SVM - 2021년 P-sat 논문스터디 1회차

권남택, 오정민, 유경민

2021년 1월 19일

A Tutorial on Support Vector Machine for Pattern Recognition에 대한 요약입니다.

1 Background

본격적인 SVM과 관련된 내용에 들어가기 전에 해당 내용들을 이해하는데 필요한 개념들 몇가지를 소개하고 시작한다.

1.1 Structural risk minimization (SRM)

일반적으로 Test error for trained machine은 아래와 같다.

$$R(\alpha) = \int \frac{1}{2}|y - f(\mathrm{x}, \alpha)|dP(\mathrm{x}, y)$$

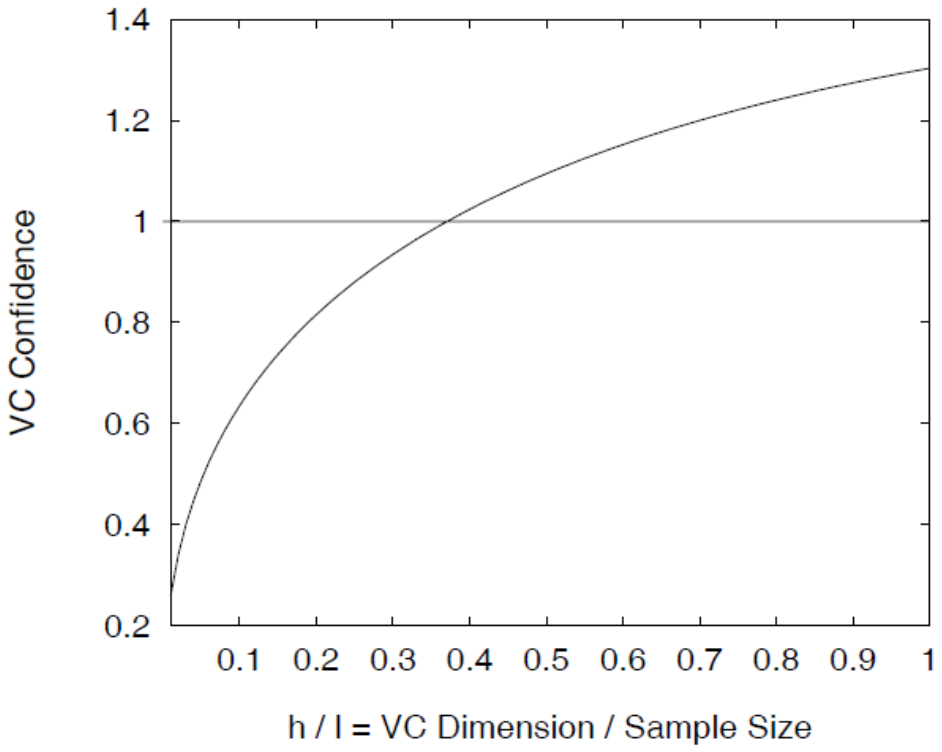
$R(\alpha)$ 값을 'expected risk', 'actual risk'라고 부르며, 우리가 궁극적으로 관심을 가지는 부분이 된다. 하지만 우리는 $R(\alpha)$ 값을 직접적으로 구할 수 없다. 따라서 우리는 직접적으로 알 수 없는 $R(\alpha)$ 값을 간접적으로나마 알 수 있도록 상한선을 지정하게 되는데 그 식은 아래와 같다.

$$R(\alpha) \leq R_{emp}(\alpha) + \sqrt{\frac{h(log(2l/h) + 1) - log(\eta/4)}{l}}$$

