기계학습

6강 커널방법론, SVM

장필훈 교수

학습목차

1 커널방법론

2 SVM(최대마진분류기)

01

커널방법론

1-1 커널방법론

- 지금까지는 훈련집합과 새로운 입력을 분리해서 사용
- 훈련데이터의 일부 혹은 전부를

  예측단계에도 사용할 수 있다.
  - 예1) K-nearest neighbor(memory-based방식)
  - 예2) 커널함수를 이용하는 방식 ← 지금부터 배움

1-2 듀얼표현

- 선형회귀모델의 제곱합 오류함수

$$J(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\left\{\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n) - t_n\right\}^2 + \frac{\lambda}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w} \qquad \lambda \geq 0$$

$\frac{\partial J(w)}{\partial w} = 0$ 놓고 풀면,

$$\mathbf{w} = -\frac{1}{\lambda}\sum_{n=1}^{N}\left\{\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n) - t_n\right\}\boldsymbol{\phi}(\mathbf{x}_n) = \sum_{n=1}^{N} a_n\boldsymbol{\phi}(\mathbf{x}_n) = \boldsymbol{\Phi}^{\mathrm{T}}\mathbf{a}$$

$$a_n = -\frac{1}{\lambda}\left\{\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n) - t_n\right\}.$$

1-2 듀얼표현

- 여기서 $\mathbf{w}$대신 $\boldsymbol{a}$를 사용하여 $J(\mathbf{w})$를 다시 적을 수 있다.

$$J(\mathbf{a}) = \frac{1}{2}\mathbf{a}^{\mathrm{T}}\boldsymbol{\Phi}\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{a} - \mathbf{a}^{\mathrm{T}}\boldsymbol{\Phi}\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{t} + \frac{1}{2}\mathbf{t}^{\mathrm{T}}\mathbf{t} + \frac{\lambda}{2}\mathbf{a}^{\mathrm{T}}\boldsymbol{\Phi}\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{a}$$

$$\mathbf{t} = (t_1, \cdots, t_n)^T$$

"듀얼표현"

$\boldsymbol{\Phi}\boldsymbol{\Phi}^{\mathbf{T}}$: 그램행렬=$\mathbf{K}$, 원소 $K_{nm} = \Phi(x_n)^T\Phi(x_m)$

1-2 듀얼표현

- 모양($K_{nm} = \Phi(x_n)^T \Phi(x_m)$)만으로 알 수 있듯,

$$K_{nm} = \boldsymbol{\phi}(\mathbf{x}_n)^{\mathrm{T}} \boldsymbol{\phi}(\mathbf{x}_m) = k(\mathbf{x}_n, \mathbf{x}_m)$$

  으로 적을 수 있다. 이것이 kernel function.

- 이것을 이용해서 J를 다시 적으면,

$$J(\mathbf{a}) = \frac{1}{2}\mathbf{a}^{\mathrm{T}}\mathbf{K}\mathbf{K}\mathbf{a} - \mathbf{a}^{\mathrm{T}}\mathbf{K}\mathbf{t} + \frac{1}{2}\mathbf{t}^{\mathrm{T}}\mathbf{t} + \frac{\lambda}{2}\mathbf{a}^{\mathrm{T}}\mathbf{K}\mathbf{a}.$$

- $\nabla J = 0$의 해는, $\mathbf{a} = (\mathbf{K} + \lambda \mathbf{I}_N)^{-1}\mathbf{t}$.

1-2 듀얼표현

- 이것을 새로운 회귀모델에 대입, 새로운 예측치를 얻음.

$$y(\mathbf{x}) = \mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}) = \mathbf{a}^{\mathrm{T}}\boldsymbol{\Phi}\boldsymbol{\phi}(\mathbf{x}) = \mathbf{k}(\mathbf{x})^{\mathrm{T}}\left(\mathbf{K} + \lambda\mathbf{I}_N\right)^{-1}\mathbf{t}$$

$$\mathbf{k}(\mathbf{x}) = k_n(\mathbf{x}) = k(\mathbf{x}_\mathrm{n}, \mathbf{x})$$

→ 해를 온전히 커널함수에 관해 표현할 수 있다.

