'margin'

convex.

Classifier: max. margin classifier = SVM

primal problem

→ dual problem → solution

comple. slackness

Inclass 09: Support Vector Machine

[SCS4049] Machine Learning and Data Science

Seongsik Park (s.park@dgu.edu)

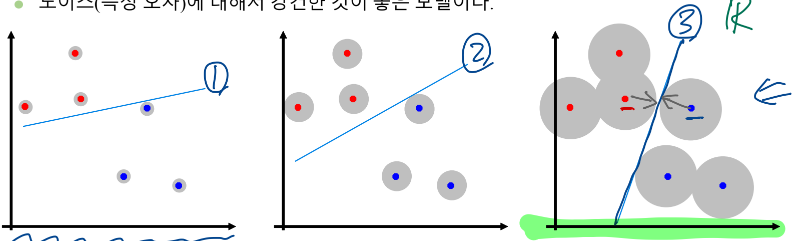
School of AI Convergence & Department of Artificial Intelligence, Dongguk University

Maximum Margin Classifier

What is a good decision boundary?

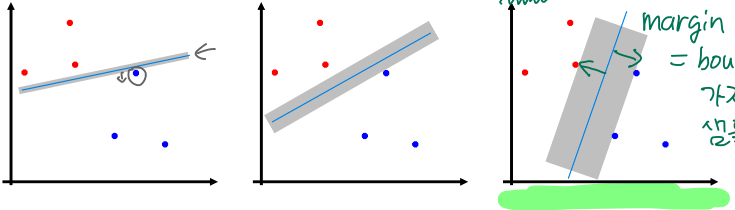
①

- 데이터 노이즈에 대한 강건성 (Robustness)
  - 노이즈(측정 오차)에 대해서 강건한 것이 좋은 모델이다.



②

- 여유로운 것이 더 강건하다 ⇒ 넓은 통로가 좋다 ⇒ Large Margin Classification



= boundary부터 가장가까운 샘플까지의 거리.

What is a good decision boundary?

샘플중의 일부. boundary를 형성, 예측할때 사용.

의사 결정은 경계의 데이터(support vectors)에 의해서 결정됨

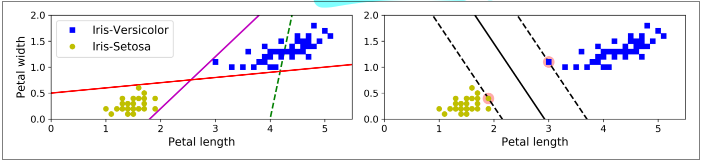


Figure 5-1. Large margin classification

거리 기반, margin

normalize 필요.

Input feature의 scale에 민감한 support vector machine

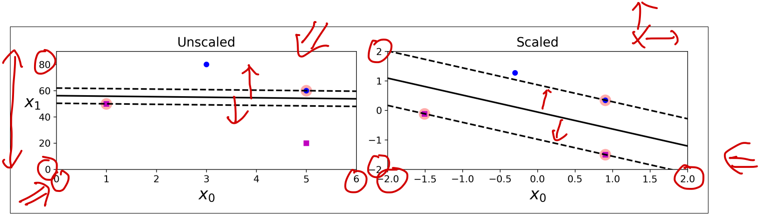


Figure 5-2. Sensitivity to feature scales

Hard margin vs. soft margin

Hard margin classification (hard-SVM)

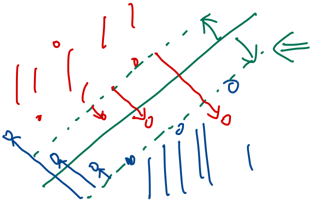
- 모든 데이터들이 margin 밖에 위치하도록 boundary를 설정
- 데이터가 linearly separable할 때만 적용 가능
- outlier에 매우 민감

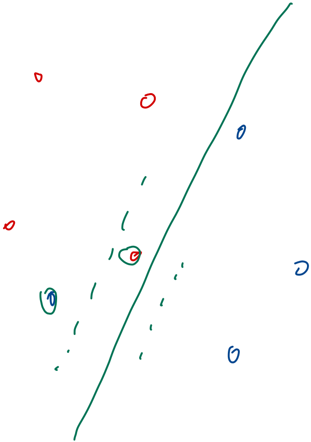
구속조건

정해진 margin안으로 못들어가는게 hard-SVM

Soft margin classification (soft-SVM)

