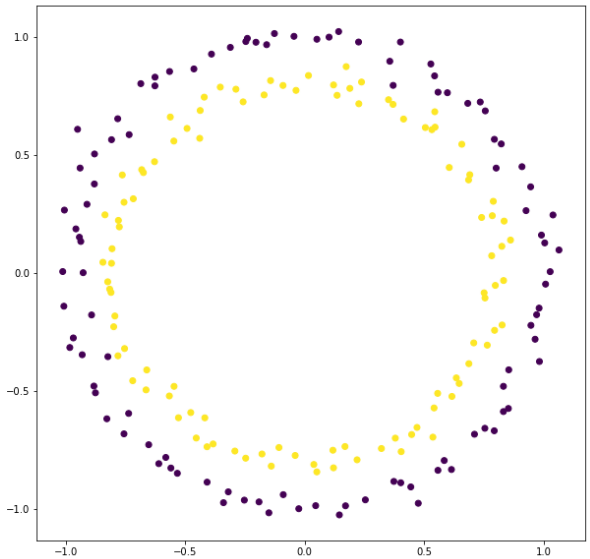
PRML6章

担当：大木

導入

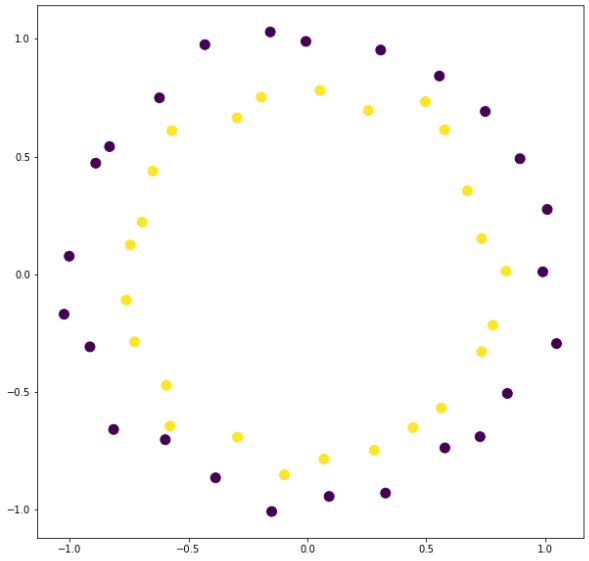
3,4章では回帰・分類問題に
対する線形手法を扱った

問：次のデータを二値分類せよ

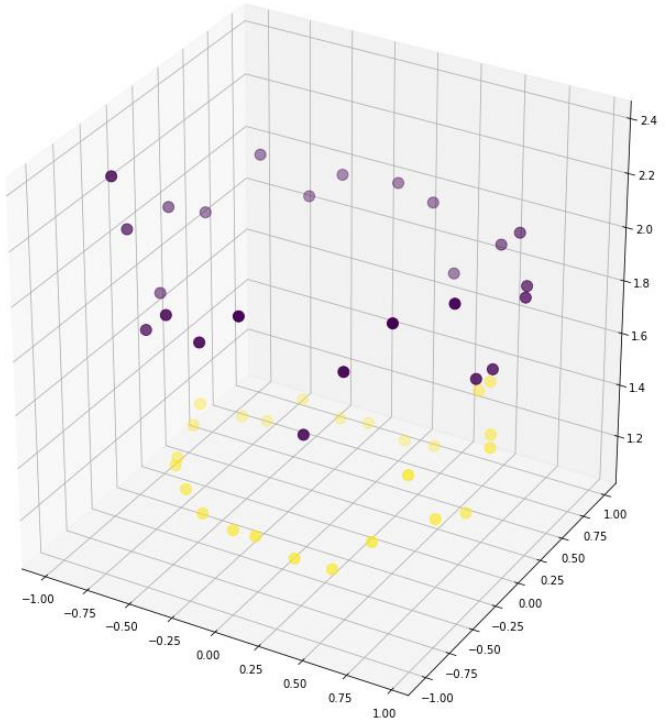


線形識別では無理そう

特徴選択



特徴変換



変換後は平面で分離できそう（線形識別可能）

アイデア

元のデータ空間を特徴空間に非線形写像して、
これまでの線形手法を適用！

しかし…

- 適切な特徴変換を見つけるのは大変

- 特徴空間はデータ次元に対して非常に高次元で、計算資源的に厳しい

cf. 200次元データに対して3次のモーメントを考慮する場合

$$ {}_{200}C_1 + {}_{200}C_2 + {}_{200}C_3 = 1333500 $$

カーネル法

データの非線形性や高次モーメントを
計算コストを抑えつつ、特徴変換を陽に決めずに
上手く取り入れる手法

カーネル法がうまくいく理由
→カーネルトリック

特徴変換関数を $\Phi(x)$ とすると、
特徴空間における内積（類似度）

$$\langle \Phi(X_i), \Phi(X_j) \rangle = k(X_i, X_j)$$

の値が得られれば変換関数の具体形が
わからなくても線形のデータ解析が行える

さらに…

カーネル関数が実数なら入力データX_iは
複素数や文字列、グラフ構造などでも
良い

→パターン認識での様々な応用

例：自然言語処理、バイオインフォマティクス. . .

ここから本題

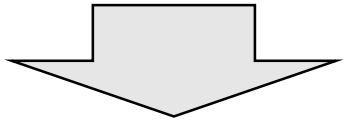
教科書(PRML)をベースに
適宜補足しながら進めていきます

目次

目次

双対表現

回帰・分類に用いられる線形モデルの多くは、
双対表現によってカーネル関数で書き直すことが可能



ということで、3章、4章でそれぞれ扱った
リッジ回帰とフィッシャー判別分析について
見ていく

カーネルリッジ回帰

学習データセット

