知的情報処理
3. 原因があって結果がある(か？)

櫻井彰人
慶應義塾大学理工学部

本日の目的

- データを生成する法則が存在すると仮定し、それを推定することを考える。その場合、
  - 推定できるのか？
  - 推定する方法はあるのか？
  - 推定しなくてもよいということはないのか？

  という問いを背景に
  - 「モデル」という概念
  - 「モデル」を推定するということ
  - 「モデル」を推定しないということ

  を知る
- なお、事例ベース学習は、丸暗記・丸暗記の拡張であった。

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良い、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 (決定論的とでもいいましょうか)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。

確率的モデルとは

- ある確率密度分布があって、その分布に従い、データが生まれてくるような、モデル
  - モデルとは「模型」。本物ではないが、その動きのある面(今一番関心があるところ)をうまく表現するであろう、もの。
- 数学的には、あっさりと、「データ x は、確率密度分布 p(x) に従って生成される」といった具合に書く。
  - p(x) を具体的に書かないことには話しにならない。
  - 確率変数 X は確率密度関数 p(X) に従う。$x_1,..,x_5$ はXのサンプルである。
- 例えば、「データ x は、正規分布 $p(x)=(1/\sqrt{(2\pi)})\exp(-x^2/2)$ によって生成される」といった具合。

確率的モデルとは　2

- 絵で書くと(あんまり変わらないが)

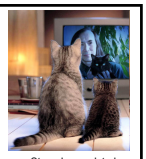
$$p(x)=\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}} \longrightarrow x_1, x_2, \ldots$$

データは複数個

- 絵の方には描いたが、統計的な生成モデルを考えるときには、データは複数個(一般にはたくさん)あるのが前提。 $x_1, x_2, \ldots$
- データー個は、仮に、正規分布に従うとしよう。次の一個も正規分布に従うとしよう。
- しかし、現実には、2番目のデータが一番目のデータの値に依存することは、よく、ある。
- それを考えるべきであろうか？

- 当然考えるべき。
- しかし、初めからそれを考えるのは、難しい。

独立性



Strongly correlated

- そこで、まず、各データは、独立に生成されるものとする。
- この独立性は、正しくは成り立っていなくても、結構良い近似になっていることが多い。
- 従って、多くの場合、データ間の独立性は暗黙に仮定する。
- なお、属性間の独立性は、一般には仮定しない。
  (しかし、従属の場合、好ましくないことが発生しがち)
  - 喉が腫れれば、熱がでる。咳がでれば、喉が腫れる。しかし、咳、喉の腫れ、熱は、それぞれ、重要な症状として、考えるのが普通である(本当に従属なら、どれか一つがあればよい)

図は http://metalogue.img.jugem.jp/20090317_572334.jpg

例: コイン投げ

- コイン投げ10回中の表の回数の1000回分
  - 表が出る確率 p のコインで、各試行は独立だとする
  - 表の回数は、二項分布 B(p,1-p) に従う

$$P(X=x)=\binom{10}{x}0.5^{10}$$

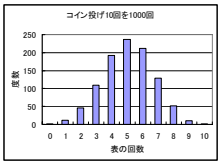
生成

$$P(X=x)=\binom{10}{x}p^x(1-p)^{10-x}$$

生成と仮定
パラメータ推定

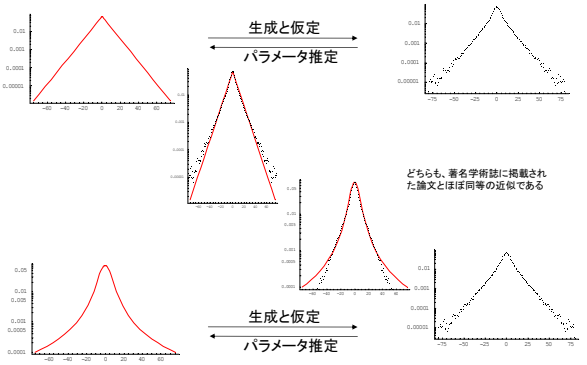
$$P(X=x)=\frac{1}{\sqrt{2\pi}\sigma}e^{\frac{(x-\mu)^2}{2\sigma^2}}$$

