機械学習
2-5

工学院情報通信系
中原 啓貴

今日の内容

- 決定的機械学習と統計的機械学習
  - 生成モデルと識別モデル
  - 最尤推定法
  - ベイズ推定
- ＋上記内容の演習

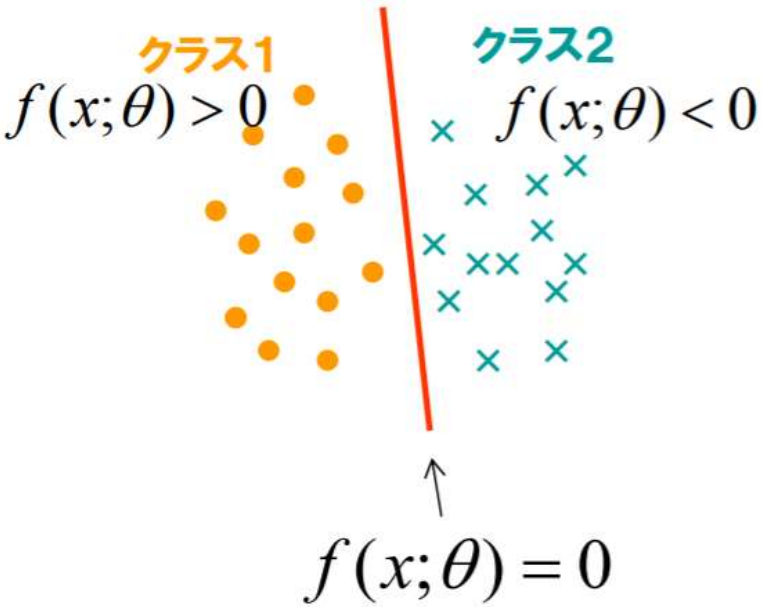
統計的機械学習

- 識別モデル（確率分布の仮定なし）
  - ニューラルネットワーク, SVM, ランダムフォレスト
  - 汎化誤差最小化問題
  - 実問題毎に適した学習法の探索
- 生成モデル（確率分布の仮定あり）
  - ベイズ学習
  - 潜在的な情報のモデル化
    →高い汎用性

