＜ 修 士 論 文 ＞

識別モデルと生成モデルの比較について

（要　旨）

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

目的

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。

方法

識別モデルとしてロジスティック回帰、SVM、深層学習の３つのモデル、そして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つの異なるデータセットに適用し、その識別精度の比較を行った。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。

結果

生成モデル（ナイーブベイズモデル）は識別モデルと遜色ない識別性能であった。「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ず、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。

生成モデルであるナイーブベイズモデルのデータセットごとの識別精度の差異については、「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」という Zhang(2005)の指摘を検証した。データセットごとの説明変数の相関、分散共分散行列のヒートマップによる可視化、Box の M 検定での等質性の確認、固有値の分布の確認を行うことで、分散共分散構造のクラス間の類似度の検証を行った。今回のデータに関してはこの指摘がある程度当てはまっている、すなわち、クラス間で分散共分散構造が似ていることが、ナイーブベイズモデルの識別性能の良さにつながった可能性がある。