No.
Date.

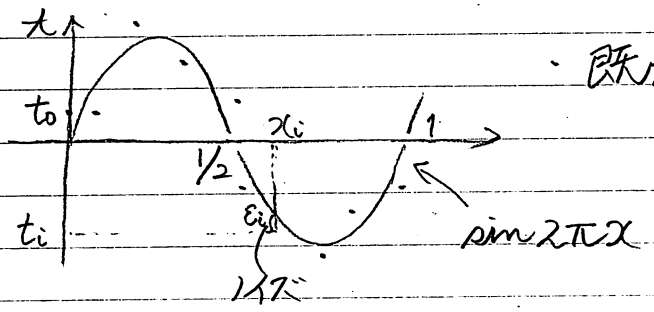
応用数理IIでやりたいこと
- 学習理論の紹介
  (線形回帰，分類問題
  サポートベクトルマシン (SVM)
- バイオインフォマンスへの学習理論の応用

参考書
- Bishop, Pattern Recognition and Machine Learning, Springer
- 阿久津達也，バイオインフォマンスの数理とアルゴリズム，共立出版

総科 C819, 6482

例題：多項式による曲線近似
N コの既知のデータ $(x_1, t_1), \cdots, (x_N, t_N)$
$x_i \in [0,1]$, $t_i \in \mathbb{R}$
の特徴を学習して未知のパラメータ $\hat{x}$ に対する出力 $\hat{t}$ を予想せよ．
(t_i は $\sin 2\pi x_i + \underbrace{\varepsilon_i}_{\text{ノイズ}}$ によって生成) … ☆

注意) 現実の問題では関係☆などは不明．
(理解を助ける為に明示している)



No.

Date.

多項式：$y(x, \underline{w}) = w_0 + w_1 x + w_2 x^2 + \cdots + w_M x^M$

で (x_i, t_i) $i=1, \cdots, N$ を近似してみよう．つまり

$w_0, w_1, \cdots, w_M$ を学習（推定）してみよう．

解法1　$y(x_i, \underline{w}) = w_0 + w_1 x_i + \cdots + w_M x_i^M = t_i \quad i=1, \cdots, N$

$$\begin{pmatrix} 1 & x_1 & x_1^2 & \cdots & x_1^M \\ 1 & x_2 & x_2^2 & \cdots & x_2^M \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_N & x_N^2 & \cdots & x_N^M \end{pmatrix} \begin{pmatrix} w_1 \\ w_2 \\ \vdots \\ w_M \end{pmatrix} = \begin{pmatrix} t_1 \\ t_2 \\ \vdots \\ t_N \end{pmatrix}$$

既知　　未知　　既知

$N=M$ の時以外はこの方法は使えない

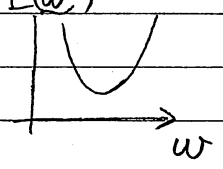
実用的ではない！

解法2　エラー関数

$E(\underline{w}) := \frac{1}{2} \sum_{n=1}^{N} (y(x_n, \underline{w}) - t_n)^2$

測定値と多項式のズレ

を最小化する．



- $E(\underline{w})$ は $\underline{w}$ に関して2次関数（下に凸）
- $\frac{\partial E}{\partial w_i}$ は $\underline{w}$ に関して1次関数

$$\Rightarrow \begin{pmatrix} \frac{\partial E}{\partial w_0}(\underline{w}) \\ \vdots \\ \frac{\partial E}{\partial w_M}(\underline{w}) \end{pmatrix} = \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix} \Big\} M+1\text{コ} \quad \cdots (\$)$$

が定める M コの連立1次方程式をとくことで

$\underline{w} = (w_1, \cdots, w_M)$ が決定できる．

レポート1．($) を具体的に表しましょう（$A\underline{w} = \underline{b}$）

問題点（解法2の）：OVER FITTING

No.
Date.

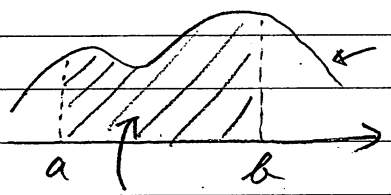
学習理論ではどのようにして Over fitting を回避しているか？
⇒ 確率の導入（ベイズ推定）

確率の復習

- 離散確率変数 x
  - $\mathcal{X} = \{x_1, \cdots, x_n\}$ x がとりうる事象
  - $P(x = x_i)$: $x = x_i$ となる確率
  - $(\sum_{i=1}^{n} P(x_i) = 1)$

例. コイン投げ
$\mathcal{X} = \{$表, 裏$\}$
$P(表) = P(裏) = \frac{1}{2}$

- 連続確率変数 x
  確率密度関数 $0 \leq p(x)$ s.t. $\int_{-\infty}^{\infty} p(x)\,dx$
  $P(a \leq x \leq b) = \int_a^b f(x)\,dx$



面積が確率に対応

例. 正規分布（ガウス分布）

$$N(x \mid \underset{\text{パラメータ}}{\mu, \sigma^2}) = \frac{1}{(2\pi\sigma^2)^{\frac{1}{2}}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right)$$

- x, y を確率変数としたとき
  $p(x, y)$: 結合確率
  ↖ x かつ y をとる確率
  $p(x \mid y)$: 条件付き確率
  （y が既知の下での x の確率）

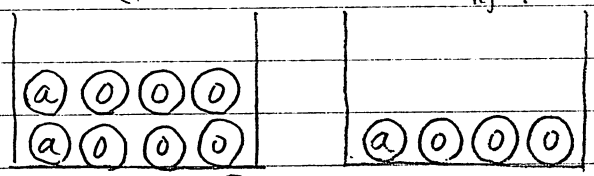
No.

Date.

重要な性質　$p(x) = \sum_y p(x,y)$　① 周辺化

・$p(x,y) = p(x|y)p(y)$　②

(例)



ⓐ：リンゴ，ⓞ：オレンジ

ルール

赤い箱は40%　}の確率で選ばれる
青い箱は60%　}

各箱のフルーツは等確率で選ばれる．

	ⓐ	ⓞ
赤	$\frac{4}{10} \times \frac{2}{8} = \frac{1}{10}$	$\frac{4}{10} \times \frac{6}{8} = \frac{3}{10}$
青	$\frac{6}{10} \times \frac{3}{6} = \frac{9}{20}$	$\frac{6}{10} \times \frac{1}{6} = \frac{3}{20}$

確率の表

$\sum_{z \in Z} P(z) = 1$　（Z：全ての事象）

$\Rightarrow \frac{1}{10} + \frac{3}{10} + \frac{9}{20} + \frac{3}{20} = 1$

・$P(赤) = \frac{4}{10}$，$P(赤,ⓐ) + P(赤,ⓞ) = \frac{1}{10} + \frac{3}{10} = \frac{4}{10}$

$\Rightarrow$ ①が成り立っている

・$P(赤,ⓐ) = \frac{1}{10}$　$P(ⓐ|赤)P(赤) = \frac{2}{8} \times \frac{4}{10} = \frac{1}{10}$

$\Rightarrow$ ②が成り立っている

ベイズの法則

$p(x,y) = p(x|y)p(y) = p(y|x)p(x)$ より

$p(y|x) = \frac{p(x|y)p(y)}{p(x)}$ ：ベイズの法則

x：観測可能なデータ　}とみなす
y：調べたい確率変数　}

No.
Date. ・ ・

$p(y)$：事前確率、$p(y|x)$：事後確率
$p(x|y)$：尤度関数（yについての関数）
$p(x)$：正規化因子
（規格化）
$p(y|x)$はyについての確率なので
$\Sigma_y p(y|x) = 1$ となる必要がある
$\Sigma_y c\,p(x|y)\,p(y) = 1 \Rightarrow c$が一意に定まる
$\frac{1}{c} = p(x)$

（例の続き）ベイズの法則を体感しましょう
$p(赤) = \frac{6}{10}$, $p(青) = \frac{4}{10}$ ← 事前確率
$p(a|赤) = \frac{1}{4}$　$p(o|赤) = \frac{3}{4}$
$p(o) = p(赤, o) + p(青, o) = \frac{3}{10} + \frac{3}{20} = \frac{9}{20}$
問　とり出されたフルーツがオレンジだとしたら　→観測データ
どちらの箱から選ばれたんだろう？

$$p(赤|o) = \frac{p(o|赤)\,p(赤)}{p(o)} = \frac{3}{4} \times \frac{4}{10} \times \frac{20}{9} = \frac{2}{3}$$

$$p(青|o) = 1 - p(赤|o) = \frac{1}{3}$$

期待値と分散

関数$f(x)$の期待値

$$E(f) = \begin{cases} \Sigma_x p(x) f(x) & 離散 \\ \int p(x) f(x)\,dx & 連続 \end{cases}$$

多変数の場合（$p(x, y)$）

$E_x(f) = \Sigma_x p(x) f(x, y)$　xに関する期待値
$E_x(f|y) = \Sigma_x p(x|y) f(x, y)$　xに関する
条件付期待値

関数$f(x)$の分散

$$var(f) = E((f(x) - E(f))^2)$$
$$= E(f(x)^2 - 2E(f)f(x) + E(f)^2)$$

No.

Date.

