SASによる機械学習入門

西野 嘉彦
SAS Institute Japan 株式会社

An Introductory Overview of Machine Learning with SAS®

Yoshihiko Nishino
SAS Institute Japan Ltd.

要旨:

機械学習の簡単な整理と代表的な手法をSASで利用する場合のアプローチ、簡単な事例のご紹介

キーワード: 機会学習、SASシステム

アジェンダ

- はじめに
- 機械学習とは
- 機械学習の応用例
- データマイニングとの違い
- 機械学習と他の分野の関係
- 学習手法について
- SASによる機械学習アルゴリズム
- ディープ・ラーニングについて

はじめに

- 本セッションの発表内容は以下の資料にもとづきます。

Machine Learning　What it is & why it matters

(English)

http://www.sas.com/en_us/insights/analytics/machine-learning.html

(日本語)

http://www.sas.com/ja_jp/insights/analytics/machine-learning.html

An Overview of Machine Learning with SAS Enterprise Miner

Hall, Patrick; Dean, Jared; Kabul, Ilknur Kaynar; Silva, Jorge; SAS Institute, Inc. 2014

http://support.sas.com/resources/papers/tnote/datamining.html

機械学習とは

- 分析モデル構築を自動化するデータ分析手法
- データから反復的に学習するアルゴリズムを利用
- 探索場所を明示的にプログラミングすることなく、隠れた洞察を発見

機械学習の応用例

- 不正検知
- Webの検索結果
- Webページやモバイルデバイスへのリアルタイムの広告掲示
- テキストベースのセンチメント分析
- 信用スコアリングとネクスト・ベスト・オファー（次に提示すべき最良オファー）
- 設備不良の予測
- 新たな価格設定モデル
- ネットワーク侵入検知
- パターン認識と画像認識
- スパムメール・フィルター

データマイニングとの違い

- 機械学習とデータマイニングには同じアルゴリズムと技術がいくつも使われますが、何を予測するのかという点で違いがあると考えます。

- データマイニング
  - 未知のパターンや知識を見つけ出すために利用

- 機械学習
  - 既知のパターンや知識を再生成するために利用され、その結果を他のデータに自動的に適用し、さらにその適用結果を意思決定や行動に活用

機械学習と他の分野の関係

- This graphic was originally created by SAS in a 1998 primer about data mining, and the new field of data science was added for this paper.