生成モデルと識別モデル

識別モデル

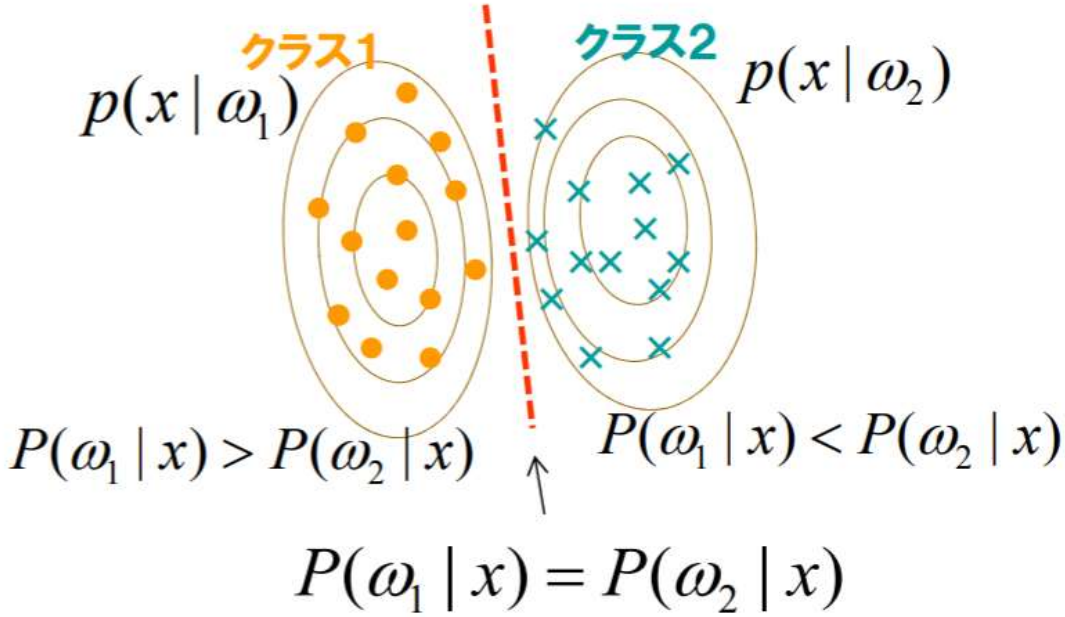
データのクラス境界
を直接学習



データの生成過程は考慮せず、
所与のデータで問題を直接解く

生成モデル

クラスごとの生成モデル
を確率分布として学習

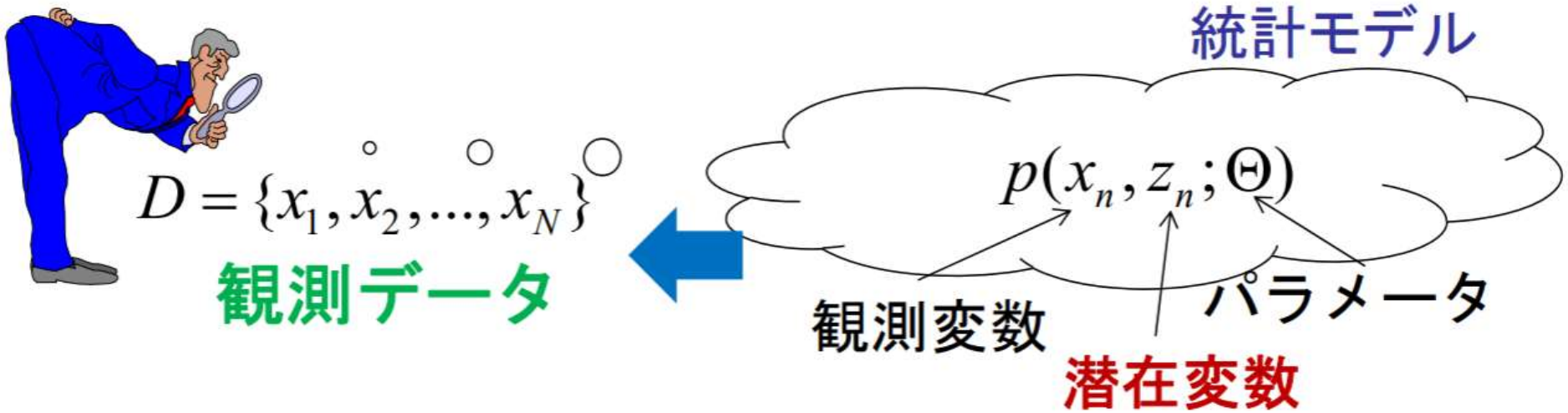


クラス事後確率により識別
（クラス境界は結果として得られる）

Source: https://www.nii.ac.jp/userdata/karuizawa/h23/111104_3rdlecueda.pdf
(上田, “統計的機械学習入門,” 2011)

生成モデルの構成要素

- 生成モデル(Generative model)
  - 観測データがどのような確率モデルとして生成されたか仮説を立てる
- 潜在変数
  - 本来観測されない隠れ変数
- パラメータ
  - 確率分布(統計モデル)を特徴づける変数



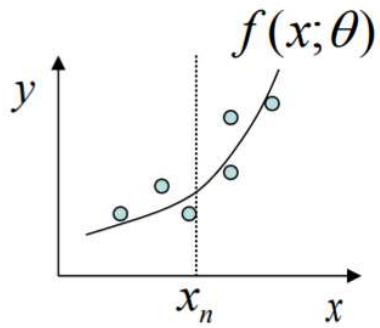
観測変数のみの統計モデル

回帰モデル $y = f(x;\theta) + \varepsilon$ ← 付加正規ノイズ

真値

平均0,分散 σ^2

観測データ $D = \{x_n, y_n\}$



統計モデル：真値を平均値とする1次元正規分布

$$p(y_n \mid x_n;\theta) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\{-\frac{1}{2\sigma^2}(y_n - f(x_n;\theta))^2\}$$

