

(83342) パターン認識特論

張 潮

zhang@u-fukui.ac.jp

0776-27-8477



本日の内容

- 第1回 パターン認識のための線形代数
- 第2回 パターン認識のための確率論
- 第3回 凸最適化の概念
- 第4回 線形回帰とLMSアルゴリズム
- 第5回 分類とロジスティック回帰
- 第6回 一般化線形モデル
- 第7回 ガウシアン判別分析と単純ベイズ分類器
- 第8回 サポートベクターマシン(1)
- 第9回 サポートベクターマシン(2)
- 第10回 正則化とモデル選択(ハイパーパラメータのチューニング)
- 第11回 k 平均法 **ガウス過程**
- 第12回 混合ガウスモデルとEMアルゴリズム
- 第13回 因子分析
- 第14回 主成分分析
- 第15回 未定

K-means

- まずは、10分くらいで添付pdf資料を理解してみてください

最終レポート & 15回目の予告について

● 最終レポート

■ 14回目終了後回収する

■ A4一面

■ フォントサイズ18

■ フォント:MSゴシック

■ 学籍番号・氏名を記入

■ 表・グラフでまとめる

■ 任意のデータ+任意のパターン認識アルゴリズム

◆ 目的

◆ 使用したアルゴリズム

◆ 結果と考察

● プレゼンテーション(15回目)

■ 30人×2分

■ 張のほうで回収したレポートをスキャンし映す(レポートを使ってプレゼンする)

目的
テキスト
テキスト
使用したアルゴリズム
テキスト
結果と考察

テキスト
テキスト
テキスト

ベイズ方法 (Bayesian methods)

- 今まで紹介した学習アルゴリズムの考え方
 - 学習データセット (母集団 (情報源) 分布からサンプリングした標本セット)
 - 学習モデルの**パラメータ最適化**を凸最適化問題に落とし込む
 - 最適化されたモデルでテストデータを予測
- ベイズ方法
 - **最適なパラメータを推定しない**
 - 与えられた学習モデルで**事後確率分布**を作る
 - さらに**予測分布**を計算しテストデータを予測

ベイズ方法 (Bayesian methods)

- 重要な概念

- ベイズ方法における学習
- 真の確率密度関数 (True probability density function)
- 事前確率密度関数 (A priori probability density function, Prior)
- 事後確率密度関数 (A posteriori probability density function)
- 予測分布 (Predictive distribution)

ベイズ方法における学習の世界観

- 情報源が確率分布で表されている
 - 情報源を表す確率密度関数: $q(x) = q(y=? | \text{特徴量 } x^*)$
 - ◆ 真の確率密度関数 (True probability density function)
 - 学習データセット: $D = \{ x_1, x_2, \dots, x_n \}$
 - ◆ n 個のサンプルを n 個の確立変数として見なす
 - ◆ それぞれの x_i が $q(x)$ に独立に従う (IID)

実世界では, 真の確率密度関数 $q(x)$ は分からないことが多い



$q(x)$ を推測したい

サンプルから情報源を推測することを 学習 (Learning) あるいは 統計的推測 (Statistical estimation) という

ベイズ方法における学習の世界観

- パラメータ w によって定まる x の確率密度関数 $p(x|w)$
 - 確率モデル, 統計モデル, 学習モデル (Probabilistic model, Statistical model, Learning machine)
- w の集合上の確率密度関数 $\varphi(w)$
 - 事前確率密度関数 (A priori probability density function, Prior)
- $p(w|D) = (1/Z) p(x_1|w) p(x_2|w) \cdots p(x_n|w) \varphi(w)$ **ベイズの定理**
 - 事後確率密度関数 (A posteriori probability density function)
 - $Z = \int p(x_1|w) p(x_2|w) \cdots p(x_n|w) \varphi(w) dw$
 - ◆ 周辺尤度 (Marginal likelihood) あるいは 証拠 (エビデンス) (Evidence)
- $p(x|D) = \int p(x|w) p(w|D) dw$
 - 予測分布 (Predictive distribution)