生成と仮定
パラメータ推定

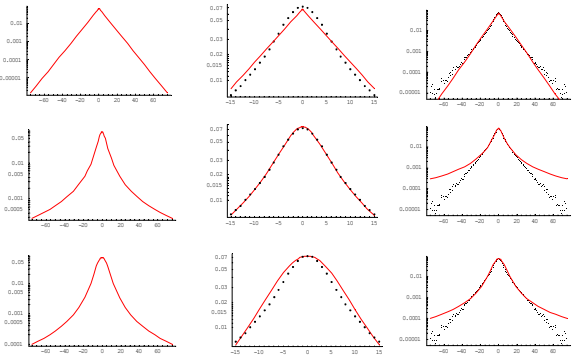


図は http://i.dailymail.co.uk/i/pix/2008/10/24/article-1080212-008026FD00000258-365_233x370.jpg

注意: ある分布

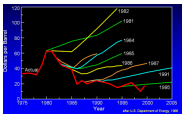


注意: ある分布

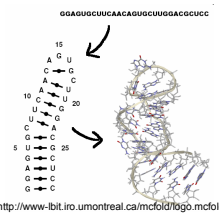


機械学習として

- さきほどの説明は、分布の近似という意味あいが強い。次に、未知データの予測という意味あいで、述べてみよう。



米国エネルギー省の1996年の資料より



http://www-lbit.iro.umontreal.ca/mcfold/logo.mcfold.png

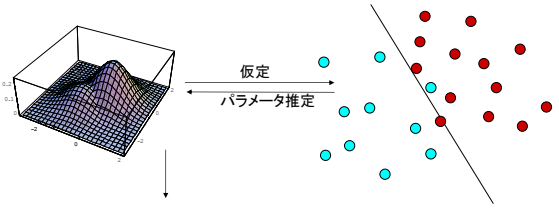
Prediction

http://rovicky.wordpress.com/2006/10/19/do-we-need-a-prediction/

右図は http://www.wpsychic.com/wp-content/uploads/2009/07/free-psychic-prediction.jpg

2個のクラス

- 学習サンプル: 属性値とクラスが分かる
  - 様々な検査値と(名医が診断した)病名
- テストサンプル: 属性値のみ。クラスは不明
  - あなたの目の前の患者さん。検査結果あり。病名不明



確率に基づき、最適な判断境界を定める

図は http://www.whizzdome.com/sciatica/diagnosis_small.jpg

確率分布の推定

- 「注意: ある分布」で示したように、確率分布の推定は難しい。
- （今回は説明しないが）次元が上がる（属性の個数が増える）と分布の推定はもっと難しくなる。
- 一般に属性の個数は多い。
  - 現在では、数個ということは少ない。
- それにも関わらず、モデルを考えることに意味があるのか？
  - 実用上、極めて意味がある。説明は、naïve Bayesの説明の中で行います。

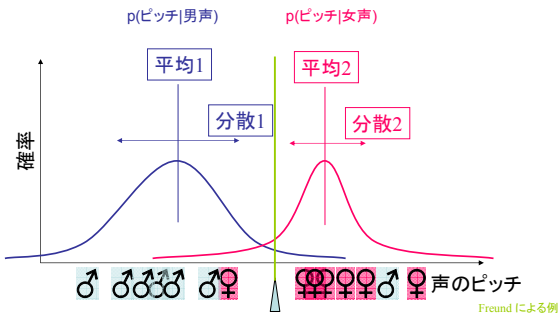
では、どう行うか

- 枠組み:
- p(m) : クラス m の生起確率
  - m は、例えば、風邪ひき、風邪ひきでない。
  - 別例: 男声 or 女声
- p(x|m) : クラス m のときに、サンプル（患者）の属性（検査値）が x である確率。
  - x の値は、体温とか咳の程度。分かっているとする。
  - 別例: 声の高さ（ピッチ）
- p(x|m) p(m) を最大とする m を求めるクラスとする
  - 例えば、p(x|風邪) p(風邪) と p(x|not風邪) p(not風邪) とを比較し、前者の方が大きければ、風邪だと結論する
  - 別例: 声のピッチから、それが男声か女声かを決める

式で書くと

- $\max_m$ p(x|m) p(m) を与える m を答えとする。
- これを、しばしば、
  argmax_m p(x|m) p(m)
  と書く

属性が一個のとき



Freund による例

なぜ、クラス確率を用いるのか？

- なぜ、p(x|m) p(m) を比較するのか？
  - つまり、なぜ単に、p(x|m) の比較で済ませないのか？
- p(m) が m ごとに異なるからである。
  - 例えば、x は咳があるかないか、m は風邪か、風邪でないか、としよう
  - p(咳|風邪)=0.9, p(¬咳|風邪)=0.1, p(咳| ¬風邪)=0.5, p(¬咳| ¬風邪)=0.5 としよう。つまり、風邪なら確率0.9で咳、風邪でなければ確率0.5で咳をするとする。
  - この場合、咳があれば必ず風邪と診断することになる。
  - しかし、実際には、風邪になる確率 p(風邪) は0.2であるとすれば、
    - 風邪で咳がある確率は、p(咳| 風邪)p(風邪) = 0.18 であるのに、
    - 風邪でないのに咳がある確率は p(咳| ¬風邪)p(¬風邪) = 0.4 となる。
  - つまり、風邪でない確率の方が高いのに、風邪だと判断していることになる。
  - これを防ぐには、p(m) を考慮するしかない。