$$(\mathbf{x}_1, t_1), \ldots, (\mathbf{x}_N, t_N)$$

$\mathbf{x}_n$：多次元説明変数

t_n：1次元目的変数

以下を最小化することを考える $(\lambda \geq 0)$

$$J(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N} \|t_n - \mathbf{w}^{\mathrm{T}}\phi(\mathbf{x}_n)\|^2 + \lambda\|\mathbf{w}\|^2$$

…(6.2)

極値を求めると

$$\mathbf{w} = -\frac{1}{\lambda}\sum_{n=1}^{N}\{\mathbf{w}^{\mathrm{T}}\phi(\mathbf{x}_n) - t_n\}\phi(\mathbf{x}_n)$$

$a_n = -\frac{1}{\lambda}\{\mathbf{w}^{\mathrm{T}}\phi(\mathbf{x}_n) - t_n\}$ とおくと

$\mathbf{w} = \sum_{n=1}^{N} a_n \phi(\mathbf{x}_n)$ という双対表現が得られる

これを用いて、目的関数をaによって
書き直すと(6.5)のようになる

式長いので参照にて勘弁してください

さらに、(6.1)のカーネル関数を行列要素に持つ

グラム行列 $\mathbf{K} = \mathbf{\Phi}\mathbf{\Phi}^{\mathrm{T}}$

を用いて書き直すと、(6.7)のようになる

また、(6.3)を用いて(6.4)からwを消去して
aについて解くと

$$\mathbf{a} = (\mathbf{K} - \lambda \mathbf{I}_N)^{-1}\mathbf{t} \qquad (6.8)$$

これを線形モデルに代入しなおせば、
新たなデータxに対する予測は

$$y(\mathbf{x}) = \mathbf{k}(\mathbf{x})^{\mathrm{T}}(\mathbf{K} - \lambda \mathbf{I}_N)^{-1}\mathbf{t} \qquad (6.9)$$

