순위인 공변량에 기반한 커널 로지스틱 회귀

서울시립대학교 통계학과

김 윤 회

Contents

연구 목적

1

연구 목적

- **공변량** (covariates)이 순위인 경우 **순열** (permutation)로 표현 가능하며 N개의 공변량에 대해 N! 개의 서로 다른 순열이 존재.

- 순열에 자료에 대한 이진 분류문제에 위해 효율적 계산이 가능한 **켄달 커널**과 **멜로우즈 커널**을 사용한 지지벡터기계 모형을 고려.

- 지지벡터기계는 로지스틱 회귀에서와 같이 분류에 대한 직접적인 **확률 추정값** (probability estimates)을 제공하지 못함.

켄달 커널과 멜로우즈 커널을 사용한 커널 로지스틱 회귀 모형을 고려

순위 자료에 대한 켄달 커널과 멜로우즈 커널

- 순위 자료와 커널
- 켄달 커널과 멜로우즈 커널

2

순위 자료와 커널

순위 자료 (Ranking data)

- 항목에 대한 선호를 표현하는 경우

  예) 선거 자료, 소비자 선호도 조사

- 값의 절대적인 크기보다 상대적 순서가 중요하다고 생각되는 경우

  예) 유전자 발현 데이터 (D. Genman et al. , 2004)

순위 자료와 커널

- 완전 순위 **(total ranking)**

$$x_{i_1} \succ x_{i_2} \succ \cdots \succ x_{i_n} \tag{2.1}$$

$$\{1, 2, \ldots, n\} =: [1, n]$$

- 순열 **(permutation)**

$$\sigma : [1, n] \to [1, n] \tag{2.2}$$

예) $x_2 \succ x_4 \succ x_3 \succ x_1$ $\quad \sigma = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 1 & 4 & 2 & 3 \end{pmatrix}$

$$\sigma(1) = 1,\ \sigma(2) = 4,\ \sigma(3) = 2,\ \sigma(4) = 3$$

- 대칭군 **(symmetric group)**

$(\sigma_1 \sigma_2)(i) = \sigma_1(\sigma_2(i))$ 이 부여된 $\mathbb{S}_n$

순위 자료와 커널

- 일치 쌍 **(concordant pair)**

$$n_c\left(\sigma,\sigma'\right)=\sum_{i<j}\left[\mathbb{1}_{\sigma(i)<\sigma(j)}\mathbb{1}_{\sigma'(i)<\sigma'(j)}+\mathbb{1}_{\sigma(i)>\sigma(j)}\mathbb{1}_{\sigma'(i)>\sigma'(j)}\right] \qquad (2.4)$$

- 불일치 쌍 **(discordant pair)**

$$n_d\left(\sigma,\sigma'\right)=\sum_{i<j}\left[\mathbb{1}_{\sigma(i)<\sigma(j)}\mathbb{1}_{\sigma'(i)>\sigma'(j)}+\mathbb{1}_{\sigma(i)>\sigma(j)}\mathbb{1}_{\sigma'(i)<\sigma'(j)}\right] \qquad (2.5)$$

켄달 커널과 멜로우즈 커널

켄달과 멜로우즈 커널 (Kendall and Mallows kernel) (Jiao and Vert, 2015)

$$\Phi : \mathbb{S}_n \to \mathbb{R}^{\binom{n}{2}}$$

$$\Phi(\sigma) = \left(\frac{1}{\sqrt{\binom{n}{2}}} \left(\mathbb{1}_{\sigma(i)>\sigma(j)} - \mathbb{1}_{\sigma(i)<\sigma(j)} \right) \right)_{1 \le i < j \le n} \tag{2.6}$$

$$\text{Kendall kernel} : K_\tau\left(\sigma, \sigma'\right) = \frac{n_c\left(\sigma, \sigma'\right) - n_d\left(\sigma, \sigma'\right)}{\binom{n}{2}} \tag{2.7}$$