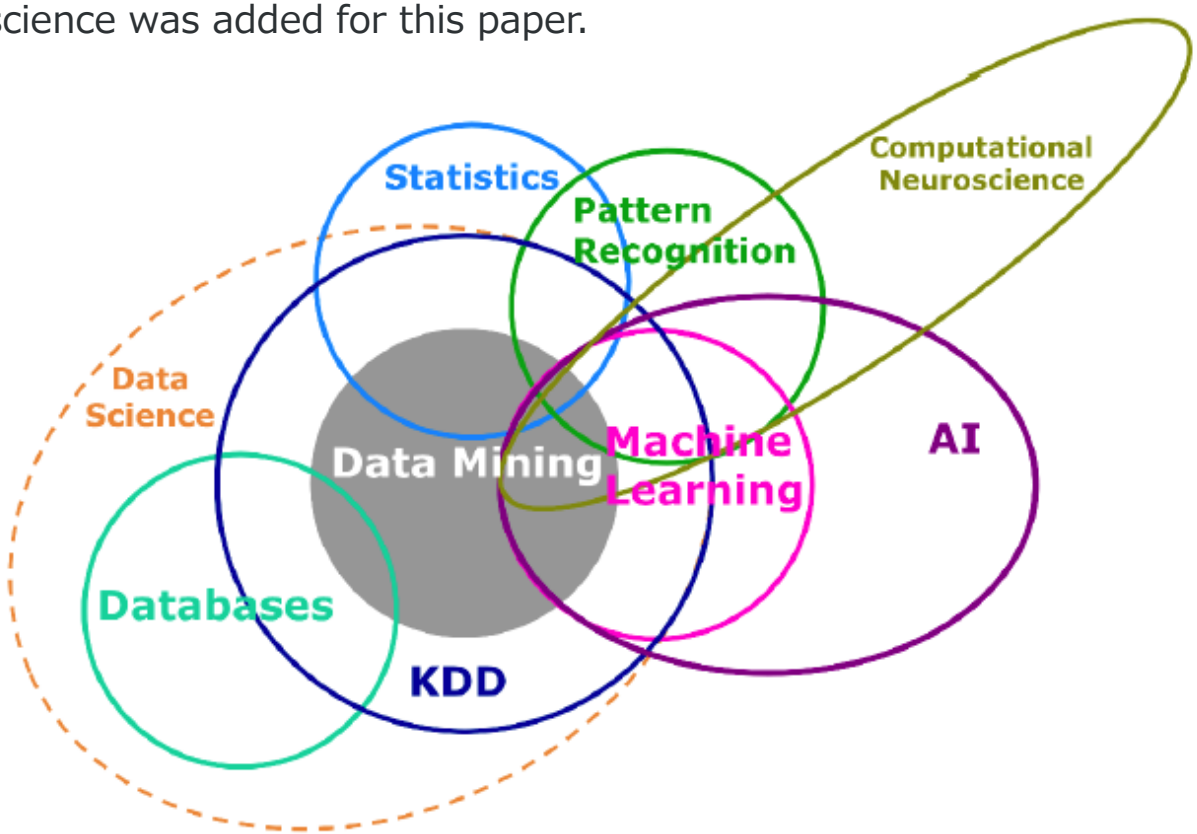

Figure 1. Multidisciplinary Nature of Machine Learning

学習手法について

- 教師あり学習では、ラベル付きのデータを使って学習を実行し、ラベルの付いていないデータのラベル値を予測するために利用される。

- 教師なし学習は、履歴ラベルが存在しないデータに対して利用され、データを探索してその内部に何らかの構造を見つけ出すために利用される。

- 半教師あり学習は、教師あり学習と同じ用途に利用されるが、ラベル付きデータとラベルなしデータの両方を使って学習を行う。少量のラベル付きデータと大量のラベルなしデータを使うケースが典型的（ラベルなしデータの方が入手にかかる費用も労力も少なくて済むため）。