以上から、リッジ回帰の最適解は
カーネル関数のみで表現できることが確認できた

双対表現によってM次元パラメータベクトルwから、
N次元パラメータベクトルaへ表記しなおすことができた

計算量は？

$M \times M$　　vs.　　$N \times N$

カーネル法の特長

特徴ベクトルを陽に扱わずに
高次元の特徴空間を考えることが出来る

おまけ：カーネルフィッシャー判別分析

フィッシャー判別分析（4.1.4～参照）

2クラス分類において、判別分析には
以下の最大化問題を解けばよかった

$$\max \mathbf{w}^{\mathrm{T}}\mathbf{S}_{\mathrm{B}}\mathbf{w} \qquad \text{s.t. } \mathbf{w}^{\mathrm{T}}\mathbf{S}_{\mathrm{W}}\mathbf{w} = 1$$

非線形化するために、1次元射影関数を
特徴空間上のものと考える

$$y = \mathbf{w}^{\mathrm{T}}\mathbf{x} \quad \Rightarrow \quad y = \mathbf{w}^{\mathrm{T}}\phi(\mathbf{x})$$

ここで、特徴ベクトルの張る部分空間に直交する
ベクトルは目的関数の値に寄与しないので最適解は

$$\mathbf{w} = \sum_{n=1}^{N} a_n \phi(\mathbf{x}_n)$$

→リプレゼンター定理できちんとやります

と書ける

これを用いてクラス内分散の和とクラス間分散を
カーネル関数で書き直すと

※途中の計算はスライドから省きます

クラス間分散　$\mathbf{a}^{\mathrm{T}}\mathbf{S}_B^{\phi}\mathbf{a}$

$$\mu_i = \sum_{n \in C_i} \mathbf{k}(X_n, X)$$

$$\mathbf{S}_B^{\phi} = (\mu_2 - \mu_1)(\mu_2 - \mu_1)^{\mathrm{T}}$$

クラス内分散の和　$\mathbf{a}^{\mathrm{T}}\mathbf{S}_W^{\phi}\mathbf{a}$

$$(K_l)_{ij} = k(X_i, X_j),\ (i \in C_l, j = 1, ..., N)$$

$$Q_N = I_N - \frac{1}{N}\mathbf{1}_N\mathbf{1}_N^{\mathrm{T}}$$

$$\mathbf{S}_W^{\phi} = K_1^{\mathrm{T}} Q_N K_1 + K_2^{\mathrm{T}} Q_N K_2$$

※実際はカーネル法の表現能力の高さから、
クラス内分散の和が0になるような解が選ばれてしまう(過学習)

正則化項を加える

$$\tilde{J}(\mathbf{a}) = \frac{\mathbf{a}^{\mathrm{T}}\mathbf{S}_B^{\phi}\mathbf{a}}{\mathbf{a}^{\mathrm{T}}(\mathbf{S}_W^{\phi}+\lambda I_N)\mathbf{a}}$$

目次

カーネル法では、特徴変換を考えなくとも
カーネル関数を与えさえすればよい

この節では、具体的なカーネル関数の
設計法について見ていく

カーネル関数は何か特徴ベクトルの内積という形になっていなければならない

$$\langle \Phi(X_i), \Phi(X_j) \rangle = k(X_i, X_j)$$

これを満たす必要十分条件は、
カーネル関数が正定値性カーネル
(グラム行列が半正定値行列)であるということ

補足

正定値カーネル

再生核ヒルベルト空間

リプレゼンター定理

正定値カーネルの定義

集合$\mathcal{X}$に対して、次の二つの条件を満たすカーネルを$\mathcal{X}$上の正定値カーネルという

対称性 $\forall x, y \in \mathcal{X}$ に対して $k(x, y) = k(y, x)$

正値性 $\forall n \in \mathbb{N}, x_1, \ldots, x_n \in \mathcal{X}, c_1, \ldots, c_n \in \mathbb{R}$ に対し

$$\sum_{i,j} c_i c_j k(x_i, x_j) \geq 0$$

※これらはグラム行列が半正定値であることと同値

$K_{ij} = k(x_i, x_j)$ を行列要素に持つ行列

再生核ヒルベルト空間

集合$\mathcal{X}$上の再生核ヒルベルト空間とは、
$\mathcal{X}$上の関数からなるヒルベルト空間$\mathcal{H}$で、
任意の $x \in \mathcal{X}$ に対して $k(\cdot, x) \in \mathcal{H}$ が存在し

