Foundation of Machine Learning

정재헌, 우지수 / 23.03.09

Computational Data Science LAB

CONTENTS

- Kernel : 관측치 간의 유사도를 의미
- Kernel Method : 비선형 문제를 해결하기 위해 관측치들을 고차원에 매핑하여 유사도 계산
- **Kernel Trick** : 실제 데이터를 고차원에 매핑시키지 않아도 저차원에서 유사도 계산 가능
  - ✓ 이유 = Hilbert Space, 정사영

01 | Representer Theorem

General Objective Function for Linear Hypothesis Space

- Featurized SVM 목적함수
  - ✓ $\min_{w \in R^d} \frac{1}{2} \| w \|^2 + \frac{c}{n} \sum_{i=1}^{n} \max(0, 1 - y_i[< w, \psi(x_i) >])$

w와 $\psi(x_1), \dots, \psi(x_n)$은 H공간에서 벡터로 나타낼 수 있음

- 일반화된 목적함수
  - ✓ $\min_{w \in H} R(\| w \|) + L(< w, \psi(x_1) >, \dots, < w, \psi(x_n) >)$

Regularization term   Loss term

Representer Theorem는 커널기반 모델의 해가 선형 조합으로 이루어진다는 이론

- Let
  - ✓ $J(w) = R(\| w \|) + L(< w, \psi(x_1) >, \dots, < w, \psi(x_n) >)$

- 이 식을 최소화 하는 w^*
  - ✓ $w^* = \sum_{i=0}^{n} \alpha_i \psi(x_i)$

01 Representer Theorem

Kernelized Regularization

- Kernelized Predictions Ex) $f_{new} = \text{sgn}(\hat{w}^T x_{new} + \hat{b})$
  - ✓ $f(x) = < w, \psi(x) > = < \sum_{i=1}^{n} \alpha_i \psi(x_i), \psi(x) >$
  - ✓ $= \sum_{i=1}^{n} \alpha_i < \psi(x_i), \psi(x)\rangle$
  - ✓ $= \sum_{i=1}^{n} \alpha_i \, k(x_i, x)$

$$w = \sum_{i=0}^{n} \alpha_i \psi(x_i)$$

- 정규화항 $R(|| w ||)$
  - ✓ $|| w ||^2 = < w, w > = < \sum_{i=1}^{n} \alpha_i \psi(x_i), \sum_{j=1}^{n} \alpha_j \psi(x_j) >$
  - ✓ $= \sum_{i,j=1}^{n} \alpha_i \alpha_j < \psi(x_i), \psi(x_j) >$
  - ✓ $= \sum_{i,j=1}^{n} \alpha_i \alpha_j k < x_i, x_j >$
  - ✓ $|| w ||^2 = \alpha^T K \alpha \rightarrow R(|| w ||) = R(\sqrt{\alpha^T \alpha})$

- Kernel Matrix
  - ✓ $K = \left(k(x_i, x_j)\right)_{i,j} = \begin{pmatrix} k(x_1, x_1) & \cdots & k(x_1, x_n) \\ \vdots & \ddots & \vdots \\ k(x_n, x_1) & \cdots & k(x_n, x_n) \end{pmatrix}$

01 | Representer Theorem

Kernelized Predictions

- Kernelized Predictions Ex) $f_{new} = \text{sgn}(\hat{w}^T x_{new} + \hat{b})$
  - ✓ $f(x) = < w, \psi(x) > = < \sum_{i=1}^{n} \alpha_i \psi(x_i), \psi(x) >$
  - ✓ = $\sum_{i=1}^{n} \alpha_i < \psi(x_i), \psi(x) \rangle$
  - ✓ = $\sum_{i=1}^{n} \alpha_i \, k(x_i, x)$

$$w^* = \sum_{i=0}^{n} \alpha_i \psi(x_i)$$

$$K = \left(k(x_i, x_j)\right)_{i,j} = \begin{pmatrix} k(x_1, x_1) & \cdots & k(x_1, x_n) \\ \vdots & \ddots & \vdots \\ k(x_n, x_1) & \cdots & k(x_n, x_n) \end{pmatrix}$$

