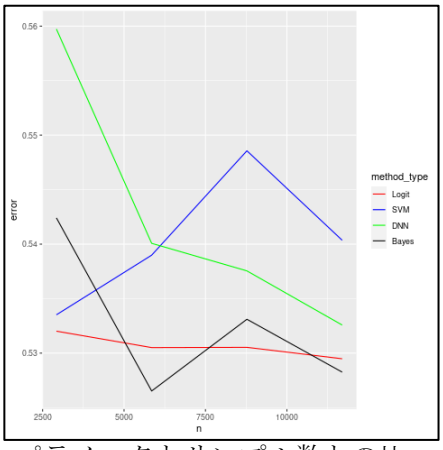


パラメータとサンプル数との比
ナイーブベイズ　：0.870

Sonar：F1 スコア

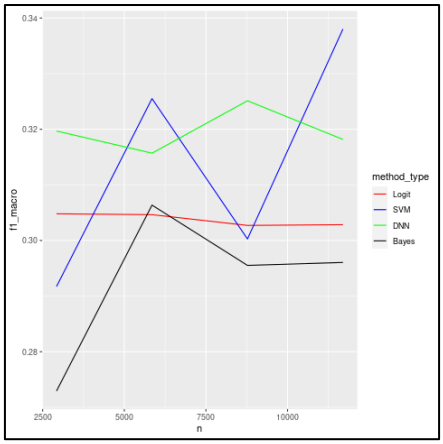


スポーツチームの顧客データ：エラー率



パラメータとサンプル数との比
ナイーブベイズ　：0.001

スポーツチームの顧客データ：F1 スコア



生成モデルであるナイーブベイズ のデータセットごとの識別精度の差異について

ナイーブベイズ はクラス間で共通の対角行列を想定している。データセットの各クラスの分散共分散行列が似ている場合、対角行列ではないもののクラスごとの分散構造が似ているという点ではナイーブベイズ の想定に近いと考えられることから、分散共分散行列の構造について「クラス間で分散共分散が似ているとナイーブベイズの性能の性能が良くなる」という指摘(Zhang(2005))がある。この指摘を検証する為、データセットごとの説明変数の相関、分散共分散行列をヒートマップ にして可視化、Box の M 検定での等質性の確認、固有値の分布の確認を行った。

Box の M 検定について

Box の M 検定は説明変数の分散共分散行列がグループ間（今回はクラス間）で等しいという帰無仮説を検定するものであり、分散共分散行列の等質性の検定と呼ばれる。 P 値が小さく有意であると、「比較したグループは同質である」とはいえないという結果になる。なお、この検定は多変量正規性に敏感であり、有意になりやすいといわれている。また、今回使った４つのデータセットは、いずれも数がそれなりに多いので、これも有意になりやすい原因となる。

