機械学習のお話

LTの目標

- 機械学習の全体像をざっくり把握する. (統計知らない人)

- 機械学習を統計学的視点からざっくり掴む.
  (統計知ってる人)

機械学習とは

経験 E , タスク T , パフォーマンス尺度 P

「(機械) 学習とは」 :

タスク Tについて, Pで測られるタスクの実行能力が
経験 Eを通じて向上すること.

例) 手書き文字認識(T) において,
書いてある数字を当てる正答率(P)が
データ(E) を通じて向上する.

最近流行りのワード

→

よく分からない

教師あり学習

分類問題

ベイジアンネットワーク

クラスタリング

SVM

深層学習

学習法での分類

- **教師あり学習**
- **教師なし学習**
- **強化学習**
- 半教師つき学習
- 転移学習
- マルチタスク学習

教師あり学習（supervied learning）

学習データに正解ラベルを付けて学習する方法.

訓練データ

$$D = \{(x_1,\ y_1),\ (x_2,\ y_2),\ \ldots,\ (x_n,\ y_n)\}$$

xが訓練用のデータベクトル. yが正解の値.
xから正しいyの値を出力する関数を作る.

(例.)

アヤメのデータからアヤメの種類を分類する.

x_i = (がく片の長さ, がく片の幅, 花びらの長さ, 花びらの幅)

y_i = (0: ヒオウギアヤメ, 1: ブルーフラッグ, 2: Virginica)

新しい x_* を与えたら 正しく種類が判定される関数を作る.

ラベル	がく片の長さ	がく片の幅	花びらの長さ	花びらの幅
ヒオウギアヤメ	3.0	1.0	5.0	10.0
ヒオウギアヤメ	3.4	1.3	4.4	1.1

・
・
・

教師なし学習

学習データに正解ラベルを与えず,
データのみからデータの本質的構造を抽出する学習法.

訓練データ

$$D = \{ x_1, x_2, \ \ldots, x_n \}$$

(例) : 生徒のテストデータから良いクラスタリングを見つける.

100人の生徒のテスト結果

x_i = (国語の点, 数学の点, 英語の点, 社会の点, 理科の点)

生徒をクラス内で実力差の少ない3つのクラスに分ける.