最尤推定

Note: 正規ノイズモデルでは、最尤推定は最小自乗推定と等価

パラメトリックモデルの学習

- パラメトリックモデル：パラメータを持った確率分布を用いたモデル
- 学習法

1. パラメータを含むモデルを設定
2. パラメータを評価する基準を定める
3. 最良の評価を与えるパラメータを決定

- 最尤推定法
  - 「あるデータが得られる確率」を設定して最良のパラメータを決定する

パラメータを含むモデルの設定

- 「M次多項式の関係+標準偏差σの誤差が含まれている」と仮定
  - 最小二乗法+誤差を仮定

$$f(x) = \sum_{m=0}^{M} w_m x^m$$

- 観測点 x_n における観測値 t は $f(x_n)$ を中心としておよそ $f(x_n)\pm\sigma$ の範囲に散らばる

  →正規分布で表現

$$N(t|f(x_n),\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}\{t-f(x_n)\}^2}$$

分布の中心が
$f(x_n)$であることに注意

- t_n: データセット, t_0: 次に観測されるデータ

観測変数のみのモデル例

- 係数w_mとσが学習したいパラメータ

$$N(t|f(x_n),\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}\{t-f(x_n)\}^2}$$

$$f(x) = \sum_{m=0}^{M} w_m x^m$$

- 正規分布以外の確率分布は可能？
  - 答えはYes, 新たな仮定を立てる必要はあり得る
  - t分布, ロジスティック分布, 一様分布など

尤度関数によるパラメータの評価

- トレーニングセットに含まれるデータが得られる確率を計算
  - 観測点 x_n におけるデータ t_n が得られる確率

$$N(t_n|f(x_n),\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}\{t_n - f(x_n)\}^2}$$

- すべて(N個)の観測点について計算
- それぞれの確率の積を求める

$$P = \prod_{n=1}^{N} N(t_n|f(x_n),\sigma^2)$$

パラメータ(w_m,σ)の関数 → 尤度関数

尤度推定法

- **仮説：観測されたデータ（データセット）は、最も発生確率が高いデータに違いない**
- **仮説が正しいものとして確率Pを最大にするパラメータを決定 → 最尤推定法**
- **尤度関数の最大値問題に帰着**

  →　解析的に解ける！

(数式は省略)中井 悦司, ITエンジニアのための機械学習理論入門 (技術評論社) pp. 94-95 を参考

結論から言うと、多項式の係数は最小二乗法と同じ

標準偏差は平方根二乗誤差に一致

演習

https://github.com/HirokiNakahara

- 最尤推定法による回帰
- 標準偏差σも計算可能→尤もらしさも推定可能

生成モデルにおける正則化

- 最大事後確率(MAP)推定
  - 事前にパラメータの分布が既知（経験則、つまり主観確率）であることを事前分布としてパラメータに加える
  - w_nの事前分布を決めた上で事後確率を最大化
  - つまり, w_n に優劣をつける→正則化
- 実際には最尤関数にwの正則化項を追加