再生性　　$\langle f, k(\cdot, x) \rangle = f(x)$

を満たすものをいう
このカーネル関数kを$\mathcal{H}$の再生核と呼ぶ

$\mathcal{X}$上の再生核ヒルベルト空間の再生核kは
$\mathcal{X}$上の正定値カーネルで、一意的

証明:

$$\sum_{i,j} c_i c_j k(x_i, x_j) = \sum_{i,j} c_i c_j \langle k(\cdot, x_i) k(\cdot, x_j) \rangle$$
$$= \langle \sum_i c_i k(\cdot, x_j), \sum_j c_j k(\cdot, x_j) \rangle \geq 0$$

対称性、一意性はスライドでは割愛

※証明はしませんが、逆に正定値カーネルから
再生核ヒルベルト空間を構成可能

特徴写像を $\phi(\mathbf{x}) = k(\mathbf{x}, \cdot)$ と定めれば、再生性から

$$\begin{aligned}\langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_j) \rangle &= \langle k(\mathbf{x}_i, \cdot), k(\mathbf{x}_j, \cdot) \rangle \\ &= k(\mathbf{x}_i, \mathbf{x}_j)\end{aligned}$$

以上から、正定値カーネルが特徴ベクトルの内積として
与えられるカーネル関数に相当することが確認できた

リプレゼンター定理

正則化付きの誤差関数

$$R(f) = R_{\mathrm{L}}(\{f(\mathbf{x}_i), y_i\}_{i=1,\ldots,N}) + \lambda R_{\mathrm{P}}(\|f\|_{\mathcal{H}}^2)$$

ただし　f : 再生核ヒルベルト空間$\mathcal{H}$の元

R_{L} : 任意の関数

R_{P} : 狭義単調増加関数

$\lambda > 0$　とする

このとき、最適解はカーネル関数の線形結合

$$f(\mathbf{x}) = \sum_{n=1}^{N} a_n k(\mathbf{x}_n, \mathbf{x})$$

でかける

証明

再生核ヒルベルト空間$\mathcal{H}$に対して、
データの張る部分空間 $\{k(\mathbf{x}_i, \cdot), i = 1, ..., N\}$ を
$\mathcal{H}_0$とし、その直交補空間を$\mathcal{H}_\perp$とする

ここで、R_{L}の最適解をfとすると、
それぞれの成分 $f_0, f_\perp$ で

$$f = f_0 + f_\perp$$

と表せる

$$\begin{aligned}\langle f, k(\mathbf{x}_i, \cdot)\rangle &= \langle f_0, k(\mathbf{x}_i, \cdot)\rangle + \underbrace{\langle f_\perp, k(\mathbf{x}_i, \cdot)\rangle}_{=0} \\ &= \langle f_0, k(\mathbf{x}_i, \cdot)\rangle\end{aligned}$$

つまり、$f = f_0$ としてよい

また、正則化項 R_{P} に関しても

$$R_{\mathrm{P}}(\|f\|^2) = R_{\mathrm{P}}(\|f_0\|^2 + \|f_\perp\|^2)$$
$$\geq R_{\mathrm{P}}(\|f_0\|^2)$$

より最適解は $f = f_0$

f_0 は $\mathcal{H}_0$ の元であったので $R(f)$ の最適解は

$$f(\mathbf{x}) = \sum_{n=1}^{N} a_n k(\mathbf{x}_n, \mathbf{x})$$

と表せる(証明終)

リプレゼンター定理から、データを高次元の特徴空間に
写像したとしても、その解はデータ数のオーダーの線形和で
与えられることが保証される！

$$f(\mathbf{x}) = \sum_{n=1}^{N} a_n k(\mathbf{x}_n, \mathbf{x})$$

※制約条件に対するリプレゼンター定理もほぼ同じ様にして証明

PRMLに戻ります

(6.13)〜(6.22)が
正定値カーネルになることを確認(板書にて)

代表的なカーネル関数

多項式カーネル

$$k(\mathbf{x}, \mathbf{x}') = (\mathbf{x}^{\mathrm{T}}\mathbf{x}' + c)^M \qquad (c > 0, M \in \mathbb{N})$$