国語	数学	理科	社会	英語

強化学習

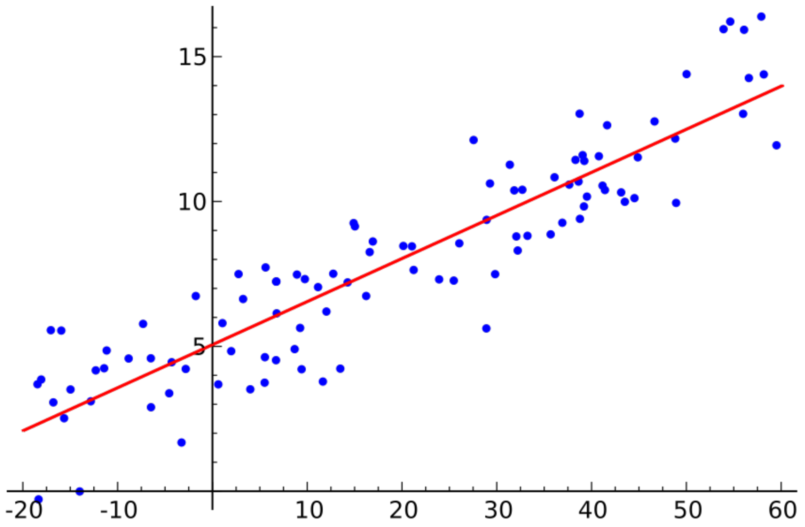
- ごめんよく知りません。

モデルでの分類

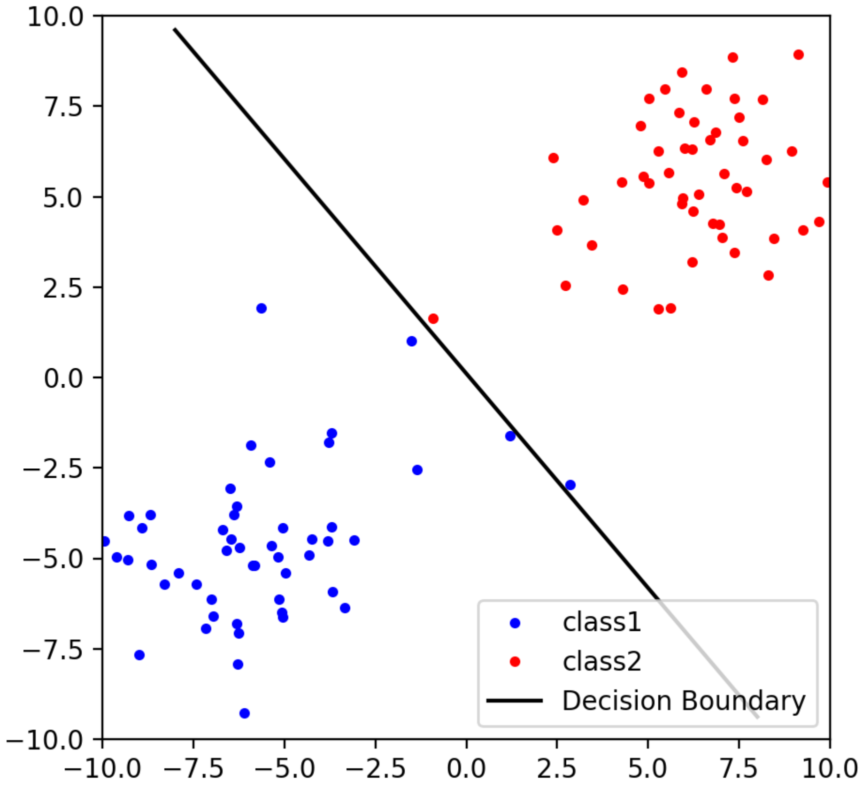
- 線形モデル
- 加法モデル　　（ +正則化 ）
- カーネルモデル
  (線形モデルと組み合わせるカーネルトリック. SVMは有名)
- 深層モデル

他に木やブースティングなどのモデルもある。

線形モデル



線形回帰



線形分類

深層モデル

https://qiita.com/hiyoko9t/items/089c3c828b2d3b04f756

「人工知能でバレンタインチョコが本命か義理かを判別する」

- 本命の確率が90%以上と予測されているもの

- 義理の確率が90%以上と予測されているもの

CNN（Convolution Neural Network）: 画像認識分野で非常に強力

学習法とモデルの関係

今流行ってるやつ

	線形	**加法**	**カーネル**	深層
教師あり	一般化線形モデル SVM	平滑化スプライン	カーネル密度	CNN, RNN
教師なし	PCA K-means			AE, VAE
強化学習				DQN

易 ⟷ 難

タスク別の分類

学習法による分類をさらに細分化する.