- margin을 가능한 넓게 하면서도 margin 안에 들어오는 것도 허용
- hyperparameter C: 클수록 좁아짐 (엄격), 작을 수록 넓어짐 (허용)





Hard margin vs. soft margin

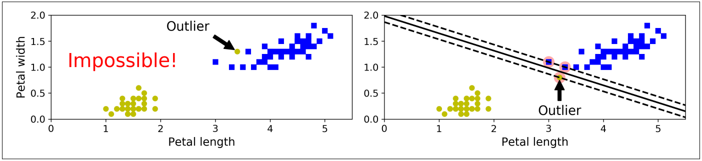


Figure 5-3. Hard margin sensitivity to outliers

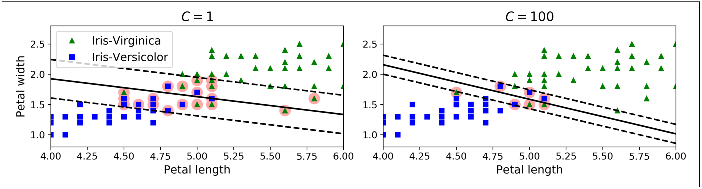


Figure 5-4. Large margin (left) versus fewer margin violations (right)

A brief history of SVM

- SVM은 1992년 Boser, Guyon and Vapnik에 의해서 소개됨
- Statistical Learning Theory에 이론적 바탕을 둔 알고리즘 ↙ convex opt.
- 손글씨 숫자 인식에서 뛰어난 성능을 보이면서 널리 쓰이게 됨
- SVM으로 1.1% Test error rate ≈ 신중히 설계된 신경망과 맞먹음
- 실용적으로 우수한 성능 ↙
- Bioinformatics, text, image recognition 등 많은 성공 사례
- 선형/비선형 분류 뿐 아니라 회귀, outlier detection 도 수행
- 복잡한 소규모/중규모 데이터셋의 분류에 특히 잘 맞음
- Kernel 방법을 사용하는 대표적 알고리즘
- Liblinear & libsvm: Scikit-Learn 에서 liblinear 및 libsvm 을 사용

어떤 아이디어. ⇄ 수학적, 최적화.

maximum margin → 최적화 정의

↓ convex

Support vector ← solution

complementary slackness

Hard SVM

Maximum margin classifier

We begin our discussion of support vector machines to the two-class classification problem using linear models of the form

↳nonlinear 확장 가능.

$$\boxed{y(\mathbf{x})} = \mathbf{w}^T\mathbf{x} + b \qquad (1)$$

~~where $\mathbf{x}$ denotes a fixed feature-space transformation, and we have made the bias parameter b expli~~cit.

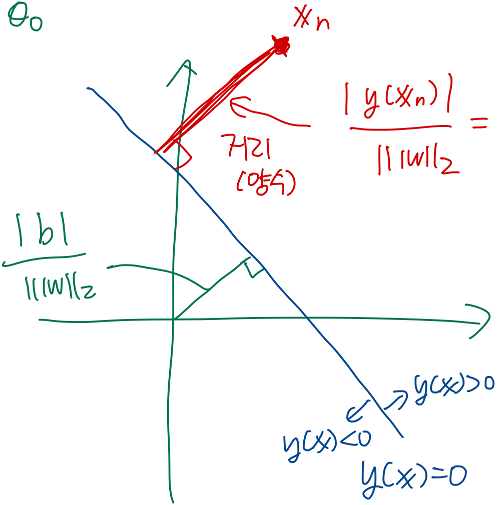
The training data set comprises N input vectors $\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N$, with corresponding target values $t_1, t_2, ..., t_N$ where $t_n \in \{-1, 1\}$, and new data points $\mathbf{x}$ are classified according to the sign of $y(\mathbf{x})$

$y(x) > 0 \longrightarrow t = 1$

$y(x) < 0 \longrightarrow t = -1$

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

$$= \theta^T \mathbf{x} + \theta_0$$



$$\frac{|y(\mathbf{x}_n)|}{\|\mathbf{w}\|_2} = \frac{|\mathbf{w}^T \mathbf{x}_n + b|}{\|\mathbf{w}\|_2}$$

$$\mathbf{x} \in \mathbb{R}^2$$

학습데이터

입력 $\underline{x}_1, x_2, x_3, \cdots, x_N$

출력 $t_1, t_2, t_3, \cdots, t_N$

boundary까지의 거리 $\underbrace{\frac{|w^T x_1 + b|}{||w||}}, \underbrace{\frac{|w^T x_2 + b|}{||w||}}, \underbrace{\frac{|w^T x_3 + b|}{||w||}}, \cdots$

