

3.5 Representer 定理

3.6 グラム行列の低ランク近似

茨城大学工学部情報工学科

08T4018Y 小幡智裕

3.5 Representer定理

定理

$$\min_{f \in H, c \in R^m} L \left(\{X_i\}_{i=1}^N, \{Y_i\}_{i=1}^N, \left\{ f(X_i) + \sum_{a=1}^m c_a h_a(X_i) \right\}_{i=1}^N \right) + \Psi(\|f\|).$$

$h_1(x), \dots, h_m(x)$: 固定された X 上の基底関数

$\Psi: [0, \infty) \rightarrow R$: 狭義の単調増加関数

の仮定のもと、この最小値を達成する f は

$$f = \sum_{i=1}^N \alpha_i k(\cdot, X_i)$$

の形をもつ。

3.5 Representer定理の証明

証明

$H_0 = \text{Span}\{k(\cdot, X_1), \dots, k(\cdot, X_N)\}$ をデータの張る有限次元部分空間とし、その直交補空間を H_0^\perp とおくと、直交分解

$$H = H_0 \oplus H_0^\perp$$

が成り立つ。仮定の最適化問題を達成する H 上の元を f とし、上の直交分解に従って

$$f = f_0 + f^\perp, \quad f_0 \in H_0, f^\perp \in H_0^\perp$$

と表すと $\langle f^\perp, k(\cdot, X_i) \rangle = 0$ より $f(X_i) = \langle f, k(\cdot, X_i) \rangle = \langle f_0, k(\cdot, X_i) \rangle = f_0(X_i)$ である。したがって f を f_0 に変えても L の値は変わらない。

3.5 Representer定理

証明(続き)

一方、直交性から $\|f\|^2 = \|f_0\|^2 + \|f^\perp\|^2$ であるので、 Ψ の仮定から $\Psi(\|f_0\|) \leq \Psi(\|f\|)$ である。以上により、

$$\begin{aligned} & L\left(\{X_i\}_{i=1}^N, \{Y_i\}_{i=1}^N, \left\{f_0(X_i) + \sum_{a=1}^m c_a h_a(X_i)\right\}_{i=1}^N\right) + \Psi(\|f_0\|) \\ & \leq L\left(\{X_i\}_{i=1}^N, \{Y_i\}_{i=1}^N, \left\{f(X_i) + \sum_{a=1}^m c_a h_a(X_i)\right\}_{i=1}^N\right) + \Psi(\|f\|) \end{aligned}$$

となり、 f_0 が最小値を達成する。

3.6 グラム行列の低ランク近似

- ・カーネル法ではデータ数 n のサイズをもつグラム行列を用いた計算が基本となるため、 n が大きい場合には計算量の問題が生じる。

⇒グラム行列を低ランクの行列で近似して計算量を減らす。

Nystrom近似、不完全Cholesky分解

Nystrom近似

- もともと積分作用素の固有関数と固有値を近似する方法。

正定値カーネル k で定まる積分作用素の固有値問題

$$\int k(y, x)\phi(x)dP(x) = \lambda\phi(y), \quad \int \phi(x)^2 dP(x) = 1$$

を考える。固有値を $\lambda_1 \geq \lambda_2 \geq \dots \geq 0$ 対応する固有ベクトルを ϕ_i とすると、上の問題を

$$\frac{1}{m} \sum_{i=1}^m k(y, X_i)\phi(X_i) \approx \lambda\phi(y), \quad \frac{1}{m} \sum_{i=1}^m \phi(X_i)^2 \approx 1$$

と近似できる。 X_1, \dots, X_m は P に従うサンプル。ここで $y = X_j$ とすると

$$\frac{1}{m} \sum_{i=1}^m K_{ij}^{(m)} \phi(X_i) \approx \lambda\phi(X_j), \quad (K_{ij}^{(m)} = k(X_i, X_j))$$

となる。

Nystrom近似(続き)

$K^{(m)}$ の固有値分解を

$$K^{(m)} = U^{(m)} \Lambda^{(m)} U^{(m)T}$$

($\Lambda^{(m)}$ は $\lambda_1^{(m)} \geq \dots \geq \lambda_m^{(m)} \geq 0$ を成分に持つ対角行列, $U^{(m)} = (u_1^{(m)}, \dots, u_m^{(m)})$ は単位固有ベクトルをなす行列)とすると、

$$\phi_i(y) \approx \sqrt{m} u_{ji}^{(m)}, \quad \lambda_i \approx \frac{\lambda_i^{(m)}}{m}$$

という近似が得られ、任意の y に対して

$$\phi_i(y) \approx \frac{\sqrt{m}}{\lambda_i^{(m)}} \sum_{j=1}^m k(y, X_j) u_{j,i}^{(m)}$$

を得る。

Nystrom近似(続き)