(듀얼공식화)

1-2 듀얼표현

- x에서의 예측값이 훈련집합에서 타겟변수들의 선형집합
- 원래의 문제는 w를 구하기 위해 M by N 행렬의 역을 구함
- 반면 듀얼문제는 a를 구하기 위해 N by N 행렬의 역을 구함
  - 보통 N>M이기 때문에 계산상의 잇점은 없지만, 공식 전체를 커널함수로 표현할 수 있고, $\Phi(x)$를 명시적으로 다루지 않아도 된다=높은차원의 특징공간을 직접 다루지 않을 수 있다.

1-3 커널의 구성

- 커널의 유효성 : 특징공간(feature space)상 스칼라곱에 해당하는가

예) 가정: $k(\mathbf{x},\mathbf{z}) = (\mathbf{x}^T,\mathbf{z})^2, \mathbf{x} = (x_1, x_2)$

$$
\begin{aligned}
k(\mathbf{x},\mathbf{z}) &= \left(\mathbf{x}^{\mathrm{T}}\mathbf{z}\right)^2 = (x_1 z_1 + x_2 z_2)^2 \\
&= x_1^2 z_1^2 + 2 x_1 z_1 x_2 z_2 + x_2^2 z_2^2 \\
&= (x_1^2, \sqrt{2} x_1 x_2, x_2^2)(z_1^2, \sqrt{2} z_1 z_2, z_2^2)^{\mathrm{T}} \\
&= \boldsymbol{\phi}(\mathbf{x})^{\mathrm{T}} \boldsymbol{\phi}(\mathbf{z}).
\end{aligned}
$$

1-3 커널의 구성

- 커널의 유효성 test

① 필요충분 조건이 알려져 있다.

② 단순한 커널들을 활용할 수 있다.

예) $ck(\mathbf{x}, \mathbf{x}')$가 유효하면, $\exp(ck(\mathbf{x}, \mathbf{x}'))$도 유효

$q(ck(\mathbf{x}, \mathbf{x}'))$도 유효. q는 계수가 0이상인 다항함수

→ 커널 엔지니어링

1-3 커널의 구성

- 가우시안 커널

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

- 생성모델로 정의하는것도 가능

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

- 시그모이드 커널

$$k(\mathbf{x}, \mathbf{x}') = \tanh\left(a\mathbf{x}^{\mathrm{T}}\mathbf{x}' + b\right)$$

1-4 RBF

- 방사기저함수(기저함수의 한 예)

$$\phi_j(\mathbf{x}) = h(\|\mathbf{x} - \boldsymbol{\mu}_j\|)$$

- 함수근사에 쓰일 수 있다

$$f(\mathbf{x}) = \sum_{n=1}^{N} w_n h(\|\mathbf{x} - \mathbf{x}_n\|).$$

w를 추정하는데,

계수숫자=제약조건 숫자 → 모든 표적값에 근사 가능

1-5 가우시안 프로세스

- 베이지안 관점하에서 어떤식으로 커널이 자연스럽게 등장하는지 관찰
- 선형회귀에서는, $y(\mathbf{x}, \mathbf{w}) = \mathbf{w}^T \phi(\mathbf{x})$에서 $\mathbf{w}$에 대한 사전분포로부터 $y(\mathbf{x}, \mathbf{w})$에 대한 사전분포를 유도했었음.
- 가우시안 프로세스는 매개변수모델을 생략하고 함수들에 대한 사전분포를 직접 정의함

1-5 가우시안 프로세스

- 선형회귀 예시 $y(\mathbf{x}) = \mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x})$ 에서,

  w의 사전분포 $p(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathbf{I})$, 정밀도 α

  $y_n = y(\mathbf{x}_n)$ 인 벡터 $\mathbf{y}$를 정의하면,

$$\mathbf{y} = \boldsymbol{\Phi}\mathbf{w}$$

Φ는 $\boldsymbol{\Phi}_{nk} = \Phi_k(\mathbf{x}_n)$

design matrix

1-5 가우시안 프로세스

- w는 가우시안 분포이고 $\mathbf{y}$는 그것들의 선형결합이므로 $\mathbf{y}$도 가우시안 분포이다.