固有値の分布の確認について

クラスごとの分散共分散行列の固有値の最小値、最大値の比を計算し、比較することでクラス間の分散共分散行列の構造の比較を行った。

各データセットの識別精度についての考察
エラー率および F1 値は $n = N$ の場合の値を記載している。

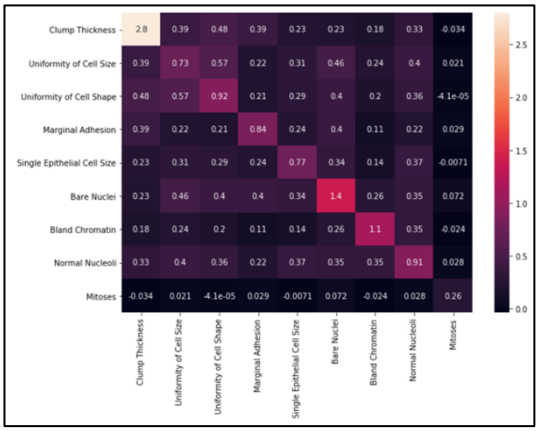
1）Breast Cancer

・各分析の値

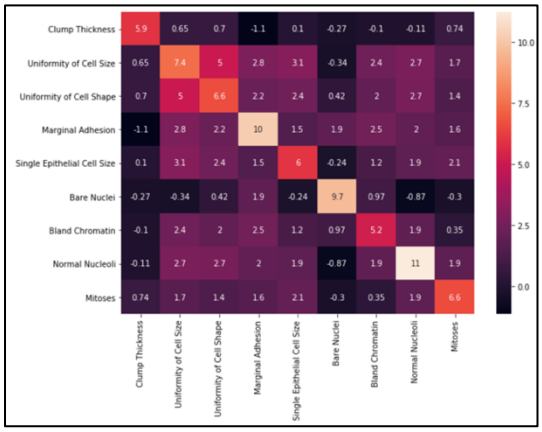
ロジスティック回帰のエラー率：0.035
ロジスティック回帰の F1 値：0.962
SVM のエラー率：0.335
SVM の F1 値：0.444
DNN のエラー率：0.045
DNN の F1 値：0.951
ナイーブベイズ のエラー率：0.039
ナイーブベイズ の F1_macro ： 0.96
ナイーブベイズのパラメータ数とサンプル数の比：0.040
Box の M 検定 P 値：0
Class 0 の分散共分散行列の固有値　最小値と最大値の比：0.060
Class 1 の分散共分散行列の固有値　最小値と最大値の比：0.088

・ナイーブベイズ生成モデルは、ロジスティック回帰や DNN と同様の識別性能を有している。
・SVM との比較では、かなり優位な識別性能を示している。
・以上の優劣の状態は、n の大小にかかわらず常に成立している。