(= Linear kernel in Euclidean space)

$$\text{Mallows kernel} : K_M^\nu\left(\sigma, \sigma'\right) = e^{-\nu n_d\left(\sigma, \sigma'\right)} \tag{2.8}$$

(= Gaussian kernel in Euclidean space)

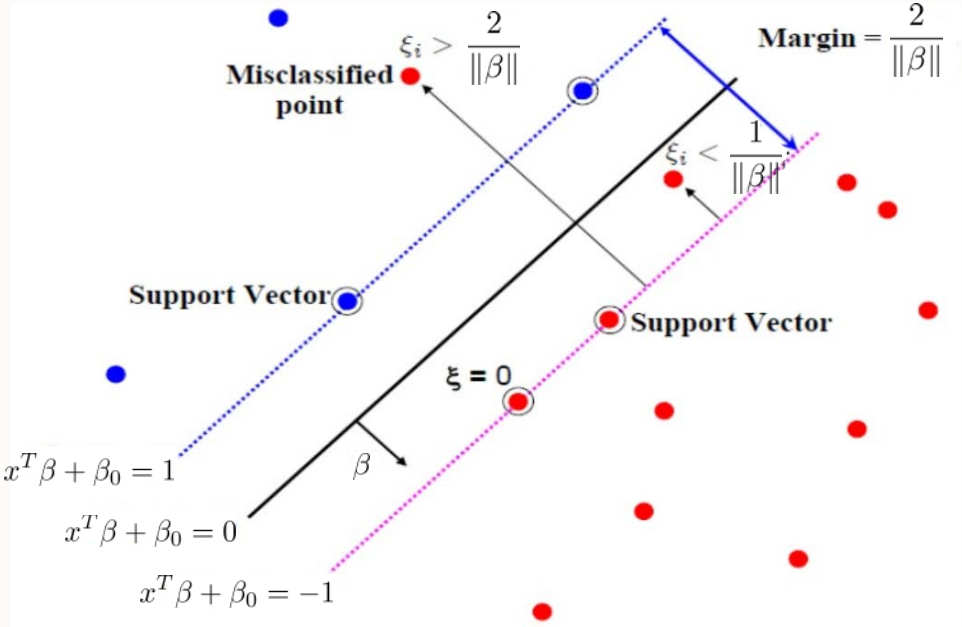
지지벡터기계와 커널 로지스틱 회귀

3

- 개요
- 정규화 함수 추정

개 요

지지벡터기계 (Support Vector Machine)



Optimization with Quadratic Programming :

$$\begin{aligned} \underset{\beta,\beta_0}{\text{minimize}} \quad & \frac{1}{2}\|\beta\|^2 + C\sum_{i=1}^{N}\xi_i \\ \text{subject to} \quad & y_i\left(x_i^T\beta + \beta_0\right) \geq 1 - \xi_i, \\ & \xi_i \geq 0, \ \forall i \end{aligned} \tag{3.1}$$

개 요

지지벡터기계 (Support Vector Machine)

- Wolfe dual problem :

$$\begin{aligned} \underset{\alpha}{\text{maximize}} \quad & \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j x_i \cdot x_j \\ \text{subject to} \quad & \alpha_i \geq 0, \\ & \sum_{i=1}^{N} \alpha_i y_i = 0, \ \forall i \end{aligned} \tag{3.2}$$

- Kernel function : $K(x_i, x_j) = \langle h(x_i), h(x_j) \rangle$ (3.3)

- Model : $f(x) = \beta_0 + \sum_{i=1}^{N} \alpha_i y_i K(x, x_i)$ (3.4)

개 요

로지스틱 회귀 (Logistic Regression)

- Model : $f(x) = \log \frac{Pr(Y=+1|x)}{Pr(Y=-1|x)} = \beta_0 + \beta^T x$,

$$Pr(Y=+1|x) = \frac{e^{f(x)}}{1+e^{f(x)}} \tag{3.5}$$

최대가능도추정 (Maximum Likelihood Estimation) with Newton-Rapshon

- Log-Likelihood : $\ell(\beta) = \sum_{i=1}^{N} \left[y_i \beta^T x_i - \log\left(1 + e^{\beta^T x_i}\right) \right]$ (3.6)

$$\beta^{\text{new}} = \beta^{\text{old}} - \left(\frac{\partial^2 \ell(\beta)}{\partial \beta \partial \beta^T} \right)^{-1} \frac{\partial \ell(\beta)}{\partial \beta}$$

