情報認識
「最尤推定法（第４，５章）」

- 担当教員：　杉山　将（計算工学専攻）
- 居室：　W8E-505
- 電子メール:　　sugi@cs.titech.ac.jp

「情報認識」の全体構成

- 識別関数のよさを測る規準
- 条件付き確率の推定
  - パラメトリック法
    - 最尤推定法，EMアルゴリズム
    - 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
    - 核密度推定法
    - 最近傍密度推定法
- 手書き文字認識の計算機実習

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機の構成

- 事後確率 $p(y \mid x)$ 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， $p(y \mid x)$ は実際には未知.
- 訓練標本(training sample) $\{x_i\}_{i=1}^{n}$: 属するカテゴリが既知のパターン
- 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 $\{x_i\}_{i=1}^{n}$ 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,

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

- 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，事前確率のとき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：
  - 核密度推定法
  - 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)
- Θ : パラメータの定義域(domain)
- 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.

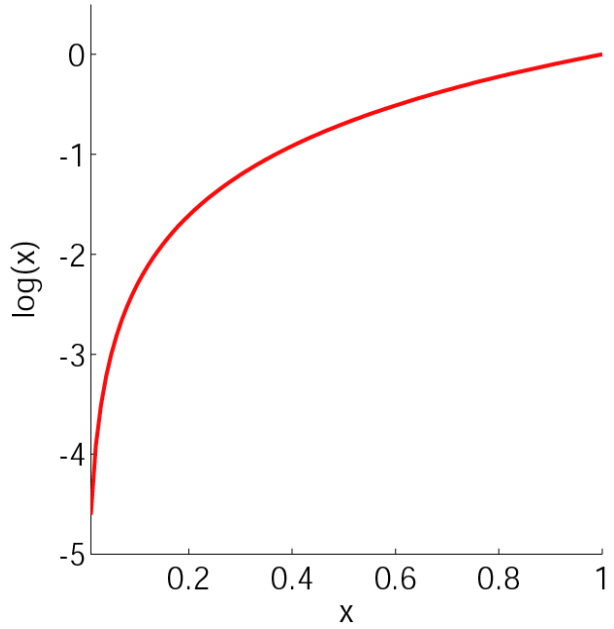
$$\hat{\theta}_{ML} = \arg\max_{\theta \in \Theta} L(\theta)$$

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

- $\hat{\theta}_{ML}$ 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．

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

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

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

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

演習

- 1次元のガウスモデル

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

に対する最尤推定量 $\hat{\mu}_{ML},\hat{\sigma}^2_{ML}$ を求めよ.

最尤推定量のよさ

- θ_{opt} :最適なパラメータ
- 一致性(consistency):

  $n \to \infty$ のとき, $\hat{\theta}_{ML} \to \theta_{opt}$

  → 標本が無限にたくさんあれば,
  最適なパラメータが求まる
- 漸近有効性(asymptotic efficiency):

  $n \to \infty$ のとき, $\hat{\theta}_{ML}$ は漸近正規推定量の中で漸近分散が最小となる推定量

  → 標本が十分にたくさんあるとき,
  推定結果は安定している

最尤推定量の振る舞い

- 漸近正規性(asymptotic normality)：訓練標本数 n が十分大きいとき，最尤推定量 $\hat{\theta}_{ML}$ は期待値が θ_{opt}，分散共分散行列が $\frac{1}{n}F^{-1}$ の正規分布に近似的に従う

$$\sqrt{n}(\hat{\theta}_{ML} - \theta_{opt}) \to N(0, F^{-1})$$

$$F_{i,j} = E\left[\frac{\partial}{\partial\theta^{(i)}}\log q(x;\theta)\frac{\partial}{\partial\theta^{(j)}}\log q(x;\theta)\right]$$

フィッシャー情報行列(Fisher information matrix)

$$\theta = (\theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(\dim\theta)})^T$$

まとめ

- 訓練標本から確率密度関数を推定したい.
- パラメトリックモデルを用いる.
- パラメータの値を最尤推定法で決定する.
- 最尤推定法のよさは理論的に保証される.

小レポート（第3回）

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

1. 上記のガウスモデルの最尤推定量は次式で与えられることを証明せよ.

$$\hat{\mu}_{ML}=\frac{1}{n}\sum_{i=1}^{n}x_i \qquad \hat{\Sigma}_{ML}=\frac{1}{n}\sum_{i=1}^{n}(x_i-\hat{\mu}_{ML})(x_i-\hat{\mu}_{ML})^T$$

- 証明のヒント:
  - ベクトルでの微分の公式

$$\frac{\partial}{\partial\mu}\mu^T\Sigma\mu=2\Sigma\mu \qquad \frac{\partial}{\partial\mu}\mu^T\Sigma x=\Sigma x$$

  - 行列での微分の公式

$$\frac{\partial}{\partial\Sigma}x^T\Sigma^{-1}x=-\Sigma^{-1}xx^T\Sigma^{-1} \qquad \frac{\partial}{\partial\Sigma}\log|\Sigma|=\Sigma^{-1}$$

小レポート（続き）

2. Octaveなどを用いて以下の実験を行え.

- 平均0，分散1の標準正規分布に独立に従う標本をn個生成せよ．そして，下記のガウスモデルのパラメータ μ, σ^2 を最尤推定法で決定せよ.

$$p(x;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2} \right)$$

- 真の標準正規分布，および最尤推定法で求めた分布の確率密度関数を図示せよ.
- 標本数nを変化させ，推定結果がどのように変化するか考察せよ.

Octaveのサンプルプログラム

ex3.m

```
clear all
n=5; mu=0; sigma=1;
xx=sigma*randn(n,1)+mu;

mu_MLE=mean(xx);
sigma_MLE=std(xx,1);

x=-4:0.1:4;
y=normal_pdf(x,mu,sigma);
y_MLE=normal_pdf(x,mu_MLE,sigma_MLE);
plot(x,y,'r-',x,y_MLE,'b-');
legend('true','estimated')
print -deps gauss1d_pdf.eps
```