Eの線形性

$$= E(f(x)^2) - 2E(f)E(f(x)) + E(f)^2$$
$$= E(f^2) - 2E(f)^2 + E(f)^2$$
$$= E(f^2) - E(f)^2$$

特に xの分散 : $var(x) = E((x - E(x))^2)$
$$= E(x^2) - E(x)^2$$

確率変数 x, y の共分散

$$var(x,y) = E_{xy}((x - E_x(x))(y - E_y(y))$$
$$= E_{xy}(xy) - E_x(x)E_y(y)$$

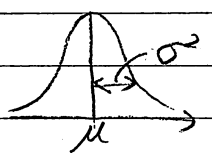
確率変数 $\underline{x}, \underline{y}$ の共分散

$$var(\underline{x}, \underline{y}) = E_{xy}((\underline{x} - E_x(\underline{x}))(\underline{y} - E_y(\underline{y}))')$$

(m, n) 行列

ガウス分布

$$N(x|\mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right)$$



$E(x) = \int_{-\infty}^{\infty} N(x|\mu, \sigma^2)\, dx = \mu$
$var(x) = \sigma^2$
レポート(示せ)

この意味で μ を平均, σ^2 を分散と呼ぶ.

ガウス分布に対する最尤推定法

ガウス分布 :

$$N(x|\mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right)$$

μ, σ^2 は未知である $N(x|\mu, \sigma^2)$ から N個の独立なデータ

$$\underline{x} = (x_1, \cdots, x_N)$$

が得られたとする. これから μ, σ^2 を推定しよう!

No.
Date. ・ ・

$P(\boldsymbol{x}|\mu,\sigma^2) = N(x_1|\mu,\sigma^2)\cdots N(x_N|\mu,\sigma^2)$ (独立だから)

最尤推定法の方針
: 尤度関数 $P(\boldsymbol{x}|\mu,\sigma^2)$ の最大点, $(\mu_{ML}, \sigma^2_{ML})$ を探す.
Maximum Likelihood.

$$\ln P(\boldsymbol{x}|\mu,\sigma^2) = \sum_{n=1}^{N} \ln N(x_n|\mu,\sigma^2)$$
$$= \sum_{n=1}^{N}\left[-\ln(2\pi\sigma^2)^{1/2} - \frac{1}{2\sigma^2}(x_n-\mu)^2\right]$$
$$= -\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n-\mu)^2 - \frac{N}{2}\ln\sigma^2 - \frac{N}{2}\ln(2\pi) \cdots ①$$

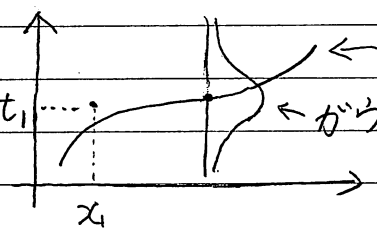
①の最大点, μ_{ML}, σ^2_{ML} を求めればよい.

$$\frac{d}{d\mu}\ln P(\boldsymbol{x}|\mu,\sigma^2) = \frac{1}{\sigma^2}\sum_{n=1}^{N}(x_n-\mu) = 0$$
$$\Leftrightarrow \sum_{n=1}^{N} x_n - N\mu = 0$$
$$\mu_{ML} = \frac{1}{N}\sum_{n=1}^{N} x_n \quad (\text{データの平均})$$

同様に
$$\frac{d}{d\sigma^2}\ln P(\boldsymbol{x}|\mu,\sigma^2) = \frac{1}{2\sigma^4}\sum_{n=1}^{N}(x_n-\mu_{ML})^2 - \frac{N}{2\sigma^2} = 0$$
$$\Leftrightarrow \sum_{n=1}^{N}(x_n-\mu_{ML})^2 = N\sigma^2$$
$$\sigma^2_{ML} = \frac{1}{N}\sum_{n=1}^{N}(x_n-\mu_{ML})^2 \quad (\text{データの分散})$$

これをもとに多項式による曲線近似(第1回)の例を再考

解法2の問題点, ⇒ over fitting ($|w| \gg 1$)
「解法2 = 最尤推定法, over fitting ⇐ ベイズ推定」



$t_i = y(x_i,w) + \varepsilon_i$ ε_iをガウス分布でモデリングする.
$P(t|x,w,\beta) = N(t|y(x|w),\beta^{-1})$
によって, データが生成されているとする.

No.

Date.

各データは独立なので

$p(\underline{t}|\underline{x}, \underline{w}, \beta) = N(t_1|y(x_1,\underline{w}), \beta_1^{-1}) \cdots N(t_N|y(x_N,\underline{w}), \beta_N^{-1})$

データに対する尤度関数

最尤推定を用いて $\underline{w}$ を推定してみる。①と同様

$\ln p(\underline{t}|\underline{x}, \underline{w}, \beta) = \sum_{n=1}^{N} N(t_n|y(x_n,\underline{w}), \beta^{-1})$

$= -\frac{\beta}{2}\sum_{n=1}^{N}(y(x_n,\underline{w}) - t_n)^2 + \frac{N}{2}\ln\beta - \frac{N}{2}\ln(2\pi)$

$\underline{w}_{ML}$ は $E(\underline{w}) = \frac{1}{2}\sum_{n=1}^{N}(y(x_n,\underline{w}) - t_n)^2$

を最小化することで得られる。

ベイズ推定を導入

方針：事後確率を最大化する。

事前確率 $p(\underline{w})$ を例えば次で導入する。

$p(\underline{w}) = N(w_0|0, \alpha^{-1}) \cdots N(w_M|0, \alpha^{-1})$

ベイズの法則より

$$p(\underline{w}|\underline{x}, \underline{t}, \beta) = \frac{p(\underline{t}|\underline{x}, \underline{w}, \beta)\, p(\underline{w})}{p(\underline{t}|\underline{x}, \beta)} \leftarrow \text{規格化因子}$$

$$\propto p(\underline{t}|\underline{x}, \underline{w}, \beta)\, p(\underline{w})$$

この事後確率を最大にする $\underline{w}$ を求める！

$p(\underline{w}|\underline{x}, \underline{t}, \beta) = C\, p(\underline{t}|\underline{x}, \underline{w}, \beta)\, p(\underline{w})$

$= C \prod_{n=1}^{N} N(t_n|y(x_n,\underline{w}), \beta^{-1}) \prod_{i=0}^{M} N(w_i|0, \alpha^{-1})$

$= C \prod_{n=1}^{N} \frac{\beta^{1/2}}{(2\pi)^{1/2}} \exp\left\{-\frac{\beta}{2}(y(x_n,\underline{w}) - t_n)^2\right\}$

$\times \prod_{i=0}^{M} \frac{\alpha^{1/2}}{(2\pi)^{1/2}} \exp\left(-\frac{\alpha}{2}w_i^2\right)$

$= C' C \exp\left\{-\left(\frac{\beta}{2}\sum_{n=1}^{N}(y(x_n,\underline{w}) - t_n)^2 + \frac{\alpha}{2}\sum_{i=0}^{M} w_i^2\right)\right\}$

（C'：$\underline{w}$ に依存しない定数）

$\Rightarrow \frac{\beta}{2}\sum_{n=1}^{N}(y(x_n,\underline{w}) - t_n)^2 + \frac{\alpha}{2}\sum_{i=0}^{M} w_i^2$ を最小にする $\underline{w}$ を探せばよい。

$\Leftrightarrow \tilde{E}(\underline{w}) = \underbrace{\frac{1}{2}\sum_{n=1}^{N}(y(x_n,\underline{w}) - t_n)^2}_{E(\underline{w})} + \underbrace{\frac{\alpha}{2\beta}\sum_{i=0}^{M} w_i^2}_{\text{regularization 項}}$ を

$\underline{w}$ が大きくなるのを防ぐ項

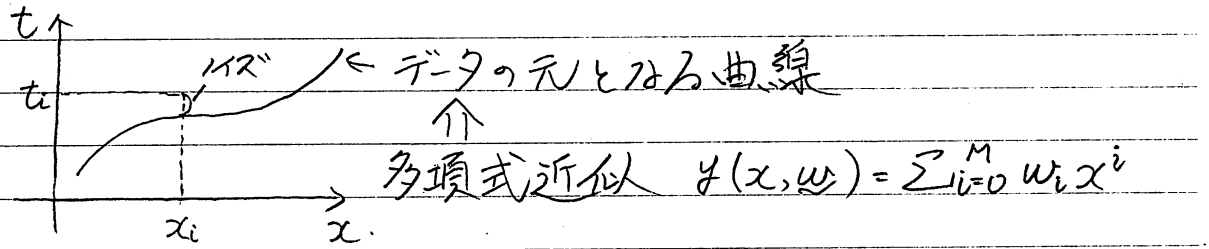
$\Rightarrow$ over fitting を防げる。

No.
Date.