クラス確率だけでよいのか？

- m_c として、確率が非常に低いクラスをとる。例えば、極めて稀なしかし致死率の高い病気であったとする。
- p(x|m) はそこそこに大きい値であっても p(x|m)p(m) は非常に小さい値になり、この推定法では、m_c が推定されることがなくなる。
- コストを考えに入れればよい！ 例えば、c(m) p(x|m)p(m) を最大化するmを求めればよい
- しかし、また問題が発生！
- x が発熱であったとする。熱が出るとすぐm_c を推定しまう。
- これが、昔（今でも！）、診断システムが成功しなかった理由。症状を入れると、とにかく、重篤な病気から日常的な病気まで、いろいろ推定してくる。医者はどうしているのだろうか？

なぜ、「確率最大」か

- 前にも議論したが、
  - 風邪か風邪でないかを判断するのに、風邪の確率 0.6, 風邪でない確率 0.4 では困るから。
    - 治療するかしないかの二者択一をしないといけないから。
    - 二者択一でなかったり、繰り返し行えるなら、この限りではない。
  - ところで、「確率最大」という言葉は少し不正確である

ところで

- 条件付確率の定義から、
  $p(x|m)\ p(m) = p(x, m)$
  である。再び、条件付確率の定義を用いると
  $p(x, m) = p(x)\ p(m|x)$
  すなわち
  $p(x)\ p(m|x) = p(x|m)\ p(m)$
  書き換えると
  $p(m|x) = p(x|m)\ p(m) / p(x)$
  となる。これはご存じ「ベイズの定理」である。
- 従って、さきほど行った推定は, x が定数であるから
  $\mathrm{argmax}_m\ p(x|m)\ p(m) = \mathrm{argmax}_m\ p(x|m)\ p(m) / p(x)$
  $= \mathrm{argmax}_m\ p(m|x)$

事後確率

- ところで、$p(m|x)$ はなんであろうか？
- これは、サンプルの属性値（検査値等）が x であると分かったとき、それを生成したモデルが m である条件付確率である。
- これをモデルm の事後確率という。「事後」というのは、サンプルが生成された後という意味である。
- 従って、$\mathrm{argmax}_m\ p(m|x)$ を求めることは事後確率を最大化することである。
  - ちなみに、xが既知のとき、$p(m|x)$ は確率である。総和も1になっている。
- 事後確率を最大化するパラメータ（今の場合、モデルm）を推定する量 $\mathrm{argmax}_m\ p(m|x)$ のことを maximum a posteriori estimator 事後確率最大化推定量（ MAP 推定量）という

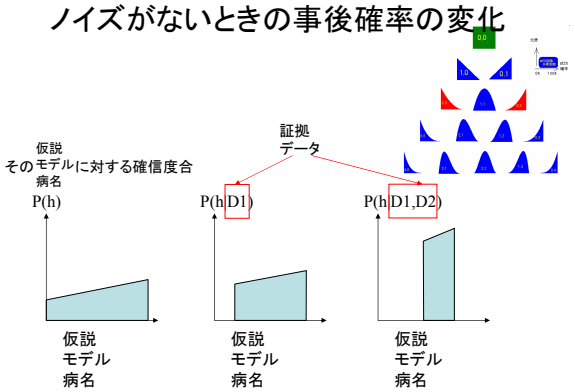
事前確率

- 事後があれば事前がある。
  $p(m|x) = p(x|m)\ p(m) / p(x)$
  の右辺に表れた $p(m)$ をモデルmの事前確率という。
- サンプルを見る前から知っている、モデルmの確率だから。

- ある患者を前にして、
- 診察も検査もしないとき、その患者が風邪である確率は p(風邪)となる。
  - 事前確率！
- 診察なり検査なりの結果x が分かると、風邪だと考えられる確率は p(風邪|x) に変化する。
  - 事後確率！