따라서,

$$\mathbb{E}[\mathbf{y}] = \mathbf{\Phi}\mathbb{E}[\mathbf{w}] = \mathbf{0}$$

$$\mathrm{cov}[\mathbf{y}] = \mathbb{E}\left[\mathbf{y}\mathbf{y}^{\mathrm{T}}\right] = \mathbf{\Phi}\mathbb{E}\left[\mathbf{w}\mathbf{w}^{\mathrm{T}}\right]\mathbf{\Phi}^{\mathrm{T}} = \frac{1}{\alpha}\mathbf{\Phi}\mathbf{\Phi}^{\mathrm{T}} = \mathbf{K}$$

1-5 가우시안 프로세스

- (cont.)

  K는 그램행렬

$$K_{nm} = k(\mathbf{x}_n, \mathbf{x}_m) = \frac{1}{\alpha}\boldsymbol{\phi}(\mathbf{x}_n)^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_m)$$

- 함수 $y(\mathbf{x}, \mathbf{w})$에 대한 분포를 바탕으로 예측분포를 도출했다.

1-5 가우시안 프로세스

- 예시

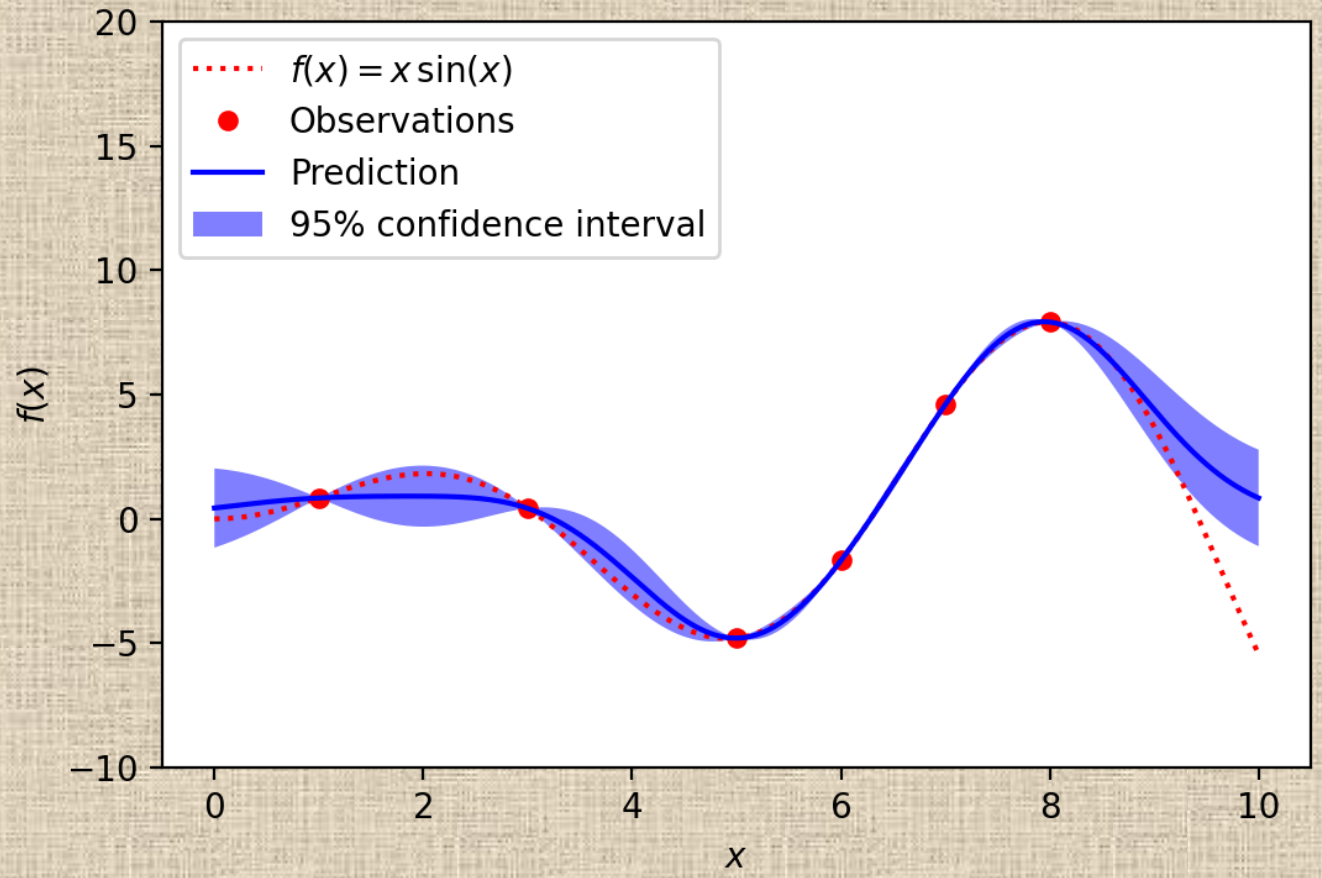
$$k(x_i, x_j) = \exp\left(-\frac{d(x_i, x_j)^2}{2l^2}\right)$$