Class 0 の説明変数（X）の分散共分散行列



Class1 の説明変数（X）の分散共分散行列



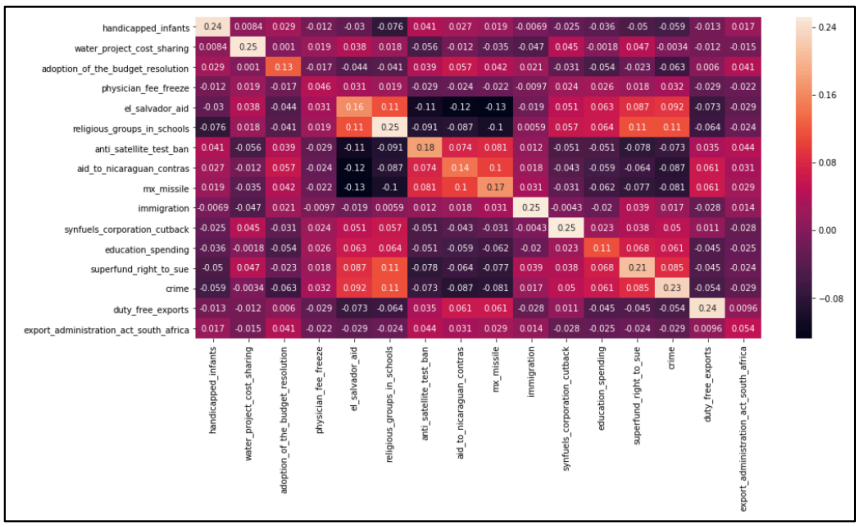
2）U.S.Voting

・各分析の値

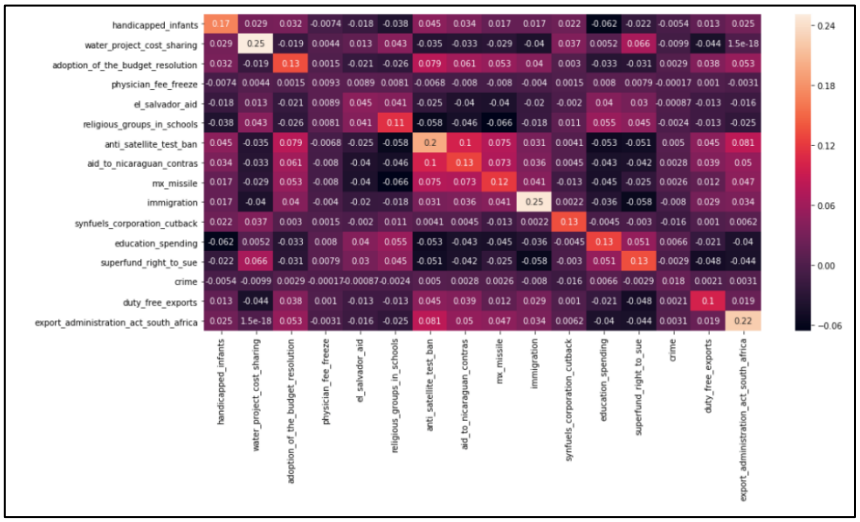
ロジスティック回帰のエラー率：0.048
ロジスティック回帰の F1 値：0.950
SVM のエラー率：0.035
SVM の F1 値：0.965
DNN のエラー率：0.055
DNN の F1 値：0.945
ナイーブベイズ のエラー率：0.055
ナイーブベイズ の F1_macro：0.94
ナイーブベイズのパラメータ数とサンプル数の比：0.211
Box の M 検定 P 値：2.13e-51
Class 0 の分散共分散行列の固有値　最小値と最大値の比：0.0189
Class 1 の分散共分散行列の固有値　最小値と最大値の比：0.0095

・このデータセットは、4 つの中で一番識別がしやすいデータであり、3 つの識別モデル、ベイズ生成モデル、すべてが低いエラー率、高い F1 値を残している。
・ベイズ生成モデルも、十分識別に使用できるレベルである。