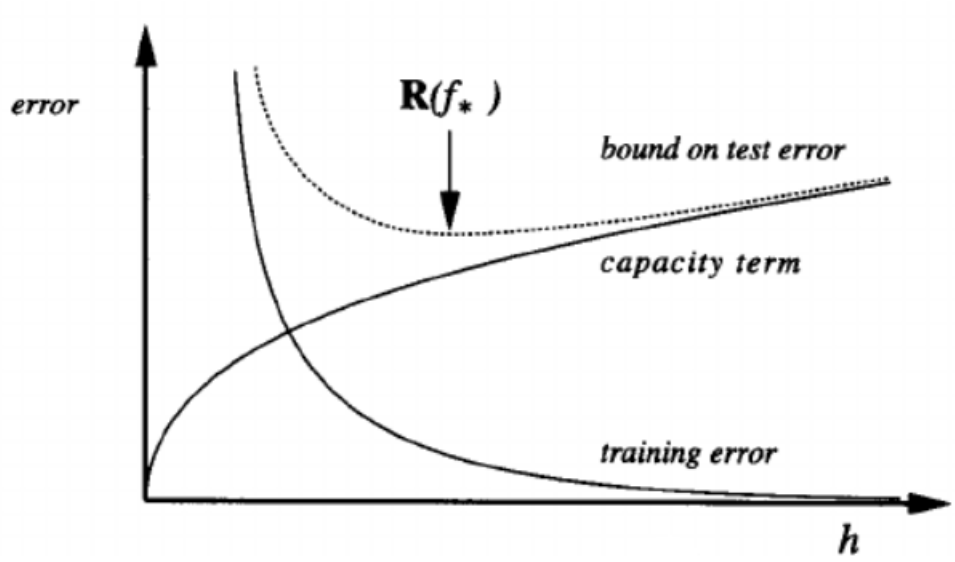
여기서 l은 관측값의 갯수, h는 VC dimension, η는 0과 1사의 값이고 $R_{emp}(\alpha)$는 'Empirical error'인데 우리에게 더 친숙한 용어인 training error로 생각을 해주면 된다. 위의 식에서 우변의 2번째 term을 'vc confidence'라고 부른다. $R_{emp}(\alpha)$에 대한 식은 다음과 같다.

$$R_{emp}(\alpha) = \sum_{i=1}^{l} \frac{1}{2l}|y - f(\mathrm{x_i}, \alpha)|$$

결국 'Structural risk minimization'의 아이디어는 우리가 관심이 있지만 직접적으로 구할수 없는 $R(\alpha)$ 값을 우리가 직접 구할수 있는 값들을 이용하여 상한선을 만들고 상한선에 해당하는 값을 최소화 함으로써 $R(\alpha)$ 값이 가장 작을것으로 추정되는 모델을 정할 수 있다라는 느낌으로 생각을 해주면 될 것 같다. 우변(상한선)을 최소화 한다는 것은 결국 우변의 첫번째 term인 'Empirical error'과 우변의 2번째 term인 'vc confidence'에서 h값 즉, 'VC dimension'을 최소로 해주는 모델을 찾는 과정이라고 생각해주면 된다.



위의 그래프는 'vc confidence'와 'vc dimension'의 관계를 나타낸 것이다. 'vc confidence'는 h에 대해 단조증가함을 알 수 있다.



그래프에서 'capacity term'이 'vc confidence', 'training error'가 'Empirical error'에 해당한다. SRM은 이러한 trade-off를 고려하여 최소화하는 지점에서의 모델을 최선의 모델로 보는 아이디어이다.

1.2 VC dimension

앞에서 나온 'VC dimension'이라는 개념을 알아보기 전에 먼저 'Shatter'라는 개념을 알아보자

특정함수 $f(X, \alpha)$에 대해서 'Shatter'된다는 의미는 개별 point들을 가능한한 모든 조합의 이진 레이블로 구분해 낼 수 있다는 것을 의미한다. 아래의 그림을 통해 쉽게 이해해보자