정규화 함수 추정

커널을 사용하는 정규화된 함수 추정 **(regularized function estimation with kernel)**

- Objective function (in RKHS) :

$$\min_{f \in \mathcal{H}_K} \left[\sum_{i=1}^{N} L\left(y_i, f\left(x_i\right)\right) + \lambda \|f\|_{\mathcal{H}_K}^2 \right] \tag{3.7}$$

by Representer theorem (Kimeldorf & Wahba, 1971)

- Model :

$$f(x) = \sum_{i=1}^{N} \alpha_i K\left(x, x_i\right) \tag{3.8}$$

- Regularization term :

$$J(f) = \sum_{i=1}^{N} \sum_{j=1}^{N} K\left(x_i, x_j\right) \alpha_i \alpha_j \tag{3.9}$$

- Objective function (in finite dim space) :

$$\min_{\alpha} L\left(\mathbf{y}, \mathbf{K}\alpha\right) + \lambda \alpha^T \mathbf{K} \alpha \tag{3.10}$$

정규화 함수 추정

지지벡터기계의 정규화 함수 추정

- Objective function : $\min_{\beta,\beta_0} \sum_{i=1}^{N} [1 - y_i f(x_i)]_+ + \frac{\lambda}{2}\|\beta\|^2$ (3.11)

- Model : $f(x) = \beta_0 + \sum_{i=1}^{N} \alpha_i K(x, x_i)$ (3.12)

- Objective function : $\min_{\beta_0,\alpha} \sum_{i=1}^{N} (1 - y_i f(x_i))_+ + \frac{\lambda}{2}\alpha^T \mathbf{K}\alpha$ (3.13)

정규화 함수 추정

로지스틱 회귀의 정규화 함수 추정 (커널 로지스틱 회귀)

- Objective function : $\min_{\beta,\beta_0} \sum_{i=1}^{N} ln\left(1+e^{-y_i f(x_i)}\right) + \frac{\lambda}{2}\|\beta\|^2$ (3.14)

- Model : $f(x) = \log \frac{Pr(Y=+1|x)}{Pr(Y=-1|x)} = \beta_0 + \sum_{i=1}^{N} \alpha_i K(x, x_i)$ (3.15)

- Objective function : $\min_{\beta_0,\alpha} \sum_{i=1}^{N} ln\left(1+e^{-y_i f(x_i)}\right) + \frac{\lambda}{2}\alpha^T \mathbf{K}\alpha$ (3.16)