- $f_\alpha(x_n) = \sum_{i=1}^{n} \alpha_i \, k(x_i, x)$을 특징공간에서 계산
  - ✓ $\begin{pmatrix} f_\alpha(x_1) \\ \vdots \\ f_\alpha(x_n) \end{pmatrix} = \begin{pmatrix} \alpha_1 k(x_1, x_1) & \cdots & \alpha_1 k(x_1, x_n) \\ \vdots & \ddots & \vdots \\ \alpha_1 k(x_n, x_1) & \cdots & \alpha_1 k(x_n, x_n) \end{pmatrix} = \begin{pmatrix} k(x_1, x_1) & \cdots & k(x_1, x_n) \\ \vdots & \ddots & \vdots \\ k(x_n, x_1) & \cdots & k(x_n, x_n) \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_n \end{pmatrix}$
  - ✓ = $K\alpha$

Kernelized Objective

- 최종 일반화 식
  - ✓ $\min_{w \in H} R(\| w \|) + L(< w, \psi(x_1) >, \dots, < w, \psi(x_n) >)$
  - ✓ $\rightarrow \min_{\alpha \in R^d} R\left(\sqrt{\alpha^T K \alpha}\right) + L(K\alpha)$

커널기반 모델의 해가 선형 조합으로 이루어져 커널트릭을 이용할 수 있게됨

- 최종 SVM 식
  - ✓ $\min_{w \in H} \frac{1}{2} \| w \|^2 + \frac{c}{n} \sum_{i=1}^{n} \max(0, 1 - y_i [< w, \psi(x_i) >]$
  - ✓ $\rightarrow \min_{\alpha \in R^d} \frac{1}{2} \alpha^T K \alpha + \frac{c}{n} \sum_{i=1}^{n} (1 - y_i (K\alpha)_i)$

Some Kernel function

- Linear Kernel
  - ✓ Input space : $X = R^d$, Kernel Function : $K(w, x) = w^T x + b$
  - ✓ w와 x간의 유사도를 계산하는데 사용
  - ✓ 관측치를 고차원 공간으로 변환하지 않음
- Polynomial Kernel
  - ✓ Input space : $X = R^d$, Kernel Function : $K(w, x) = (1 + < w, x >)^M$
  - ✓ M: 차수 매개변수
  - ✓ Ex) $wx + (wx)^2 + \cdots + (wx)^M$
- RBF Kernel(가우시안 커널)
  - ✓ Input space : $X = R^d$, Kernel Function : $K(w, x) = exp(-\frac{||w-x||^2}{2\sigma^2})$

- RBF Kernel(가우시안 커널)

1. $k(w,x) = e^{-\frac{1}{2}(w-x)^2} = e^{-\frac{1}{2}(w+x)^2}e^{wx}$

2. $e^{wx} = 1 + \frac{1}{1!}wx + \frac{1}{2!}(wx)^2 + \cdots + \frac{1}{\infty!}(wx)^\infty$

3. $wx + (wx)^2 + \cdots + (wx)^\infty$

4. $e^{wx} = \left(1, \sqrt{\frac{1}{1!}}w, \sqrt{\frac{1}{2!}}w^2, \dots, \sqrt{\frac{1}{\infty!}}w^\infty\right)\left(1, \sqrt{\frac{1}{1!}}x, \sqrt{\frac{1}{2!}}x^2, \dots, \sqrt{\frac{1}{\infty!}}x^\infty\right)$

5. $e^{-\frac{1}{2}(w+x)^2}\left[\left(1, \sqrt{\frac{1}{1!}}w, \sqrt{\frac{1}{2!}}w^2, \dots, \sqrt{\frac{1}{\infty!}}w^\infty\right)\left(1, \sqrt{\frac{1}{1!}}x, \sqrt{\frac{1}{2!}}x^2, \dots, \sqrt{\frac{1}{\infty!}}x^\infty\right)\right]$

✓ Ex) $k(0,0) = e^0(1,0,0,\dots)(1,0,0,\dots) = 1(1,0,0,\dots)(1,0,0,\dots) = 1$

Positive Semidefinite

- Positive SemiDefinite(양의 준정부호성)
  - ✓ Mercer이론은 psd를 만족하는 조건이 존재해야 내적이 가능
  - ✓ psd는 대칭행렬 $M \in R^d$이 있을 때, 임의의 x에 대해서 $x^T Mx \geq 0$을 만족해야 한다.
  - ✓ 또한 M은 제곱근을 가지며 고윳값(λ)이 0보다 크거나 같아야 한다.