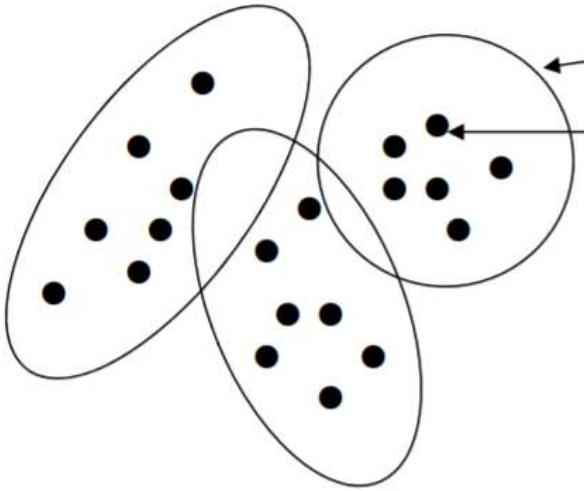
潜在変数を含む場合

混合分布モデル

Mixture models

$$p(x;\Theta) = \sum_{i=1}^{m} \alpha_i p_i(x;\theta_i)$$

混合比　　要素分布



x_nはどの要素分布から生成されたかは観測不可能

$z_n \in \{1,\cdots,m\}$: **潜在変数**

x_nを生成した要素分布のインデックス

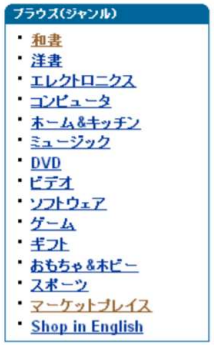
トピックモデル

- 潜在クラスを導入した統計モデル
  - データマイニングの実応用
  - テキストマイニング，協調フィルタリング

The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services," Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center's share will be $200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School, where music and the performing arts are taught, will get $250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.

"Arts"	"Budgets"	"Children"	"Education"
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI

D. Blei, A. Ng, M. Jordan, Latent Dirichlet Allocation, JMLR2003



テレビ番組　映画　CD / DVD　VIDEO / DVD

代表的なトピックモデル

- **PLSA: Probabilistic Latent Semantic Analysis**

$$P(\mathbf{w}) = \prod_{n=1}^{N} \sum_{z} P(z \mid \theta) P(w_n \mid z)$$

- **LDA: Latent Dirichlet Allocation**
  - **PLSAにおいて, トピック比率にディリクレ事前分布を仮定**

$$P(\mathbf{w}) = \int P(\theta) \prod_{n=1}^{N} \sum_{z} P(z \mid \theta) P(w_n \mid z) d\theta$$

ディリクレ事前分布

トピック比率

単語分布

協調フィルタリング

既に分かっている評点を元に，□ 部分を予測する．

アイテム（ 数千個 ～ ）

ユーザ（数万人～）

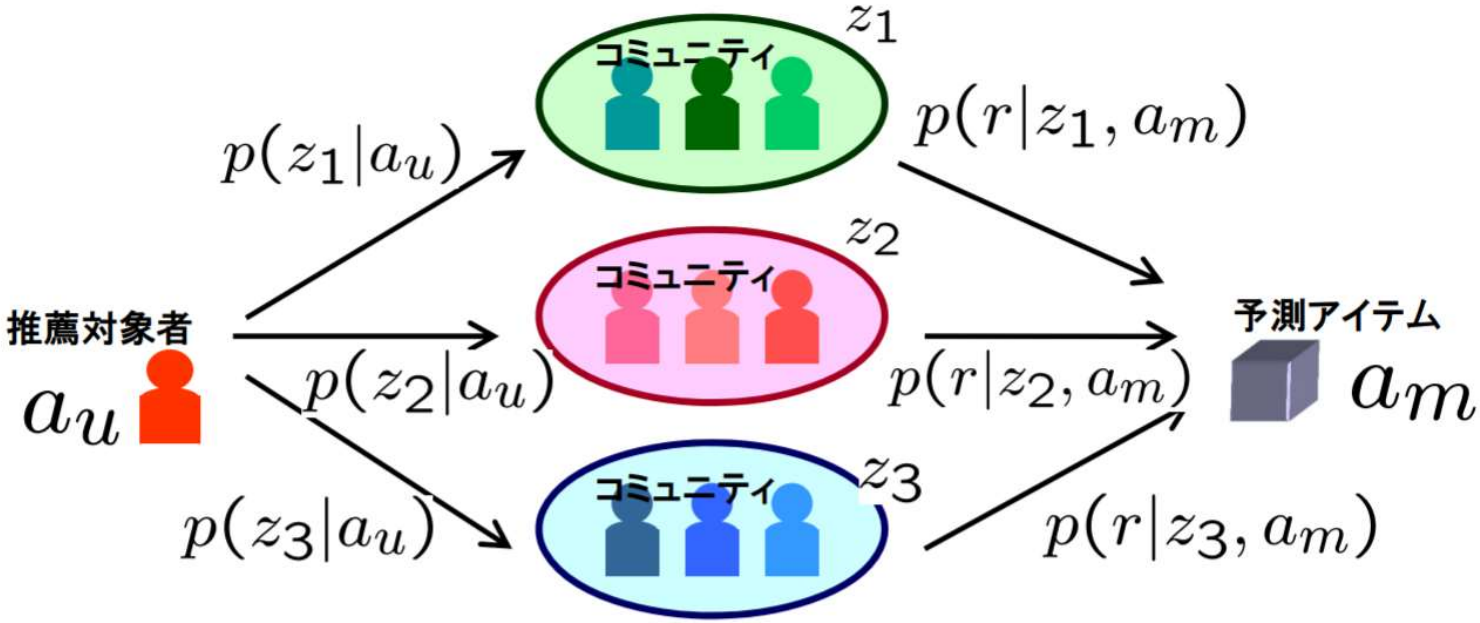
	ランボー	男はつらいよ	釣りバカ日誌	リング
A		4		3
B	3		2	5
C				2
D	4	4	2	
E			3	4
F	4		2	
G	4			2