教師あり学習 : 回帰, 分類

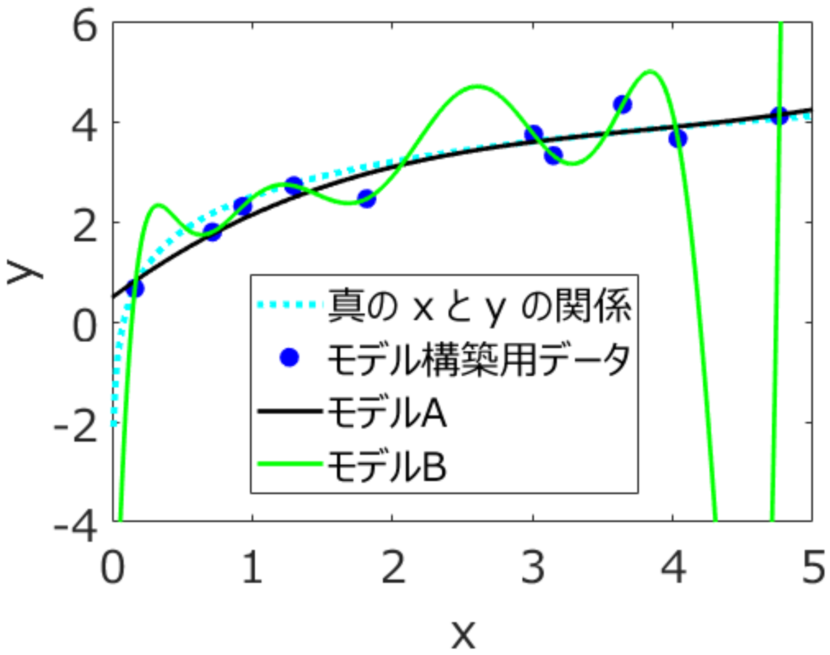
教師なし学習 : クラスタリング, 次元圧縮, 密度推定

回帰（Regression）

データから

$$\boldsymbol{x} \in \mathbb{R}^M,\ y \in \mathbb{R},\ y = f(\boldsymbol{x})$$

を満たす関数fを発見する. yは連続値をとる.



適切なモデルを作らないと過学習

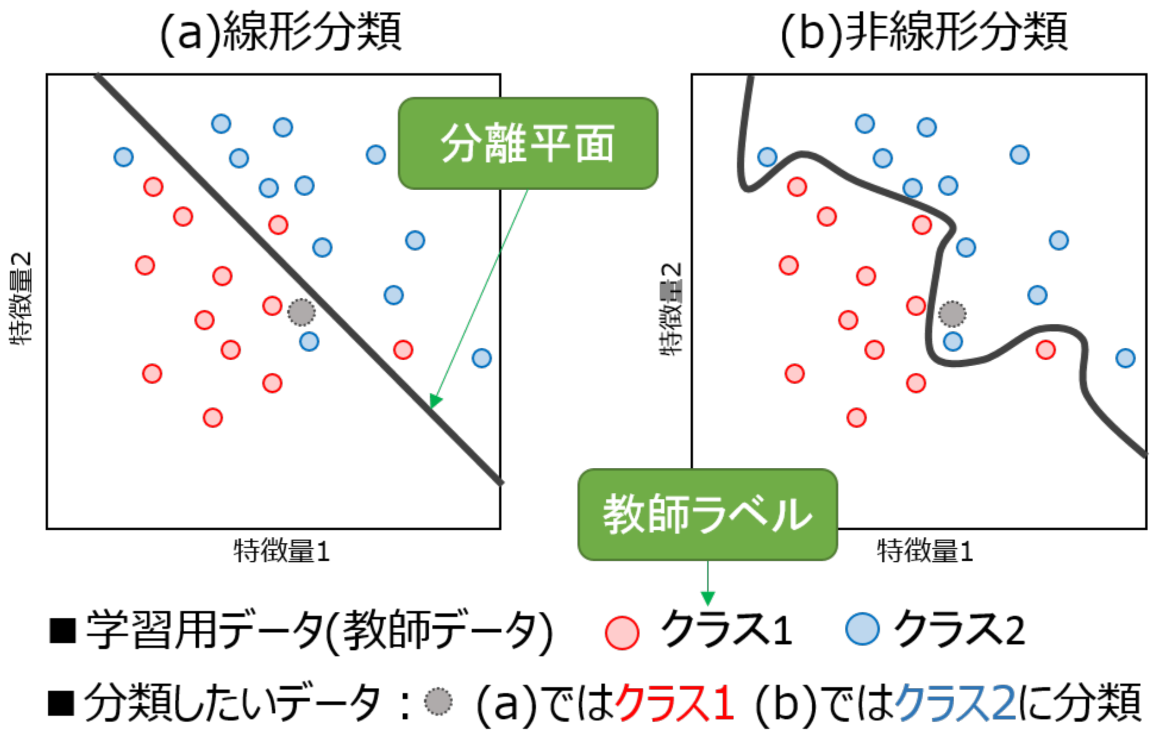
分類 (Classification)

データから

$$\boldsymbol{x} \in \mathbb{R}^M,\ y \in \{1,\ 2,\ \ldots,\ n\},\ y = f(\boldsymbol{x})$$

を満たす関数fを発見する. yは離散値をとる.