Class0(民主党)の説明変数（X）の分散共分散行列



Class1(共和党)の説明変数（X）の分散共分散行列



3）Sonar

・各分析の値

ロジスティック回帰のエラー率：0.240

ロジスティック回帰の F1 値：0.755

SVM のエラー率：0.138

SVM の F1 値：0.860

DNN のエラー率：0.156

DNN の F1 値：0.840

ナイーブベイズ のエラー率：0.301

ナイーブベイズ の F1_macro：0.69

ナイーブベイズのパラメータ数とサンプル数の比：0.870

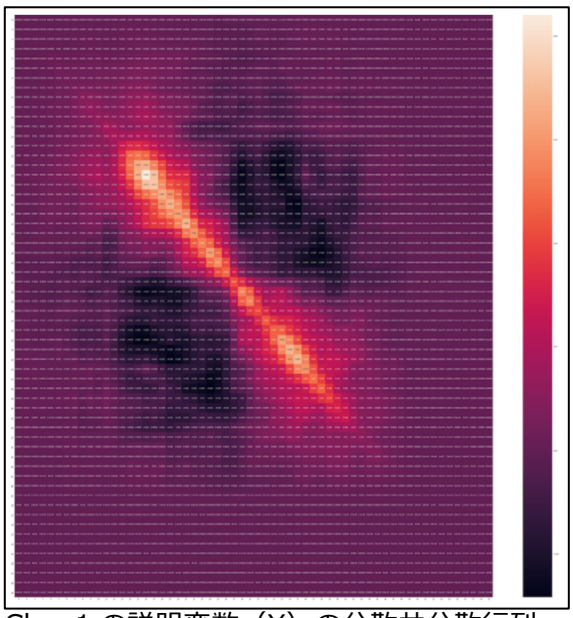
Box の M 検定 P 値：1.94e-72

Class 0 の分散共分散行列の固有値　最小値と最大値の比：0.0000032

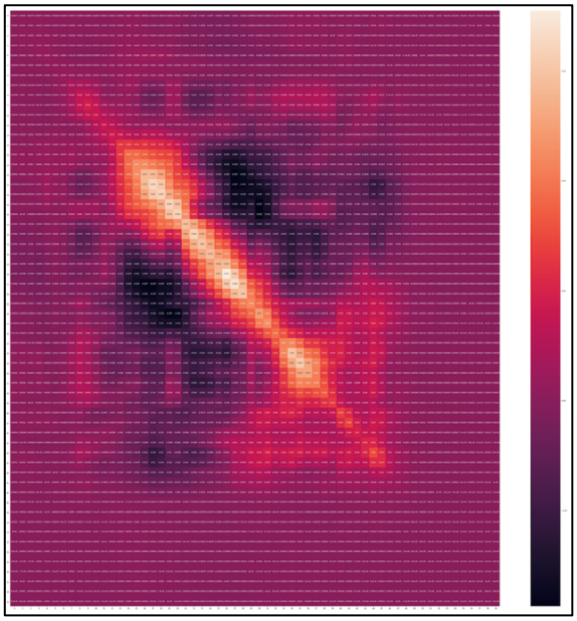
Class 1 の分散共分散行列の固有値　最小値と最大値の比：0.0000058

・このデータセットでは、明らかにナイーブベイズの識別能力が、識別モデルよりも劣っていることが分かる。

Class0 の説明変数（X）の分散共分散行列



Class1 の説明変数（X）の分散共分散行列



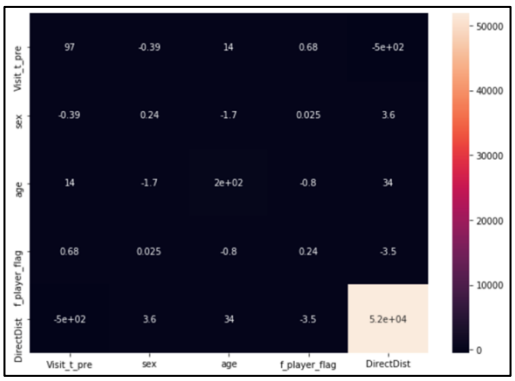
4）スポーツチームの顧客データ

・各分析の値

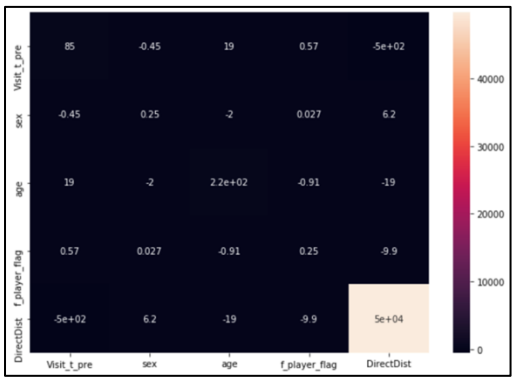
ロジスティック回帰のエラー率：0.529
ロジスティック回帰の F1 値：0.302
SVM のエラー率：0.540
SVM の F1 値：0.338
DNN のエラー率：0.533
DNN の F1 値：0.318
ナイーブベイズ のエラー率：0.528
ナイーブベイズ の F1_macro：0.30
ナイーブベイズのパラメータ数とサンプル数の比：0.001
Box の M 検定 P 値：3.94e-101
Class 1 の分散共分散行列の固有値　最小値と最大値の比：0.0000040
Class 2 の分散共分散行列の固有値　最小値と最大値の比：0.0000043
Class 3 の分散共分散行列の固有値　最小値と最大値の比：0.0000038

・ナイーブベイズ生成モデルは、エラー率に関して、３つの識別モデルと同程度、あるいは若干だが優れた識別性能を有している。
・Ｆ１値に関しては、識別モデルにやや劣った結果となっているが、それほど大きな差ではない。
・ Yのクラスが３つになっていることから、クラス間の分散共分散行列の類似が、２クラスの場合よりみたされにくくなっている。