정규화 함수 추정

- **Hinge Loss**

$$[1 - y_i f(x_i)]_+$$

- **Binomial Negative Log Likelihood**

$$ln\left(1 + e^{-y_i f(x_i)}\right)$$

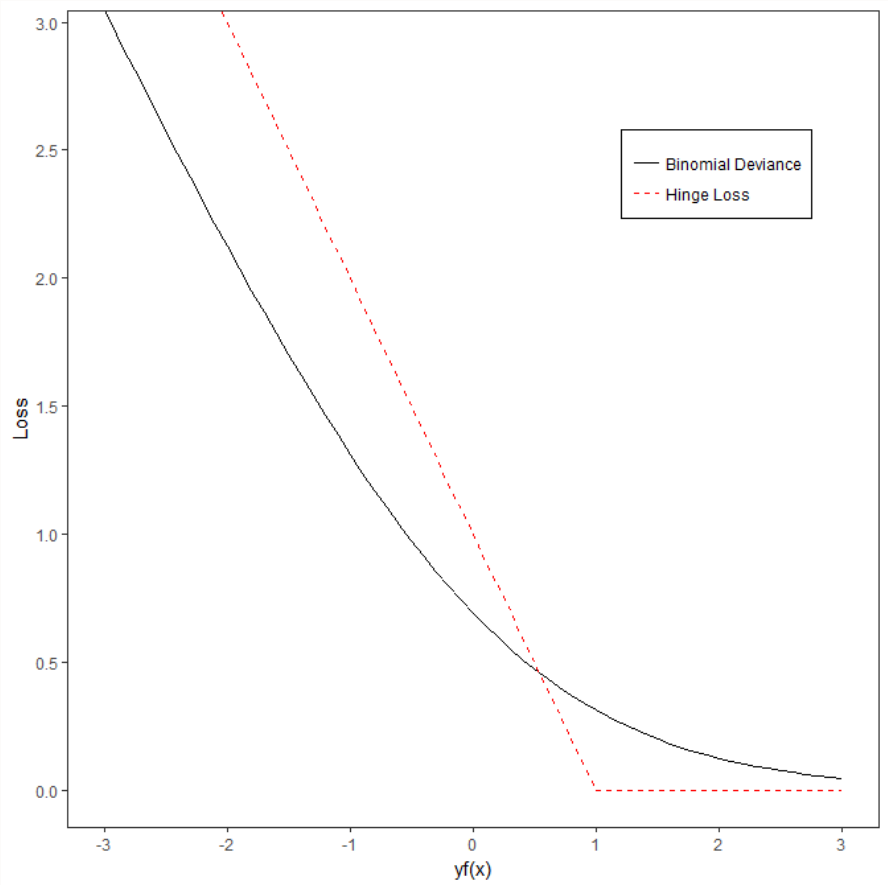


그림 1 지지벡터기계(Hinge loss)와 로지스틱 회귀 손실함수, $y \in \{-1, 1\}$

데이터 분석

- 모의 실험
- 실제 데이터

4

모의 실험

데이터 생성 방법

- μ^+ : $\underbrace{(\underbrace{D, \cdots, D}_{q\%}, \underbrace{0, \cdots, 0}_{(1-q)\%})}_{p}$ $\quad (\mu^- = -\mu^+)$

- Σ : (k,l) 원소가 $r^{|k-l|}$ 인 분산-공분산 행렬 $(\, r \in [0,1) \,)$

$n/2$ 개의 데이터 $\mathcal{X}_1 \sim N_p(\mu^+, \Sigma)$ $y=1$ 할당 $\quad$ $n/2$ 개의 데이터 $\mathcal{X}_2 \sim N_p(\mu^-, \Sigma)$ $y=-1$ 할당

- 데이터 : $\mathcal{X} = \mathcal{X}_1 + \mathcal{X}_2$

- 데이터 변환 : $\mathbb{R}^p \rightarrow \mathbb{S}_p$

모의 실험

실험 과정

$$p : 10,\ 50,\ 250 \ / \ D = 3 \ / \ r = 0 \ / \ q : 10\%,\ 30\%,\ 50\%$$

- 훈련 데이터 : $n = 100$
- 시험 데이터 : $n = 10000$
- λ & ν 튜닝 : 5-fold 교차 검증(cross validation)
- 평가 기준 : 오분류율 (표준오차)
- 반복 수 : 200 회

모의 실험

표 1 모의실험 결과 : 평균 오분류율 (표준오차)

n	p	Significant Variable	SVMkdt	KLRkdt	SVMmlws	KLRmlws
100	10	10%	0.0905 (0.0007)	0.1145 (0.0015)	**0.0884** **(0.0004)**	0.1001 (0.0022)
		30%	0.1270 (0.0010)	0.1521 (0.0019)	**0.1252** **(0.0007)**	0.1443 (0.0030)
		50%	0.1738 (0.0011)	0.2032 (0.0020)	**0.1690** **(0.0005)**	0.1870 (0.0021)
	50	10%	0.1134 (0.0007)	0.1203 (0.0007)	**0.1096** **(0.0005)**	0.1148 (0.0007)
		30%	0.1546 (0.0008)	0.1619 (0.0008)	**0.1470** **(0.0004)**	0.1539 (0.0008)
		50%	0.2054 (0.0010)	0.2139 (0.0010)	0.2085 (0.0026)	**0.2036** **(0.0010)**
	250	10%	0.1959 (0.0011)	0.1934 (0.0008)	0.1976 (0.0023)	**0.1909** **(0.0009)**
		30%	0.2429 (0.0014)	0.2382 (0.0011)	0.2443 (0.0020)	**0.2359** **(0.0012)**
		50%	0.3065 (0.0012)	0.2988 (0.0011)	0.3122 (0.0026)	**0.2979** **(0.0012)**