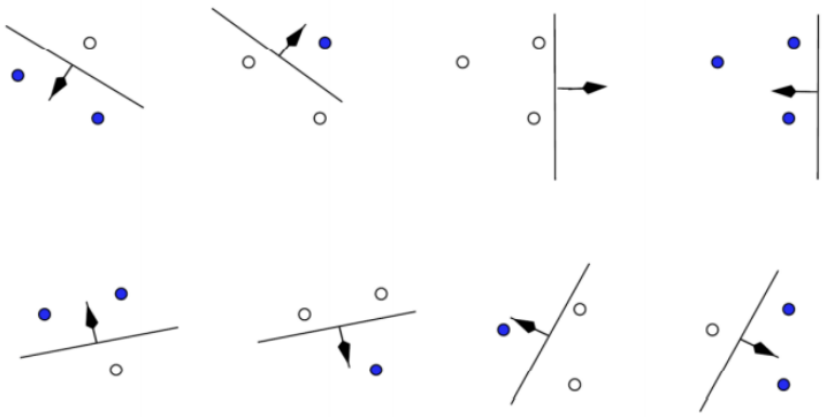


Figure 1: 2차원 3point case

$f(X, \alpha)$가 선형 분류기일 때, 2차원의 공간에서 3개의 점이 있는 경우 가능한 경우의 수는 2^3으로 8가지 경우이다. 위의 그림은 해당 8가지 경우를 완벽하게 분류하고 있다.

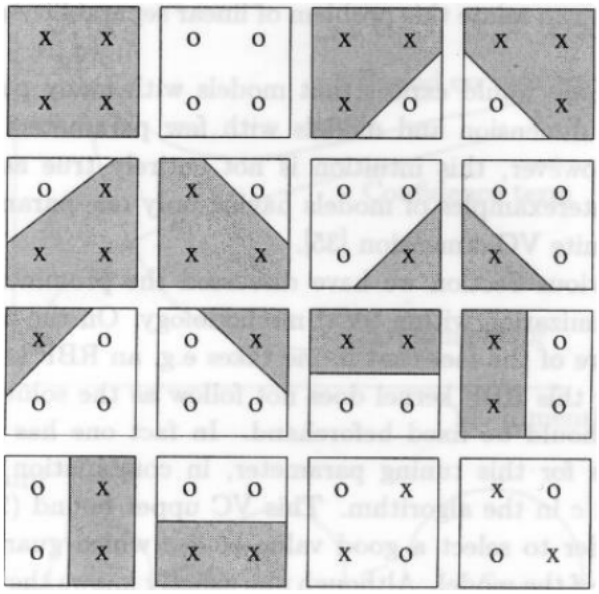


Figure 2: 2차원 4point case

$f(X, \alpha)$가 선형 분류기일 때, 2차원의 공간에서 4개의 점이 있는 경우 가능한 경우의 수는 2^4으로 16가지 경우이다. 하지만 위의 그림을 보면 알 수 있듯이 16가지 경우 중 14가지 경우에 대해서만 분류가 가능하다. 즉, 2차원에서 선형분류기는 XOR문제를 풀지 못한다. 따라서 선형분류기는 2차원 공간에서 점 4개부터는 shatter를 할 수 없다. (선형분류기는 '차원수+1'개까지의 point들을 shatter 가능하다.)

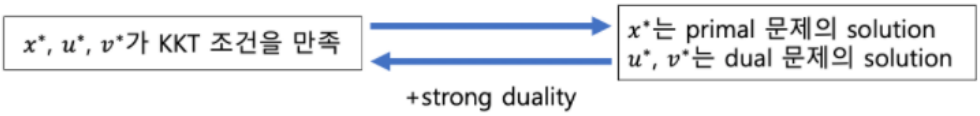
VC Dimension은 어떤 가설 공간의 Capacity를 측정한 지표이다. Capacity란 특정 모델의 복잡도와 같다. VC Dimension은 특정 분류기에 의해서 최대로 'Shatter'될 수 있는 점의 수로 정해진다. 위에서 알아본 바와 같이 2차원 상에 있는 선형 분류기의 VC Dimension, h=3 이라고 할 수 있다.

1.3 KKT condition

본격적인 svm을 다루는 내용들에서 Lagrange Primal, Dual에 관한 문제를 보게 될텐데, 해당 문제를 해결하는데 있어 필요한 개념인 'KKT condition'에 대해 소개하고 가겠다. 이 부등식(inequality) 제한조건이 있는 최적화 문제를 풀 때 사용되는 개념이 'KKT condition'이다.