ガウスカーネル

$$k(\mathbf{x}, \mathbf{x}') = \exp(-\|\mathbf{x} - \mathbf{x}'\|^2 / 2\sigma^2)$$

※ガウスカーネルは無限次元の特徴ベクトルに対応する(演習6.11)

カーネル法は入力空間 $\mathcal{X}$ が実数のみならず、
文字列、グラフ構造、集合といった場合にも
カーネル関数さえ定義できれば適用可能

→例：演習6.12にて紹介

生成モデルを用いてカーネルを定義することも可能

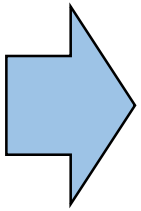
$$k(\mathbf{x}, \mathbf{x}') = p(\mathbf{x})p(\mathbf{x}')$$

カーネル関数の構成法から、上のカーネルは(6.29)～(6.31)のように拡張可能

他にも、フィッシャーカーネル(6.33)などもある

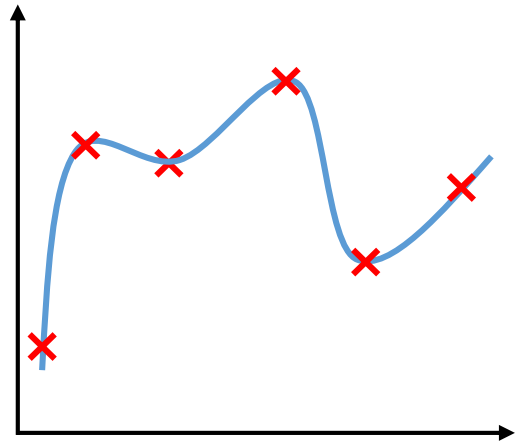
目次

３章では、固定された基底関数の
線形結合を用いた回帰を考えた



・RBF(radical basis function:動径基底関数)を基底関数に用いることが多い
・ということで、RBFに着目してみていく

RBFは、関数補間で初めて導入された



$$f(\mathbf{x}) = \sum_n w_n h(\|\mathbf{x} - \mathbf{x}_n\|)$$

最小二乗法で求解

→応用上、過学習を抑える必要(正則化)

※正則化項に微分値を用いた二乗和誤差関数では、
最適解はグリーン関数の線形和で表される

入力変数にノイズが含まれる場合にもRBFは利用される

入力$\mathbf{x}$のノイズをξとすると、二乗和誤差関数は

分布$\nu(\xi)$に従う確率変数

$$E = \frac{1}{2}\sum_{n=1}^{N}\int \{y(\mathbf{x}_n + \xi) - t_n\}^2 \nu(\xi) d\xi \quad \textbf{(6.39)}$$

変分法によって最適な$y(\mathbf{x})$は

$$y(\mathbf{x}) = \sum_{n=1}^{N} t_n h(\mathbf{x} - \mathbf{x}_n) \quad \textbf{(6.40)}$$

RBF

となる

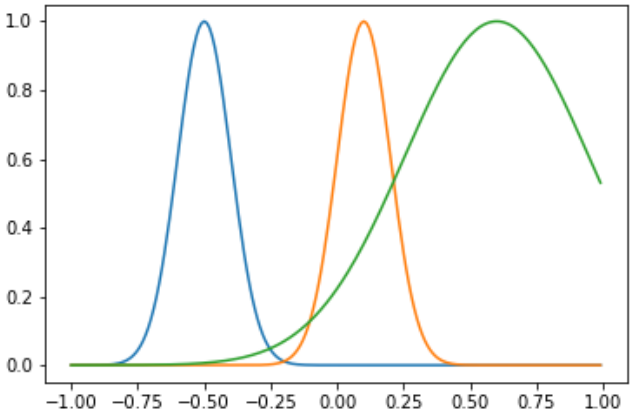
ここでのRBFは

$$h(\mathbf{x}-\mathbf{x}_n)=\frac{\nu(\mathbf{x}-\mathbf{x}_n)}{\sum_{n=1}^{N}\nu(\mathbf{x}-\mathbf{x}_n)} \qquad (6.41)$$

