早稲田大学大学院 先進理工学研究科

博士論文審査報告書

論　文　題　目

データ分布空間における距離構造の学習に関する研究

A Study on Distance Metric Learning of Data Distribution Spaces

申　請　者

日野	英逸
Hideitsu	Hino

電気・情報生命専攻
情報学習システム研究

2010 年　7 月

統計的機械学習やデータマイニング手法はその適用範囲を広げ続けており、多種多様なデータから有用な情報を抽出する課題が日々生まれてきている。多くの学習アルゴリズムの性能は、入力データから抽出する情報の性質とデータの間に定義される距離構造に大きく依存している。これまで多くの判別手法が提案され、実問題に適用されて成果をあげているが、精度良く所望の結果を得られる手法は、与えられた問題・データから有用な情報を抽出するための適切な距離を定めることに成功している手法であると考えられる。本論文では、教師付き学習の枠組みの中で、データ間の距離構造の学習問題を扱っている。距離構造の学習としては、データの変換により適切な特徴空間を直接学習するアプローチの他に、データ同士の内積を学習するアプローチや、あるいはデータから学習される確率的な生成モデルに基づいて自然な距離構造を導出するというアプローチがある。本論文では、特徴空間の学習及びデータ間の内積の学習を統一する手法として、情報理論に基づく条件付きエントロピー最小化基準による方法を提案している。これにより、従来の距離構造の学習問題では十分に論じられていなかった、学習対象の情報論的な意味を明確に議論している。また、確率モデルに基づきデータ間に自然な距離構造を学習するアプローチの一つとして、近年重要性を増している映画や書籍等の評価データに対する生成モデルを提案し、その効率的な推定法と提案モデルに基づくデータ間の距離を導出している。以下に本論文の構成を示す。

まず第１章では導入として、本論文で対象とする距離構造の学習という問題の背景と動機を述べている。特に、データが連続量として与られる場合と離散量として与えられる場合それぞれについて、医療における病気の診断と、個人の嗜好に応じたアイテムの推薦サービスという問題を例としてあげている。上述の２例に代表される連続的あるいは離散的な観測データに対して、情報論的観点からの距離構造学習という立場からそれぞれ「データ同士の内積を目的に応じて適切に学習・定義するアプローチ」及び「データが発生する分布、すなわち生成モデルを学習し、モデルに基づく自然な距離構造を学習するアプローチ」をとることを説明している。

第２章では本論文で重要な役割を果たす Shannon の情報エントロピーをはじめとして、相互情報量、Kullback-Leibler ダイバージェンスといった基本的な量を導入している。また、観測データに基づくエントロピーの推定に関する従来の研究をまとめている。エントロピー推定は古くから研究が行われてきた重要な問題であり、これまでにも多くの手法が提案されているが、ここでは代表的な手法である Leave-One-Out 法、k 近傍法と、近年提案されたより効率的な手法である Mean-Nearest-Neighbor 法を紹介し、数値実験により Mean-Nearest-Neighbor

法の有効性を示している。

第3章では、情報論的な観点から教師付きの次元削減及び距離構造の学習に関する研究を紹介している。エントロピーを目的関数とした情報論的な距離構造の学習に関する研究は既に幾つか行われているが、その多くは Shannon のエントロピーを扱うことが困難であることから、その近似として Renyi エントロピーと呼ばれる量を最適化することが多いことを指摘している。また、データ同士の距離の学習と特徴空間における内積の学習が等価であることを述べ、特徴空間における内積を間接的に定めるカーネル関数を用いた手法の総称であるカーネル法の理論的な背景を説明している。更に、カーネル関数の設計に関する二つのアプローチとして、多数のカーネル関数の凸結合を考えてその結合係数を最適化する Multiple Kernel Learning と、データの確率的な生成モデルに基づいて定められるカーネル関数である Fisher カーネルを紹介している。