보통 Primal보다는 Dual로서 문제를 다루는 것이 더 편하다. 그런데 Primal은 최소화문제이고, Primal을 변형한 Dual은 최대화문제인데, 두 문제의 해가 동일하다는 보장이 없다. 기본적으로 Primal Solution $\geq$ Dual Solution 인데, 이 등호를 만족하게 하는 조건이 바로 KKT condition이라고 이해하면 된다.

'KKT condition'과 primal, dual solution의 관계는 다음과 같다.



위의 관계를 보면 알 수 있듯이 x^*, u^*, v^*가 'KKT condition'을 만족한다면, x^*, u^*, v^*는 Primal, Dual문제에 대한 해가 됨을 알 수 있다. 이와 같은 충분조건을 바탕으로 뒤에서 우리는 Primal, Dual문제에 대한 해를 찾는 과정을 접근할 것이다.

다음과 같은 일반적인 최적화 문제가 주어졌을때

$$\begin{aligned} \min_{x} \quad & f(x) \\ \text{subject to} \quad & h_i(x) \leq 0,\ i = 1, ..., m \\ & l_j(x) = 0,\ j = 1, ..., r \end{aligned}$$

해당 문제에 대한 'KKT condition'은 다음과 같다.

1. $0 \in \partial(f(x) + \sum_{i=1}^{m} \lambda_i h_i(x) + \sum_{j=1}^{r} v_j l_j(x))$ (Stationarity): 최적화하려는 미지수(여기서는 x)로 편미분한 식이 0이 되는 값이 있음을 의미한다.
2. $\lambda_i h_i(x) = 0 \, for all \, i$ (Complementary Slackness): λ_i와 h_i중 적어도 하나의 값은 0을 가짐을 의미한다.
3. $h_i(x) \leq 0, l_j(x) = 0 \, for all \, i, j$ (Primal Feasibility): Primal problem의 제약조건들에 대한 만족여부를 나타낸다.
4. $\lambda_i \geq 0 \, for \, all \, i$ (Dual Feasibility): Dual problem의 제약조건들에 대한 만족여부를 나타낸다.

위와 같은 'KKT condition'을 만족하는 x^*는 결국 $\min_{x} f(x)$에 대한 해가 된다.

2 Linear separable case

2.1 Hyperplane

hyperplane은 p차원에서 한 점을 통과하는 해의 집합으로, p-1차원의 공간을 형성한다.

- (ex.2차원: line, 3차원: flat 2-dim subspace, 4차원: flat 3-dim subspace)

두 범주가 overlap 되지 않고, separable한 case에서, 2개의 그룹을 나누는 hyperplane은 무수히 많이 존재할 수 있다.

그렇다면 무수히 많은 초평면 중, 최적의 hyperplane은 어떻게 정할 수 있을까?

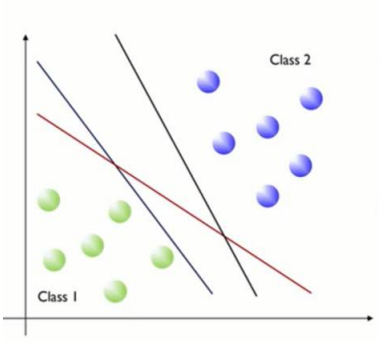


Figure 3: Separating hyperplane

2.2 Margin

Margin은 $d_+ + d_-$ ($d_+(d_-)$: hyperplane과 +(-) 관측치 사이의 최단거리)으로 계산되며, plus plane과 minus plane사이의 거리를 의미한다. 우리가 찾고자 하는 최적의 hyperplane은 margin을 최대로 하는 hyperplane이다.