データ X_1, \dots, X_n に対し、積分作用素の固有値問題の式で $P = \frac{1}{n} \sum_{i=1}^n \delta_{X_i}$ とすると積分作用素はグラム行列で表わされ、 X_1, \dots, X_n から一様にサンプルされた部分データによって K を近似することになる。近似に用いる部分データを X_1, \dots, X_m ($m \leq n$) とし、部分データに対する近似を合わせると

$$u_{j,i}^{(n)} \approx \sqrt{\frac{m}{n}} \frac{1}{\lambda_i^{(m)}} \sum_{a=1}^m K_{j,a} u_{a,i}^{(m)} \equiv \tilde{u}_{j,i}^{(n)}, \quad \lambda_i^{(n)} \approx \frac{n}{m} \lambda_i^{(m)} \equiv \tilde{\lambda}_i^{(n)}$$

を得る。このとき K の近似が $K \approx \sum_{i=1}^m \tilde{\lambda}_i^{(n)} \tilde{u}_i^{(n)} \tilde{u}_i^{(n)T}$ により得られるが、さらに大きい p 個 ($p \leq m$) の固有値だけを用いれば

$$K \approx \sum_{i=1}^p \tilde{\lambda}_i^{(n)} \tilde{u}_i^{(n)} \tilde{u}_i^{(n)T} \equiv \tilde{K}_p$$

という近似を得る。 \tilde{K}_p をグラム行列 K の Nystrom 近似という。

不完全Cholesky分解

- Cholesky分解

$n \times n$ 半正定値行列AのCholesky分解は、非負の対角成分をもつ下半三角行列Rによる $A = RR^T$ という分解である。Aのランクがrのとき、ある置換行列Pがあつて、

$$PAP = RR^T, \quad R = \begin{pmatrix} R_{11} & 0 \\ R_{21} & 0 \end{pmatrix} \quad (R_{11} \text{は対角成分が正の} r \times r \text{下半三角行列})$$

と分解することができる。

不完全Cholesky分解(続き)

Cholesky分解はガウスの掃き出し法の別表現である。Aが非特異の場合を考え、

$$A = \begin{pmatrix} \alpha & b^T \\ b & K \end{pmatrix} \quad (\alpha \in R, b = R^{n-1})$$

とすると、ガウスの掃き出し法の第一段階は

$$A = \begin{pmatrix} \sqrt{\alpha} & 0 \\ b/\sqrt{\alpha} & I_{n-1} \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & K - \frac{bb^T}{\alpha} \end{pmatrix} \begin{pmatrix} \sqrt{\alpha} & b^T/\sqrt{\alpha} \\ 0 & I_{n-1} \end{pmatrix}$$

とみなせる。この操作を繰り返すと、 $A = RI_nR^T$ においてRは下半三角となる。

不完全Cholesky分解は、Cholesky分解を途中まで行い行列Aの低ランク近似を得る方法である。

不完全Cholesky分解(続き)

$A: n \times n$ 半正定値行列 ε_{tol} : 終了条件のしきい値

(1) 初期化: $i = 1$. R は空行列. $A' = A$. $P = I_n$. $R_{jj} = A_{jj}$ ($1 \leq j \leq n$)

(2) もし $\sum_{j=i}^n R_{jj} < \varepsilon_{tol}$ であれば終了. そうでなければ(3)へ

(3) $j^* = \arg \max_{j=i, \dots, n} R_{jj}$ とおく

(4) 置換を行う: $P_{ii=0}, P_{j^*j^*} = 0, P_{ij^*} = 1, P_{j^*i} = 1$. A' の i 行と j^* 行を置換.
 i 列と j^* 列を置換. $R_{i,1i}$ と $R_{j^*,1i}$ を置換.

(5) $R_{ii} = \sqrt{A'_{ii}}$.

(6) R の第 i 列の計算:

$$R_{i+1:n,i} = \frac{1}{R_{ii}} \left(A'_{i+1:n,i} - \sum_{k=1}^{i-1} R_{i+1:n,k} R_{ik} \right).$$

(7) 対角成分の更新: $R_{jj} = A'_{jj} - \sum_{k=1}^i R_{jk}^2$ ($i+1 \leq j \leq n$)

(8) $i = i+1$ として(2)へ戻る

$n \times (i-1)$ 行列 R と置換行列 P を返す

不完全Cholesky分解（続き）

このアルゴリズムでは掃き出し法とは異なる原理に基づいて計算を行っている。 $A = RR^T$ のとき

$$A_{ij} = \sum_{k=1}^j R_{ik} R_{jk} \quad (1 \leq j \leq i \leq n)$$

であるので

$$R_{ij} = \frac{A_{ij} - \sum_{k=1}^{j-1} R_{ik} R_{jk}}{R_{jj}} \quad (j+1 \leq i \leq n)$$

を得る。