

知的情報処理 5. 原因がなくても結果がある(か?)

櫻井彰人
慶應義塾大学理工学部

再登場: 生成モデル

- データがある、ということは、そのデータを生成する原因がある、と考える
 - のちほど、原因を考えることはしない方がよい、という主張を紹介する。どちらがよいかは、神のみぞ知る
 - deterministic (決定論的とでもいしましょうか) なモデルであれば、(状況が同じであれば) 結果は一個。
 - ところが、データは複数個ある。
 - 風邪なら、「体温38度、咳が1時間に20回、喉の腫れは5mm。ほかには目立った症状なし」なんて綺麗に症状が記述できればよい。世の中そうではない。
 - ということは、決定論的モデルでは不便である(不適当だと言っているわけではない)
- そこで、確率論的モデルを考えよう。

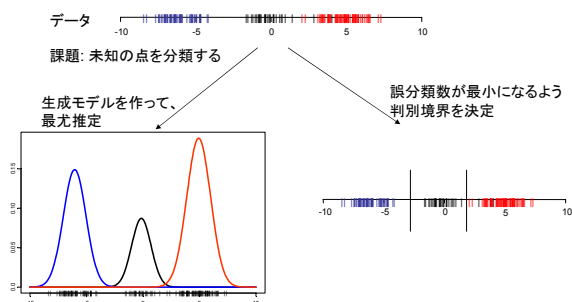
分類器のモデル

- 機械学習で、まず、考える(つまり、まあ、一番やさしい)問題は、分類問題。
- 分類するだけのために、統計的モデルを考える必要はあるのか?
- もし「必要はない」のであれば、無理することはないでしょう?
- 通常は、データ数が少ないので、モデルを推定しようとすると、どうしても精度が悪くなる。
- そんな推定精度の悪いモデルを使って、分類が(高い精度で)できるとは思えない。

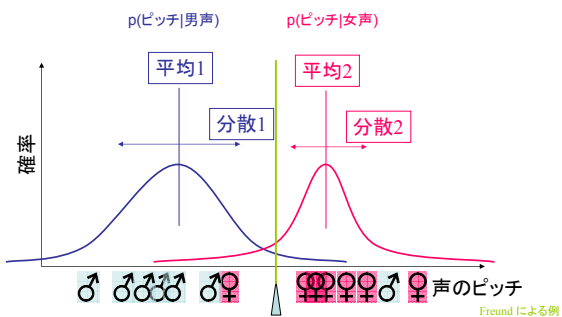
分類器のモデル

- (分類すべきクラスの)分布があるの、ないの、という話ではなく、あるかもしれないし、ないかもしれないが、とにかく、分布は推定しない方がよいのでは? ということ
- 分布を推定しないで、分類はできるのか?
- 機械学習では、もともと、そう(つまり分布は想定しないと)考えていた。
- どうするのか?
- 話は簡単。
- 行いたいことは、 $\arg \max P(\text{Class} \mid \text{instance})$ を知ることであった。
であれば、 $f(\text{instance}) = \arg \max P(\text{Class} \mid \text{instance})$ となる関数 f の近似関数を(近似プログラムでもよい)求めればよいでしょう?
- こういうモデルを、判別モデルといったりする

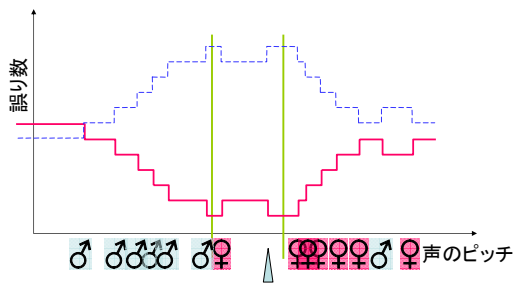
生成モデル vs. 判別モデル



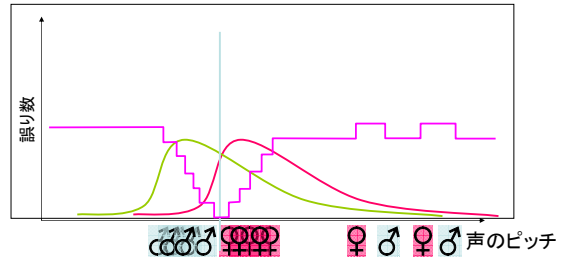
生成モデル



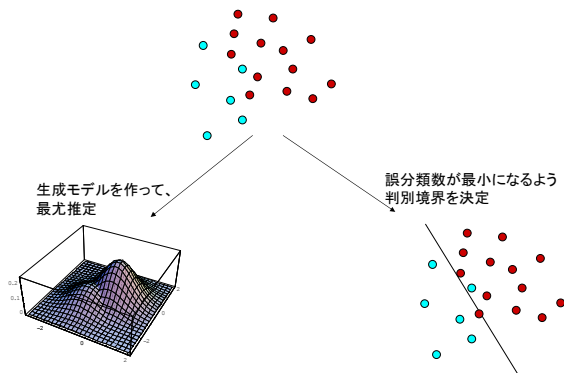
判別モデル



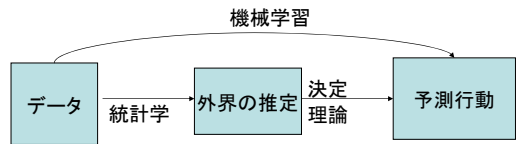
偏ったデータ



生成モデル vs. 判別モデル



統計学 vs. 機械学習 かも



生成・判別モデルの比較

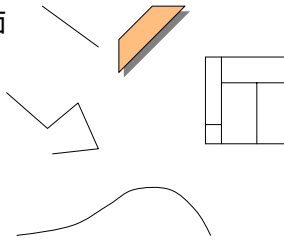
	生成モデル	判別モデル
目的	確率の推定	分類規則の作成
性能の尺度	尤度	誤分類率
誤りの原因	はずれ値	クラスラベル誤り

判別モデルと境界

- 判別モデルでは、「境界」が最重要
- 次を決める必要がある
 - 境界の形
 - 最適であるべき基準
 - 実際のアルゴリズム
 - (相互に依存関係はある)

境界の形

- 直線、平面、超平面
- 超平面の組合せ
- 多項式
- その他いろいろ



最適性の基準

- 誤分類個数最小
- 誤分類点の、境界からの距離和最小
 - 「距離」はいろいろ
- 境界を $f(x)=0$ とし、 $f(x)>0$ がクラス1、 $f(x)<0$ がクラス2 とし、 $\sum (f(x_i) - C_i)^2$ を最小化
- その他いろいろ

アルゴリズム

- 最小化すればよいのだが、、、
- 微分=0 が解析的にとければ、OK
- とけないときは、逐次近似
 - ORで得られた結果を使用
 - ときには、naïve な方法のほうがよいことも。

その他

- 「境界は(数学的な)関数で(簡単に)表現できる」とは考えない方がよい。
- 実際、その方が簡便で、よい結果が得られることが、しばしば、ある。
- 従って、この前のスライドの記述は、頭の整理だと思ってください。
 - 現実には、「誤差」を減らすために、あらゆる努力をしている、のです