d:euclidean l:length scale

```
from sklearn.gaussian_process import GaussianProcessRegressor
from sklearn.gaussian_process.kernels import RBF
kernel = RBF(10, (1e-2, 1e2))
gp = GaussianProcessRegressor(kernel=kernel, n_restarts_optimizer=9)
gp.fit(X, y)

y_pred, sigma = gp.predict(x, return_std=True)
```

1-5 가우시안 프로세스



Gaussian Processes regression: basic introductory example
https://scikit-learn.org/stable/auto_examples/gaussian_process/plot_gpr_noisy_targets.html

02

SVM(1)

2-1 희박한 커널머신

- 커널에 기반한 알고리즘들은 커널함수 $k(\mathbf{x}, \mathbf{x}_n)$의 값을 훈련집합의 모든 $\mathbf{x}_m$, $\mathbf{x}_n$짝에 대해 계산해야 함
- 따라서, 더 희박한(sparse) 해를 가지는 방법에 대한 연구가 이루어짐.

2-2 Support Vector Machine

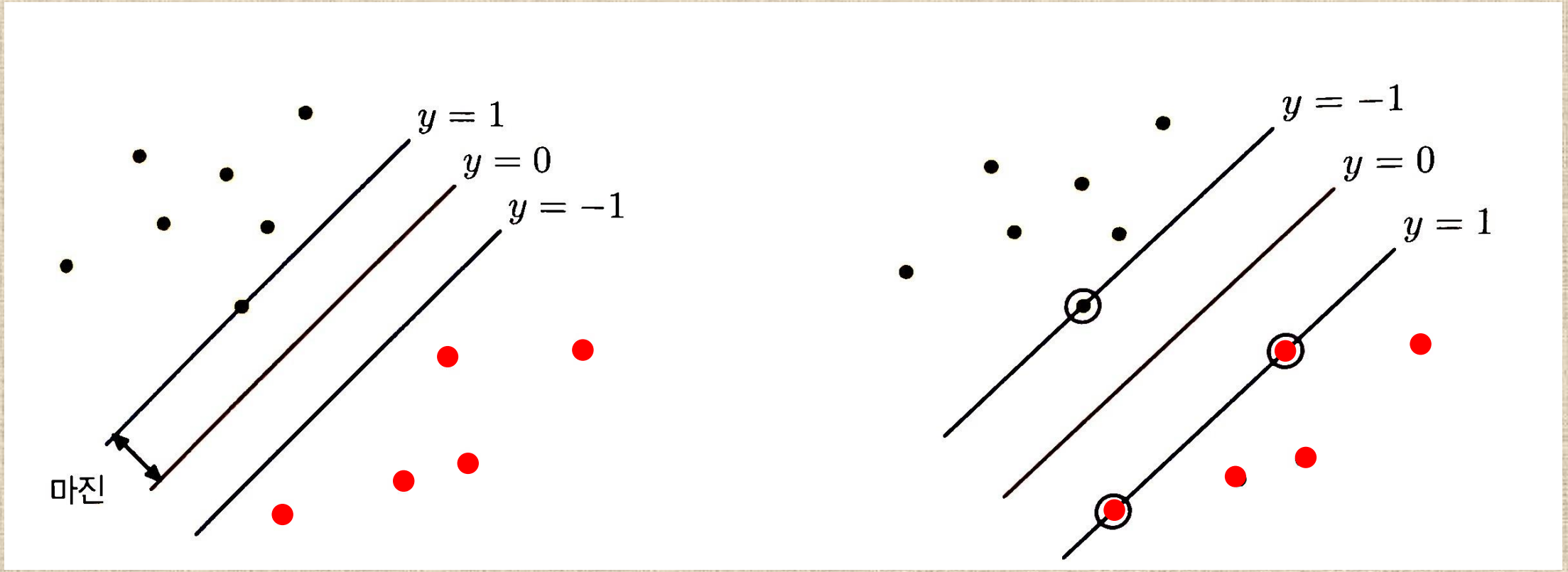
- 최대마진 분류기
- 2클래스 분류기부터 시작해보면,

$$y(\mathbf{x}) = \mathbf{w}^T \boldsymbol{\phi}(\mathbf{x}) + b$$

선형분류 가능하다고 가정,

→최대마진을 찾는 것이 목표

2-2 Support Vector Machine



Bishop, Christopher M. Pattern Recognition and Machine Learning. New York :Springer, 2006. fig7.1

2-2 Support Vector Machine

- $y(\mathbf{x}) = 0$에 해당하는 초평면으로부터

  포인트 $\mathbf{x}$까지 수직거리는 $\dfrac{|y(\mathbf{x})|}{\|\mathbf{w}\|}$

  모두 올바르게 분류된 경우 $t_n y(\mathbf{x}_n) > 0$

  따라서 $\mathbf{x}_n$으로부터 결정표면까지의 거리는