まとめると

最尤推定	ベイズ推定
over fitting が生じる	over fitting は抑制できる
$E(\boldsymbol{w})$ の最小点,	$\tilde{E}(\boldsymbol{w})$ の最小点,

今週と来週の内容：線形回帰
(linear Regression)
多項式による曲線（ノイズ）近似の一般化
前回まで



多項式近似 $y(x,\boldsymbol{w}) = \sum_{i=0}^{M} w_i x^i$

keyは $\boldsymbol{w}$ に関して線形計算のみ

$$y(\boldsymbol{x},\boldsymbol{w}) = w_0 + \sum_{j=1}^{M-1} w_j \phi_j(\boldsymbol{x}) \in \mathbb{R} \quad \cdots ①$$
$$= \boldsymbol{w}'\boldsymbol{\phi}(\boldsymbol{x})$$

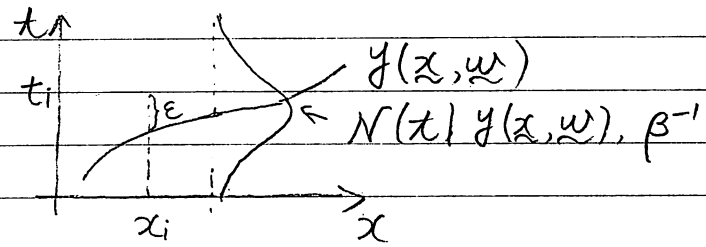
ここで $\boldsymbol{\phi}(\boldsymbol{x}) = (\phi_0(\boldsymbol{x}), \cdots, \phi_{M-1}(\boldsymbol{x}))$
$\phi_0(\boldsymbol{x}) = 1$, $\boldsymbol{x} \in \mathbb{R}^D$, $\boldsymbol{w} \in \mathbb{R}^M$
と一般化して考える.
つまり近似を行う基底を
$\{x^j\} \to \{\phi_j(\boldsymbol{x})\}$
と一般化
線形回帰問題：入力 $\boldsymbol{x}$ に対して出力 t がノイズを
伴って観測されるとき
$t = y(\boldsymbol{x},\boldsymbol{w}) + \varepsilon$
の形で対応関係をモデル化すること.
$\phi_0(\boldsymbol{x}), \cdots, \phi_{M-1}(\boldsymbol{x})$：基底関数

No.

Date.

注意① 多項式による曲線近似は、
$D=1$, $\phi_3(x)=x^3$ に対応
② 基底関数として多項式以外でも
ガウス関数，三角関数 等々.

前回同様，tをガウス分布でモデリング



$p(t|x,w,\beta)=N(t|y(x,w),\beta^{-1})$
既知データ：$(x_1,t_1),\cdots(x_N,t_N)$
を用いて，w,βを学習したい.
各データは独立 ⇒ データに対する尤度関数は.
$p(\underline{t}|X,w,\beta)=N(t_1|y(x_1,w),\beta^{-1})\times\cdots\times N(t_N|y(x_N,w),\beta^{-1})$
$=N(t_1|w'\phi(x_1),\beta^{-1})\times\cdots\times N(t_N|w'\phi(x_N),\beta^{-1})$
ここで $\underline{t}=(t_1,\cdots,t_N)$, $X=(x_1,\cdots,x_N)$

解法1：最尤推定法
$\Leftrightarrow$ $p(\underline{t}|X,w,\beta)$の最大点，$w_{ML}$, β_{ML}を求める
$\Leftrightarrow$ $\log p(\underline{t}|X,w,\beta)$の 〃
以後Xは省略
$\log p(\underline{t}|w,\beta)=\sum_{n=1}^{N}\log N(t_n|w'\phi(x_n),\beta^{-1})$
$=\sum_{n=1}^{N}\log\frac{\beta^{1/2}}{(2\pi)^{1/2}}\exp\{-\frac{\beta}{2}(t_n-w'\phi(x_n))^2\}$
$=\frac{N}{2}\log\beta-\frac{N}{2}\log(2\pi)-\beta E(w)$
ここで $E(w)=\frac{1}{2}\sum_{n=1}^{N}(t_n-w'\phi(x_n))^2$
○ w_{ML}を求める.
$\frac{\partial}{\partial w}\log p(\underline{t}|w,\beta)=0 \Leftrightarrow \frac{\partial}{\partial w_i}E(w)=0$, $i=0,\cdots,M-1$

より) $\sum_{n=1}^{N}(t_n - \underline{w}'\underline{\phi}(\underline{x}_n))\phi_i(\underline{x}_n) = 0 \quad i=0,\cdots,M-1$

$\Rightarrow \sum_{n=1}^{N} t_n\phi_i(\underline{x}_n) - \underline{w}'\sum_{n=1}^{N}\underline{\phi}(\underline{x}_n)\phi_i(\underline{x}_n) = 0$

横ベクトル表示 $i=0,\cdots,M-1$

$\Rightarrow \sum_{n=1}^{N} t_n\underline{\phi}(\underline{x}_n)' - \underline{w}'\sum_{n=1}^{N}\underline{\phi}(\underline{x}_n)\underline{\phi}(\underline{x}_n)' = \underline{0}$

$\Rightarrow \left(\sum_{n=1}^{N}\underline{\phi}(\underline{x}_n)\underline{\phi}(\underline{x}_n)'\right)\underline{w} = \sum_{n=1}^{N}\underline{\phi}(\underline{x}_n)t_n \quad (\$)$

$M\times M$ 行列

レポート

$$\Phi = \begin{pmatrix} \phi_0(\underline{x}_1) & \cdots & \phi_{M-1}(\underline{x}_1) \\ \vdots & \ddots & \vdots \\ \phi_0(\underline{x}_N) & \cdots & \phi_{M-1}(\underline{x}_N) \end{pmatrix}$$ としたとき

- $\underline{\Phi}'\underline{\Phi} = \sum_{n=1}^{N}\underline{\phi}(\underline{x}_n)\underline{\phi}(\underline{x}_n)'$
- $\underline{\Phi}'\underline{t} = \sum_{n=1}^{N}\underline{\phi}(\underline{x}_n)t_n$

} $(\star)$

を示しましょう

($\$$)と($\star$)より

$\underline{\Phi}'\underline{\Phi}\,\underline{w}_{ML} = \underline{\Phi}'\underline{t}$

$\Rightarrow \underline{w}_{ML} = (\underline{\Phi}'\underline{\Phi})^{-1}\underline{\Phi}'\underline{t}$

○ β_{ML} を求めよ.

$\frac{\partial}{\partial\beta}\log p(\underline{t}|\underline{w},\beta) = 0$

$\Rightarrow \frac{N}{2\beta_{ML}} - E(\underline{w}_{ML}) = 0$

$\frac{1}{\beta_{ML}} = \frac{2}{N}E(\underline{w}_{ML}) = \frac{1}{N}\sum_{n=1}^{N}(t_n - \underline{w}'\underline{\phi}(\underline{x}_n))^2$ //

$\underline{x}\in\mathbb{R}^D$ に対するガウス分布

$$N(\underline{x}|\underline{\mu},\underline{\Sigma}) = \frac{1}{(2\pi)^{D/2}}\frac{1}{|\underline{\Sigma}|^{1/2}}\exp\left\{-\frac{1}{2}(\underline{x}-\underline{\mu})'\underline{\Sigma}^{-1}(\underline{x}-\underline{\mu})\right\}$$

$\underline{\mu}\in\mathbb{R}^D$, $\underline{\Sigma}$: $D\times D$ 行列

平均　　共分散行列

($D=1$ では通常のガウス分布)

前回同様 X は省略

No.

Date. ・ ・

ベイズ推定法

事前確率を導入：　　c単位行列

☆ $p(\underset{\sim}{w}|\alpha) = N(\underset{\sim}{w}|\underset{\sim}{0}, \alpha^{-1}I)$ … 多次元ガウス分布

($\underset{\sim}{w}$がベクトルだから)

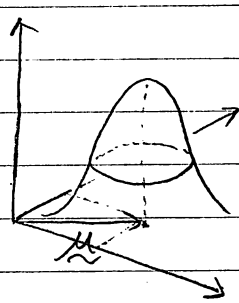
一般に $\underset{\sim}{x} \in \mathbb{R}^D$ に対する多次元ガウス分布は、

$$N(\underset{\sim}{x}|\underset{\sim}{\mu}, \Sigma) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(\underset{\sim}{x}-\underset{\sim}{\mu})^T \Sigma^{-1} (\underset{\sim}{x}-\underset{\sim}{\mu})\right\}$$

$\underset{\sim}{\mu} \in \mathbb{R}^D$. Σ : $D \times D$ 行列

平均　　共分散行列

で与えられる。当然 $D=1$ では通常のガウス分布



実は $\underset{\sim}{\mu} = \underset{\sim}{0}$, $\Sigma = \alpha^{-1}I$ より

$$p(\underset{\sim}{w}|\alpha) = \prod_{i=0}^{M-1} \frac{\alpha^{1/2}}{(2\pi)^{1/2}} \exp\left\{-\frac{\alpha}{2} w_i^2\right\}$$

