統計的機械学習
第3回　最尤推定法

2013/06/11

表現工学科　尾形研究室　ゼミ
野田　邦昭

最大事後確率則

- 最大事後確率則(maximum a posteriori probability rule): 入力パターンが属する可能性が最も高いカテゴリを選ぶ

- これは, x を事後確率が最大になるカテゴリに分類することに対応する.

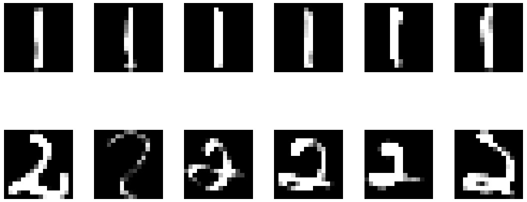
$$\arg\max_{y} \; p(y \mid x)$$

訓練標本からの識別器の構成

- 事後確率 $p(y|x)$ が分かれば，最大事後確率則によってパターンを分類できる．
- しかし，$p(y|x)$ は実際には未知．
- 訓練標本(training sample) $\{(x_i, y_i)\}_{i=1}^{n}$ ：属するカテゴリが既知のパターン

$$x_i \in \Re^d \qquad y_i \in \{1, 2, \ldots, c\}$$

- 手持ちの訓練標本を用いて事後確率を推定することにする．



訓練標本

- 訓練標本は次のように生成されたと仮定：
  - カテゴリを事前確率 $p(y)$ に従ってランダムに選ぶ
  - 選んだカテゴリに対して，パターンを条件付き確率 $p(x \mid y)$ に従ってランダムに取り出す
- 訓練標本 $\{(x_i, y_i)\}_{i=1}^{n}$ は，独立に同一な分布 $p(x, y)$ に従う(independent and identically distributed; i.i.d.)

事後確率の推定

- 事後確率 $p(y \mid x)$を直接推定するのは難しい
- ベイズの定理を用いれば,

$$p(y \mid x) = \frac{p(x \mid y)p(y)}{p(x)} \propto \underbrace{p(x \mid y)}_{\text{条件付き確率}}\underbrace{p(y)}_{\text{事前確率}}$$

- 条件付き確率と事前確率を推定することにする

事前確率の推定

- n_y :カテゴリ y に属する訓練標本の数
- 事前確率は離散的な確率分布なので,単純にそのカテゴリに含まれる標本の割合で推定する.

$$\hat{p}(y) = \frac{n_y}{n}$$

条件付き確率の推定

- 条件付き確率は連続的な確率分布なので，事前確率のように単純には推定できない．
- パラメトリック法：
  - 最尤推定法
  - ベイズ推定法
  - 最大事後確率推定法
- ノンパラメトリック法：
  - カーネル密度推定法
  - 最近傍密度推定法

条件付き確率の推定(続き)

- 以後，簡単のため，条件付きでない確率密度関数 $p(x)$ を全訓練標本 $\{x_i\}_{i=1}^{n}$ から推定する問題を考える.
- カテゴリ y に関する条件付き確率 $p(x \mid y)$ を推定するときは，y に属する n_y 個の標本のみを用いればよい.

パラメトリック法

- θ :パラメータ(parameter)
- パラメトリックモデル(parametric model)
  $q(x;\theta)$:有限次元のパラメータで記述された確率密度関数の族
- パラメトリック法(parametric method):
  パラメトリックモデルを用いて確率密度関数を推定する方法

例：ガウスモデル（正規分布）

- d次元の確率ベクトル： $x = (x^{(1)}, x^{(2)}, \ldots, x^{(d)})^T$
- ２つのパラメータ：
  - d 次元ベクトル μ
  - d 次元正値対称行列 Σ

$$q(x;\mu,\Sigma) = \frac{1}{(2\pi)^{d/2} \det(\Sigma)^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu) \right)$$

- ガウス分布の期待値，分散共分散行列

$$E[x] = \mu \qquad V[x] = \Sigma$$

最尤推定法

- 最尤推定法(maximum likelihood estimation):
手元にある訓練標本が最も生起しやすいようにパラメータ値を決める方法
- “最も尤もらしいようにパラメータの値を決める”
- 訓練標本 $\{x_i\}_{i=1}^{n}$ がモデル $q(x;\theta)$ から生起する確率:

$$p(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} q(x_i;\theta)$$

- 尤度(likelihood):これを θ の関数とみたもの

$$L(\theta) = \prod_{i=1}^{n} q(x_i;\theta)$$

最尤推定法(続き)

- 尤度を最大にするようにパラメータの値を決定.