「予測分布 $p(x|D)$ は, きっと, 真の密度関数 $q(x)$ に近いであろう」

ベイズ方法における学習の世界観

● 疑問

- 真の確率密度関数 $q(x)$ が分からないのに、なぜ学習モデル $p(x|w)$ と事前確率密度関数 $\varphi(w)$ を勝手に作るの？
- $p(x|w)$ を作る際には、 $q(x)$ と関係なくていい
- $\varphi(w)$ も、 $q(x)$ と関係なくていい
- 「 $p(x|w)$ と $\varphi(w)$ 」のペアが作られたと思えばいい
- 事後確率密度関数 $p(w|D)$ で両者の関連性が評価される
- $p(w|D)$ が分かれば、予測分布 $p(x|D)$ が分かる

$p(x|w)$ と $\varphi(w)$ を設計

↓

ベイズ定理で $p(w|D)$ を計算する

↓

$p(x|D)$ で予測する

ベイズ方法における学習例

● 多変量正規分布を用いた線形回帰分析

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

$$x \sim \mathcal{N}(\mu, \Sigma)$$

$p(y|\theta)$ と $\varphi(\theta)$ を設計

$$P(y^{(i)} | x^{(i)}, \theta) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y^{(i)} - \theta^T x^{(i)})^2}{2\sigma^2}\right).$$

$$\theta \sim \mathcal{N}(0, \tau^2 I)$$

ベイズ定理で $p(\theta|S)$ を計算する

$$p(\theta | S) = \frac{p(\theta)p(S | \theta)}{\int_{\theta'} p(\theta')p(S | \theta')d\theta'} = \frac{p(\theta) \prod_{i=1}^m p(y^{(i)} | x^{(i)}, \theta)}{\int_{\theta'} p(\theta') \prod_{i=1}^m p(y^{(i)} | x^{(i)}, \theta')d\theta'}.$$

$p(y|S)$ で予測する

$$p(y_* | x_*, S) = \int_{\theta} p(y_* | x_*, \theta)p(\theta | S)d\theta.$$

ベイズ方法における学習例

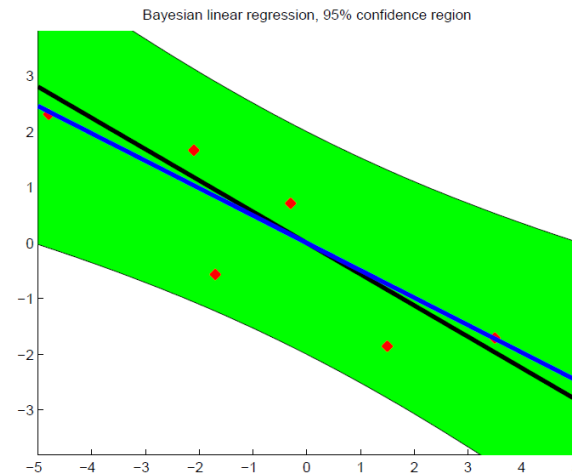


Figure 1: Bayesian linear regression for a one-dimensional linear regression problem, $y^{(i)} = \theta x^{(i)} + \epsilon^{(i)}$, with $\epsilon^{(i)} \sim \mathcal{N}(0,1)$ i.i.d. noise. The green region denotes the 95% confidence region for predictions of the model. Note that the (vertical) width of the green region is largest at the ends but narrowest in the middle. This region reflects the uncertainty in the estimates for the parameter θ . In contrast, a classical linear regression model would display a confidence region of constant width, reflecting only the $\mathcal{N}(0, \sigma^2)$ noise in the outputs.