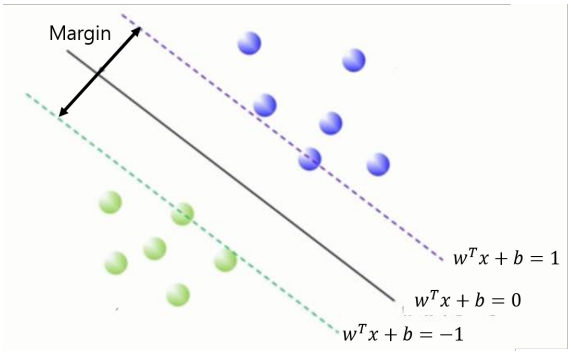


Figure 4: Margin

2.3 최적의 hyperplane 구하기

2.3.1 가정: 모든 관측치들은 다음과 같은 제약을 만족한다고 가정한다.

$$\begin{aligned} x_i \cdot w + b &\geq +1 \quad for \quad y_i = +1 \\ x_i \cdot w + b &\leq -1 \quad for \quad y_i = -1 \end{aligned}$$

이는 $y_i(x_i \cdot w + b) \geq +1$ 이라는 하나의 제약식으로 요약될 수 있다.

2.3.2 Margin

plus plane 위의 점 x_+와 minus plane 위의 점 x_-은 $x^+ = x^- + \lambda w$의 관계를 가진다. 이는 x_+가 x_-에서 w 방향으로 λ만큼 평행이동한 것 임을 의미한다. 이 식을 이용하여 λ를 유도할 수 있다.

$$\begin{aligned} w^\top x^+ + b &= 1 \\ w^\top (x^- + \lambda w) + b &= 1 \\ w^\top x^- + \lambda w^\top w + b &= 1 \\ -1 + \lambda w^\top w &= 1 \end{aligned}$$

$$\lambda = \frac{2}{\mathsf{w}^\top \mathsf{w}}$$

Margin은 plus plane과 minus plane사이의 거리이고, 이는 x^+와 x^-의 거리를 구하는 것과 같다. 따라서 다음과 같이 Margin을 구할 수 있다.

$$\begin{aligned} distance(\mathsf{x}^+, \mathsf{x}^-) &= ||\mathsf{x}^+ - \mathsf{x}^-||_2 \\ &= ||\lambda \mathsf{w}||_2 \\ &= \lambda \sqrt{\mathsf{w}^\top \mathsf{w}} \\ &= \frac{2}{\mathsf{w}^\top \mathsf{w}} \sqrt{\mathsf{w}^\top \mathsf{w}} \\ &= \frac{2}{\sqrt{\mathsf{w}^\top \mathsf{w}}} \\ &= \frac{2}{||\mathsf{w}||_2} \end{aligned}$$

2.3.3 Original Problem

우리의 목적은 margin을 최대화 하는 hyperplane을 찾는 것이다. 따라서 우리는 앞서 구한 margin인 $\frac{2}{||\mathsf{w}||_2}$ 을 최대화 하는 문제를 풀어야 한다. 하지만 L2 norm은 제곱근을 포함하고 있기에 계산상의 편의를 위하여 이 문제를 margin 역수의 제곱을 최소화하는 문제로 바꾸어서 풀 것이다.

$$\begin{aligned} & max \quad \frac{2}{||\mathsf{w}||_2} \\ \rightarrow \quad & min \quad \frac{||\mathsf{w}||_2}{2} \\ \rightarrow \quad & min \quad \frac{{||\mathsf{w}||_2}^2}{2} \end{aligned}$$

제약식과 목적식을 정리하면 다음과 같다.

$$min \quad \frac{{||\mathsf{w}||_2}^2}{2} \quad subject\ to\ \mathsf{y}_\mathsf{i}(\mathsf{x}_\mathsf{i} \cdot \mathsf{w} + b) \geq +1$$

목적식은 2차이고, 제약식은 1차식이므로 이를 이차계획법 convex optimization 문제로 풀 수 있다. 하지만 이렇게 제약식을 포함하는 형태로 최적화하는 것은 쉽지않으므로 라그랑주 승수법을 이용하여 목적함수를 간단하게 바꾸어 준다.

