MachineLearning and Statistics

ari

2018/1201

Contents

1 Introduction

世の中最新手法も様々出ていますが、基礎知識が全くない状況での最新手法を調べるのは理解の上で非効率だと思っているので、機械学習、統計の基本を一人で振り返るアドベントカレンダーをする。これだけできれば機械学習の数理の基本はできているという状態を目指す。

2 Basic of Machine Learning

2.1 機械学習とは何か

そもそも機械学習とは何かについて説明する。私の理解では機械学習とは、与えられた"データ"を用い、未知の"データ"についても推測できるような"モデル"を作成することである。例えば、文字認識の場合、画像と文字の組み合わせを大量に与えられている状態で、未知の画像に対してどういう文字が書かれているかを予測することになる。これがうまく言った場合に機械がデータの正解不正解だけでなく、"パターン"を学習し、人がルールを定義せずとも答えが与えられたという意味で機械学習と呼んでいる。

2.1.1 機械学習の問題の類別

機械学習もどういう問題を考えるかによって以下のように類別されている。

- 教師あり学習
  1. 分類
  2. 回帰
- 教師なし学習
  1. クラスタリング (k-means 等)
  2. その他

この分類を具体的に説明すると教師あり、教師なし学習の違い与えられているデータに正解が与えられているかどうかである。先程例にあげた文字認識の場合は画像と文字の組み合わせをもらっていれば正解となる文字ももらっているため教師あり学習である。教師なし学習では正解は与えられていない。例えば購買情報をもとにユーザーを分類する場合などで使われる。これは予め特定の種類のユーザーが何を買うかわかっていないことが多いので教師なし学習の枠組みで考えられる事が多い。

教師あり学習の 2 つの分け方であるが, 類と回帰の違いは正解の値が"離散的"かどうかである。離散的な場合は分類であり、そうでない場合は回帰と呼ぶ。典型的な分類は 2 値分類である。回帰は値の変化も順序的な意味を持つことが多い。例えば、家を買うかどうかは二値分類であるが、家の値段がいくらかは連続的に変化すると考える。家の値段も現実には 1 円単位で取引されるので、整数値しか取らないが、値の大小に意味があり、なおかつ数千万に対して、1 円は十分小さいから 0.5 円と 1 円にまとめても実際の結果に対して差がないだろうという意味で連続値であると考えることがある。

Remark. 連続値の定義は正直不明である。数学的に定義できるものではないだろう。離散的な写像は連続であるし、有限の値しか取らない以上像の濃度ははるかに小さい。基本的には知りたいオーダーと変化する最小単位に乖離があり、最小単位での変化が誤差の範疇で数値的に無視できる場合に連続値ということが多い。

教師なし学習の 2 つの分け方は適当で申し訳ないがクラスタリング以外は何というか不明だったので、その他とした。教師ありと同様にクラスタリングは有限の値に分けるものを指す。

Remark. 機械学習は教師あり学習、教師なし学習、強化学習の三種類で与えられるという考え方もあるが、単純さを重要視し、教師あり学習と教師なしに限定した。

2.1.2 機械学習の問題設定

問題設定を考える. 機械学習では実際に予測をする関数を作る操作を学習という. 学習方法を定める際は以下を考える。

- 損失関数: 目指すべき予測をする関数の良さを測るもの
- パラメータ: 学習時に変更できるもの
- パラメータ変更アルゴリズム

これらを定めた後

1. データを設定する
2. パラメータの初期値を定める
3. パラメータ変更アルゴリズムを一定回数動かす

とするのが、機械学習の学習処理である。

Remark. 正直私が機械学習の本を読んでいてもパラメータ変更アルゴリズム等が曖昧でよくわからないことが多い。そもそも無限和を取っているが現実には有限和を取る場合や, 数値計算の精度誤差を意識して変更しているアルゴリズム等もあるように思う. そこでここでは私はわからないこととわかることを明確にしながら書きたいと思う。

Remark. パラメータという書き方も誤解を招きやすいかもしれない。*Gradeint Boosting* などの関数空間上で動作させるものは *non parametric* と言われうるためである。ただ *Gradient Boosting* であっても決定木から定める場合は関数空間自体は制限されるし、その極限でどこまでかけるかはデータ次第なので、ひとまずパラメータと書くことにした.

2.1.3 機械学習の評価

学習では与えられたデータを最適化することに特化したが、時おり、学習したデータに対してだけ、異常に精度がよい関数になってしまう場合がある。そこで機械学習では作った関数が実際に精度がよいかを測る評価のプロセスを採用する事が多い。典型的には k-fold cross validation と呼ばれるデータの分割手法である。

ここまでで機械学習の概要を説明した。これから教師あり学習を中心に具体的に名前のついた手法 (Parameter, 学習アルゴリズム) を考えるかを説明する。

2.2 線形回帰

線形回帰は回帰という名前からもわかるように線形関数による回帰の問題である。上で説明した枠組みにのっとり, 損失関数、parameter,parameter 変更アルゴリズムについて説明する。

2.2.1 Parameter

今回考えるパラメーター全体, つまり、考える関数全体の集合を以下で定める。

$$\mathcal{F} = \{f : \mathbb{R}^n \to \mathbb{R} \mid \exists a \in \mathbb{R}^n, \exists b \in \mathbb{R}\ s.t\ f = a \cdot x + b\}$$

$a = (a_1, \ldots, a_n)$ と表す.

2.2.2 損失関数

損失関数は定め方はいろいろあるが、ここではひとまず典型的な平均二乗誤差とする。つまり、

$$L\colon \mathbb{R}^{n+1} \to \mathbb{R}_{\geq 0}(a_1, \ldots, a_n, b) \mapsto \sum_{i=1}^{n} l(y^{(i)}, ax^{(i)} + b)/n$$

となる。ここで $l(x, y) = (x - y)^2$ である。他にも root をとった $RMSE$ や絶対値で評価する MAE 等が有名な損失関数として存在する.

2.2.3 paramete 変更アルゴリズム

ロス関数の最小値を求めるように Parameter を変更するが、最小値を求めるアルゴリズムは, 典型的には勾配降下法を用いる。

Remark. 二次計画問題なので、この問題に特化した解法もあると思われるが。汎用的な方法をひとまず書くことにした。

Remark. 現実的にはただの線形回帰だけでなく、リッジ回帰やラッソ回帰のように 正則化項をいれることが多い。これは使われるデータに特化し、未知のデータに対して精度が低くなる場合があり、正則化項を入れたほうが精度が高い可能性が高いためである。リッジやラッソのどちらが良いかなどは場合による。ただし、データの分布的に正規分布等平均が重要な値であれは リッジ回帰を, 中央値が必要な場合はラッソ回帰を使う事が多い.