$$\frac{t_n y(\mathbf{x}_n)}{\|\mathbf{w}\|} = \frac{t_n(\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n) + b)}{\|\mathbf{w}\|}.$$

2-2 Support Vector Machine

- 최대마진해

$$\arg\max_{\mathbf{w},b}\left\{\frac{1}{\|\mathbf{w}\|}\min_{n}\left[t_n\left(\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n)+b\right)\right]\right\}$$

- $\mathbf{w}$와 b에 상수배해도 데이터포인트부터 결정경계까지 거리는 변하지 않는다. 따라서, 표면에 가장 가까운 점 $\mathbf{x}$에 대해 $t_n\left(\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n)+b\right)=1$ 라고 둘 수 있고,

2-2 Support Vector Machine

- (cont.) 모든 점이 다음을 만족한다.

$$t_n \left(\mathbf{w}^{\mathrm{T}} \boldsymbol{\phi}(\mathbf{x}_n) + b \right) \geqslant 1, \qquad n = 1, \ldots, N.$$

- 이 조건을 만족시키면서, $\|\mathbf{w}\|^2$를 최소화한다.

(∵ $\|\mathbf{w}\|^{-1}$ 의 최대화)

$$\arg \min_{\mathbf{w}, b} \frac{1}{2} \|\mathbf{w}\|^2$$

2-2 Support Vector Machine

- 제약조건 있는 최적화문제 → 라그랑주승수법

$$L(\mathbf{w}, b, \mathbf{a}) = \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{n=1}^{N} a_n \left\{ t_n(\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n) + b) - 1 \right\}$$

제약조건마다 하나의 승수 a_n을 도입

$$\mathbf{a} = (a_1, \ldots, a_N)^{\mathrm{T}}$$

- $L(\mathbf{w}, b, \mathbf{a})$ 의 $\mathbf{w}$ 에 대한 미분 $= \mathbf{0}$ 으로 두고

2-2 Support Vector Machine

- (cont.)