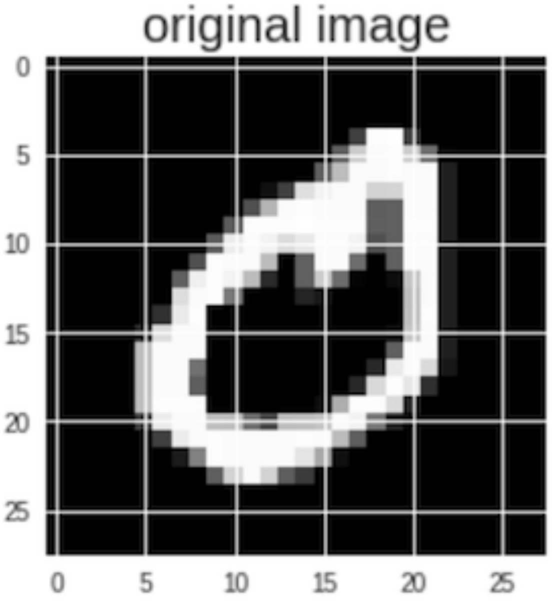
例：



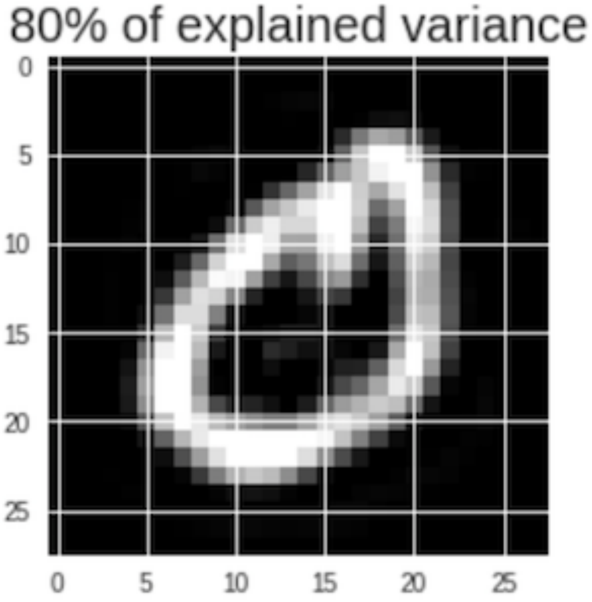
次元削減 (Dimensionality reduction)

データの次元数を減らし, データの圧縮や潜在構造の分析を行う.

例 :　PCA (主成分分析)



784個の変数



150 個の変数

統計的機械学習

「 機械学習のうち、データの確率的な生成規則を学習するもの 」

統計と機械学習の違いって？(諸説あり)

1. 統計学はデータを説明することを重視

2. 機械学習はデータから予測を行うことを重視

確率を用いた機械学習
(統計を知ってる人向け)

確率を用いる機械学習

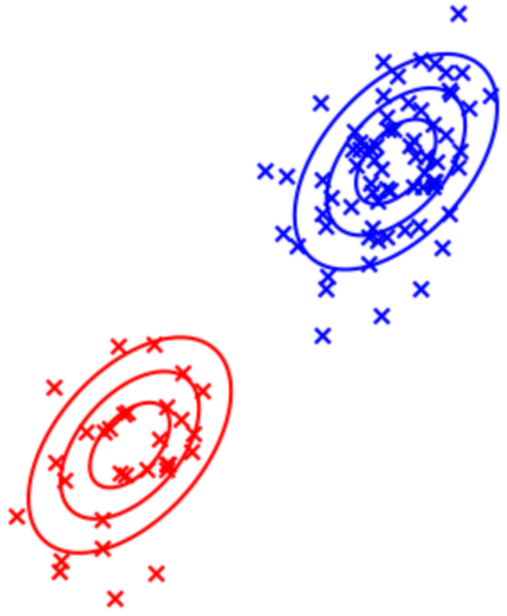
x : データ
y : (ex.) 回帰曲線の値, 属する分類, 属するクラスタ

$$\boldsymbol{x} \in \mathbb{R}^M, \; y = f(\boldsymbol{x})$$

- 関数モデル (確率を考えない)
- 機械学習に対する確率的アプローチ

1. 識別モデル

   事後確率　$P(y \mid \boldsymbol{x})$ を直接モデル化



2. 生成モデル

   $P(\boldsymbol{x} \mid y)$ と $P(y)$　をモデル化.

$$P(y \mid \boldsymbol{x}) = \frac{P(\boldsymbol{x} \mid y)P(y)}{P(\boldsymbol{x})}$$

ベイズの公式を使う.