모의 실험

$n = 10$ $n = 50$ $n = 250$

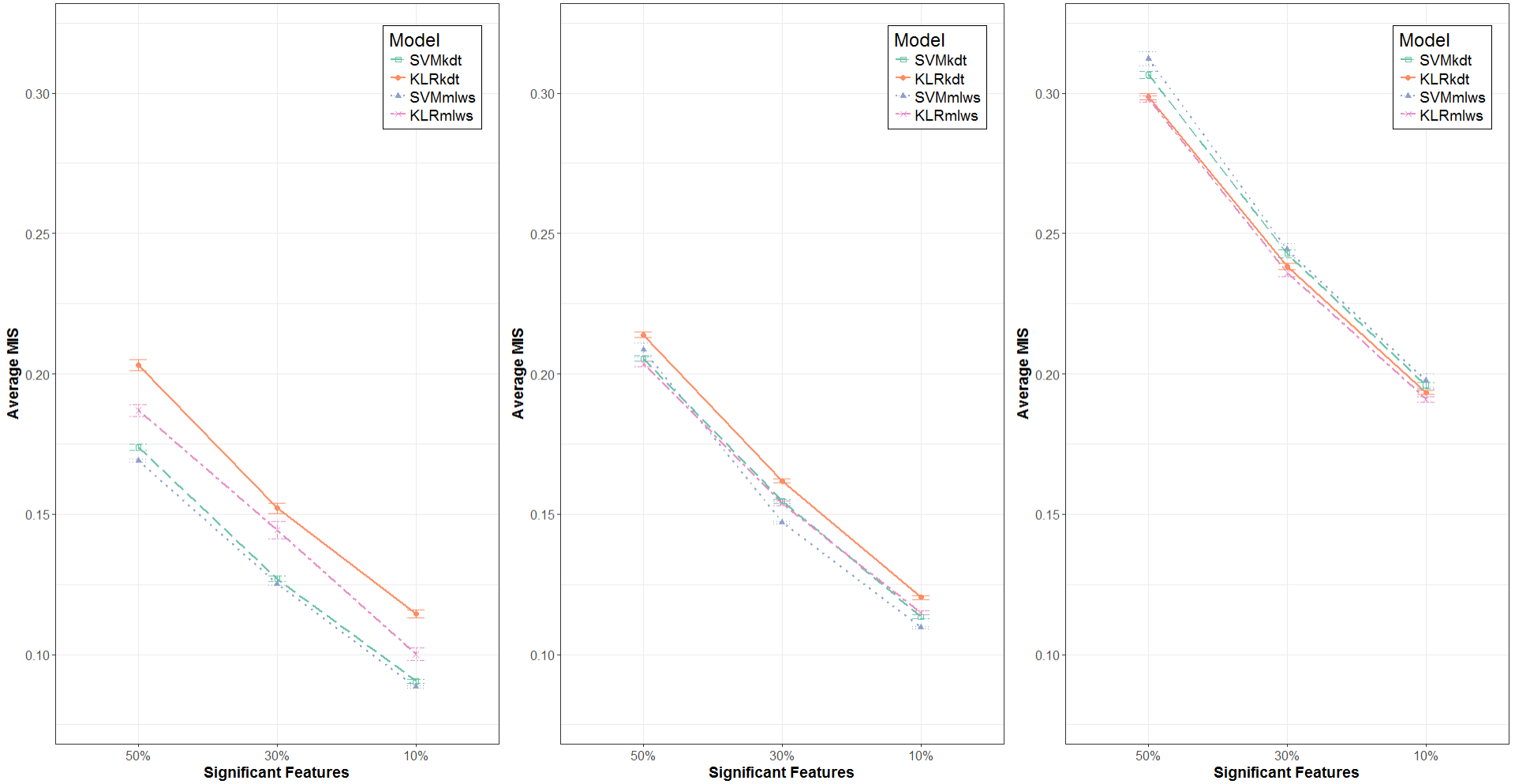


그림 2 모의실험 모형 성능 비교

실제 데이터

Eurobarometer 55.2 (2012)

No. of samples : 12527

Respondent	Gender	Age	Ranking of news sources
1	F	32	TV > Radio > School/University > Newspapers/Mags. > Web > Sci. Mags.
2	F	84	TV > Radio > Newspapers/Mags. > School/University > Sci. Mags. > Web
3	F	65	TV > Newspapers/Mags. > Sci. Mags. > Radio > School/University > Web
4	M	29	Web > Radio > Newspapers/Mags. > TV > Sci. Mags. > School/University

$$Age \geq 40, \quad y = 1 \quad / \quad Age < 40, \quad y = -1$$

- 훈련 데이터 : $n = 500$
- 시험 데이터 : $n = 12027$
- 평가 기준 : 오분류율 (표준오차)
- 반복 수 : 200 회

실제 데이터

- 유전자 발현 (gene expression) 데이터

Dataset	No. of features	No. of samples		Reference
		Class 1	Class 2	
결장 종양 (CT)	2000	40 (종양)	22 (정상)	(Alon et al., 2005)
폐 암 (LC)	12533	150 (암)	31 (정상)	(Gordon et al., 2005)
수모세포종 (MB)	7129	39 (실패)	21 (생존)	(Pomeroy et al., 2005)
전립선 암 (PC)	15154	77 (종양)	59 (정상)	(Singh et al., 2005)