1：嫌い ⟷ 5：好き

ただし，表の90％以上は欠損

probabilistic Latent Semantic Analysis (pLSA)

嗜好が似通った 潜在コミュニティ（潜在クラス）の存在を仮定



$$\text{評点} = \sum_{i=1}^{3} \underbrace{p(z_i|a_u)}_{\text{帰属度}} \underbrace{p(r|z_i, a_m)}_{\text{評点分布（正規分布を仮定）}}$$

演習

https://github.com/HirokiNakahara

・LDA法による協調フィルタリング

ベイズ推定モデル

- 最尤推定法との違い
- 最尤推定法
  - パラメータ（確率）を決定的に推定する
  - wは１つに定まる
- ベイズ推定モデル
  - パラメータを確率的に予測する
  - wの確率分布を求める
  - パラメータの更新にベイズの定理を利用

ベイズの定理

posterior　　prior

$$P(H \mid E, c) = \frac{P(E \mid H, c)P(H \mid c)}{P(E \mid c)}$$

hypothesis　effect　context

ベイズの定理は以下の積則から導出される

$$P(A, B \mid I) = P(A \mid B, I)P(B \mid I) = P(B \mid A, I)P(A \mid I)$$

条件付き確率：

$$P(A \mid B) = \frac{P(A, B)}{P(B)} = \frac{P(B \mid A)P(A)}{P(B)}$$

機械学習的には、

観測データ　潜在変数

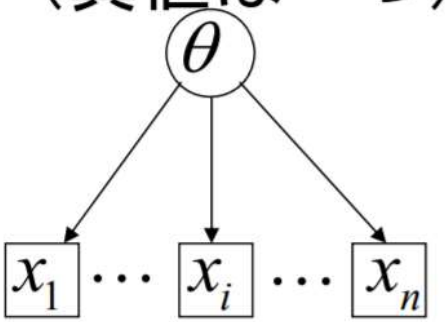
$$P(\theta \mid D, Z) = \frac{P(D \mid \theta, Z)P(\theta \mid Z)}{P(D \mid Z)}$$

