統計的機械学習
第4回　モデル選択

2013/06/18

表現工学科　尾形研究室　ゼミ
野田　邦昭

最大事後確率則に基づくパターン認識

- 事後確率:

$$p(y \mid x) \propto p(x \mid y)p(y)$$

- 最尤法による事前確率の推定:

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

- 最尤法による条件付き確率の推定:

$$\hat{p}(x \mid y) = q(x; \hat{\theta}_{ML})$$

- 最尤推定法:あらかじめ定めたパラメトリックモデルの中から最も尤もらしい確率密度関数を選ぶ

最尤推定法とモデルの選択

- モデルはどうやって定めればよいか？
- 例えば，一口にガウスモデルといっても，分散共分散行列が
  - 任意の正値対称行列の場合（自由度 $d(d+1)/2$ ）
  - 対角行列で対角成分が異なる場合（自由度 d ）
  - 対角行列で対角成分が等しい場合（自由度 1 ）

  などいろいろなものがある.
- どれを選べばよいか？

多次元ガウスモデル

- 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:
  - d 次元ベクトル μ
  - d 次元正値行列 Σ

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

- 正規分布の期待値, 分散共分散行列

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

多次元ガウスモデル（続き）

- 共分散がゼロ（即ち $\Sigma = diag(\sigma_i^2)$ ）のとき

$diag(\sigma_i^2)$: $\sigma_1^2, \sigma_2^2, \ldots, \sigma_d^2$ を対角成分に持つ対角行列

$$q(x;\mu,\{\sigma_i^2\}_{i=1}^d) = \frac{1}{(2\pi)^{d/2}\prod_{i=1}^d \sigma_i}\exp\left(-\sum_{i=1}^d \frac{(x^{(i)}-\mu^{(i)})^2}{2\sigma_i^2}\right)$$

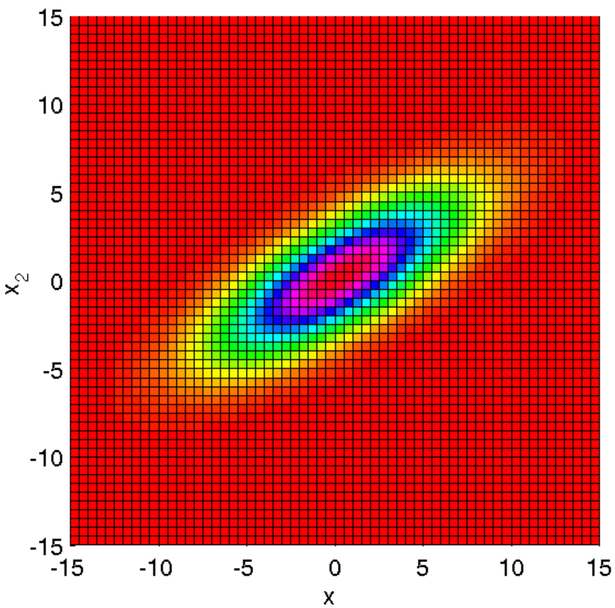
- さらに分散が等しい（即ち $\Sigma = \sigma^2 I$ ）のとき

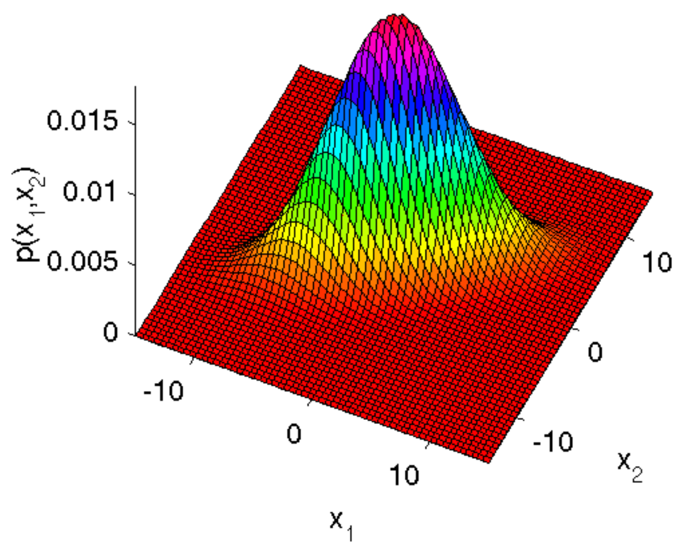
I :単位行列

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

多次元正規分布の例(1)