2.3.4 Lagrange Primal

목적식에 제약식과 라그랑주 승수를 곱한 항을 더하여 제약이 없는 Lagrange Primal 문제로 변환하여준다. 우리의 새로운 목적함수는 원래 문제의 lower bound가 된다.

$$\min_{\mathsf{w},b} \mathcal{L}(\mathsf{w}, b, \alpha) = \frac{{||\mathsf{w}||_2}^2}{2} - \sum_{i=1}^{n} \alpha_i (\mathsf{y}_\mathsf{i}(\mathsf{x}_\mathsf{i} \cdot \mathsf{w} + b) - 1) \quad subject\ to\ \alpha_i \geq 0$$

2.3.5 Lagrange Dual

Lagrange Primal은 원래 문제의 lower bound가 되므로 원래 문제의 최적해를 찾기 위해서는 lower bound를 최대화 시켜야 한다.

먼저 $\mathcal{L}(\mathsf{w}, b, \alpha)$를 최소화하는 w, b를 찾기 위해 w, b에 대해 미분한다.

$$\frac{\partial \mathcal{L}}{\partial \mathsf{w}} = \mathsf{w} - \sum_{i=1}^{n} \alpha_i y_i x_i = 0 \rightarrow \mathsf{w} = \sum_{i=1}^{n} \alpha_i y_i x_i$$
$$\frac{\partial \mathcal{L}}{\partial b} = \sum_{i=1}^{n} \alpha_i y_i = 0$$

이 식을 Lagrange Primal의 목적식에 대입하면 다음과 같이 α에 대한 식으로 간단하게 정리된다.

$$\frac{1}{2}(\sum_{i=1}^{n} \alpha_i y_i x_i)(\sum_{j=1}^{n} \alpha_j y_j x_j) - (\sum_{i=1}^{n} \alpha_i y_i x_i)(\sum_{j=1}^{n} \alpha_j y_j x_j) - b\sum_{i=1}^{n} \alpha_i y_i + \sum_{i=1}^{n} \alpha_i$$
$$= -\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \alpha_j x_i x_j y_i y_j + \sum_{i=1}^{n} \alpha_i$$
$$where \quad \sum_{i=1}^{n} \alpha_i y_i = 0$$

모든 α에 대하여 이 식은 원래 문제의 lower bound가 되고, 그 중에서 이 식을 최대로 하는 값이 정확히 최적값과 일치하게 된다. 이제 Primal Problem은 다음과 같이 lower bound를 최대화 하는 문제인 Dual Problem으로 변화하였다.

$$\max_{\alpha} \; -\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \alpha_j x_i x_j y_i y_j + \sum_{i=1}^{n} \alpha_i \quad subject\ to\ \alpha_i \geq 0$$

2.4 Support Vector

앞선 과정에서 우리는 $\mathsf{w}^* = \sum_{i=1}^{n} \alpha_i y_i x_i$
임을 구했다. 이를 통해 α_i가 0이 아닌 관측치들만 hyperplane을 구하는데 영향을 준다는 것을 알 수 있다. 더불어 KKT condition의 조건 중 complete slackness 조건은 α_i가 0이 아닌 관측치들이 곧 plus/minus plane 위에 있는 점, 즉 suppor vecor 임을 보여준다. complete slackness 조건식은 다음과 같다.

$$\alpha_i(\mathsf{y_i}(\mathsf{x_i} \cdot \mathsf{w} + b) - 1) = 0 \quad \forall i$$

- α_i = 0 인 경우, $\mathsf{y_i}(\mathsf{x_i} \cdot \mathsf{w} + b) \neq 1$ 이므로 $\mathsf{x_i}$는 plus/minus plane 위의 점이 아니다.
- α_i > 0 인 경우 $\mathsf{y_i}(\mathsf{x_i} \cdot \mathsf{w} + b) = 1$ 이므로 $\mathsf{x_i}$는 plus/minus plane 위의 점인 support vector이다.

정리하자면, svm의 hyperplane은 plus/minus plane 위의 관측치인 support vector만을 이용하여 구해진다.

3 Linear non-separable case

분류 경계면이 명확하게 나타나기 어려운 경우를 확인해보자.