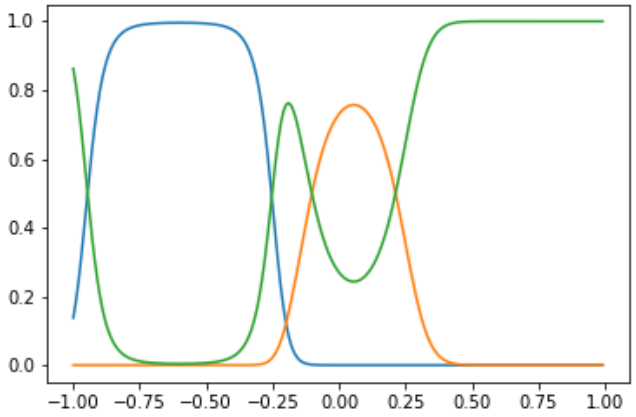
で与えられ、全てのデータ点を
基底として持ったモデルとなっている

→Nadaraya-Watsonモデル(後述)

(6.41)は正規化されているため、
基底によって入力空間を埋め尽くすことができている

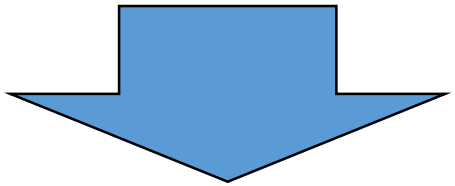


正規化なし



正規化あり

RBFでは基底の計算量がデータ数に相当するため、
データが多いときは予測にかかる計算コストが大きい



・直交最小二乗法
・クラスタリングアルゴリズム
等を用いた基底に用いるデータ点の選別

Nadaraya-Watsonモデル(カーネル回帰)

(6.41)をカーネル密度推定から再導出する

テストデータセット $\{\mathbf{x}_n, t_n\}_{n=1}^{N}$

Parzen推定法を用いて、同時分布 $p(\mathbf{x}, t)$を推定することを考える

密度関数

$$p(\mathbf{x}, t) = \frac{1}{N} \sum_{n=1}^{N} f(\mathbf{x} - \mathbf{x}_n, t - t_n) \qquad \textbf{(6.42)}$$

1.5.5節で議論したように、回帰関数 $y(\mathbf{x})$ を求めるには $E[t|\mathbf{x}]$ を考えればよい。このとき、(6.42)を用いて

$$E[t|\mathbf{x}] = \frac{\sum_n \int t f(\mathbf{x}-\mathbf{x}_n, t-t_n)dt}{\sum_m \int f(\mathbf{x}-\mathbf{x}_m, t-t_m)dt} \qquad (6.43)$$

ここで、$f(\mathbf{x},t)$ の t についての平均を0とすると

Nadaraya-Watsonモデル

$$y(\mathbf{x}) = \frac{\sum_n g(\mathbf{x}-\mathbf{x}_n)t_n}{\sum_m g(\mathbf{x}-\mathbf{x}_m)} = \sum_n k(\mathbf{x},\mathbf{x}_n)t_n \qquad (6.45)$$

$k(\mathbf{x},\mathbf{x}_n), g(\mathbf{x})$ は(6.46),(6.47)で与えられる

目次

ガウス過程の考え方を動機づけるために、
まず線形回帰について再考する

$$y(\mathbf{x}) = \mathbf{w}^{\mathrm{T}}\phi(\mathbf{x}) \qquad \textbf{(6.49)}$$

$\mathbf{w}$ の事前分布として、

$$p(\mathbf{w}) = \mathcal{N}(\mathbf{w}|0, \alpha^{-1}\mathbf{I}) \qquad \textbf{(6.50)}$$

を設定

ここで、$y(\mathbf{x})$ は $\mathbf{w}$ の1次結合であることから
$\mathcal{N}(0, \alpha^{-1}\phi(\mathbf{x})^{\mathrm{T}}\phi(\mathbf{x}))$ に従う