第4章では、条件付きエントロピー最小化基準に基づく教師付き距離構造学習の新たな枠組みを提案している。この枠組みは、代表的な教師付き線型次元削減手法である Fisher の判別分析が学習データの条件付きエントロピーの上界を最小化していることに着目し、情報論的な観点から教師付き距離構造学習を捉えた非常に一般的なものである。この基準を用いることによって、低次元空間における距離構造の学習手法として、線型演算による次元削減手法を具体的に構成している。また、カーネル関数に付随する非線型特徴空間における距離構造の学習を近年盛んに研究されている Multiple Kernel Learning のための一手法として定式化している。提案する線型次元削減手法及び Multiple Kernel Learning 手法を標準的な 2 クラス判別問題のベンチマークデータセットに適用して、両手法ともに従来の手法と同等以上の性能を示すことを確認している。

第5章では離散観測データに対して、データの生成モデルに基づく類似度及びモデルのパラメタ空間へのデータ配置の学習手法を考察している。まず、映画や書籍、レストランなどへの評価データの新しい生成モデルを提案している。多数のアイテムに対して多数のユーザが比較をおこなったりランキングを与えたりすることで得られるデータを解析して、一つ一つのデータが有する本質的な価値を推定する問題は心理学や経済学の分野で古くから研究が行われているが、近年機械学習の分野でも注目されている。従来、こうした比較データやランキングデータは Bradley-Terry モデルあるいは Plackett-Luce モデルと呼ばれる確率モデルによってモデル化されてきた。本章では、ランキングデータの生成モデルである Plackett-Luce モデルの自然な一般化として、グループ化ランキングモデルを提案している。このモデルはアイテムの持つ価値パラメタによって特徴付けられるが、尤度関数の直接評価が困難である。そこで、効率的に評価可能な尤度関数の

近似を与え、さらに情報幾何学的な考察を通してモデルのパラメタ推定方法を提案している。また、提案する確率モデルを現実の映画評価データ及び書籍評価データに適用し、Fisherカーネルと呼ばれるカーネル関数を用いてユーザ同士の類似度を定義し、これを、協調フィルタリングにおけるユーザ間類似度として用いたアイテム推薦システムを提案している。従来のアイテム推薦手法では、ユーザによるアイテム評価値をベクトル表現し、その内積や相関を類似度とするが、こうしたデータの扱いは理論的妥当性が十分とは言い難い。一方、提案手法では生成モデルのパラメタ空間にデータを埋め込んだ上で、モデルに基づく自然な内積を定義しているので、ユーザの類似度に対して統計的・情報理論的な解釈を与えることが可能となる。本論文で提案された手法による推薦の精度は従来の手法と同等であるが、評価データの生成モデルに基づく手法であることから、ユーザ同士・アイテム同士の関係性を確率分布のパラメタ空間において可視化することが可能であり、推薦の根拠が自然に提示できるという特長を持つ。

第６章では本研究の内容をまとめ、今後の課題及び展望について述べている。

以上、本論文では観測データの背後にある距離構造の学習問題を統計学･情報理論･情報幾何を背景とする情報論的な観点から統一的に論じ、学習により獲得された距離構造に確率論的な意味付けを与えることができる枠組みを提案している。連続確率変数に対しては条件付きエントロピーの最小化原理に基づいて次元縮約とMultiple Kernel Learningにおける新たな手法を具体的に構成しており、またランキングデータを代表とする離散確率変数に対しては新たな生成モデルを提案し、その実用的な推定手法を導出している。これらは現在発展著しいウェブなどの大規模･多次元のデータの中から効率良く情報を抽出するための技術としても寄与するところが大きいと考えられる。よって本論文は博士（工学）の学位論文として価値あるものと認める。

２０１０年７月

審査員

（主査）	早稲田大学教授	博士（工学）東京大学	村田　昇
	早稲田大学教授	工学博士（早稲田大学）	松本　隆
	早稲田大学教授	工学博士（早稲田大学）	内田　健康
	早稲田大学准教授	博士（医学）京都大学	井上　真郷
	産業技術総合研究所研究員	博士（工学）東京大学	赤穂　昭太郎