単回帰（関数モデル）

問題　データ $D = \{(x_1, y_1),\ (x_2,\ y_2),\ \ldots,\ (x_n,\ y_n)\}$　から

$$y = ax + b$$

$a,\ b \in \mathbb{R}$ はパラメータ. 良い $a,\ b$ を見つける.

どういう関数を定義したら良さそうか？

$$Error(\boldsymbol{x}) = \sum_{i=1}^{N} \{y_i - (ax_i + b)\}^2$$

$$Error(\boldsymbol{x}) = \sum_{i=1}^{N} \{y_i - (ax_i + b)\}^4$$　とか採用しても一見良さそう

単回帰の確率的考察

（準備1）　ベクトル表現に直す.

$$\boldsymbol{x_i} = \begin{bmatrix} 1 \\ x_i \end{bmatrix}, \ \boldsymbol{w} = \begin{bmatrix} b \\ a \end{bmatrix}$$

$$minimize \sum_{i=1}^{N} \{y_i - (ax_i + b)\}^2 \Leftrightarrow minimize \sum_{i=1}^{N} y_i - \boldsymbol{w}^T \boldsymbol{x_i}$$

単回帰の確率的考察

（準備2）　最小二乗法を最尤推定として解釈する.

$$minimize \sum_{i=1}^{N}(y_i - \boldsymbol{w}^T\boldsymbol{x_i})^2 \Leftrightarrow maximize \sum_{i=1}^{N}\log\frac{1}{\sqrt{2\pi\sigma^2}}\exp\frac{-(y_i-\boldsymbol{w}^T\boldsymbol{x_i})^2}{2\sigma^2}$$

σは何かしらの値。

yが正規分布に従っているとした時の

平均の最尤推定量が　$\boldsymbol{w^T x}$

単回帰の識別モデルアプローチ

（識別モデルとしての定式化）

仮定:

$$P(y_n \mid \boldsymbol{x_n},\ \boldsymbol{w}) = N(y_n \mid \boldsymbol{w}^{\boldsymbol{T}}\boldsymbol{x_n},\ (\sigma^2)^{-1})$$

$$P(\boldsymbol{w}) = N(\boldsymbol{w} \mid \boldsymbol{\mu},\ \boldsymbol{\Sigma}^{-1})$$