- 病名の候補がたくさんあるとき、データが増えれば増えるほど、一般には、病名に対する確信度合いが高くなる。

ノイズがないときの事後確率の変化



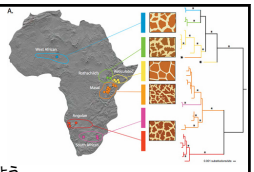
右上図は http://dev.team-lab.com/index.php?itemid=165 より

ベイズ推定

- これまで述べてきたように、（例えば、サンプル x を生成したモデル m を推定するにあたって）、
  - モデルmの事前分布を考え、
  - xを観測した後の事後分布を考え、
  - この事後分布に従って推定を行うことを

  ベイズ推定という。
- 特に、モデルm を推定するのは、その典型である。
- モデルmの事前分布が分からない、または、哲学として、知らない、知りようがない、存在しないと考え、$p(x|m)$を最大化するm を求める手法がある。それを最尤推定法と呼ぶ

最尤推定



- ベイズ推定で行うことは、
  $\mathrm{argmax}_m\ p(x|m)\ p(m)$
  を求めることであった。
- クラス確率 p(m) が全て等しい場合を考えてみよう
  - ちょっとずるいが、情報不足でクラス確率 p(m) が分からない場合は、p(m)は全部等しいと仮定してしまうことがある。この場合も含む
- その場合、行うことは
  $\mathrm{argmax}_m\ p(x|m)$
  を求めることになる。
- p(m)が何であってもこの公式を使うことも考えられる（前に不適当だといったが）。
- この場合、最大化しているのは、$p(x|m_1)$, $p(x|m_2)$,… であるが、その和は一般に1ではない。つまり、確率ではない。
  - 正規化（総和=1）にすればよいかというと、そもそも、総和してよいか、甚だ議論である。
- 確率と区別するために、これ（例えば、$p(x|m_1)$ ）を m_1 の尤度（likelihood）という
- 尤度を最大にするもの（今の場合 m ）を推定するので、この方法を最尤推定という（maximum likelihood estimation）

図は http://farm3.static.flickr.com/2245/2128695311_dedf0a67f8.jpg

ベイズ推定を実行するには

条件付き確率　事前確率

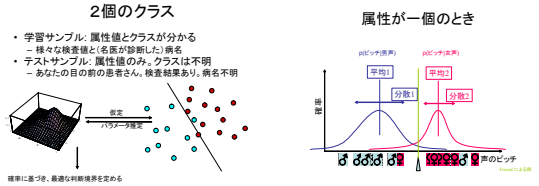
$$\underbrace{p(m|x)}_{\text{事後確率}} = \frac{p(x,m)}{p(x)} = \frac{p(x|m)p(m)}{p(x)}$$

であるから、ベイズ推定を行うには、事前確率 p(m) と条件付確率 p(x|m) を知る必要がある。

- p(m) はクラスmの頻度で推定すればよい
- では、p(x|m) はどうしたら推定できるだろうか？

p(x|m) の推定

- 以前のスライドを思い出してください



- m ごとに、分布を推定すればよいわけです。

簡単か？

- 考えてみると
  - 分布の形（つまり関数形）が分かっていて、
  - それが簡単（正規分布とか二項分布とか）かつ
  - 次元が低い（属性の個数が少ない）なら、

  確かに簡単である。
- しかし、世の中そんなに甘くない。
  - 分布の形なぞ分かりようがない
  - 正規分布のように綺麗なわけがない
  - 属性は山ほどある

  というのが普通である。

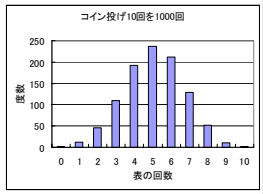
しかし、簡単化してみよう

- 多くの場合、正規分布や（離散変数: サイコロの目、コインの裏表の場合には）多項分布で近似できるから、分布は、
  - 連続値なら正規分布
  - 離散値なら多項分布

  で考えよう
- しかし、属性数が問題。

なぜ属性数が問題か？

- 前のスライドの図を思い出してください。
- コイン投げをして、コインの表が出る確率を推定する問題と考えてください。
- 正解は 0.5 です。
- しかし、10回投げたうち、3回以下しか表が出ない場合が169回、7回以上出てしまう場合が192回もある。
- つまり、2値属性のパラメータを1個推定するにも、サンプル10個では不足だということである。
- 独立な属性が10個あれば、それらのパラメータをまあまあの精度で推定するには、10 2^{10} =1万個のサンプルが必要になる。
- 一般にはなかなか難しい。



では、どうするか？