ガウス過程

- 有限集合 $\mathcal{X} = \{x_1, \dots, x_m\}$
- $h(\cdot) \in \mathcal{H}$ を使って \mathcal{X} を実数空間へ写像することができる
- 例

$$h_0(x_1) = 5, \quad h_0(x_2) = 2.3, \quad h_0(x_2) = -7, \quad \dots, \quad h_0(x_{m-1}) = -\pi, \quad h_0(x_m) = 8.$$

- m次元のベクトルを使って任意のhを表現することができる

$$\vec{h} = [h(x_1) \quad h(x_2) \quad \dots \quad h(x_m)]^T$$

- hのベクトル表現で関数における確率分布が定義される

ガウス過程

- h のベクトル表現で関数における確率分布

$$p(h) = \prod_{i=1}^m \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(h(x_i) - \mu_i)^2\right).$$

$$\vec{h} \sim \mathcal{N}(\vec{\mu}, \sigma^2 I)$$

- 多変量正規分布で h 関数の確率分布を表せる

ガウス過程

● 確率過程

- stochastic process

h自体が変わるので, 要注意

- 時間とともに変化する確率変数 (株価や為替の変動)

$$\{h(x) : x \in \mathcal{X}\}$$

● ガウス過程

- Gaussian process

- 多変量正規分布に従う性質をもつ確率過程

$$\begin{bmatrix} h(x_1) \\ \vdots \\ h(x_m) \end{bmatrix} \sim \mathcal{N} \left(\begin{bmatrix} m(x_1) \\ \vdots \\ m(x_m) \end{bmatrix}, \begin{bmatrix} k(x_1, x_1) & \cdots & k(x_1, x_m) \\ \vdots & \ddots & \vdots \\ k(x_m, x_1) & \cdots & k(x_m, x_m) \end{bmatrix} \right).$$

$$h(\cdot) \sim \mathcal{GP}(m(\cdot), k(\cdot, \cdot)).$$

ガウス過程

$$\begin{bmatrix} h(x_1) \\ \vdots \\ h(x_m) \end{bmatrix} \sim \mathcal{N} \left(\begin{bmatrix} m(x_1) \\ \vdots \\ m(x_m) \end{bmatrix}, \begin{bmatrix} k(x_1, x_1) & \cdots & k(x_1, x_m) \\ \vdots & \ddots & \vdots \\ k(x_m, x_1) & \cdots & k(x_m, x_m) \end{bmatrix} \right)$$

hのベクトル表現が多変量正規分布に従っている

共分散関数（半
正定値行列が前
提条件）

平均関数（任意の実数関数でいい）

$$\begin{aligned} m(x) &= E[x] \\ k(x, x') &= E[(x - m(x))(x' - m(x'))]. \end{aligned}$$

関数mと関数kが決まれば，関数hの分布が決まる
→ $h(x_1) \dots h(x_m)$ の値が分かる（サンプリングすることができる）

ガウス過程

- 例: ゼロ平均ガウス過程

- 平均関数を0に

$$h(\cdot) \sim \mathcal{GP}(0, k(\cdot, \cdot)).$$

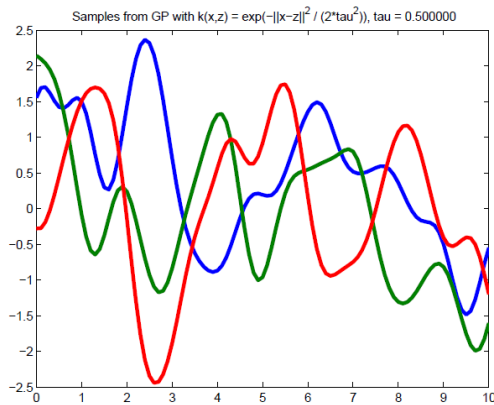
- 共分散行列をガウスカーネルにする (SVMの話, Mercerの定理)

