



ビッグデータに隠れた構造を見つけ出す 確率的潜在変数モデルに基づく知識発見

Discovering hidden structure in big data

Knowledge discovery based on probabilistic latent variable models



上田特別研究室

岩田 具治 Tomoharu Iwata

プロフィール

2001年慶應義塾大学環境情報学部卒業。2003年東京大学大学院総合文化研究科修士課程修了。同年日本電信電話株式会社入社。2008年京都大学大学院情報学研究科博士課程修了 博士(情報学)。2012年から2013年にかけてケンブリッジ大学客員研究員。現在、NTT コミュニケーション科学基礎研究所 主任研究員(特別研究員)。データマイニング、機械学習の研究に従事。2014年長尾真記念特別賞等受賞。情報処理学会、電子情報通信学会各会員。

インターネットやセンサの普及とともに膨大なデータを容易に入手、蓄積できるようになってきています。しかしながら、得られたデータをそのまま眺めたとしても、どのような性質をもつのか、どのような価値があるのかを知ることは困難です。データから有用な知識を自動的に発見することはビッグデータ解析の大きな課題となっています。

様々な場面で得られたデータを見てみると、隠れた構造を持っていることが多々あります。例えば、新聞記事では同じ記事中に同時に現れやすい単語の集まりがあります。政治記事では「首相」と「国会」という単語が同時に現れやすいですし、スポーツ記事では「サッカー」と「ゴール」が同時に現れやすいです。また、画像の場合、包丁が写っていたら、まな板も写っている可能性が高いです。ソーシャルネットワークでは友達の友達は友達であることが多いですし、SF小説を買った人は次もSF小説を買う可能性が高いです。

このような隠れた構造は有用な知識となります。例えば、たくさんさんの研究論文から、同時に現れやすい代表的な単語のまとまりをトピックとして抽出できれば、どのような研究が話題となっているかを知ることができます。またそのトピックを利用することで、単純なキーワード検索では得られない関連論文を探し出すことができます。購買データから同じような商品を購入している顧客グループを見つけ出せれば、グループ毎に特化したマーケティング戦略を練ることができます。

隠れた構造を自動的に見つける方法として、潜在変数モデルが注目されています。潜在変数モデルでは、観測データが生成される過程において隠れた構造が重要な役割を果たしていると考えます。具体的には、隠れた構造を観測することができない変数である「潜在変数」として表します。そして、その潜在変数に依存してデータが生成されるモデルを考えます。例えば、新聞記事が生成されるモデルの場合、まず新聞記事の潜在変数であるトピックが決められ(政治トピック)、そのトピックに従ってその記事の単語が生成される(首相、国会、政策)、と考えます。



潜在変数モデルを使うことによって潜在変数からデータが生成される過程を表現できますが、実際に与えられるのはデータなので、データから潜在変数を推定する逆問題を解く必要があります。この逆問題を解くためにベイズ則という確率の基本的な公式が利用できます。ベイズ則を利用することにより、潜在変数からデータを生成するモデルをもとに、与えられたデータに関する潜在変数を推定できます。ここで確率を基盤にすることによって、不確実性を扱うことが可能になり、データに含まれるノイズにも頑健な推定が可能になります。

機械学習は「教師あり学習」と「教師なし学習」に大別できます。確率的潜在変数モデルによる隠れた構造の推定は「教師なし学習」になります。教師あり学習では、入力と出力のペアを大量に教師データとして用意して、ある入力に対応する出力を予測できるように学習します。学習データ作成のコストが高い場合や、プライバシー保護等の理由で学習データが得られない場合は、教師あり学習は利用できません。一方、教師なし学習は、教師データなしで学習できるという利点があります。その反面、タスクに適したモデルを使わなければ、望ましい有用な結果を得ることができないため、モデルの設計が重要となります。

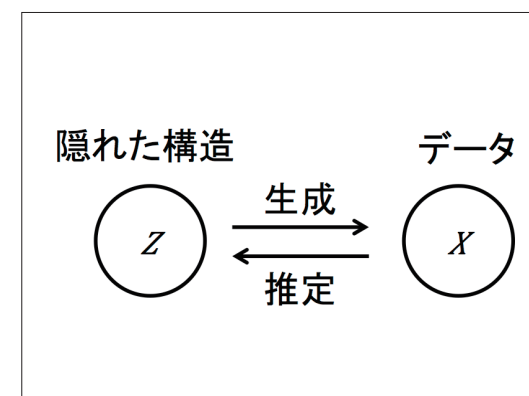


図1: 隠れた構造に依存したデータの生成と、データに基づく隠れた構造の推定

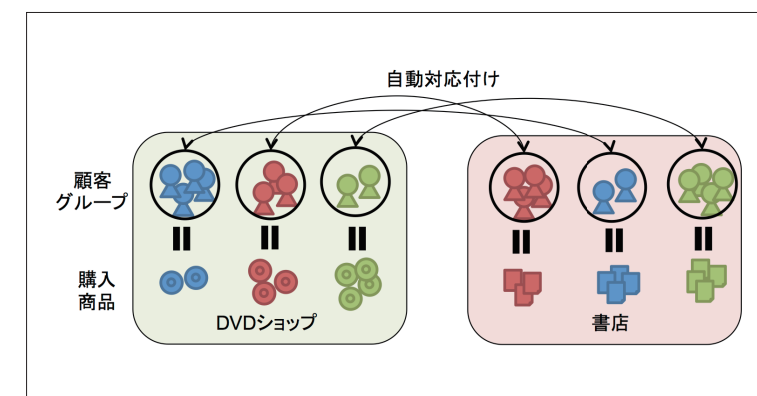


図2: 異なる店の顧客グループを対応付けるオブジェクトマッチング

関連文献

- [1] T. Iwata, T. Yamada, Y. Sakurai, N. Ueda, "Online multiscale dynamic topic models," in *Proc. 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pp. 664-672, 2010.
- [2] T. Iwata, J. Lloyd, Z. Ghahramani, "Unsupervised many-to-many object matching for relational data," *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, Vol. 38, No. 3, pp. 607-619, 2016.
- [3] T. Iwata, T. Yamada, N. Ueda, "Probabilistic latent semantic visualization: Topic model for visualizing documents," in *Proc. 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pp. 363-371, 2008.
- [4] T. Iwata, A. Shah, Z. Ghahramani, "Discovering latent influence in online social activities via shared cascade Poisson processes," in *Proc. 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pp. 266-274, 2013.