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

$$0 = \sum_{n=1}^{N} a_n t_n$$

- 이들을 이용해서, $\mathbf{w}$를 구하고 L에 대입

2-2 Support Vector Machine

- (cont.)

$$\widetilde{L}(\mathbf{a}) = \sum_{n=1}^{N} a_n - \frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{N} a_n a_m t_n t_m k(\mathbf{x}_n, \mathbf{x}_m)$$

최대마진문제의 dual representation을 얻음.

이 식을 $\mathbf{a}$에 대해 최대화한다.

$\mathbf{a}$를 찾는 방법: 2차계획법

2-2 Support Vector Machine

- 최대마진해를 구한 뒤, 새 데이터포인트의 분류

$$y(\mathbf{x}) = \sum_{n=1}^{N} a_n t_n k(\mathbf{x}, \mathbf{x}_n) + b.$$

위 식에 대입한다. 슬라이드22에 $\mathbf{w}$를 대입한것.

- 이때, $a_n\{t_n y(\mathbf{x}_n) - 1\} = 0$이어야 한다.(KKT조건, 과정생략)

2-2 Support Vector Machine

- $a_n = 0$이면 $y(\mathbf{x})$에 반영되지 않을 것이므로 $t_n y(\mathbf{x}_n) = 1$ 이다. $t_n y(\mathbf{x}_n) = 1$인 데이터포인트가 support vector
  - 특징공간의 최대마진 초평면상 포인트
  - 한번 모델이 훈련되면 이 서포트벡터만 중요하다.
    - 나머지 데이터포인트는 예측에 쓰지 않음.

2-2 Support Vector Machine

- $\mathbf{a}$를 찾은 뒤에는 b를 찾음
  - 모든 서포트벡터 $\mathbf{x}_n$은 $t_n y(\mathbf{x}_n) = 1$이므로,

$$t_n \left(\sum_{m \in \mathcal{S}} a_m t_m k(\mathbf{x}_n, \mathbf{x}_m) + b \right) = 1$$

(p30대입한것)

  - $t_n^2 = 1$이므로 양변에 t_n곱한 후 정리.

$$b = \frac{1}{N_{\mathcal{S}}} \sum_{n \in \mathcal{S}} \left(t_n - \sum_{m \in \mathcal{S}} a_m t_m k(\mathbf{x}_n, \mathbf{x}_m) \right)$$

서포트벡터 개수

2-3 클래스 분포간 중첩

- 특징공간 $\phi(\mathbf{x})$에서 선형분리 가능하지 않고

  클래스 분포간에 중첩이 존재할 때

  → 몇몇 포인트들의 오분류를 허용하는 것이 아이디어.

  (지금까지는 오분류가 아예 없을 때를 가정했다)

2-3 클래스 분포간 중첩

- slack variable(ξ)의 도입
  - 오분류에 대한 불이익을 경계면으로부터 거리에 대한 선형함수로 만듦.
  - 마진경계에 존재하면 $\xi = 1$

    그 외의 경우는 $\xi_n = |t_n - y(\mathbf{x}_n)|$
  - 결정경계와 마진경계의 거리=1 이므로 오분류되면 $\xi > 1$

2-3 클래스 분포간 중첩

- 이 경우 분류제약조건은 다음과 같다.