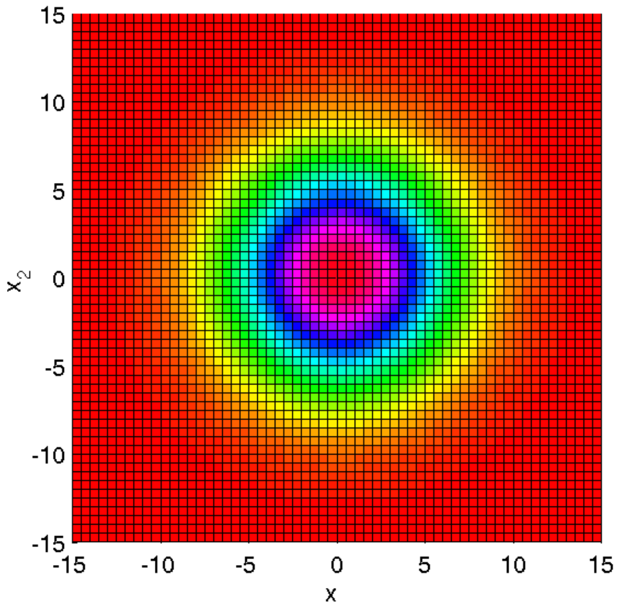
$d = 2$ $\mu = (0,0)^T$ $\Sigma = \begin{pmatrix} 20 & 10 \\ 10 & 9 \end{pmatrix}$

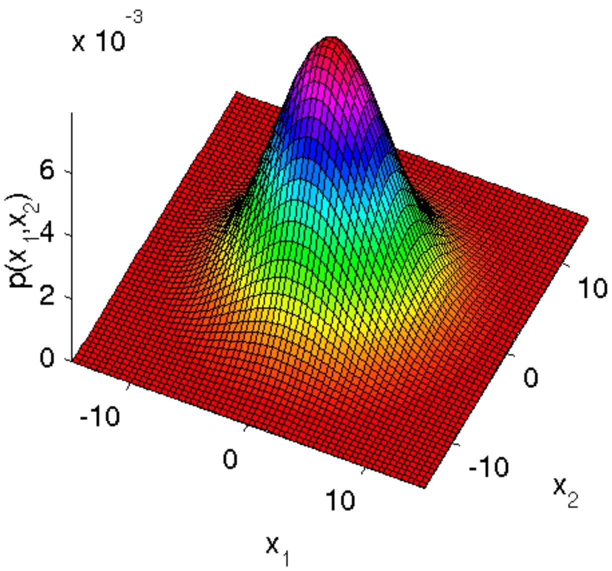




多次元正規分布の例(2)

$d = 2$ $\mu = (0,0)^T$ $\Sigma = \begin{pmatrix} 20 & 0 \\ 0 & 20 \end{pmatrix}$





複雑なモデルがよい理由

- パラメトリック法では，モデルの中に真の確率密度関数を良く近似するものが含まれていなければ，そもそもよい結果は得られない．

↓

- 真の確率密度関数を含むよう，パラメータ数の多い表現力の豊かなモデルを選ぶべき．

単純なモデルがよい理由

- 訓練標本数がパラメータ数と比べてそれほど多くない場合，最尤推定法のよさは，理論的には保証されない．

- 推定量の分散が十分小さくなるよう，パラメータ数の少ないモデルを選ぶべき．

モデル選択

- 訓練標本を用いて適切なモデルを選ぶことを，モデル選択(model selection)という．

1. いくつかのパラメトリックモデルを用意する．

$$\{q_i(x;\theta)\}_i$$

2. 各々のモデルに対して最尤推定量$\hat{\theta}_{ML_i}$を求める．
3. それぞれのモデルから得られた確率密度関数の推定量を次のように定める．

$$\hat{p}_i(x) = q_i(x;\hat{\theta}_{ML_i})$$

4. $\{\hat{p}_i(x)\}_i$ から真の確率密度関数 $p(x)$ に最も「近い」ものを選ぶ．

確率密度関数の近さを測る規準

- カルバック・ライブラー情報量(Kullback-Leibler information):

$$KL(p \,\|\, \hat{p}) = \int_D p(x) \log \frac{p(x)}{\hat{p}(x)} dx$$

- KL情報量は常に非負で, $\hat{p}(x) = p(x)$ のときだけゼロになる.
- 従って, KL情報量が小さければ, $\hat{p}(x)$ は「よい」といえる.

距離

- 数学的な距離(distance)の定義
  1. $d(x,y) \geq 0$
  2. $d(x,y) = d(y,x)$
  3. $d(x,y) = 0 \Leftrightarrow x = y$
  4. $d(x,y) + d(y,z) \geq d(x,z)$

- KL情報量は2と4を満たさないため，厳密には距離ではないことに注意．

$$KL(p \| \hat{p}) \neq KL(\hat{p} \| p)$$

KL情報量の推定

- KL情報量には未知の確率密度関数 $p(x)$ が含まれているため，直接計算できない.
- 訓練標本からKL情報量を推定する.

$$KL(p \| \hat{p}) = \underbrace{\int_D p(x) \log p(x) dx}_{\text{エントロピー(entropy)}} - \int_D p(x) \log \hat{p}(x) dx$$

- エントロピーは定数なので，第二項目のみを推定すればよい.