$$k_{SE}(x, x') = \exp\left(-\frac{1}{2\tau^2}\|x - x'\|^2\right)$$

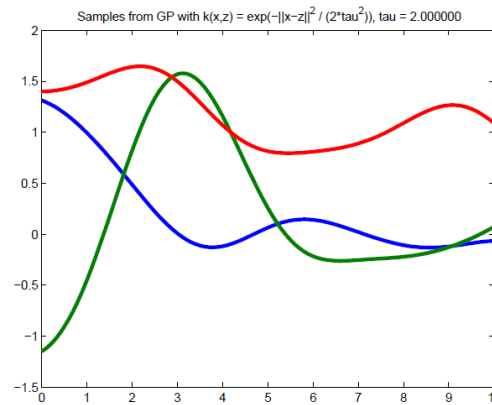
- 特性

- $h(x)$ and $h(x')$ will tend to have high covariance if x and x' are “nearby” in the input space (i.e., $\|x - x'\| = \|x - x'\| \approx 0$, so $\exp(-\frac{1}{2\tau^2}\|x - x'\|^2) \approx 1$).
 - $h(x)$ and $h(x')$ will tend to have low covariance when x and x' are “far apart” (i.e., $\|x - x'\| \gg 0$, so $\exp(-\frac{1}{2\tau^2}\|x - x'\|^2) \approx 0$).

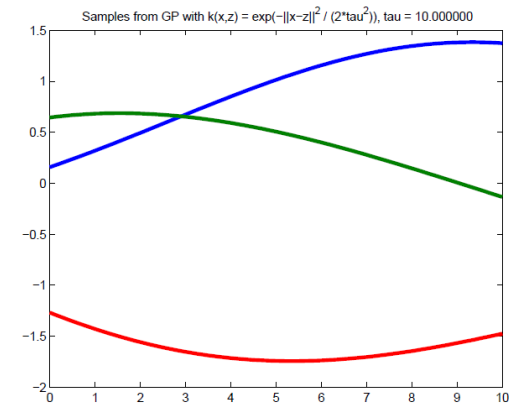
ガウス過程



(a)



(b)



(c)

Figure 2: Samples from a zero-mean Gaussian process prior with $k_{SE}(\cdot, \cdot)$ covariance function, using (a) $\tau = 0.5$, (b) $\tau = 2$, and (c) $\tau = 10$. Note that as the bandwidth parameter τ increases, then points which are farther away will have higher correlations than before, and hence the sampled functions tend to be smoother overall.

<http://www.tmpl.fi/gp/>

ガウス過程を用いた回帰モデル

- 線形回帰モデルとはちょっと違って...

$$y^{(i)} = h(x^{(i)}) + \varepsilon^{(i)}, \quad i = 1, \dots, m$$

事前分布

$$h(\cdot) \sim \mathcal{GP}(0, k(\cdot, \cdot))$$

$k(\cdot, \cdot)$ カーネルはユーザが決める

予測分布を求めたい



事後確率を求める

ガウス過程を用いた回帰モデル

- より簡単な求め方がある

~~予測分布を求めたい~~



~~事後確率を求める~~

- 多変量正規関数の周辺分布の特性

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

周辺分布

$$p(x_A) = \int_{x_B} p(x_A, x_B; \mu, \Sigma) dx_B$$

$$p(x_B) = \int_{x_A} p(x_A, x_B; \mu, \Sigma) dx_A$$

$$x_A \sim \mathcal{N}(\mu_A, \Sigma_{AA})$$

$$x_B \sim \mathcal{N}(\mu_B, \Sigma_{BB}).$$

ガウス過程を用いた回帰モデル

- 多変量正規関数の条件付き分布の特性

$$p(x_A | x_B) = \frac{p(x_A, x_B; \mu, \Sigma)}{\int_{x_A} p(x_A, x_B; \mu, \Sigma) dx_A}$$
$$p(x_B | x_A) = \frac{p(x_A, x_B; \mu, \Sigma)}{\int_{x_B} p(x_A, x_B; \mu, \Sigma) dx_B}$$

$$x_A | x_B \sim \mathcal{N}(\mu_A + \Sigma_{AB} \Sigma_{BB}^{-1} (x_B - \mu_B), \Sigma_{AA} - \Sigma_{AB} \Sigma_{BB}^{-1} \Sigma_{BA})$$
$$x_B | x_A \sim \mathcal{N}(\mu_B + \Sigma_{BA} \Sigma_{AA}^{-1} (x_A - \mu_A), \Sigma_{BB} - \Sigma_{BA} \Sigma_{AA}^{-1} \Sigma_{AB}).$$