これを用いてベイズの法則から事後確率を求める：

$$p(\underset{\sim}{w}|t, \alpha, \beta) = \underset{\text{規格化定数}}{c}\, p(t|\underset{\sim}{w}, \beta)\, p(\underset{\sim}{w}|\alpha)$$

$$= c \prod_{n=1}^{N} \frac{\beta^{1/2}}{(2\pi)^{1/2}} \exp\left\{-\frac{\beta}{2}(\underset{\sim}{w}^T \underset{\sim}{\phi}(x_n) - t_n)^2\right\} \times \prod_{i=0}^{M-1} \frac{\alpha^{1/2}}{(2\pi)^{1/2}} \exp\left\{-\frac{\alpha}{2} w_i^2\right\}$$

$$= c' \exp\left\{-\frac{\beta}{2}\sum_{n=1}^{N}(\underset{\sim}{w}^T \underset{\sim}{\phi}(x_n) - t_n)^2 - \frac{\alpha}{2}\underset{\sim}{w}^T \underset{\sim}{w}\right\}$$

$$= c' \exp \varphi(\underset{\sim}{w})$$

$$\varphi(\underset{\sim}{w}) = -\frac{\beta}{2}\sum_{n=1}^{N}(\underset{\sim}{w}^T \underset{\sim}{\phi}(x_n) - t_n)^2 - \frac{\alpha}{2}\underset{\sim}{w}^T \underset{\sim}{w}$$

$$= -\frac{1}{2}(\underset{\sim}{w} - \underset{\sim}{a})^T B^{-1} (\underset{\sim}{w} - \underset{\sim}{a})$$

とおいて $\underset{\sim}{a}$, B を求めてみよう

(展開すると)

$$= -\frac{1}{2}(\underset{\sim}{w}^T B^{-1} - \underset{\sim}{a}^T B^{-1})(\underset{\sim}{w} - \underset{\sim}{a})$$

$$= -\frac{1}{2}(\underset{\sim}{w}^T B^{-1} \underset{\sim}{w} - \underset{\sim}{a}^T B^{-1} \underset{\sim}{w} - \underset{\sim}{w}^T B^{-1} \underset{\sim}{a} + \underset{\sim}{a}^T B^{-1} \underset{\sim}{a})$$

⇒ ① $\underset{\sim}{w}$ の2次の係数から B^{-1} を求める。

② $\underset{\sim}{w}$ の1次の係数と B^{-1} より $\underset{\sim}{a}$ を求める

① $\varphi(\underset{\sim}{w})$ の $\underset{\sim}{w}$ に関して2次の項を比較すると

$$-\frac{1}{2}\left(\beta\sum_{n=1}^{N}(\underset{\sim}{w}^T \underset{\sim}{\phi}(x_n))^2 + \alpha \underset{\sim}{w}^T \underset{\sim}{w}\right) = -\frac{1}{2}\underset{\sim}{w}^T B^{-1} \underset{\sim}{w}$$

No.
Date. ・ ・

$\Rightarrow A^{-1} = \alpha I + \beta \Phi' \Phi$, ここで $\Phi = \begin{pmatrix} \phi_0(x_1) & \cdots & \phi_{M-1}(x_1) \\ \vdots & & \vdots \\ \phi_0(x_N) & \cdots & \phi_{M-1}(x_N) \end{pmatrix}$

② $\varphi(w)$ の w に関する1次の項を比較すると

$\beta \sum_{n=1}^{N} t_n w' \phi(x_n) = \frac{1}{2}(a' A^{-1} w + w' A^{-1} a)$

$\Rightarrow a = \beta A \Phi' t$

レポート：上の①，②を確かめよう．

結局 事後確率は

$p(w|t, \alpha, \beta) = c' \exp \varphi(w)$

$= c' \exp\{-\frac{1}{2}(w-a)' A^{-1} (w-a)\}$

多次元ガウス分布（c'：規格化定数）!!

より 推定値は，

$w_{MAP} = a$

Maximum a Posteriori （最大事後）

で与えられる

コメント：最尤推定に比べて，ベイズ推定の方が over fitting が生じにくい．

時系列ベイズ推定

N コのデータ (x_n, t_n) $n=1, \cdots, N$ が順次手に入るとき その都度 w を学習（推定）していく方法

アイデア：事後確率を次のステップの事前確率に使う．

$p(w)$ … 事前 p.

$p(t_1|w)$ ⟶ $p(w|t_1) = c_1 p(t_1|w) p(w)$ … ステップ1

(x_1, t_1) の尤度関数　事後 ↘ 事前確率として使う

$p(t_2|w)$ ⟶ $p(w|t_1, t_2) = c_2 p(t_2|w) p(w|t_1)$ … ステップ2

(x_2, t_2) の尤度関数

⋮　　⋮

k-1ステップ後に得られる事後確率

$p(t_k|w)$ ⟶ $p(w|t_1, \cdots, t_k) = c_k p(t_k|w) p(w|t_1, \cdots, t_{k-1})$ … ステップk

No.

Date. ・ ・

今日の話題：分類問題 (Classification)

「入力 $\underline{x} \in \mathbb{R}^D$ に対して K 個の異なるクラス C_k, $k=1,\cdots,K$ のどれか1つを割り当てたい」

記号：$\underline{t} (= \underline{t}(\underline{x})) = (t_1, \cdots, t_K)'$ $\Leftrightarrow$ $\underline{x}$ は C_j のクラスに属する。
$t_j = 1$, $t_k = 0$ $(k \neq j)$

問題設定：既知の分類分けデータ $(\underline{x}_i, \underline{t}_i)$, $i=1,\cdots,N$ を用いて未知入力データの分類分けを行う

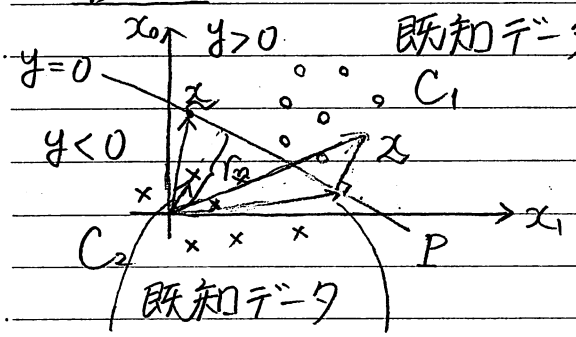
方針：既知データを用いて $\mathbb{R}^D$ を K コの部分集合 C_k, $k=1,\cdots,K$ (同じ記号を使う) に分割
$\mathbb{R}^D = \bigcup_{k=1}^{K} C_k$
できればよい

非確率的方法

① $K=2$.

$y(\underline{x}) = \underline{w}'\underline{x} + w_0 = 0$ ⇐ $(D-1)$ 次元超平面を定める。
P とする。

$D=2$.



$\underline{w}$

$\underline{x}_\perp$

既知データ ○ × から $(\underline{w}, w_0)$ を推定 (学習) できればよい

未知データ $\underline{x}$ に対しては
$y(\underline{x}) \geq 0 \Rightarrow \underline{x} \in C_1$
$y(\underline{x}) < 0 \Rightarrow \underline{x} \in C_2$
とする。

幾何的考察

・ $\underline{x}_A, \underline{x}_B \in P \Rightarrow y(\underline{x}_A) = y(\underline{x}_B)$
$\Rightarrow \underline{w}'(\underline{x}_A - \underline{x}_B) = 0$

No.
Date.

よって w は P に直交するベクトル

・原点から P までの最短距離 r_n は $x \in P$ をとると

$$r_n = \frac{w'x}{\|w\|} = \frac{-w_0}{\|w\|}$$

で与えられる.

・$x \in \mathbb{R}^D$ に対して $x_\perp \in P$ を x の P 上への射影とすると

$$x = x_\perp + r\frac{w}{\|w\|}, \quad r \in \mathbb{R} \text{ とかける.}$$

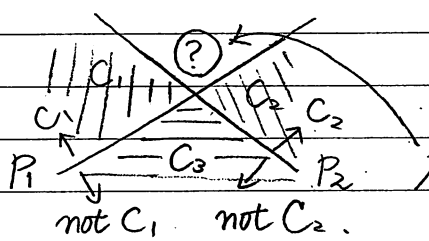
$$y(x) = w'x_\perp + r\frac{w'w}{\|w\|} + w_0 = r\|w\| \quad (w'x_\perp + w_0 = 0) \quad \therefore r = \frac{y(x)}{\|w\|}$$

② $K > 2$ の場合.

悪い分類の仕方1

$K-1$ 個の超平面 P_k, $k=1,\cdots,K-1$ $(\Leftrightarrow y_k(x)=0)$ を導入し

☆ $y_k(x) \geq 0 \Leftrightarrow x \in C_k$, $k=1,\cdots,K-1$ とする.



$K=3$ の場合

不確定領域

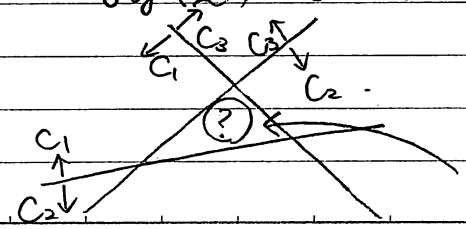
悪い分類の仕方2

C_i, C_j の分類を行う超平面 P_{ij} $(i \neq j)$ を

${}_K C_2 = \binom{K}{2} = \frac{K(K-1)}{2}$ コ導入

☆ $y_{ij}(x) \geq 0 \Leftrightarrow x \in C_i$

$y_{ij}(x) < 0 \Leftrightarrow x \in C_j$



$K=3$ の場合.