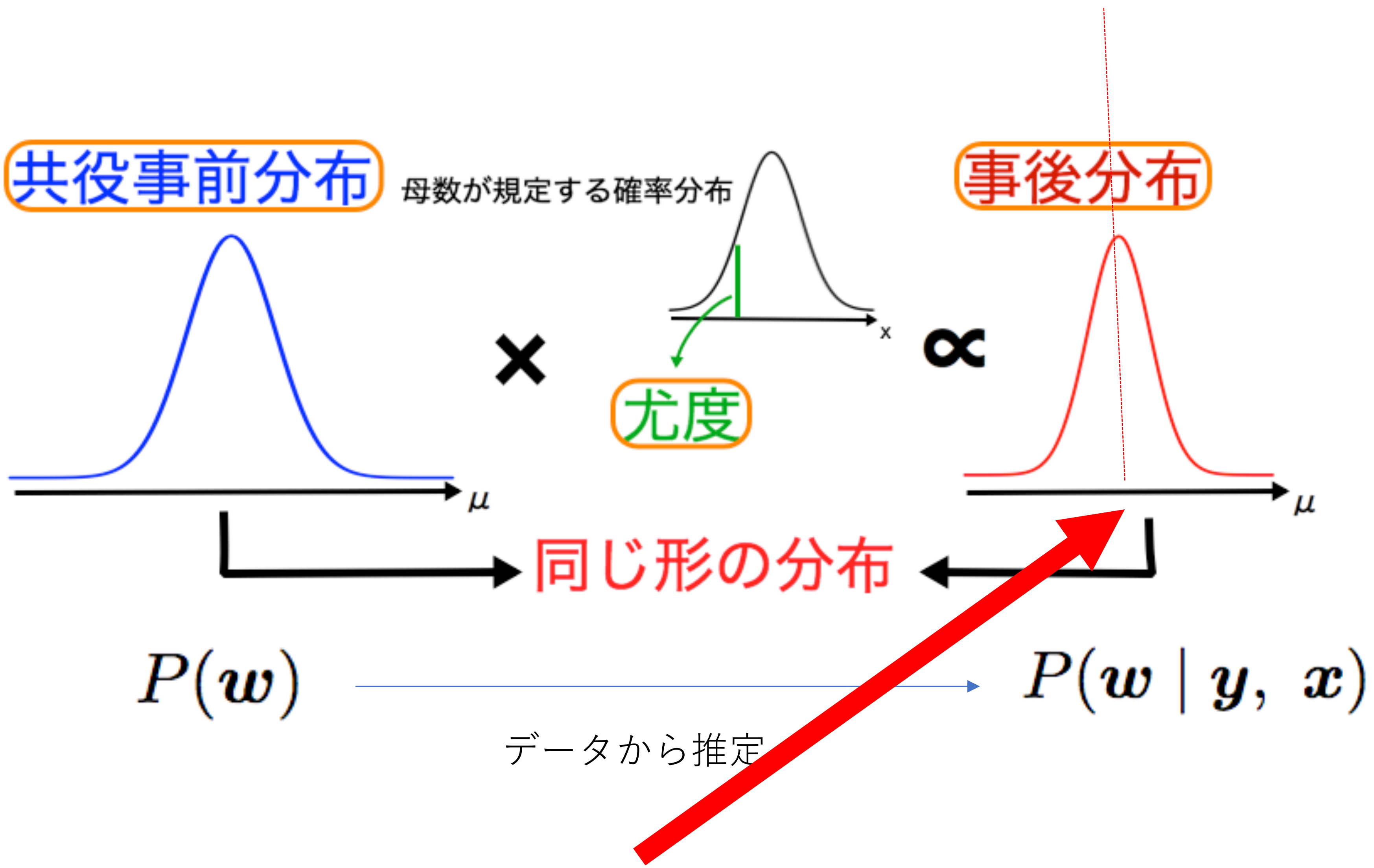
$\boldsymbol{\mu},\ \boldsymbol{\Sigma}$ も適当に仮定する.

パラメータ　$\boldsymbol{w} = \begin{bmatrix} b \\ a \end{bmatrix}$　の分布を考える発想はベイズ統計学

パラメータ $\boldsymbol{w}$ をデータから推定.

$$\begin{aligned} P(\boldsymbol{w} \mid \boldsymbol{y},\ \boldsymbol{x}) &= \frac{P(\boldsymbol{w})P(\boldsymbol{y} \mid \boldsymbol{x},\ \boldsymbol{w})}{P(\boldsymbol{y} \mid \boldsymbol{x})} \\ &\propto P(\boldsymbol{w})P(\boldsymbol{y} \mid \boldsymbol{x},\ \boldsymbol{w}) \\ &= N(\boldsymbol{w} \mid \boldsymbol{\mu},\ \boldsymbol{\Sigma}^{-1}) \prod_{i=1}^{N} N(y_i \mid \boldsymbol{w}^{\boldsymbol{T}} \boldsymbol{x_i},\ (\sigma^2)^{-1}) \\ &= \cdots = N(\boldsymbol{w} \mid \hat{\boldsymbol{\mu}},\ \hat{\boldsymbol{\Sigma}}^{-1}) \end{aligned}$$