モデル
パラメータ

ベイズモデルの特徴

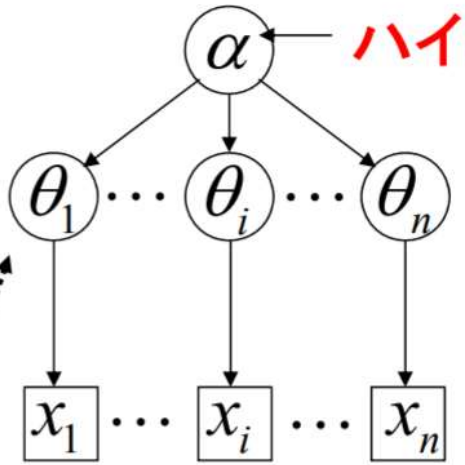
非ベイズ

パラメータは数学的な変数
（真値は一つ）



$p(x;\theta)$

ベイズモデル



ハイパーパラメータ

$p(\theta;\alpha)$

$p(x|\theta)$

パラメータは事前分布の一実現値

$\theta_i \in \{\theta_{(1)},\ldots,\theta_{(K)}\}$

K(分割)の事前分布は？

⇩

ノンパラメトリックベイズモデル (Dirichlet Process)

ベイズ推定によるパラメータの推定

- 正規分布を仮定
  - 分散σ^2は既知, 平均μを推定
- 最尤関数を下記の式に変更

$$N(t_n|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}\{t_n-\mu\}^2}$$

- したがって, N個の観測データに対する確率は

$$P(\boldsymbol{t}|u) = \prod_{n=1}^{N} N(t_n|\mu,\sigma^2)$$

ベイズの定理を適用

- ここで$P(\boldsymbol{t}|u)$は条件付き確率
- (ある特定の値の下で確率を考えている, ここではu
- ベイズの定理を適用

$$P(u|\boldsymbol{t}) = \frac{P(\boldsymbol{t}|u)}{\int_{-\infty}^{\infty} P(\boldsymbol{t}|\mu')P(\mu')d\mu'} P(\mu),$$

- $P(u|\boldsymbol{t})$は事後分布, トレーニングセット$\boldsymbol{t}$ uが観測された場合の確率
- ただし, 事前分布$P(\mu)$を正規分布として仮定

$$P(\mu) = N(\mu|\mu_0, \sigma^2)$$

ベイズ推定の性質

- 結果をまとめると, 平均μ_N, 分散β^{-1}を学習する

$$P(\mu|\boldsymbol{t}) = N\left(\mu\middle|\mu_N, \beta_N^{-1}\right)$$

- ここで, $\beta_N = N\beta + \beta_0$, $\beta = \frac{1}{\sigma^2}$
- したがって

$$\mu_N = \frac{\beta \sum_{n=1}^{N} t_n + \beta_0 \mu_0}{N\beta + \beta_0}$$

$\beta \to 0, (\sigma^2 \to \infty)$ をとると標本平均

N(データ数)を増やすと分散が小さくなることを示す

→このとき, 事前分布の影響(第2項)が小さくなる

観測値の分布の推定

- 平均μを取り得る確率が求まった→一意でない
- 未来の観測データtを求めたい
- 様々なμに対する正規分布$N(t|\mu,\sigma^2)$をそれぞれの確率の重みで足し合わせる

$$P(t) = \int_{-\infty}^{\infty} P(\mu|\boldsymbol{t})N(t|\mu,\sigma^2)d\mu$$

- 従って

$$P(t) = \int_{-\infty}^{\infty} N(\mu|\mu_N,\beta_N^{-1})N(t|\mu,\beta^{-1})d\mu$$

観測値の分布の推定

・積分を行うと

$$P(t) = N(t|\mu_N, \beta_N^{-1} + \beta^{-1})$$

ここで，$\beta_N = N\beta + \beta_0$より，トレーニングデータセット数を増やすと分散$\beta_N^{-1}$は小さくなる (真の平均に近づく)

演習

https://github.com/HirokiNakahara

・Naïve Bayes 分類

まとめ

- 統計的機械学習
  - 生成的モデルを学習
- 最尤頻度推定
- パラメトリックモデル
- トピックモデル
- 協調フィルタリング
- ベイズ推定