$\underline{\text{margin}}$

$$\boxed{\min_{\underline{n}\in}} \left\{ \frac{\overset{|y(x_n)|}{|w^T x_n + b|}}{||w||_2} \right\} = \frac{t_n y(x_n)}{||w||_2} \quad \oplus$$

제일 가까운애.

margin

$$\max_{w,b}\left(\min_{n}\ \underbrace{\frac{t_n y(x_n)}{\|w\|} = \frac{t_n(w^T x_n + b)}{\|w\|}}_{\text{margin}}\right)$$

max margin

$$\max_{w,b}\left\{\frac{1}{\|w\|}\min_{n}\ t_n y(x_n)\right\}$$

$\Rightarrow$ objective

Maximum margin classifier

We shall assume that the training data set is linearly separable in feature space, so that by definition there exists at least one choice of the parameters $\mathbf{w}$ and b such that a function satisfies $y(\mathbf{x}_n) > 0$ for points having $t_n = +1$ and $y(\mathbf{x}_n) < 0$ for points having $t_n = -1$, so that $t_n y(\mathbf{x}_n) > 0$ for all training data points.

$$\underline{t_n y(x_n) > 0}_{\,n=1,2,\cdots,N} \implies$$

모든 sample이 제대로 분류.

(linearly separable)

Maximum margin classifier: optimality criterion

Thus the distance of a point $\mathbf{x}_n$ to the decision surface is given by

$$\frac{t_n y(\mathbf{x}_n)}{\|\mathbf{w}\|} = \frac{t_n(\mathbf{w}^T\mathbf{x}_n + b)}{\|\mathbf{w}\|}. \tag{2}$$

The margin is given by the perpendicular distance to the closest point $\mathbf{x}_n$ from the data set, and we wish to optimize the parameters $\mathbf{w}$ and b in order to maximize this distance. Thus the maximum margin solution is found by solving

$$\Longrightarrow \quad \underset{\mathbf{w},b}{\arg\max} \left\{ \frac{1}{\|\mathbf{w}\|} \min[t_n(\mathbf{w}^T\mathbf{x}_n + b)] \right\} \tag{3}$$

where we have taken the factor $1/\|\mathbf{w}\|$ outside the optimization over n because $\mathbf{w}$ does not depend on n.

Dual problem for convex optimization

Primal optimization problem for hard SVM

margin을 가능한 최대로 만들자.

objective

$$\boxed{\text{maximize}\quad \frac{1}{\|\mathbf{w}\|}\min[t_n(\mathbf{w}^T\mathbf{x}_n+b)]} \qquad (4)$$

constraint

$$\boxed{\text{subject to}\quad \underline{t_n(\mathbf{w}^T\mathbf{x}_n+b)\geq 0}} \qquad (5)$$

→ 모든 sample을 제대로 판별해줘

(6)

Equivalently,

$$\text{minimize}\quad \frac{1}{2}\|\mathbf{w}\|^2 \qquad (7)$$

objective

$$\text{subject to}\quad t_n(\mathbf{w}^T\mathbf{x}_n+b)\geq 1 \qquad (8)$$

constraint

Direct solution of this optimization problem would be very complex, so we shall convert it into an equivalent problem that is much easier to solve.

$$\min_{w,b} \ \frac{1}{2}\|w\|^2 \longrightarrow \mathcal{L}(w, b, \underline{a_1}, \underline{a_2}, \underline{a_3}, \cdots, \underline{a_N})$$

$$\text{s.t.} \quad \underline{t_1(w^T x_1 + b) \overline{\ge 1}} \leftarrow \boxed{a_1}$$

$$t_2(w^T x_1 + b) \ge 1 \leftarrow a_2$$

$$\vdots$$

$$t_N(w^T x_N + b) \ge 1 \leftarrow a_N$$

$$= \frac{1}{2}\|w\|^2 + \boxed{a_1}\big(1 - t_1(w^T x_1 + b)\big) + \boxed{a_2}\big(1 - t_2(w^T x_2 + b)\big)$$

$$\vdots$$

$$f_i(\ ) \le 0.$$

$$1 - t_n(w^T x_n + b) \le 0.$$

$$\max \frac{1}{\|\mathbf{w}\|} \min \; t_n(\mathbf{w}^T\mathbf{x}_n+b)$$

$$\text{s.t.} \quad t_n(\mathbf{w}^T\mathbf{x}_n+b) \geq 0.$$

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

Scalar배.