不確定領域

No.

Date.

良い分類の仕方.

Kコの線形関数を導入:

$y_k(x) = w_k' x + w_{k0}$, $k = 1, \cdots, K$

$x \in \mathbb{R}^D$ に対して

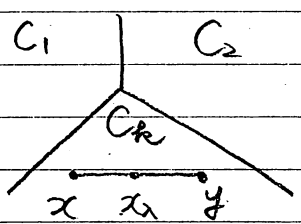
☆ $y_k(x) \geq y_j(x)$ (すべての $k \neq j$) $\Leftrightarrow$ $x \in C_k$

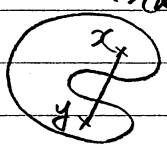
で分類分けを考える.

補題

各部分集合 $C_k \subset \mathbb{R}^D$ は凸集合

(任意の2点, $x, y \in C_k$ に対して
それらを結ぶ直線も C_k に入っている)





: 凸じゃない例

(証明) $z_A, z_B \in C_k$ とする. このとき z_A と z_B を結ぶ直線は.

$\{ z_\lambda = \lambda z_A + (1-\lambda) z_B \mid \lambda \in [0,1] \}$

と表される. 仮定から

$y_k(z_A) \geq y_j(z_A)$ $(\forall j \neq k)$

$y_k(z_B) \geq y_j(z_B)$

が成り立つ.

示したい事: $\forall \lambda \in [0,1]$

$z_\lambda \in C_k$ $(\Leftrightarrow y_k(z_\lambda) \geq y_j(z_\lambda)$ $(\forall j \neq k))$

$$\begin{aligned} y_k(z_\lambda) &= w'(\lambda z_A + (1-\lambda) z_B) + w_{k0} \\ &= \lambda w' z_A + (1-\lambda) w' z_B + \lambda w_{k0} + (1-\lambda) w_{k0} \\ &= \lambda y_k(z_A) + (1-\lambda) y_k(z_B) \\ &\geq \lambda y_j(z_A) + (1-\lambda) y_j(z_B) \\ &= \lambda w_j' z_A + (1-\lambda) w_j' z_B + w_{k0} = y_j(z_\lambda) \end{aligned}$$

//

No.
Date. ・ ・

よって既知データ $(\underset{\sim}{x}_i, \underset{\sim}{t}_i)$ $i=1,\cdots,N$ を用いて

$$\tilde{W} = \begin{pmatrix} w_{10} & w_{20} & \cdots & w_{K0} \\ \underset{\sim}{w}_1 & \underset{\sim}{w}_2 & \cdots & \underset{\sim}{w}_K \end{pmatrix} = (\tilde{w}_1, \cdots, \tilde{w}_K) \} D+1\text{コ}$$

を学習する. $\tilde{w}_k' = (w_{k0}, w_{k1}, \cdots, w_{kD})$

$\tilde{x}' = (1, x_1, \cdots, x_D)$ とすると

$y_k(\underset{\sim}{x}) = \tilde{w}_k' \underset{\sim}{x}$ と書ける.

$(\underset{\sim}{x}_i, \underset{\sim}{t}_i)$ $i=1,\cdots,N$ に対して

$$\tilde{X} = \begin{pmatrix} \tilde{x}_1' \\ \vdots \\ \tilde{x}_N' \end{pmatrix}, \quad T = \begin{pmatrix} \underset{\sim}{t}_1' \\ \vdots \\ \underset{\sim}{t}_N' \end{pmatrix}$$

$$\tilde{X}\tilde{W} = \begin{pmatrix} \tilde{x}_1' \\ \vdots \\ \tilde{x}_N' \end{pmatrix} (\tilde{w}_1, \cdots, \tilde{w}_K) = \begin{pmatrix} y_1(\underset{\sim}{x}_1) & \cdots & y_K(\underset{\sim}{x}_1) \\ \vdots & \ddots & \vdots \\ y_1(\underset{\sim}{x}_N) & \cdots & y_K(\underset{\sim}{x}_N) \end{pmatrix}$$

$$\tilde{X}\tilde{W} - T = \begin{pmatrix} y_1(\underset{\sim}{x}_1) - t_{11} & \cdots & y_K(\underset{\sim}{x}_1) - t_{1K} \\ \vdots & \ddots & \vdots \\ y_1(\underset{\sim}{x}_N) - t_{N1} & \cdots & y_K(\underset{\sim}{x}_N) - t_{NK} \end{pmatrix}$$

補題

$(y_1, \cdots, y_K)$ に対して

$$y_k > y_i \ (\forall i \neq k) \iff \underbrace{\sum_{j \neq k} y_j^2 + (y_k - 1)^2}_{A_k} < \underbrace{\sum_{j \neq i} y_j^2 + (y_i - 1)^2}_{A_i} \ (\forall i \neq k)$$

(証) $x > y \iff (x-1)^2 + y^2 < x^2 + (y-1)^2$ ①

$A_i - A_k = y_k^2 + (y_i - 1)^2 - y_i^2 - (y_k - 1)^2$

①より $A_i - A_k > 0 \iff y_k > y_i$ //

任意の i 行目に関して

$\forall i$ $\sum_{j=1}^{K} (y_j(\underset{\sim}{x}_i) - t_{ij})^2$ が最小

になるような $\tilde{W}$ は既知データの正しい分類を与える.

No.

Date.

エラー関数：

$$E(\tilde{W}) = \frac{1}{2} Tr\{(\tilde{X}\tilde{W} - T)(\tilde{X}\tilde{W} - T)'\}$$

$$\left(= \frac{1}{2} \sum_{i=1}^{N} \text{☆}_i\right)$$

の最小点 $\tilde{W}$ を求める．

Remark：$E(\tilde{W})$ は $\tilde{W}$ に関して2次

$\frac{\partial}{\partial \tilde{W}} E(\tilde{W}) = O$ なる $\tilde{W}$ ？

A, B を $n \times n$ 行列

$$\left.\begin{aligned} &\frac{\partial}{\partial A} Tr(A'BA) = (B + B')A \\ &\frac{\partial}{\partial A} Tr(A'B) = B \end{aligned}\right\} (\$)$$

($) を使うと

$$\tilde{W}_* = (\tilde{X}'\tilde{X})^{-1}\tilde{X}'T$$

よって この $\tilde{W}_*$ を用いることで K コの分類 C_k $k=1, \cdots, K$ を定める線形関数

$$\begin{pmatrix} y_1^*(x) \\ \vdots \\ y_K^*(x) \end{pmatrix} = \tilde{W}_*'\tilde{x}$$

が決定できた．未知データ $x \in R^D$ に対しては

$y_k^*(x) \geq y_j^*(x)$，全ての $j \neq k$

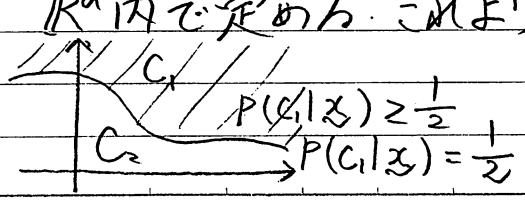
なる k を探して $x \in C_k$ とすればよい．

今日の話題：分類問題（確率的手法）

話を簡単にする：$K=2$（2種類のクラス C_1, C_2）

要は…

$p(C_1|x)$ を求めて $p(C_1|x) = 1/2$ の曲面を R^d 内で定める．これより $x \in C_1 \Leftrightarrow p(C_1|x) \geq 1/2$



No.
Date. ・ ・

手順：

① $p(x|C_1) = \frac{1}{(2\pi)^{D/2}|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(x-\mu_k)'\Sigma^{-1}(x-\mu_k)\right\}$ $(k=1,2)$

$= N(x|\mu_k, \Sigma)$ 多次元ガウス分布

$p(C_1) = \pi$, $p(C_2) = 1-\pi$ $0 \le \pi \le 1$

と仮定し，既知データ (x_i, t_i) $i=1,\cdots,N$ を用いて

パラメータ $\mu_1, \mu_2, \Sigma, \pi$ を学習

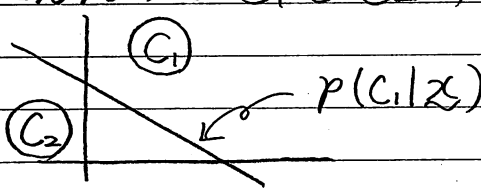
ここで $t_i = \begin{cases} 1 & x_i \in C_1 \\ 0 & x_i \in C_2 \end{cases}$

② ベイズの法則を用いて

$$p(C_1|x) = \frac{p(x|C_1)p(C_1)}{p(x)} = \frac{p(x|C_1)p(C_1)}{p(x|C_1)p(C_1)+p(x|C_2)p(C_2)}$$

を求める．(当然，$p(C_2|x) = 1 - p(C_1|x)$)

③ $p(C_1|x) = 1/2$ を定める超曲面（実は超平面になる）が C_1 と C_2 の境界として得られる．

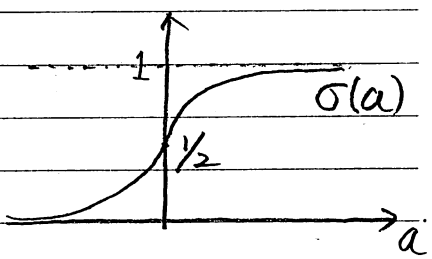


②，③ を先に説明（パラメータが推定できたとする）

$$p(C_1|x) = \frac{p(x|C_1)p(C_1)}{p(x|C_1)p(C_1)+p(x|C_2)p(C_2)} = \frac{1}{1+\exp(-a)} =: \sigma(a)$$

ここで $a = \log \frac{p(x|C_1)p(C_1)}{p(x|C_2)p(C_2)}$

$\sigma(a)$：logistic sigmoid 関数



No.