Positive Semidefinite Example

- 예시($x^T Mx$ 조건)
  - ✓ $M = \begin{bmatrix} 4 & 2 \\ 2 & 5 \end{bmatrix}, x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$
  - ✓ $x^T Mx = [x_1\ x_2]\begin{bmatrix} 4 & 2 \\ 2 & 5 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = 4x_1^2 + 4x_1x_2 + 5x_2^2$
  - ✓ 이 식에서 $x^T Mx$가 항상 0보다 큰 값을 가진다는 말은 M이 psd를 만족
- 예시($M = R^T R$ 조건)
  - ✓ $\begin{bmatrix} 4 & 2 \\ 2 & 5 \end{bmatrix} = \begin{bmatrix} 2 & 1 \\ 0 & 2 \end{bmatrix}^T \begin{bmatrix} 2 & 1 \\ 0 & 2 \end{bmatrix}$
  - ✓ 이 식에서 M의 제곱근 R을 가짐을 알 수 있고 이러한 M이 psd를 만족
- 예시($\lambda \geq 0$ 조건)
  - ✓ $\det(\mathrm{A} - \lambda\mathrm{I}) = 0$
  - ✓ $\begin{bmatrix} 4-\lambda & 2 \\ 2 & 5-\lambda \end{bmatrix} \rightarrow det(A - \lambda I) = (4-\lambda)(5-\lambda)\ 2 \times 2$
  - ✓ $= \lambda^2 - 9\lambda + 16$
  - ✓ $\therefore \lambda = 3{,}6$

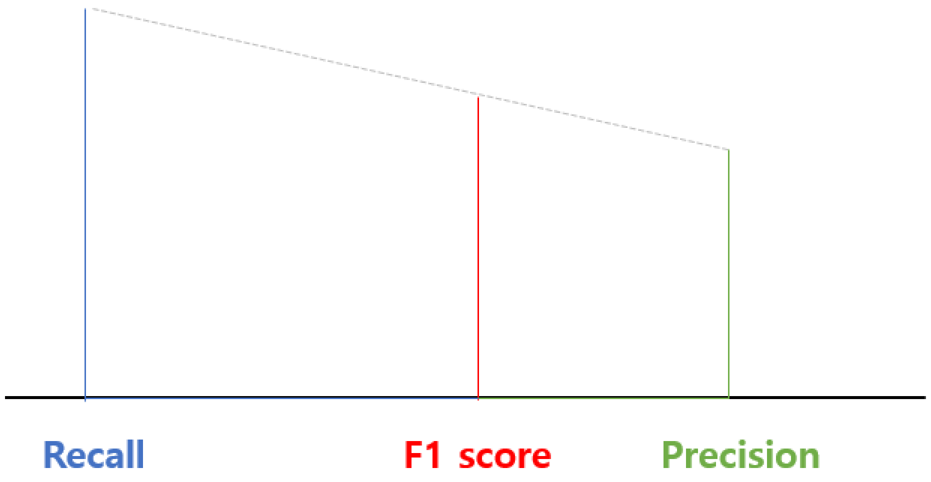
Performance Statistics

- 분류모형에서 성능평가(confusion matrix)
  - ✓ 정확도 : $\frac{TP+TN}{TP+FN+FP+TN}$
  - ✓ 오분류율 : $\frac{FP+FN}{TP+FN+FP+TN}$
  - ✓ Recall, Sensitivity(재현율, 민감도) : $\frac{TP}{TP+FN}$
  - ✓ Precision(정밀도) : $\frac{TP}{TP+FP}$
  - ✓ Specificity(특이도) : $\frac{TN}{FP+TN}$
  - ✓ False negative rate(위음성률) : $\frac{FN}{FN+TP}$
  - ✓ False positive rate(위양성률) : $\frac{FP}{FP+TN}$

		Actual	
		Class +	Class −
Predicted	Class +	TP	FP
	Class −	FN	TN

F_1-Score

- F_1-Score
  - ✓ $\frac{1}{\frac{\left(\frac{1}{recall}\right)+\left(\frac{1}{precision}\right)}{2}}$ →recall과 precision의 조화평균
  - ✓ F1 score는 데이터 label이 불균형 구조일 때, 모델의 성능을 정확하게 평가할 수 있으며, 성능을 하나의 숫자로 표현할 수 있음
  - ✓ 데이터가 불균형 구조일 때 조화평균을 사용하면 산술평균보다 이상치나 큰 값의 영향을 덜 받아서 신뢰성 높은 평균값을 구할 수 있음