① $\mathbf{w} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$ $\quad b=1$ $\quad y(\mathbf{x}) = 1 + x_1 + x_2 = 0$

$$\Downarrow$$

② $\mathbf{w} = \begin{bmatrix} 2 \\ 2 \end{bmatrix}$ $\quad b=2$ $\quad y(\mathbf{x}) = 2 + 2x_1 + 2x_2 = 0$

x_n 거리

$$\Rightarrow \frac{1}{\|w\|_2} t_n(w^T x_n + b)$$

$w = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$ $b = 1$ $\rightarrow$ ① $\frac{1}{\sqrt{3}} t_n(1 + x_1 + x_2)$

$w = \begin{bmatrix} 2 \\ 2 \end{bmatrix}$ $b = 2$ $\rightarrow$ ② $\frac{1}{2\sqrt{3}} t_n(2 + 2x_1 + 2x_2)$

×2

×2

0/1

$$\max \frac{1}{\|w\|} \underbrace{\boxed{\min} \; t_n(w^T x_n + b)} \rightarrow \boxed{1}$$

$$\text{s.t.} \quad \underline{t_n(w^T x_n + b)} \geq \underbrace{\boxed{\min} \; t_n(w^T x_n + b)}_{\hookrightarrow \boxed{1}} \geq 0$$

$$\Rightarrow \quad \underline{\max \frac{1}{\|w\|}} \; 0 \Rightarrow \underline{\min \frac{1}{2} \|w\|_2^2}$$

$$\underline{\text{s.t.} \quad t_n(w^T x_n + b) \boxed{\geq 1}}$$

Lagrangian function with constraint

Setting the derivatives of $L(\mathbf{w}, b, \mathbf{a})$ with respect to $\mathbf{w}$ and b equal to zero, we obtain the following two conditions

$$\mathbf{w} = \sum_{n=1}^{N} a_n t_n \mathbf{x}_n = a_1 t_1 x_1 + a_2 t_2 x_2 + a_3 t_3 x_3 + \vdots \qquad (9)$$

$$0 = \sum_{n=1}^{N} a_n t_n \qquad (10)$$

① $a_n = 0$ $\quad t_n(w^T x_n + b) > 1$ ⇒ 얘 x_n는 margin위가 아니고 un 영역에 안전함.

② $a_n > 0$ $\quad t_n(w^T x_n + b) = 1$ ⇒ 이 x_n은 margin을만든 거!

↓

Support vector.

Lagrangian function with constraint

Eliminating $\mathbf{w}$ and b from $L(\mathbf{w}, b, \mathbf{a})$ using these conditions then gives the *dual representation* of the maximum margin problem in which we maximize

$$\tilde{L}(\mathbf{a}) = \sum_{n=1}^{N} a_n - \frac{1}{2} \sum_{n=1}^{N} \sum_{m=1}^{N} a_n a_m t_n t_m k(\mathbf{x}_n, \mathbf{x}_m) \tag{11}$$

with respect to $\mathbf{a}$ subject to the constraints

$$a_n \geq 0, \qquad n = 1, ..., N \tag{12}$$

$$\sum_{n=1}^{N} a_n t_n = 0. \tag{13}$$

Here the kernel function is defined by $k(\mathbf{x}, \mathbf{x}') = \mathbf{x}^T \mathbf{x}'$.

Prediction for a new sample: support vector machine

In order to classify new data points using the trained model, we evaluate the sign of $y(\mathbf{x})$. This can be expressed in terms of the parameter $\{a_n\}$ and the kernel function by substituting for $\mathbf{w}$ to give

$$\mathrm{w}^T x + b = \sum_n \underbrace{a_n}_{>0} t_n \underbrace{x_n^T x}_{\text{두벡터 얼마나비슷해?}} + b$$

$$y(\mathbf{x}) = \sum_{n=1}^{N} a_n t_n \mathbf{x}_n^T \mathbf{x} + b \qquad (14)$$

$$= \sum_{n=1}^{N} a_n t_n k(\mathbf{x}, \mathbf{x}_n) + b \qquad (15)$$

$$x \longrightarrow \begin{array}{ll} x^T x_1 & a_1 > 0 \\ x^T x_7 & a_7 > 0 \\ x^T x_{10} & a_{10} > 0 \end{array}$$