学習手法について

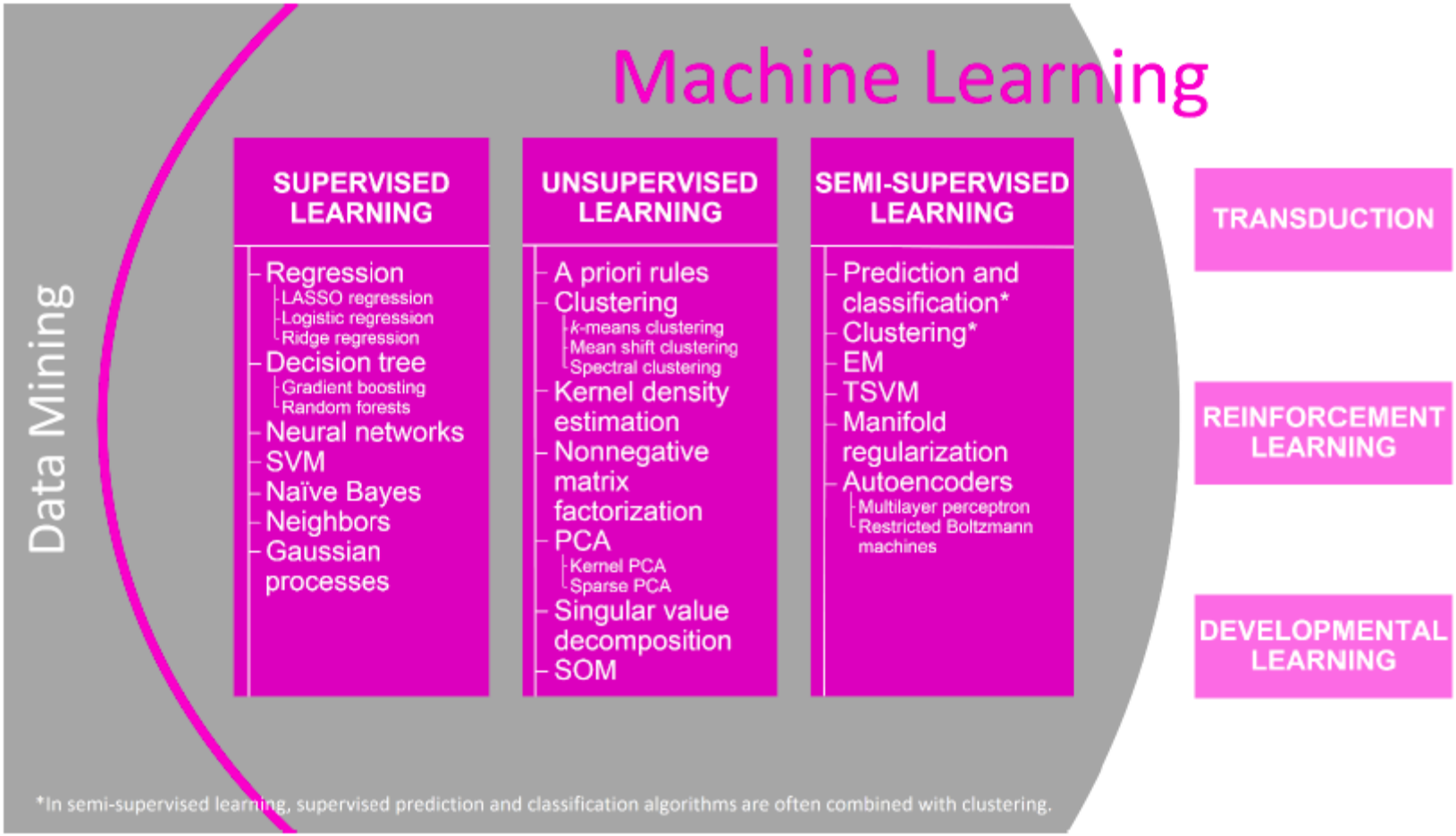

Figure 2. Machine Learning Taxonomy

学習手法について

- ニューラル・ネットワーク
- 決定木
- ランダムフォレスト
- アソシエーションとシーケンスの発見
- 勾配ブースティングとバギング
- サポート・ベクター・マシン（SVM）
- 近傍法マッピング
- k平均法クラスタリング
- 自己組織化マップ（SOM）
- 局所探索最適化手法（遺伝的アルゴリズムなど）
- 期待値最大化法
- 多変量適応型回帰スプライン法
- ベイジアン・ネットワーク
- カーネル密度推定
- 主成分分析
- 特異値分解
- ガウス混合モデル
- 逐次的カバーリング・ルールの構築

SASによる機械学習アルゴリズム（教師あり学習）

Algorithm	SAS Enterprise Miner Nodes	SAS Procedures	References
Regression	High Performance Regression LARS Partial Least Squares Regression	ADAPTIVEREG GAM GENMOD GLMSELECT HPGENSELECT HPLOGISTIC HPQUANTSELECT HPREG LOGISTIC QUANTREG QUANTSELECT REG	Panik 2009
Decision tree	Decision Tree High Performance Tree	ARBORETUM HPSPLIT	de Ville and Neville 2013
Random forest	High Performance Forest	HPFOREST	Breiman 2001b
Gradient boosting	Gradient Boosting	ARBORETUM	Friedman 2001
Neural network	AutoNeural DMNeural High Performance Neural Neural Network	HPNEURAL NEURAL	Rumelhart, Hinton, and Williams 1986
Support vector machine	High Performance Support Vector Machine	HPSVM	Cortes and Vapnik 1995
Naïve Bayes		HPBNET*	Friedman, Geiger, and Goldszmidt 1997
Neighbors	Memory Based Reasoning	DISCRIM	Cover and Hart 1967
Gaussian processes			Seeger 2004

Table A.1. Supervised Learning Algorithms

*PROC HPBNET can learn different network structures (naïve, TAN, PC, and MB) and automatically select the best model.

SASによる機械学習アルゴリズム（教師なし学習）

Algorithm	SAS Enterprise Miner Nodes	SAS Procedures	References
A priori rules	Association Link Analysis		Agrawal, Imielinski, and Swami 1993
k-means clustering	Cluster High Performance Cluster	FASTCLUS HPCLUS	Hartigan and Wong 1979
Mean shift clustering			Cheng 1995
Spectral clustering		Custom solution through Base SAS and the DISTANCE and PRINCOMP procedures	Von Luxburg 2007
Kernel density estimation		KDE	Silverman 1986
Nonnegative matrix factorization			Lee and Seung 1999
Kernel PCA		Custom solution through Base SAS and the CORR, PRINCOMP, and SCORE procedures	Scholkopf, Smola, and Muller 1997
Sparse PCA			Zou, Hastie, and Tibshirani 2006
Singular value decomposition		HPTMINE IML	Golub and Reinsch 1970
Self organizing maps	SOM/Kohonen Node		Kohonen 1984

Table A.2. Unsupervised Learning Algorithms

SASによる機械学習アルゴリズム（半教師あり学習）

Algorithm*	SAS Enterprise Miner Node	SAS Procedure	References
Denoising autoencoders		HPNEURAL NEURAL	Vincent et al. 2008
Expectation maximization			Nigam et al. 2000
Manifold regularization			Belkin, Niyogi, and Sindhwani 2006
Transductive support vector machines			Joachims 1999

Table A.3. Semi-Supervised Learning Algorithms

*In semi-supervised learning, supervised prediction and classification algorithms are often combined with clustering. The algorithms noted here provide semi-supervised learning solutions directly.

ディープ・ラーニングについて

- とくに会話／テキスト／画像の認識に画期的な進歩をもたらした研究分野(学習アプローチ)
- 多数の中間レイヤーを持つニューラル・ネットワークを用いて、コンピューターが自律的にタスクの習得、情報の整理、パターン検出などを行えるようにします。
- Webセミナー「考えるマシン： ディープ・ラーニングの実験」（英語）でSASを使った事例をご覧いただけます。

http://www.sas.com/en_us/webinars/deep-learning/register.html