機械学習輪読 I

電気通信大学大学院 修士 2 年
山岡 勇太

2019/09/21

目次

目次

背景

1990 年代以降,

- インターネットの加速度的進展
- 計算機の処理速度
- ハードディスクなどの外部記憶容量の増大

によって, 処理対象のデータが爆発的に増加した. この大規模なデータ (ビッグデータ) に対し, データマイニングを行うことが求められるようになった.

データマイニング

(簡単に言えば) 統計的処理などにより, 与えられたデータから有用な情報を抽出すること.

背景

データマイニングの数理的モデルとして位置付けられるようになったのが機械学習という学問分野.

目次

モデルとは何か？

モデルとは, 入力から出力を得る過程を数理的に定式化したものである.
以降, 具体的に定義する.

準備

とある観測者は何らかのデータ $\boldsymbol{x}$ を観測する．ここで, $\boldsymbol{x}$ は何らかのクラス y に属すものとする．クラス全体を情報源と呼ぶ．
ここで, $\boldsymbol{x}$ はベクトルとする．

具体例

情報源 : 新聞記事全体

y : 政治の記事, 経済の記事, スポーツの記事, etc.

$\boldsymbol{x}$: 実際に書かれた記事

予測過程

予測過程とは？

観測者が新たに観測したデータ $\boldsymbol{x}$ に対応する情報源のクラス y を知りたい場合に行う, ($\boldsymbol{x}$ を何らかの方法で逆変換して) y の値を予測する計算のこと.

予測過程の結果得られた値を $\hat{y}$ とする.

学習過程

学習過程とは？

逆変換は y の条件付き確率 $p(y|\boldsymbol{x})$ によって行われるとする. この確率密度関数 p を求める計算のこと.

以後, 学習といえば, この学習過程を指す. また, 確率密度関数は単に確率分布と呼ぶ.

ラベル付きデータ

ラベル付きデータとは？

観測されたデータ $\boldsymbol{x}$ に対応する y が知られているとする. このような $\boldsymbol{x}$ と y の組の集合 $\mathcal{D} = \{(\boldsymbol{x}_1, y_1), \cdots, (\boldsymbol{x}_n, y_n)\}$ のこと.

教師データや訓練データって言ったりもする.

前提

$\mathcal{D}$ の各要素は, ある確率分布に従って独立に生成されたものとする (このとき $\mathcal{D}$ は独立同一分布に従うという). 独立同一分布でない場合も機械学習の対象だが, テキストでは扱わない.

ラベル付けは人間によって行われることが多い (らしい).

予測過程のお話

観測したデータ $\boldsymbol{x}$ から未知の y を推定する問題を考える.
この問題は, $\boldsymbol{x}$ が知られているという条件のもとでの y の確率, すなわち $p(y|\boldsymbol{x})$ を最大化するような y の予測値 $\hat{y}$ を求めること.
従って, 次の式で表される.

$$\hat{y} = \underset{y}{\operatorname{argmax}}\, p(y|\boldsymbol{x}) \tag{1}$$

学習過程のお話

y についての事前情報がある, すなわち $\mathcal{D}$ が与えられている場合について考えてみる.
$\mathcal{D}$ から推定される $p(\boldsymbol{x}|y)$ は, $\mathcal{D}$ の要素数が少ないと信頼性が低い. また, 情報源の性質である $p(y)$ が使われていない (式 1 を参照).

学習過程のお話：例

以下のようなラベル付きデータ $\mathcal{D}$ がある.

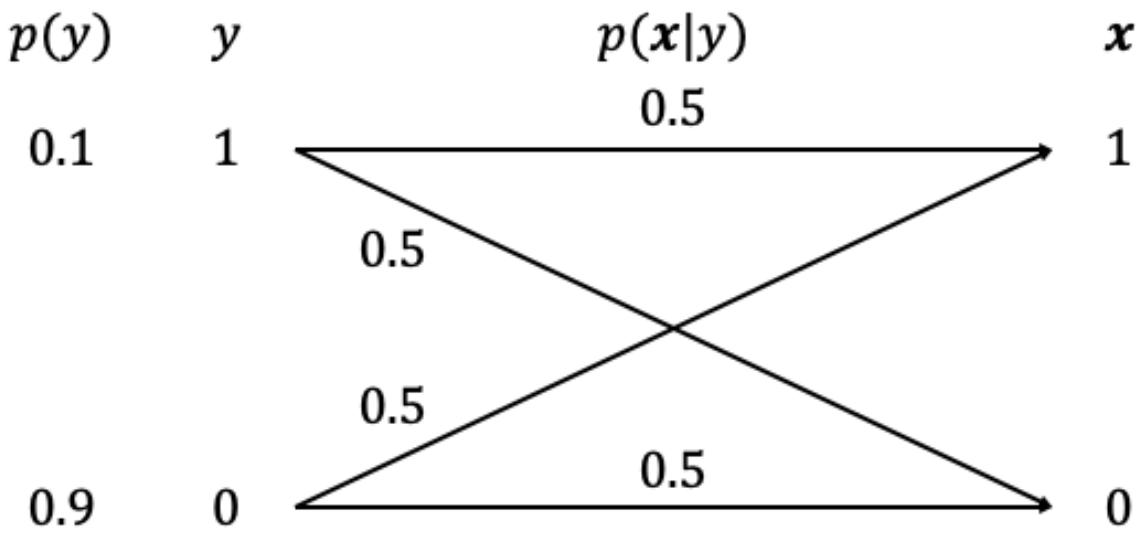
$$\mathcal{D} = \{(\boldsymbol{x}_a, y_1), (\boldsymbol{x}_b, y_2), (\boldsymbol{x}_c, y_3), (\boldsymbol{x}_a, y_2), (\boldsymbol{x}_b, y_3), (\boldsymbol{x}_a, y_3)\}$$

このとき...