KL情報量の単純な推定

- 負の対数尤度はKL情報量の近似:

  $n \to \infty$ のとき，一致する（大数の法則）

$$-\frac{1}{n}\sum_{j=1}^{n}\log \hat{p}(x_j) \to -\int_D p(x)\log \hat{p}(x)dx$$

- 尤度を最大にするモデルを選べばよい？
- 複雑なモデルほど尤度は大きいので，尤度最大のモデルを選ぶと，常に最も複雑なモデルが選ばれてしまう.
- もう少し精密なKL情報量の近似が必要！

赤池の情報量規準(AIC)

- 赤池の情報量規準(Akaike's information criterion):

$$AIC = \underbrace{-\sum_{j=1}^{n} \log \hat{p}(x_j)}_{\text{負の最大対数尤度}} + \underbrace{\dim \theta}_{\text{パラメータ数}}$$

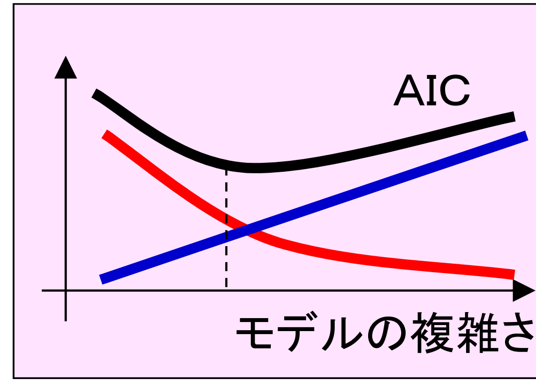
- 訓練標本が十分に多いとき

$$\frac{1}{n} AIC \approx -\int_D p(x) \log \hat{p}(x) dx$$

- AICを最小にするモデルを選ぶ.

AICの直感的解釈

$$AIC = \underbrace{-\sum_{j=1}^{n} \log \hat{p}(x_j)}_{\text{負の最大対数尤度}} + \underbrace{\dim\theta}_{\text{パラメータ数}}$$



- モデルが複雑な場合，負の最大対数尤度は小さいがパラメータ数が大きいためAICは大きい．
- モデルが単純な場合，パラメータ数は小さいが負の最大対数尤度が大きいためAICは大きい．
- モデルが程よく複雑な場合，二つの項がバランスよく小さくなり，AICは小さい．

オッカムのかみそり

- オッカムのかみそり(Occam’s Razor): 14世紀の哲学者オッカムによる「不必要に実体の数を増やしてはならない」という提言
- 現代でも科学理論を構築する上での基本的な指針としてよく用いられる.
- 「現象を同程度うまく説明する仮説があるなら，単純な方を選べ」
- 「けちの原理(principle of parsimony)」ともよばれる.

オッカムのかみそり(続き)

$$AIC = \underbrace{-\sum_{j=1}^{n} \log \hat{p}(x_j)}_{\text{負の最大対数尤度}} + \underbrace{\dim \theta}_{\text{パラメータ数}}$$

- 「現象を同程度うまく説明する仮説」: 尤度が等しい二つのモデル
- 「単純な方」: パラメータ数が少ない方
- AICはオッカムのかみそりの妥当性を理論的に裏付けている！

AICの精密化

- 理論的には，AICよりも次の竹内の情報量規準(Takeuchi’s information criterion)の方がより精密．

$$TIC = -\sum_{j=1}^{n} \log q(x_j;\hat{\theta}_{ML}) + \mathrm{trace}(JH^{-1})$$

$$J_{j,k} = \frac{1}{n}\sum_{i=1}^{n} \frac{\partial}{\partial \theta^{(j)}} \log q(x_i;\theta)\Big|_{\theta=\hat{\theta}_{ML}} \frac{\partial}{\partial \theta^{(k)}} \log q(x_i;\theta)\Big|_{\theta=\hat{\theta}_{ML}}$$

$$H_{j,k} = -\frac{1}{n}\sum_{i=1}^{n} \frac{\partial^2}{\partial \theta^{(j)} \partial \theta^{(k)}} \log q(x_i;\theta)\Big|_{\theta=\hat{\theta}_{ML}}$$

AICの精密化（続き）

■ TICの近似性能

$$E[\tfrac{1}{n}TIC] = E[-\int_D p(x)\log\hat{p}(x)dx] + o(\tfrac{1}{n})$$

- $f(n) = o\big(g(n)\big)$: $n \to \infty$ のとき $|f(n)| < Cg(n)$
- $f(n) = O\big(g(n)\big)$: $n \to \infty$ のとき $f(n)/g(n) \to 0$

■ TICは $1/n$ のオーダまで不偏

■ モデルが真の確率密度関数を含むとき，即ち $p(x) = q(x;\theta_{true})$ のとき，TIC=AIC.