$$t_n y(\mathbf{x}_n) \geq 1 - \xi_n, \qquad \xi_n \geq 0$$

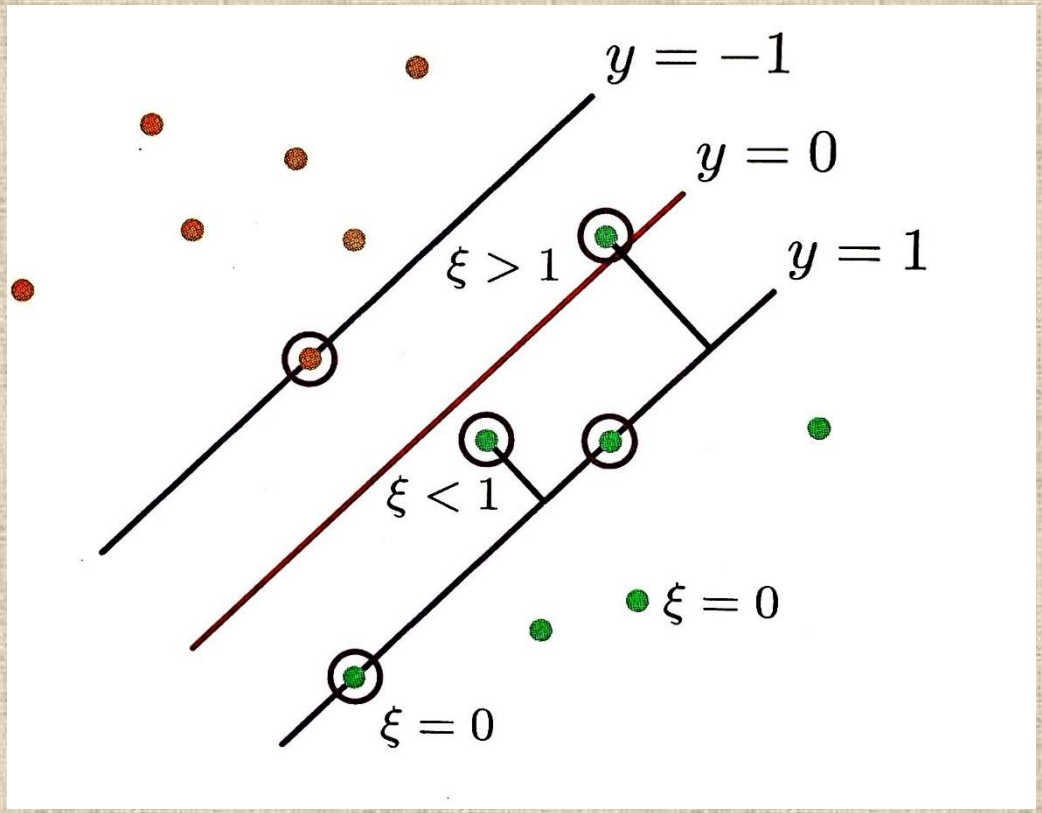
- $\xi = 0$ 인 경우 올바르게 분류됨(마진경계 혹은 외부)

  $0 < \xi \leq 1$ 마진 내부에 존재하지만 결정경계의 올바른쪽

  $\xi > 1$ 오분류

2-3 클래스 분포간 중첩

- soft margin 제약조건



Bishop, Christopher M. Pattern Recognition and Machine Learning. New York :Springer, 2006. fig7.3

2-3 클래스 분포간 중첩

- 오분류에러가 ξ에 대해 선형이라서 outlier에 민감함.
- 다음을 최소화 하는 것이 목표

$$C\sum_{n=1}^{N}\xi_n + \frac{1}{2}\|\mathbf{w}\|^2$$

C가 둘 사이의 균형을 조절함(트레이드 오프)

=정규화계수가 하는 역할을 하게 된다.

$C = \infty$의 경우 원래 SVM을 다시 얻는다.

2-3 클래스 분포간 중첩

- 최소화하기 위해, 라그랑주 승수법 이용

$$L(\mathbf{w}, b, \mathbf{a}) = \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{n=1}^{N}\xi_n - \sum_{n=1}^{N} a_n \{t_n y(\mathbf{x}_n) - 1 + \xi_n\} - \sum_{n=1}^{N}\mu_n\xi_n$$

$a_n \geq 0, \mu_n \geq 0$ 라그랑주 승수

KKT조건들 $\left(\frac{\partial L}{\partial \mathbf{w}} = \frac{\partial L}{\partial b} = \frac{\partial L}{\partial \xi_n} = 0\right)$이용해서 듀얼 라그랑주

표현식을 얻는다.

2-3 클래스 분포간 중첩

- dual Lagrangian

$$\widetilde{L}(\mathbf{a}) = \sum_{n=1}^{N} a_n - \frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{N} a_n a_m t_n t_m k(\mathbf{x}_n, \mathbf{x}_m)$$

- 제약조건들을 제외하면 선형분리 가능한 경우와 동일함.

다음시간

7강

- SVM(2)