ただし, $\hat{\boldsymbol{\Sigma}} = (\sigma)^2 \sum_{i=1}^{N} \boldsymbol{x}_n \boldsymbol{x}_n^T + \boldsymbol{\Sigma},\ \hat{\boldsymbol{\mu}} = \hat{\boldsymbol{\Sigma}}^{-1} \{(\sigma)^2 \sum_{i=1}^{N} y_n \boldsymbol{x}_n + \boldsymbol{\Sigma}\boldsymbol{\mu}\}$



$\boldsymbol{w}^* = \arg\max_{\boldsymbol{w}} P(\boldsymbol{w} \mid \boldsymbol{y},\ \boldsymbol{x})$ が最適なパラメータ.(MAP 推定)

$\boldsymbol{w}_*$ が求まったので, 未知の入力 $\boldsymbol{x}_*$ に対する出力 $\boldsymbol{y}_*$ がどうなるか予測する.

$$
\begin{aligned}
P(\boldsymbol{w}_* \mid y_*,\ \boldsymbol{x}_*) &= \frac{P(\boldsymbol{w}_*)P(y_* \mid \boldsymbol{x}_*,\ \boldsymbol{w}_*)}{P(y_* \mid \boldsymbol{x}_*)} \\
P(y_* \mid \boldsymbol{x}_*) &= \frac{P(\boldsymbol{w}_*)P(y_* \mid \boldsymbol{x}_*,\ \boldsymbol{w}_*)}{P(\boldsymbol{w}_* \mid y_*,\ \boldsymbol{x}_*)} \\
&\propto \frac{P(y_* \mid \boldsymbol{x}_*,\ \boldsymbol{w}_*)}{P(\boldsymbol{w}_* \mid y_*,\ \boldsymbol{x}_*)} \\
&= \frac{N(y_* \mid \boldsymbol{x}_*,\ \boldsymbol{w}_*)}{N(\boldsymbol{w}_* \mid \hat{\boldsymbol{\mu}}|_{y_n=y_*,\ x_n=x_*},\ \hat{\boldsymbol{\Sigma}}^{-1}|_{\boldsymbol{x}_n=\boldsymbol{x}_*})} \\
&= \cdots = N(y_* \mid \mu_*,\ (\sigma_*^2)^{-1})
\end{aligned}
$$

ただし, $\mu_* = \boldsymbol{\mu}^T \boldsymbol{x}_*$, $(\sigma_*^2)^{-1} = (\sigma^2)^{-1} + \boldsymbol{x}_*^T \Sigma^{-1} \boldsymbol{x}_*$

線形回帰モデルの比較

線形回帰も奥が深い!!

- 関数モデル

$$\boldsymbol{w}_* = \begin{bmatrix} \bar{y} - \dfrac{\sigma_{xy}}{\sigma_x^2}\bar{x} \\ \dfrac{\sigma_{xy}}{\sigma_x^2} \end{bmatrix}$$

$$f(\boldsymbol{x}) = \boldsymbol{w}_*^{\boldsymbol{T}} \boldsymbol{x}$$

- 識別モデル

$$\boldsymbol{w}_* = \arg\max_{\boldsymbol{w}} N(\boldsymbol{w} \mid \hat{\boldsymbol{\mu}},\ \hat{\boldsymbol{\Sigma}}^{-1})$$

$$\hat{\boldsymbol{\Sigma}} = (\sigma)^2 \sum_{i=1}^{N} \boldsymbol{x}_n \boldsymbol{x}_n^T + \boldsymbol{\Sigma},\ \hat{\boldsymbol{\mu}} = \hat{\boldsymbol{\Sigma}}^{-1}\{(\sigma)^2 \sum_{i=1}^{N} y_n \boldsymbol{x}_n + \boldsymbol{\Sigma}\boldsymbol{\mu}\}$$

$$f(\boldsymbol{x}) = N(y_* \mid \mu_*,\ (\sigma_*^2)^{-1})$$

$$\mu_* = \boldsymbol{\mu}^T \boldsymbol{x}_*,\ (\sigma_*^2)^{-1} = (\sigma^2)^{-1} + \boldsymbol{x}_*^T \Sigma^{-1} \boldsymbol{x}_*$$

- 一般に確率的なアプローチを行うと数学が難しい.
- 特に生成モデルは難しい.