난소 암 (OC)	12600	162 (암)	91 (정상)	(Petricoin et al., 2005)

- 훈련 & 시험 데이터 : 7 : 3
- 반복 수 : 200 회
- 평가 기준 : 오분류율 (표준오차)

실제 데이터

표 2 Eurobarometer & 유전자 발현 데이터 결과: 평균 오분류율 (표준오차)

Data	n	p	SVMkdt	KLRkdt	SVMmlws	KLRmlws
EURO	12527	10	0.3747 (0.0012)	0.3715 (0.0006)	0.3722 (0.0012)	**0.3664** **(0.0011)**
CT	62	2000	0.16 (0.0056)	**0.1264** **(0.0045)**	0.2378 (0.0097)	0.1453 (0.0052)
LC	181	12533	0.0059 (0.0007)	**0.0054** **(0.0006)**	0.124 (0.0119)	0.0064 (0.0006)
MB	60	7129	0.3615 (0.0029)	0.365 (0.0039)	**0.3588** **(0.0029)**	0.3638 (0.0031)
OC	253	15154	**0.0065** **(0.0008)**	0.0069 (0.0009)	0.36 (0)	0.64 (0)
PC	136	12600	**0.0962** **(0.0035)**	0.0988 (0.0033)	0.5 (0)	0.5 (0)

결 론

4

결 론

- 자료수에 비해 차원이 작은 경우 멜로우즈 커널을, 자료수에 비해 차원이 큰 경우 켄달 커널을 사용한 모형이 좋은 성능을 보임.

- 데이터에 따라서 차이가 있으나 전반적으로 지지벡터기계와 커널 로지스틱 회귀 모형의 성능은 비슷함.

- 커널 로지스틱 회귀 모형의 경우 지지벡터기계에 비해 계산 시간이 오래 걸리므로 근사적 방법을 고려해 볼 수 있음 (J. Zhu and T. Hastie, Import Vector Machine , 2005).

감사합니다