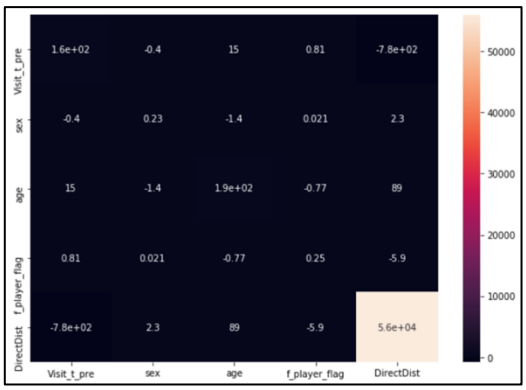
Class1 の説明変数（X）の分散共分散行列



Class2 の説明変数（X）の分散共分散行列



Class3 の説明変数（X）の分散共分散行列



Zhang の指摘の検証結果

指摘：クラス間で分散共分散が似ているとナイーブベイズの性能の識別性能は、識別モデルに近い。
１）Box の M 検定の結果では P 値は低く、仮説検定からは「分散共分散が同じ」という仮説は、すべてのデータで棄却される。
２）しかしながら、分散共分散をヒートマップで見ると、ある程度似ている。
３）また、固有根の構造（最小/最大）は、U.S.Voting データセット以外、クラス間でかなり似ている。
　今回使用した４つのデータセットでは、説明変数の分散共分散がクラス間で似ているため、ナイーブベイズが識別モデルと遜色ない識別能力をもたらした可能性はある。

４）ナイーブベイズモデルが識別モデルとの比較で一番識別能力で劣っていたのは、Sonar のデータセットの場合である。説明変数の数 60 が他のデータセットに比べて非常に多いため、各説明変数の(条件つき)独立性を仮定する、すなわち（$\binom{60}{2}$）通りの相関がすべてゼロであると仮定することに無理があることが一つの原因であると考えられる。

まとめ

１）生成モデル（ナイーブベイズ）でも、識別モデルと遜色ない識別性能を満たすことは、珍しくない。「学習サンプル数が少ない間は、生成モデルの方がよい識別成績を示し、サンプル数が徐々に増えると、識別モデルの方がパフォーマンスで上回る」という現象を Ng and Jordan（2001）が理論面（かなり大雑把な議論）・実証面で提示しているが、本研究では特にそのような傾向は見られなかった。サンプル数とは無関係に、生成モデルが良い識別能力を発揮する場合が、４つのデータセットいずれでも確認された。

２）クラスが 3 値以上となった場合では、生成モデルと識別モデルの優劣をエラー率でみるか、Ｆ１値でみるかによって、結果が違ってくる可能性が高いようにみえる。この部分については様々な条件でさらに比較検証をしていく必要がある。

３）クラス間で分散共分散が似ているとナイーブベイズの性能の性能が良いという Zhang(2005)の指摘は、今回のデータに関してはある程度当てはまっている。

参考文献

Andrew Y. NgI. JordanMichael. (2001). On Discriminative vs. Generative classifiers: A comparison of logistic regression and naive Bayes. University of California.

ChawlaV. et al.N. (2002). SMOTE: Synthetic Minority Over-Sampling TEchnique. Journal of Artificial Intelligence Research.

G.JamesT.Hastie, R.Tibshirani [訳]落海 治, 首藤 信通 D.Witten,. (2022). R による統計的学習入門 [原題]An Introduction to Statistical Learning with Applications in R. 朝倉書店.

Harry Zhang. (2005). Exploring Conditions For The Optimality Of Naïve Bayes. University of New Brunswick Fredericton, Faculty of Computer Science. nternational Journal of Pattern Recognition and Artificial Intelligence (2005).

William H. WolbergL. MangasarianOlivi. (1990). Multisurface method of pattern separation for medical diagnosis applied to breast cytology. Proc, Natl, Acad, Sci, USA.

康毅斎藤. (2018). ゼロから作る Deep Learning-Python で学ぶディープラーニングの理論と実装. ティム・オライリー.

馬場真哉. (2020). R と Stan ではじめる　ベイズ統計モデリングによるデータ分析入門. 株式会社　講談社.

本川 哲哉太郎手塚. (2019). ニューラルネットワークにおける適応的二次最適化手法. DEIM Forum 2019 A4-2.