- $p(y_1) = 1/6$
- $p(y_2) = 2/6 = 1/3$
- $p(y_3) = 3/6 = 1/2$
- $p(\boldsymbol{x}_b|y_2) = (1/6)/(1/3) = 1/2$
- $p(\boldsymbol{x}_a|y_3) = (1/6)/(1/2) = 1/3$
- $p(\boldsymbol{x}_c|y_3) = (1/6)/(1/2) = 1/3$

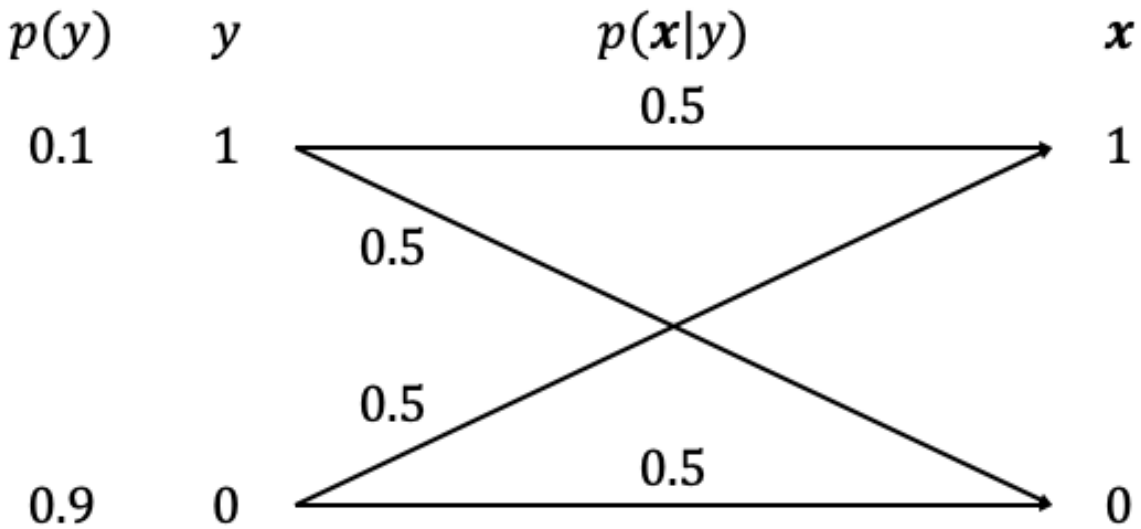
具体例

次のような状況を考える．この例では, 情報源のクラスは $0, 1$ の２つで, 観測データも $0, 1$ の２値を取るとする．ここでは, 情報源出力の 0 を 1, 1 を 0 に変換する確率は各々0.5 とする．



具体例

この状況下では, $p(y|\boldsymbol{x})$ だけを考慮した式 1 を用いると...
0 を観測 $\Rightarrow \hat{y} = p(1|0) = p(0|0) = 0.5$. しかし $y = 0$ を出力する確率が高いので $\hat{y} = 0$っぽい？



具体例

ここでベイズの定理を使ってみる.

定理 1.1 ベイズ (Bayes) の定理

$$p(y|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|y)p(y)}{\sum_y p(\boldsymbol{x}|y)p(y)}$$

$\sum_y p(\boldsymbol{x}|y)p(y) = p(\boldsymbol{x})$ より, 次のようにも書ける.

ベイズの定理 (変形後)

$$p(y|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|y)p(y)}{p(\boldsymbol{x})}$$

具体例

ベイズの定理より, 式 1 は次のように書ける.

$$\hat{y} = \underset{y}{\operatorname{argmax}}\, p(\boldsymbol{x}|y)p(y) \tag{2}$$

$\boldsymbol{x}$ は観測された値であり, y に argmax に関与しないので上記の式になる.
式 2 を用いることで, $p(y)$ を事前知識として利用できる.
この式で 0 を観測した場合の推定値を計算すると,

$$y = 0 : \; p(0|0)p(0) = 0.5 \times 0.9 = 0.45$$
$$y = 1 : \; p(0|1)p(1) = 0.5 \times 0.1 = 0.05$$

よって, $\hat{y} = 0$.

まとめ

- 機械学習とは, データマイニングの数理的モデル.
- 機械学習は, 学習過程と予測過程からなる.
- ラベル付きデータは独立同一分布に従う (そうでないものもあるが, 今回は扱わない).
- (当たり前だけど) モデル化の仕方によって, 得られる結果が変わる.