ガウス過程を用いた回帰モデル

●テストセット

$$T = \{(x_*^{(i)}, y_*^{(i)})\}_{i=1}^{m_*}$$

学習セット

$$S = \{(x^{(i)}, y^{(i)})\}_{i=1}^m$$

$$\begin{aligned}
 X &= \begin{bmatrix} - & (x^{(1)})^T & - \\ - & (x^{(2)})^T & - \\ & \vdots & \\ - & (x^{(m)})^T & - \end{bmatrix} \in \mathbf{R}^{m \times n} & \vec{h} &= \begin{bmatrix} h(x^{(1)}) \\ h(x^{(2)}) \\ \vdots \\ h(x^{(m)}) \end{bmatrix}, & \vec{\varepsilon} &= \begin{bmatrix} \varepsilon^{(1)} \\ \varepsilon^{(2)} \\ \vdots \\ \varepsilon^{(m)} \end{bmatrix}, & \vec{y} &= \begin{bmatrix} y^{(1)} \\ y^{(2)} \\ \vdots \\ y^{(m)} \end{bmatrix} \in \mathbf{R}^m, \\
 X_* &= \begin{bmatrix} - & (x_*^{(1)})^T & - \\ - & (x_*^{(2)})^T & - \\ & \vdots & \\ - & (x_*^{(m_*)})^T & - \end{bmatrix} \in \mathbf{R}^{m_* \times n} & \vec{h}_* &= \begin{bmatrix} h(x_*^{(1)}) \\ h(x_*^{(2)}) \\ \vdots \\ h(x_*^{(m_*)}) \end{bmatrix}, & \vec{\varepsilon}_* &= \begin{bmatrix} \varepsilon_*^{(1)} \\ \varepsilon_*^{(2)} \\ \vdots \\ \varepsilon_*^{(m)} \end{bmatrix}, & \vec{y}_* &= \begin{bmatrix} y_*^{(1)} \\ y_*^{(2)} \\ \vdots \\ y_*^{(m_*)} \end{bmatrix} \in \mathbf{R}^{m_*}.
 \end{aligned}$$

ガウス過程を用いた回帰モデル

- h は 0 平均ガウス過程からサンプリングした関数
- h における周辺分布はガウス分布である

$$\begin{bmatrix} \vec{h} \\ \vec{h}_* \end{bmatrix} \Big| X, X_* \sim \mathcal{N} \left(\vec{0}, \begin{bmatrix} K(X, X) & K(X, X_*) \\ K(X_*, X) & K(X_*, X_*) \end{bmatrix} \right)$$

$$\vec{h} \in \mathbf{R}^m \text{ such that } \vec{h} = [h(x^{(1)}) \ \dots \ h(x^{(m)})]^T$$

$$\vec{h}_* \in \mathbf{R}^{m_*} \text{ such that } \vec{h}_* = [h(x_*^{(1)}) \ \dots \ h(x_*^{(m_*)})]^T$$

$$K(X, X) \in \mathbf{R}^{m \times m} \text{ such that } (K(X, X))_{ij} = k(x^{(i)}, x^{(j)})$$

$$K(X, X_*) \in \mathbf{R}^{m \times m_*} \text{ such that } (K(X, X_*))_{ij} = k(x^{(i)}, x_*^{(j)})$$

$$K(X_*, X) \in \mathbf{R}^{m_* \times m} \text{ such that } (K(X_*, X))_{ij} = k(x_*^{(i)}, x^{(j)})$$

$$K(X_*, X_*) \in \mathbf{R}^{m_* \times m_*} \text{ such that } (K(X_*, X_*))_{ij} = k(x_*^{(i)}, x_*^{(j)})$$