F_β-Score

- F_β-Score
  - ✓ $(1+\beta^2)\frac{precision*recall}{(\beta^2 precision)+recall}$
  - ✓ F1 기반 평가산식 중 하나로 Beta를 매개변수로 사용해 Precision과 Recall 사이의 균형에 가중치를 부여하는 방법
  - ✓ Beta 값이 1.0보다 크면 Recall에 비중을 두고 계산
  - ✓ Beta 값이 1.0보다 작으면 Precision에 비중을 두고 계산

	Precision	Recall	F_1	$F_{0.5}$	F_2
1	0.01	0.99	0.02	0.01	0.05
2	0.20	0.80	0.32	0.24	0.50
3	0.40	0.90	0.55	0.45	0.72
4	0.60	0.62	0.61	0.60	0.62
5	0.90	0.95	0.92	0.91	0.94

04 Performance Evaluation

type 1 error vs type 2 error

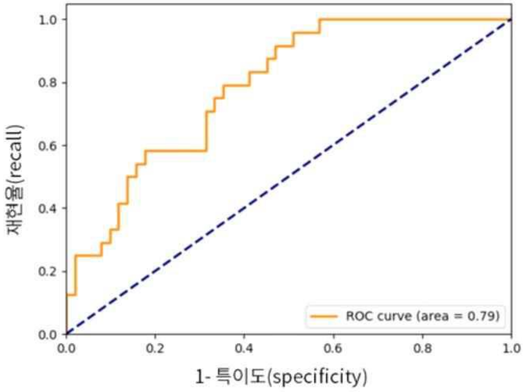
- 경우의 수
  - ✓ 잘 예측했는데 기각하기
    - ➢ type 1 error (유의수준으로 정의)
    - ➢ 옳은 귀무가설을 기각하고 대립가설은 채택하는 오류
    - ➢ 실제로는 참인데 거짓이라고 잘못 판단하는 경우
  - ✓ 잘못 예측했는데 채택하기
    - ➢ type 2 error (검정력으로 정의)
    - ➢ 옳은 대립가설을 기각하고 귀무가설을 채택하는 오류
    - ➢ 실제로는 거짓인데 참이라고 잘못 판단하는 경우

		Actual	
		Class +	Class −
Predicted	Class +		Type 1 Error
	Class −	Type 2 Error	

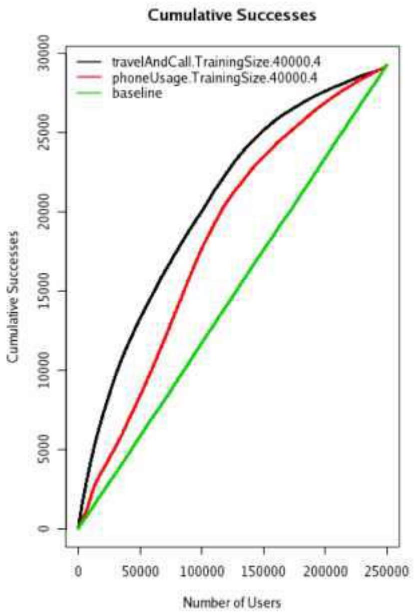
05 | The Performance Curve

ROC Curve

- ROC Curve
  - ✓ 가로축은 (1-특이도), 세로축은 재현율
  - ✓ 모델이 쓸모 없으면 파란색에 가까워짐
  - ✓ 모델이 1에 가깝게 볼록해지면 좋은 모델임
  - ✓ 따라서 ROC curve의 면적인 AUC는 0~1의 값을 가지는데 1에 가까울수록 좋음



- Lift chart
  - ✓ Lift curve = 무작위로 예측한 것에 비해 해당 알고리즘을 사용했을 때 어느정도 예측력이 향상되었는지를 측정
  - ✓ x축 : 추정확률값을 기준으로 내림차순으로 정리된 사례들의 누적개수
  - ✓ y축 : 진양성(True Positive) 누적 레코드 수



Q&A

감사합니다.