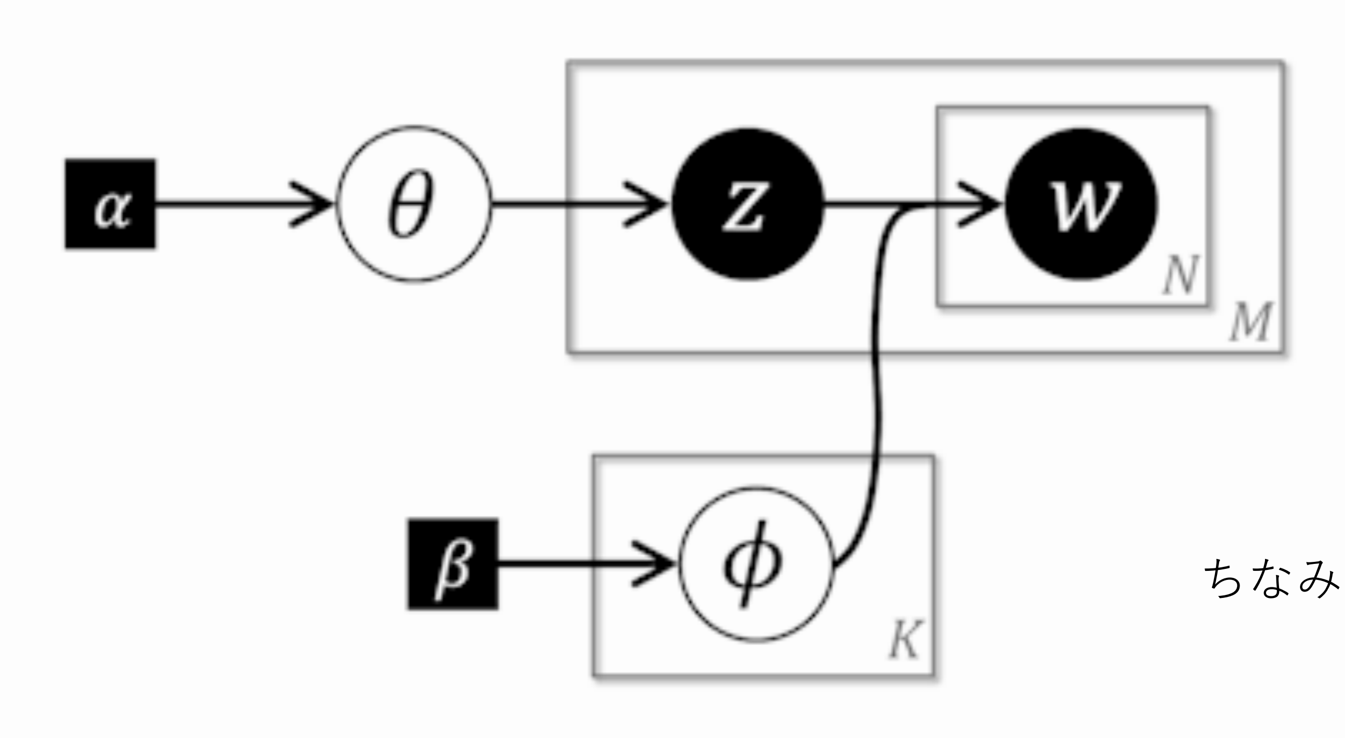
- 一般に事前分布から事後分布を求める方法：

1. 共役事前分布を使う.
2. MCMC（マルコフ連鎖モンテカルロ法）を使う.
3. 変分近似を行う.

2, 3の方法は近似的に確率分布を求める手法

グラフィカルモデル（おまけ）

- グラフによって，確率変数間の条件付き依存構造を示す.



ちなみにDAGになっている.

最後に

線形代数, 解析学, 基礎統計学 は最低勉強しましょう.

ちゃんと理論をやりたい人

集合位相, 複素解析, 測度論, ルベーグ積分, 数値解析
関数解析, 位相幾何学, 微分幾何学.