$$a_1 t_1 x^T x_1 \leftarrow$$

$$+ a_7 t_7 x^T x_7 \leftarrow$$

$$+ a_{10} t_{10} x^T x_{10} \leftarrow$$

KKT condition: complementary slackness

We show that a constrained optimization of this form satisfies the *Karush-Kuhn-Tucker* (KKT) conditions, which in this case require that the following three properties hold

$$a_n \geq 0 \qquad \text{dual const.} \tag{16}$$

$$t_n y(\mathbf{x}_n) - 1 \geq 0 \qquad \text{primal const} \tag{17}$$

$$a_n\{t_n y(\mathbf{x}_n) - 1\} = 0 \qquad \text{com. Slackness} \tag{18}$$

Thus for every data point, either $a_n = 0$ or $t_n y(\mathbf{x}_n) = 1$. Any data point for which $a_n = 0$ will not appear in the sum and hence plays no role in making predictions for new data points. The remaining data points are called *support vectors*, and because they satisfy $t_n y(\mathbf{x}_n) = 1$, they correspond to points that lie on the maximum margin hyperplanes in feature space.

KKT condition: complementary slackness

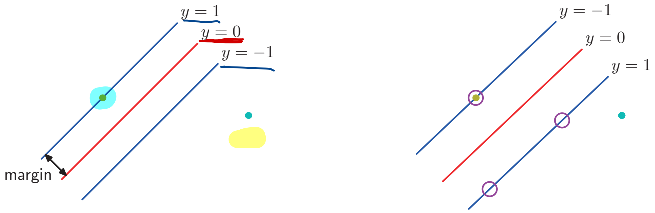


Figure 7.1 The margin is defined as the perpendicular distance between the decision boundary and the closest of the data points, as shown on the left figure. Maximizing the margin leads to a particular choice of decision boundary, as shown on the right. The location of this boundary is determined by a subset of the data points, known as support vectors, which are indicated by the circles.

KKT condition: complementary slackness

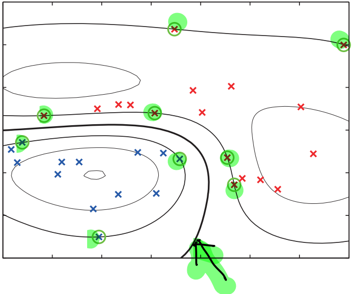


Figure 7.2 Example of synthetic data from two classes in two dimensions showing contours of constant $y(\mathbf{x})$ obtained from a support vector machine having a Gaussian kernel function. Also shown are the decision boundary, the margin boundaries, and the support vectors.

SVM: strength and limitations

장점

Convex opt.

- 강력하고 우수한 이론을 바탕으로 함
- 많은 블랙박스 알고리즘과는 대조적으로 비교적 직관적인 해석과 이해가 가능
- 학습이 상대적으로 쉬움 ← convex opt → global 보장.
- 신경망처럼 지역 최적값에 빠지는 일이 없음
- 학습 시간이 차원에 의존하지 않으며 kernel trick 덕분에 고정된 입력에만 의존
- 과적합이 잘 조절되는 경향
- 많은 분야에서 신경망 및 기타 알고리즘과 필적하는 성능
- 데이터가 작은 조건이나 고차원 공간에서도 잘 일반화

SVM: strength and limitations

단점

구속조건 : boundary, input.

- 노이즈에 민감 hard SVM, soft SVM
- 비교적 적은 수의 잘못된 label로 성능이 심각하게 악화
- 커널 함수를 선택하는 방법에 대한 정리된 원칙이 없음
- hyperparameter C의 적정값을 정하기 위한 원칙이 없음
- 컴퓨터의 메모리와 계산 시간 측면에서 비용이 높은 편이며 multiclass에서 더욱더 심화됨

① convex opt. 풀어야.

② 예측. 데이터셋↑

Appendix

Reference and further reading

- “Chap 7 | Sparse Kernel Machines” of C. Bishop, Pattern Recognition and Machine Learning
- “Chap 5 | Support Vector Machines” of A. Geron, Hands-On Machine Learning with Scikit-Learn, Keras & TensorFlow
- “Chap 4 | Convex Optimization Problems”, “Chap 5 | Duality” of S. Boyd, Convex Optimization
- “Lecture 6 | Support Vector Machines” of Kwang Il Kim, Machine Learning (2019)