ガウス過程を用いた回帰モデル

- ノイズ項

$$\begin{bmatrix} \vec{\varepsilon} \\ \vec{\varepsilon}_* \end{bmatrix} \sim \mathcal{N}\left(\vec{0}, \begin{bmatrix} \sigma^2 I & \vec{0} \\ \vec{0}^T & \sigma^2 I \end{bmatrix}\right)$$

$$\begin{bmatrix} \vec{y} \\ \vec{y}_* \end{bmatrix} \Big| X, X_* = \begin{bmatrix} \vec{h} \\ \vec{h}_* \end{bmatrix} + \begin{bmatrix} \vec{\varepsilon} \\ \vec{\varepsilon}_* \end{bmatrix} \sim \mathcal{N}\left(\vec{0}, \begin{bmatrix} K(X, X) + \sigma^2 I & K(X, X_*) \\ K(X_*, X) & K(X_*, X_*) + \sigma^2 I \end{bmatrix}\right)$$

ガウス過程を用いた回帰モデル

- 条件付き分布

$$\vec{y}_* | \vec{y}, X, X_* \sim \mathcal{N}(\mu^*, \Sigma^*)$$

$$\mu^* = K(X_*, X)(K(X, X) + \sigma^2 I)^{-1} \vec{y}$$

$$\Sigma^* = K(X_*, X_*) + \sigma^2 I - K(X_*, X)(K(X, X) + \sigma^2 I)^{-1} K(X, X_*).$$

ガウス過程を用いた回帰モデル

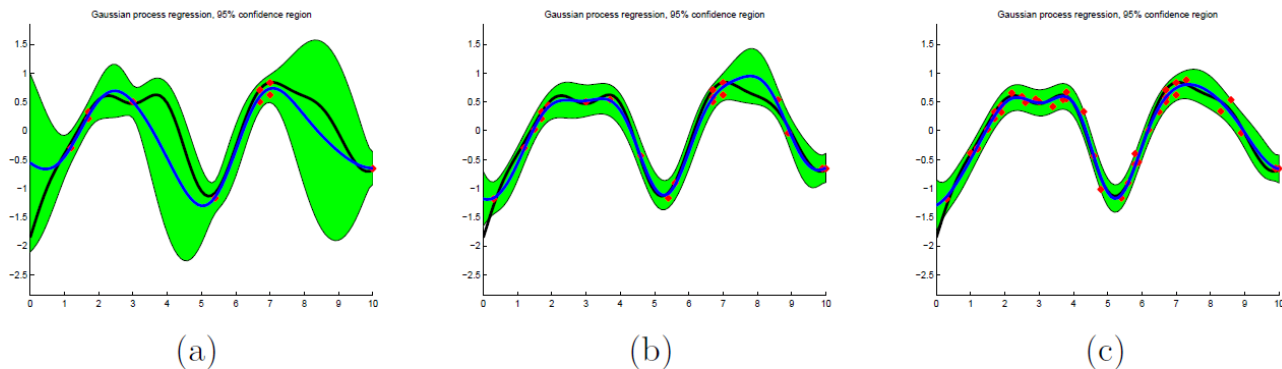


Figure 3: Gaussian process regression using a zero-mean Gaussian process prior with $k_{SE}(\cdot, \cdot)$ covariance function (where $\tau = 0.1$), with noise level $\sigma = 1$, and (a) $m = 10$, (b) $m = 20$, and (c) $m = 40$ training examples. The blue line denotes the mean of the posterior predictive distribution, and the green shaded region denotes the 95% confidence region based on the model's variance estimates. As the number of training examples increases, the size of the confidence region shrinks to reflect the diminishing uncertainty in the model estimates. Note also that in panel (a), the 95% confidence region shrinks near training points but is much larger far away from training points, as one would expect.

本日の内容

- 第1回 パターン認識のための線形代数
- 第2回 パターン認識のための確率論
- 第3回 凸最適化の概念
- 第4回 線形回帰とLMSアルゴリズム
- 第5回 分類とロジスティック回帰
- 第6回 一般化線形モデル
- 第7回 ガウシアン判別分析と単純ベイズ分類器
- 第8回 サポートベクターマシン(1)
- 第9回 サポートベクターマシン(2)
- 第10回 正則化とモデル選択(ハイパーパラメータのチューニング)
- 第11回 k 平均法ガウス過程
- 第12回 混合ガウスモデルとEMアルゴリズム
- 第13回 因子分析
- 第14回 主成分分析
- 第15回 未定