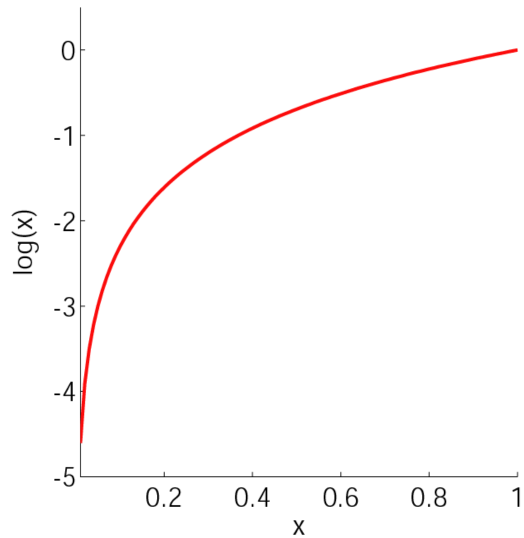
$$\hat{\theta}_{ML} = \mathop{\arg\max}_{\theta \in \Theta} L(\theta)$$

$$L(\theta) = \prod_{i=1}^{n} q(x_i; \theta)$$

- $\hat{\theta}_{ML}$ を最尤推定量(maximum likelihood estimator)とよぶ.

対数関数

- 対数(log)は，単調増加の関数.
- 対数をとってもその大小関係は変わらない.



対数尤度

- 対数をとれば積が和になることから，実際に最尤推定量を計算するときは，対数をとった尤度（対数尤度，log-likelihoodとよぶ）を用いた方が計算しやすいことが多い．

$$\hat{\theta}_{ML} = \underset{\theta \in \Theta}{\arg\max} \log L(\theta)$$

$$\log L(\theta) = \sum_{i=1}^{n} \log q(x_i;\theta)$$

- 最尤推定量は次の尤度方程式（likelihood equation）を満たす．

$$\left. \frac{\partial}{\partial \theta} \log L(\theta) \right|_{\theta = \hat{\theta}_{ML}} = 0$$

カテゴリの対数事後確率

- 対数をとっても大小関係は変わらないため，事後確率の対数をとったものを用いる.
- ベイズの定理より

$$p(y \mid x) = \frac{p(x \mid y)\,p(y)}{p(x)}$$

- 従って,

$$\log p(y \mid x) = \log p(x \mid y) + \log p(y) + C$$

$$C = -\log p(x)$$ ：定数

推定方法

- $\{(x_i, y_i)\}_{i=1}^{n}$:訓練標本 $(x_i, y_i) \overset{i.i.d.}{\sim} p(x, y)$

  $x_i \in D\left(\subset \Re^d\right)$ $y_i \in \{1,2,\ldots,c\}$

- n_y :カテゴリ y に属する訓練標本数
- 事前確率 $p(y)$:カテゴリ y に含まれる標本の割合で推定

$$\hat{p}(y) = \frac{n_y}{n}$$

- 条件付き確率 $p(x \mid y)$:ガウスモデルに対する最尤推定

$$\hat{p}(x \mid y) = q(x; \hat{\mu}_y, \hat{\Sigma}_y)$$

ガウスモデル(正規分布)

- d次元の確率ベクトル: $x = (x^{(1)}, x^{(2)}, \dots, x^{(d)})^T$
- 2つのパラメータ:
  - カテゴリ y の平均ベクトル　μ_y
  - カテゴリ y の分散共分散行列 Σ_y

$$q(x; \mu_y, \Sigma_y) = \frac{1}{(2\pi)^{d/2} \det(\Sigma_y)^{1/2}} \exp\left(-\frac{1}{2}(x - \mu_y)^T \Sigma_y^{-1} (x - \mu_y) \right)$$

$\det(\Sigma)$: Σの行列式

ガウスモデルにおける最尤推定

■ガウスモデルの最尤推定量

$$\hat{\mu}_y = \frac{1}{n_y} \sum_{i:y_i=y} x_i$$

$$\hat{\Sigma}_y = \frac{1}{n_y} \sum_{i:y_i=y} (x_i - \hat{\mu}_y)(x_i - \hat{\mu}_y)^T$$

$\sum_{i:y_i=y}$ ： $y_i = y$ を満たす i に関する和

カテゴリの対数事後確率の計算

- 分類したい入力パターンを x とすれば,

$$\log \hat{p}(y \mid x)$$

$$= -\frac{d}{2}\log 2\pi - \frac{1}{2}\log\det(\hat{\Sigma}_y) - \frac{1}{2}(x-\hat{\mu}_y)^T \hat{\Sigma}_y^{-1}(x-\hat{\mu}_y) + \log\frac{n_y}{n} + C$$

$$= -\frac{1}{2}\underbrace{(x-\hat{\mu}_y)^T \hat{\Sigma}_y^{-1}(x-\hat{\mu}_y)}_{\text{マハラノビス距離 (Mahalanobis distance)}} - \frac{1}{2}\log\det(\hat{\Sigma}_y) + \log n_y + C'$$