Date.

ここで $p(\boldsymbol{x}|C_k)$, $p(C_k)$ を代入すると

$$a = \log p(\boldsymbol{x}|C_1) - \log p(\boldsymbol{x}|C_2) + \log p(C_1)/p(C_2)$$
$$= -\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_1)'\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_1) + \frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_2)\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_2) + \log\frac{\pi}{1-\pi}.$$

$\boldsymbol{x}$ の2次の項は打ち消される.

$$= \boldsymbol{x}'\Sigma^{-1}\boldsymbol{\mu}_1 - \boldsymbol{x}'\Sigma^{-1}\boldsymbol{\mu}_2 - \frac{1}{2}\boldsymbol{\mu}_1'\Sigma^{-1}\boldsymbol{\mu}_1 + \frac{1}{2}\boldsymbol{\mu}_2'\Sigma^{-1}\boldsymbol{\mu}_2 + \log\frac{\pi}{1-\pi}$$
$$= \boldsymbol{w}'\boldsymbol{x} + w_0$$

ここで $\boldsymbol{w} = \Sigma^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$

$$w_0 = -\frac{1}{2}\boldsymbol{\mu}_1'\Sigma^{-1}\boldsymbol{\mu}_1 + \frac{1}{2}\boldsymbol{\mu}_2'\Sigma^{-1}\boldsymbol{\mu}_2 + \log\frac{\pi}{1-\pi}$$

よって $p(C_1|\boldsymbol{x}) = \sigma(\boldsymbol{w}'\boldsymbol{x} + w_0)$

手順③から

$$p(C_1|\boldsymbol{x}) = \sigma(\boldsymbol{w}'\boldsymbol{x} + w_0) = \frac{1}{2}$$

つまり $\boldsymbol{w}'\boldsymbol{x} + w_0 = 0$ ：超平面!!

が分類分けを定める境界

あとは手順①のパラメータ推定を考えればよい

既知データ：$\{\boldsymbol{x}_n, t_n\}$ $n=1,\cdots,N$ $t_n = \begin{cases} 1 & (\boldsymbol{x}_n \in C_1) \\ 0 & (\boldsymbol{x}_n \in C_2) \end{cases}$

$\boldsymbol{x}_n \in C_1 \Rightarrow p(\boldsymbol{x}_n, t_n = 1) = p(C_1)p(\boldsymbol{x}|C_1)$
$= \pi N(\boldsymbol{x}_n|\boldsymbol{\mu}_1, \Sigma)$

$\boldsymbol{x}_n \in C_2 \Rightarrow p(\boldsymbol{x}_n, t_n = 0) = p(C_2)p(\boldsymbol{x}|C_2)$
$= (1-\pi) N(\boldsymbol{x}_n|\boldsymbol{\mu}_2, \Sigma)$

よって $\boldsymbol{t} = (t_1, \cdots, t_N)$, $X = (\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N)$ とすると尤度関数は.

$$p(\boldsymbol{t}, X|\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \Sigma)$$
$$= \prod_{n=1}^{N} [\pi N(\boldsymbol{x}_n|\boldsymbol{\mu}_1, \Sigma)]^{t_n} [(1-\pi)N(\boldsymbol{x}_n|\boldsymbol{\mu}_2, \Sigma)]^{1-t_n}$$

とみる.

最大点 $(\pi^*, \boldsymbol{\mu}_1^*, \boldsymbol{\mu}_2^*, \Sigma^*)$ を求める. (最尤推定)

$\log p(\boldsymbol{t}, X|\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \Sigma)$

No.

Date.

$$= \sum_{n=1}^{N} [t_n \{\log \pi + \log N(x_n|\mu_1, \Sigma)\} + (1-t_n)\{\log(1-\pi) + \log N(x_n|\mu_2, \Sigma)\}]$$

$$\frac{\partial}{\partial \pi} \log p(*|\star) = \sum_{n=1}^{N} \left(\frac{t_n}{\pi} - \frac{1-t_n}{1-\pi}\right) = 0$$

$$(1-\pi)\sum_{n=1}^{N} t_n = \pi \sum_{n=1}^{N} (1-t_n)$$

$$\sum_{n=1}^{N} t_n = N\pi \quad \Rightarrow \quad \pi^* = \frac{1}{N}\sum_{n=1}^{N} t_n$$

$$\frac{\partial}{\partial \mu_1} \log p(*|\star) = \frac{\partial}{\partial \mu_1} \sum_{n=1}^{N} t_n \log N(x_n|\mu_1, \Sigma)$$

$$= \frac{\partial}{\partial \mu_1} \sum_{n=1}^{N} t_n \left(-\frac{1}{2}(x-\mu_1)'\Sigma^{-1}(x-\mu_1) + \log(\mu_1 \text{以外})\right)$$

$$= \sum_{n=1}^{N} t_n \left[\frac{\partial}{\partial \mu_1}(\mu_1'\Sigma^{-1}x_n) - \frac{\partial}{\partial \mu_1}\left(\frac{1}{2}\mu_1'\Sigma^{-1}\mu_1\right)\right]$$

$$= \sum_{n=1}^{N} t_n \Sigma^{-1} x_n - \sum_{n=1}^{N} t_n \Sigma^{-1}\mu_1 = 0$$

$$\sum_{n=1}^{N} t_n x_n = \left(\sum_{n=1}^{N} t_n\right)\mu_1$$

$$\mu_1^* = \frac{1}{\sum_{n=1}^{N} t_n} \sum_{n=1}^{N} t_n x_n$$

同様に

$$\mu_2^* = \frac{1}{\sum_{n=1}^{N}(1-t_n)} \sum_{n=1}^{N} (1-t_n) x_n.$$

最後に Σ^* は $\frac{\partial}{\partial \Sigma} p(*|\star) = 0$ より

$$\Sigma^* = \frac{N_1}{N} S_1 + \frac{N_2}{N} S_2$$

ここで S_1, S_2 (行列) は

$$S_1 = \frac{1}{N_1}\sum_{n \in C_1} (x_n - \mu_1^*)(x_n - \mu_1^*)'$$

$$S_2 = \frac{1}{N_2}\sum_{n \in C_2} (x_n - \mu_2^*)(x_n - \mu_2^*)'$$

$$N_1 = \sum_{n=1}^{N} t_n, \quad N_2 = \sum_{n=1}^{N} (1-t_n)$$

No.

Date.

Support Vector Machine (SVM) 入門

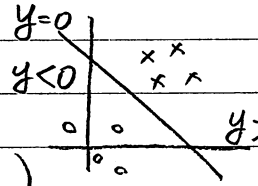
復習 (2分類問題)

既知データを用いて超平面

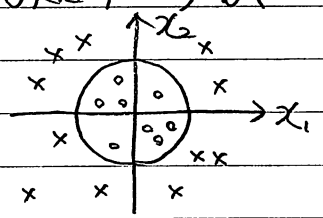
$y(x) = w'x + w_0 = 0 \quad x \in \mathbb{R}^d$

を学習させて (つまり) w, w_0 を推定)

未知データの分類を行った.



既知データが



$\Rightarrow$ 打開策

$y(x) = x_1^2 + x_2^2 - r^2$

$= w'\phi(x) + w_0$

$w' = (1, 1), \ \phi(x) = (x_1^2, x_2^2)$

$w_0 = -r^2$

とすることで既知データの分類は可能になる.

の場合には平面での分類は不可能

今後一般に.

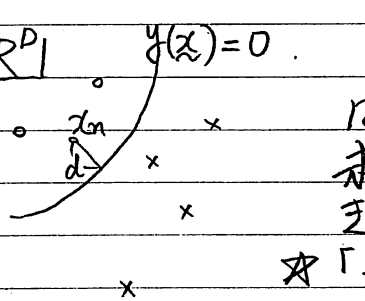
(1) $y(x) = w'\phi(x) + w_0$

$\phi(x) = (\phi_1(x), \cdots, \phi_M(x))'$ はある非線形変換の形で分類問題を考える. (つまり曲面によって $\mathbb{R}^D$ の分割)

既知データを

$$t_n = \begin{cases} 1 & y(x_n) > 0 \Leftrightarrow C_1 \text{の領域} \\ -1 & y(x_n) < 0 \Leftrightarrow C_2 \quad \text{〃} \end{cases}$$

としよう $|y(x_n)| = t_n y(x_n)$ に注意.



○と×を分類する曲面 $y(x) = 0$ は連続的に存在するが, 今後の未知データを最も正確に分類できそうなのは.