さらに、訓練データ集合 $\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N$ から
得られる関数 $y(\mathbf{x}_1), y(\mathbf{x}_2), ..., y(\mathbf{x}_N)$ を
ベクトル $\mathbf{y}$ とすると、これは $\mathcal{N}(0, \alpha^{-1}\Phi\Phi^{\mathrm{T}})$ に従う
確率変数ベクトル

カーネル法の手引きに則り、分散共分散行列
$\alpha^{-1}\Phi\Phi^{\mathrm{T}}$ の行列要素をカーネル関数で書く

$$(\alpha^{-1}\Phi\Phi^{\mathrm{T}})_{m,n} = k(\mathbf{x}_m, \mathbf{x}_n)$$

・このモデルは、ガウス過程の一つの例
(離散時間モデル)になっている

・カーネルは正定値であれば
データ構造などによって自由に決めてよい

例：指数カーネル

$$k(\mathbf{x}, \mathbf{x}') = \exp(-\theta|\mathbf{x} - \mathbf{x}'|) \qquad (6.56)$$

目標変数にガウスノイズが含まれる場合を考える
このノイズはデータ点ごとに独立に決まるものとする
このとき、$\mathbf{y} = (y_1, ..., y_N)^{\mathrm{T}}$ が与えられた上での
目標値 $\mathbf{t} = (t_1, ..., t_N)^{\mathrm{T}}$ の同時分布は

$$p(\mathbf{t}|\mathbf{y}) = \mathcal{N}(\mathbf{t}|\mathbf{y}, \beta^{-1}\mathbf{I}) \qquad \textbf{(6.59)}$$

また、周辺分布 $p(\mathbf{y})$ は(6.60)のガウス分布となる
よって周辺分布 $p(\mathbf{t})$ は

$$p(\mathbf{t}) = \int p(\mathbf{t}|\mathbf{y})p(\mathbf{y})d\mathbf{y} = \mathcal{N}(\mathbf{t}|0, \mathbf{C}) \quad \textbf{(6.61)}$$

ここで、分散共分散行列$\mathbf{C}$は

$$C(\mathbf{x}_m, \mathbf{x}_n) = k(\mathbf{x}_m, \mathbf{x}_n) + \beta^{-1}\delta_{mn} \quad \textbf{(6.62)}$$

を行列要素に持つ

ガウス過程による回帰に用いるカーネル関数は
(6.63)で与えられる

続いて、予測分布 $p(t_{N+1}|\mathbf{t})$ を導出する

同時分布 $p(\mathbf{t}_{N+1})$ は

$$p(\mathbf{t}_{N+1}) = \mathcal{N}(\mathbf{t}_{N+1}|\mathbf{0}, C_{N+1}) \quad (6.64)$$

で与えられる

ここで、$\mathbf{t}_{N+1}$ を $\mathbf{t}_N$ と t_{N+1}に分けて考える

(2.3.1)節の結果を用いれば、

$$p(t_{N+1}|\mathbf{t}_N) = \mathcal{N}(t_{N+1}|m(\mathbf{k}, \mathbf{t}_N), \sigma^2(\mathbf{k}))$$

平均、分散は(6.66)(6.67)で与えられる

カーネル関数が正定値カーネルならば、
分散共分散行列は正定値となるため
有効なカーネルは再び正定値カーネルであることがわかる

→ガウス過程による回帰

通常のパラメトリックな回帰に必要であったパラメータ推定は
ガウス過程による回帰には不必要
ただし、カーネル関数に含まれるハイパーパラメータを
チューニングしてやる必要がある
→(対数)周辺尤度を最大化するものを選択
　(参考)3.5節

関連度自動決定(ARD)によって入力変数の数を削減できる
→予測の効率化に有用

続いて、ガウス過程による分類を見ていく

ロジスティック回帰といった確率的な分類アプローチでは
入力変数に対して目標変数の事後分布が与えられ、
それによって分類していた
ガウス過程でも同様のアプローチをする
ただし、ガウス過程の出力は0～1とは限らないので
出力層に活性化関数を用いる

→ガウス過程による分類