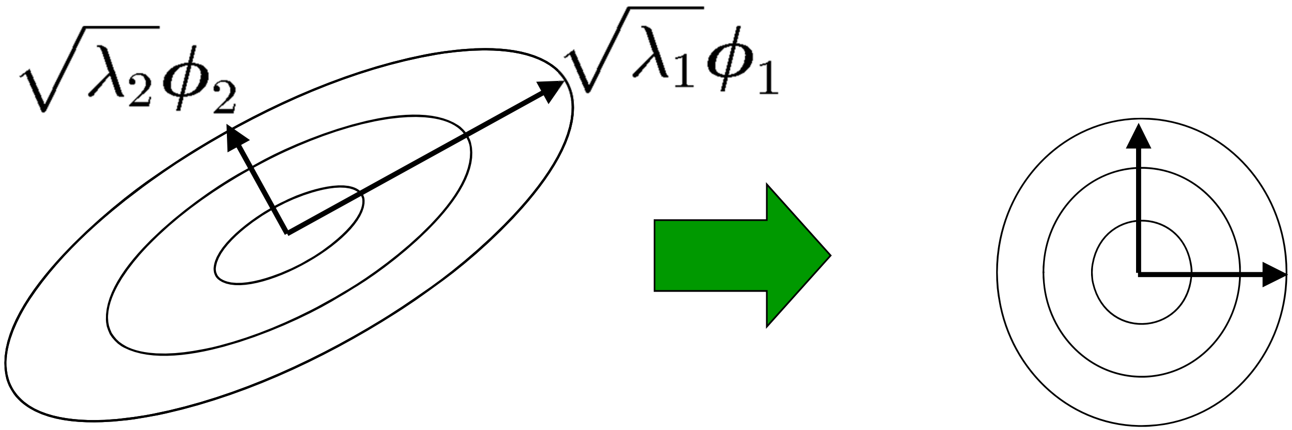
$$C' = -\frac{d}{2}\log 2\pi - \log n + C$$

- カテゴリの対数事後確率は x の二次形式

マハラノビス距離

- "楕円を正円に変換した距離"

$$(x-\mu)^T \Sigma^{-1}(x-\mu) = \left\| \Sigma^{-1/2}(x-\mu) \right\|^2$$



$$\mathbf{\Sigma} = \lambda_1 \phi_1 \phi_1^\top + \lambda_2 \phi_2 \phi_2^\top$$

- $\Sigma^{-1/2}(x-\mu)$ の変換のことを球状化(sphering), あるいは, 白色化(whitening)という.

共分散行列が共通のとき

- 各カテゴリの分散共分散行列が等しいという前提知識があるときを考える:

$$\Sigma_1 = \Sigma_2 = \cdots = \Sigma_c = \Sigma$$

- 共通の分散共分散行列 Σ の最尤推定量は

$$\hat{\Sigma} = \frac{1}{n}\sum_{y=1}^{c}\sum_{i:y_i=y}(x_i - \hat{\mu}_y)^T(x_i - \hat{\mu}_y)$$

$$= \sum_{y=1}^{c}\frac{n_y}{n}\hat{\Sigma}_y$$

各カテゴリの分散共分散行列の重み付き平均

共分散行列が共通のとき(続き)

- 各カテゴリの分散共分散行列が等しい時,

$$\log \hat{p}(y|x)$$

$$= -\frac{1}{2}x^T\hat{\Sigma}^{-1}x + \hat{\mu}_y^T\hat{\Sigma}^{-1}x - \frac{1}{2}\hat{\mu}_y^T\hat{\Sigma}^{-1}\hat{\mu}_y - \frac{1}{2}\log\det(\hat{\Sigma}) + \log n_y + C'$$

$$= \hat{\mu}_y^T\hat{\Sigma}^{-1}x - \frac{1}{2}\hat{\mu}_y^T\hat{\Sigma}^{-1}\hat{\mu}_y + \log n_y + C''$$

$$C'' = -\frac{1}{2}x^T\hat{\Sigma}^{-1}x - \frac{1}{2}\log\det(\hat{\Sigma}) + C'$$

- カテゴリの対数事後確率は x の一次形式

決定境界の計算

- カテゴリ数が2のとき，決定境界は

$$\hat{p}(y=1\,|\,x)=\hat{p}(y=2\,|\,x)$$

- ガウスモデルと最尤推定を用いたとき，決定境界は二次形式.
- 更に分散共分散行列が共通のとき，決定境界は一次形式，即ち，超平面(hyper-plane).

$$a^T x+b=0$$

$$a=\hat{\Sigma}^{-1}(\hat{\mu}_1-\hat{\mu}_2)$$

$$b=-\tfrac{1}{2}(\hat{\mu}_1^T\hat{\Sigma}^{-1}\hat{\mu}_1-\hat{\mu}_2^T\hat{\Sigma}^{-1}\hat{\mu}_2)+\log(n_1/n_2)$$

- この場合を特に，フィッシャーの線形判別分析(Fisher's linear discriminant analysis)という.