☆「最も近い既知データとの距離が最大になる曲面」

No.
Date.

$$d = \frac{|y(x_n)|}{\|w\|} = \frac{t_n y(x_n)}{\|w\|} = \frac{t_n(w'\phi(x_n)+w_0)}{\|w\|}$$ より ☆ は

($) $$\underset{w,w_0}{\operatorname{argmax}} \left\{ \frac{1}{\|w\|} \min_n [\, t_n(w'\phi(x_n)+w_0) \,] \right\}$$

で与えられる.

{*} は $w \to \kappa w$, $w_0 \to \kappa w_0$ $(0 \neq \kappa \in \mathbb{R})$ で不変なので

最も曲面に近い点 x_n で

$$t_n(w'\phi(x_n)+w_0) = 1$$

と規格化しておく. すると

$$t_i(w'\phi(x_n)+w_0) \geq 1 \quad i=1,\cdots,N$$

これより ($) は

$$\underset{w,w_0}{\operatorname{argmax}} \left\{ \frac{1}{\|w\|} \min_n [\quad] \right\} = \underset{w,w_0}{\operatorname{argmax}} \frac{1}{\|w\|},$$

つまり

$$\underset{w,w_0}{\operatorname{argmin}} \frac{1}{2}\|w\|^2$$

の解として曲面が求まる.

まとめると 既知データ $\{x_n, t_n\}$ の曲面 (1) による分類は

制約条件: $t_i(w'\phi(x_i)+w_0) \geq 1 \quad i=1,\cdots,N$

の下で

$$\underset{w,w_0}{\operatorname{argmin}} \frac{1}{2}\|w\|^2$$

を求めれば良い.

制約条件付き極値問題 ⇐ ラグランジュの未定乗数法

復習 ラグランジュ未定乗数法

① 制約条件: $g(x)=0 \quad x \in \mathbb{R}^D$
の下で関数 $f(x) \in \mathbb{R}^D$ を最大にする
点 x^* を求めたい } Q1

No.

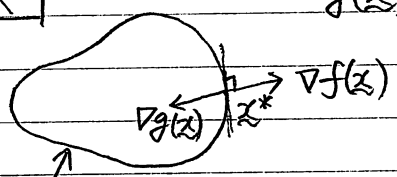
Date.

$\mathbb{R}^D$

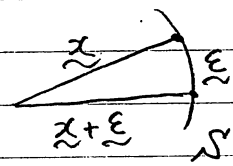
$g(\underset{\sim}{x}) = 0$ ← $\mathbb{R}^D$の$(D-1)$次元曲面を定める

↓

Sとする



$S = \{g(\underset{\sim}{x}) = 0\}$



補題：$\underset{\sim}{x} \in S$に対して

$$\nabla g(\underset{\sim}{x}) = \left(\frac{\partial g}{\partial x_1}, \cdots, \frac{\partial g}{\partial x_D}\right)$$ ← gが最も増える方向

はSに直交する

証）$\underset{\sim}{x}+\underset{\sim}{\varepsilon}$を$S$上の$\underset{\sim}{x}$に十分近い点とするとき

テイラー展開より

$$g(\underset{\sim}{x}+\underset{\sim}{\varepsilon}) = g(\underset{\sim}{x}) + \underset{\sim}{\varepsilon}'\nabla g(\underset{\sim}{x}) + O(\underset{\sim}{\varepsilon}^2)$$

$$0 \qquad 0$$

$$\underset{\sim}{\varepsilon}' g(\underset{\sim}{x}) = 0$$

$\underset{\sim}{\varepsilon}$は$\underset{\sim}{x}$での$g(\underset{\sim}{x}) = 0$の接ベクトルなので

$\nabla g(\underset{\sim}{x})$は$S$に直交する。

一方 $f(\underset{\sim}{x})$が$\underset{\sim}{x}^* \in S$で極大になるとき

$\nabla f(\underset{\sim}{x}^*)$は$S$に直交

よって

$$\nabla f(\underset{\sim}{x}^*) + \lambda \nabla g(\underset{\sim}{x}^*) = 0$$

となる $0 \neq \lambda \in \mathbb{R}$が存在する（$\because \nabla f$と$\nabla g$は平行）

$L(\underset{\sim}{x}, \lambda) = f(\underset{\sim}{x}) + \lambda g(\underset{\sim}{x})$ ：ラグランジュ関数

ラグランジュの未定乗数

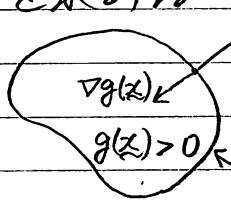
Q1が極大値を持つ必要条件

「$\nabla_{\underset{\sim}{x}} L = 0$, $\frac{\partial}{\partial \lambda} L = 0$」

これを満たす$(\underset{\sim}{x}^*, \lambda^*)$を求める方法

ラグランジュ未定乗数法

② 制約条件：$g(\boldsymbol{x}) \geq 0$，$\boldsymbol{x} \in \mathbb{R}^D$
の下で $f(\boldsymbol{x}) \in \mathbb{R}$ を極大にする点 $\boldsymbol{x}^*$
を求める　} Q2



ラグランジュ関数
$L(\boldsymbol{x}, \lambda) = f(\boldsymbol{x}) + \lambda g(\boldsymbol{x})$

場合分け
(A) $g(\boldsymbol{x}^*) > 0$
この場合 制約条件は無意味
$\Rightarrow \lambda = 0$，設 $L(\boldsymbol{x}, \lambda) = 0$
(B) $g(\boldsymbol{x}^*) = 0$，Q1と同じ
よって $0 \neq \lambda^* \in \mathbb{R}$ に対して $\nabla_{\boldsymbol{x}} L(\boldsymbol{x}^*, \lambda^*) = \frac{\partial}{\partial \boldsymbol{x}} L(\boldsymbol{x}^*, \lambda^*) = 0$
ただし $\nabla f(\boldsymbol{x}^*) = -\lambda^* \nabla g(\boldsymbol{x}^*)$ より $\lambda^* > 0$ でなければならない．
よってまとめると：　← 場合分けに対応
$g(\boldsymbol{x}) \geq 0$，$\lambda \geq 0$，$\lambda g(\boldsymbol{x}) = 0$ の下で
$L(\boldsymbol{x}, \lambda) = f(\boldsymbol{x}) + \lambda g(\boldsymbol{x})$
の極値を与える $(\boldsymbol{x}^*, \lambda^*)$ を求める方法が Q2 に対するラグランジュ未定乗数法．

複数個の制約条件 $g_j(\boldsymbol{x}) = 0$，$j = 1, \cdots, J$
$h_k(\boldsymbol{x}) \geq 0$　$k = 1, \cdots, K$ の場合
$L(\boldsymbol{x}, \{\lambda_j\}, \{\mu_k\}) = f(\boldsymbol{x}) + \sum_{j=1}^{J} \lambda_j g_j(\boldsymbol{x}) + \sum_{k=1}^{K} \mu_k h_k(\boldsymbol{x})$
（ただし $\mu_k \geq 0$，$\mu_k h_k(\boldsymbol{x}) = 0$ ）
とすればよい．

解こう
① $L(\boldsymbol{w}, w_0, \boldsymbol{a}) = \frac{1}{2}\|\boldsymbol{w}\|^2 - \sum_{n=1}^{N} a_n \{t_n(\boldsymbol{w}^T \boldsymbol{\phi}(\boldsymbol{x}_n) + w_0) - 1\}$
として　$a_n \geq 0$，$t_n y(\boldsymbol{x}_n) \geq 1$　$a_n(t_n y(\boldsymbol{x}_n) - 1) = 0$

No.

Date.

の下で $L(w, w_0, a)$ の極値を探す

$\frac{\partial L}{\partial w} = w - \sum_{n=1}^{N} a_n t_n \phi(x_n) = 0 \Rightarrow w^* = \sum_{n=1}^{N} a_n t_n \phi(x_n)$

$\frac{\partial L}{\partial w_0} = -\sum_{n=1}^{N} a_n t_n = 0$

$L(w, w_0, a)$ に代入すると

$$\tilde{L}(a) = \sum_{n=1}^{N} a_n + \frac{1}{2}\left(\sum_{n=1}^{N} a_n t_n \phi(x_n)'\right)\left(\sum_{n=1}^{N} a_n t_n \phi(x_n)\right) - \sum_{n=1}^{N} a_n t_n \phi(x_n)'\left(\sum_{n=1}^{N} a_n t_n \phi(x_n)\right)$$

$$= \sum_{n=1}^{N} a_n - \frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{N} a_n a_m t_n t_m k(x_n, x_m) \quad ②$$

ここで $k(x, x') = \phi(x)'\phi(x)$: カーネル

$a_n \geq 0$, $\sum_{n=1}^{N} a_n t_n = 0$

の下で極値 $\tilde{L}(a^*)$ が求まれば $w^* = \sum_{n=1}^{N} a_n^* t_n \phi(x_n)$ として完了

→ 2次計画問題

(非線形)最適化問題

多くの数値解法が知られている. (MATLABの関数もある)

⇒ a^* を数値的に求めることは可能!

学習解 w^* の性質

分類を定める曲面 S は

$$y(x) = w^{*\prime}\phi(x) + w_0$$

$$= \sum_{n=1}^{N} a_n t_n \phi(x_n)'\phi(x) + w_0$$

$$= \sum_{n=1}^{N} a_n t_n k(x_n, x) + w_0 = 0$$

で与えられる.

ここで全ての既知データに対して

$a_n = 0$ or $t_n y(x_n) = 1$

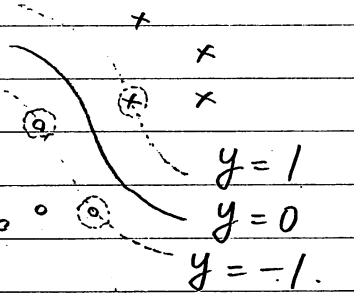
が満たされている. (∵① $a_n \geq 0$, $a_n(t_n y(x_n) - 1) = 0$)

よって

$$y(x) = \sum_{n \in S} a_n t_n k(x_n, x) + w_0$$

$S := \{n \mid t_n y(x_n) = 1\}$

分類曲面 $y(x) = 0$ に最も近くにある既知データ



⇒ 未知データの分類の計算量が大幅に減少
(左の例だと
3つの○でOK)
○：サポート

まとめると
サポートベクトルマシンの特徴
「$n \in \mathcal{S}$ なる x_n (サポート) のみを用いて
未知データの分類を行う」

資料1（参考書 p7～8）

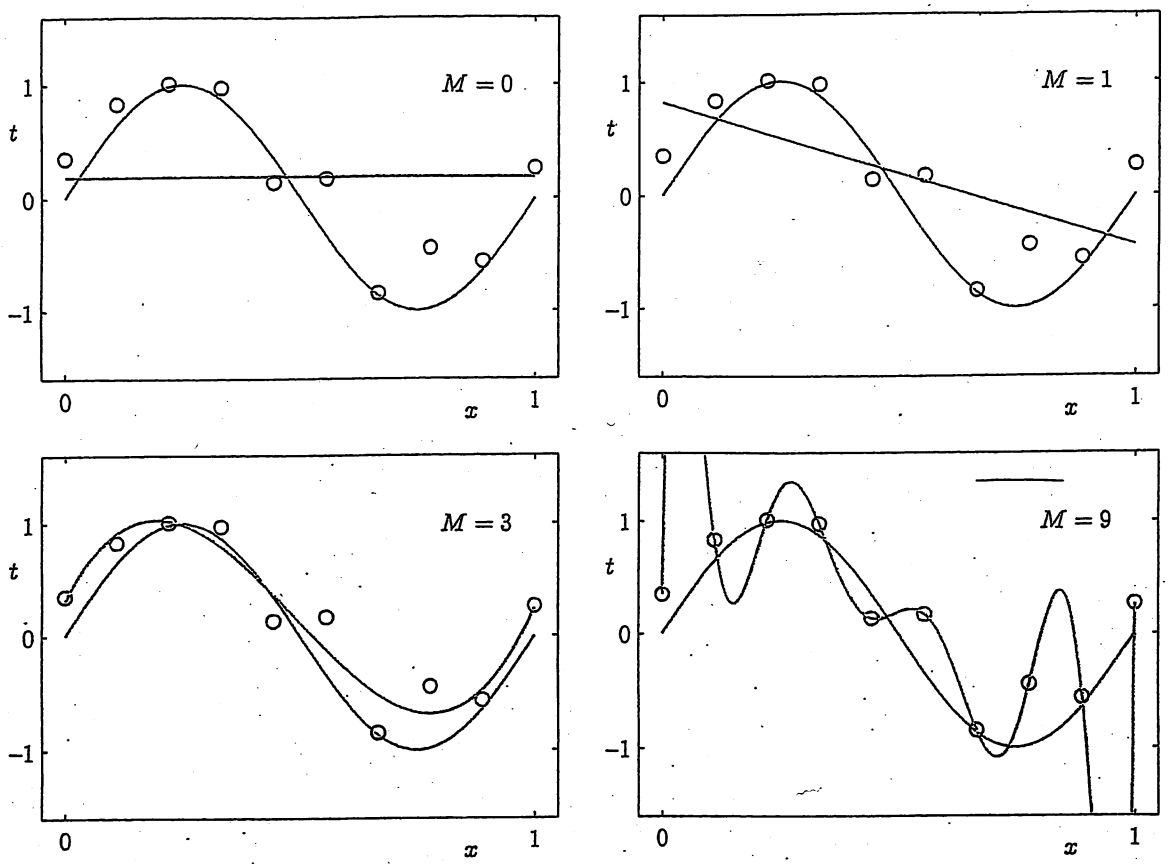


Figure 1.4 Plots of polynomials having various orders M, shown as red curves, fitted to the data set shown in Figure 1.2.

Table 1.1 Table of the coefficients w* for polynomials of various order. Observe how the typical magnitude of the coefficients increases dramatically as the order of the polynomial increases.

	$M = 0$	$M = 1$	$M = 6$	$M = 9$
$w_0^\star$	0.19	0.82	0.31	0.35
$w_1^\star$		-1.27	7.99	232.37
$w_2^\star$			-25.43	-5321.83
$w_3^\star$			17.37	48568.31
$w_4^\star$				-231639.30
$w_5^\star$				640042.26
$w_6^\star$				-1061800.52
$w_7^\star$				1042400.18
$w_8^\star$				-557682.99
$w_9^\star$				125201.43

資料2

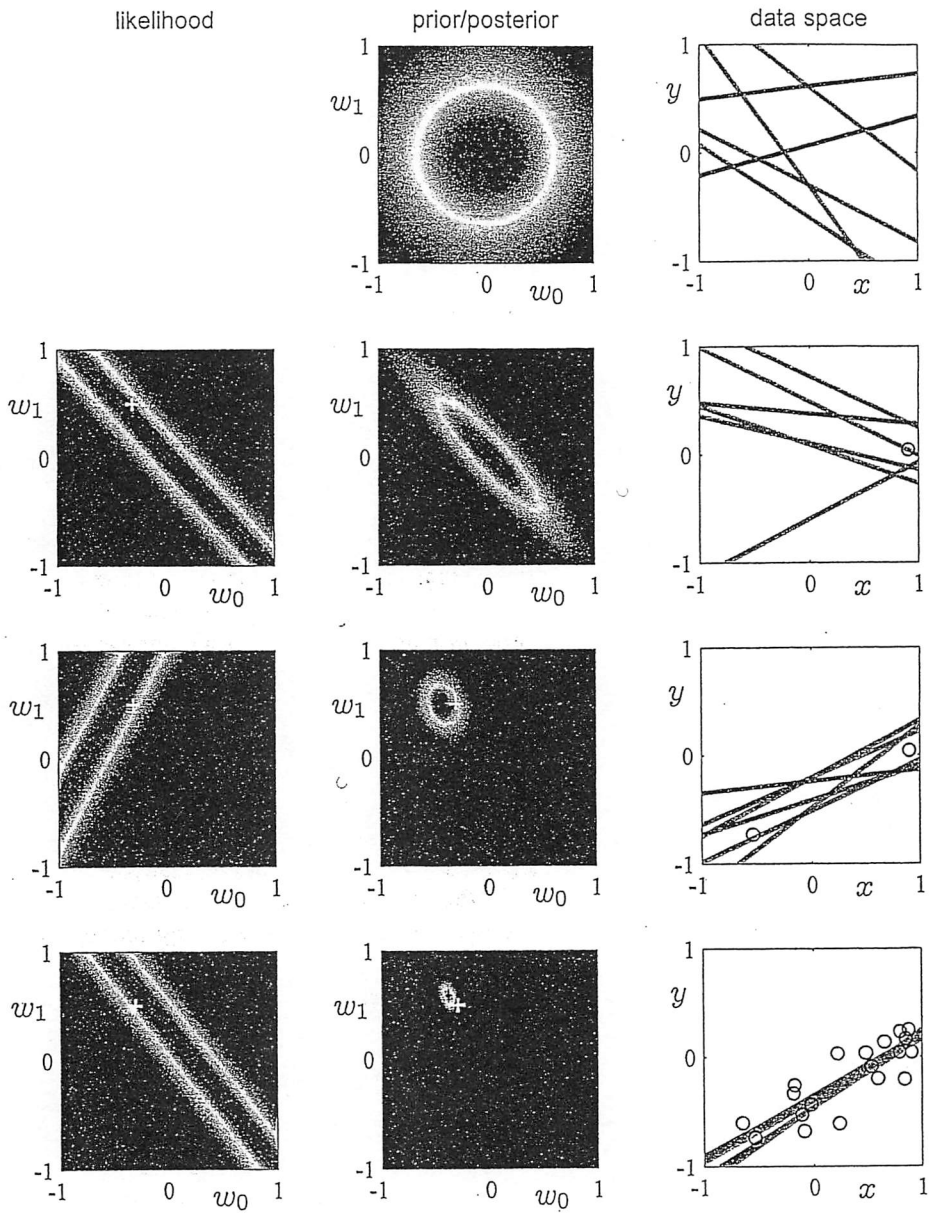


Figure 3.7 Illustration of sequential Bayesian learning for a simple linear model of the form $y(x, \mathbf{w}) = w_0 + w_1 x$. A detailed description of this figure is given in the text.