3.1 Slack Variable의 도입

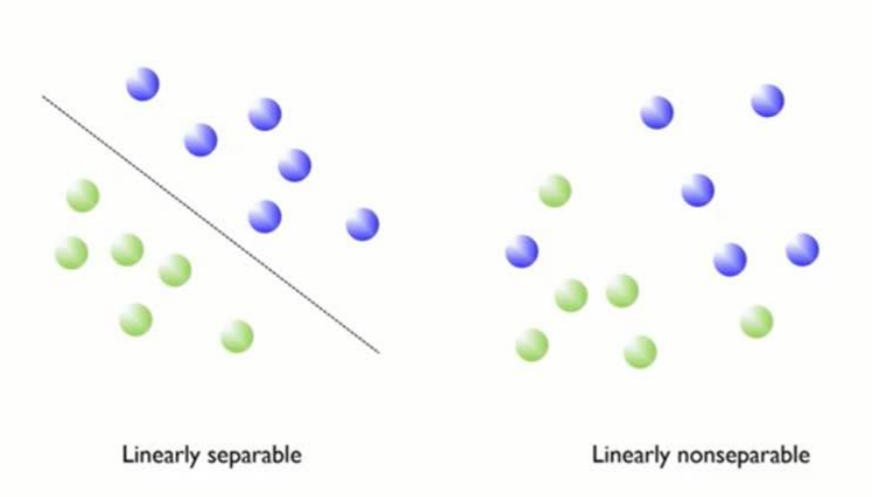


Figure 5: Separating vs Non-separable

그렇다면 오른쪽 그림과 같이 어떤 선을 그리더라도 완벽하게 분리할 수 없는 경우에는 어떻게 해야할까? 이러한 경우 Linear decision boundary를 이용하여 완벽하게 나누는 것은 불가능하므로 error를 허용한다. 제약식에서 이 error는 slack variable ξ_i로 표현되고, 제약식은 다음과 같이 변한다.

$$\begin{aligned} \mathrm{x_i} \cdot \mathrm{w} + b &\geq +1 - \xi_i \quad for \quad \mathrm{y_i} = +1 \\ \mathrm{x_i} \cdot \mathrm{w} + b &\leq -1 + \xi_i \quad for \quad \mathrm{y_i} = -1, \qquad \xi_i \geq 0 \quad \forall i \end{aligned}$$

이는 $\mathrm{y_i}(\mathrm{x_i} \cdot \mathrm{w} + b) \geq 1 - \xi_i$ 으로 요약된다.

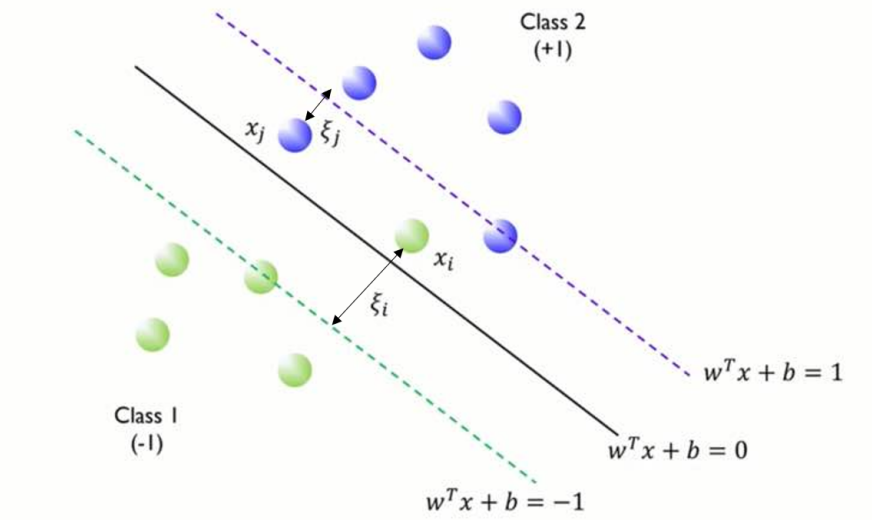


Figure 6: Non-separable case

3.2 최적의 hyperplane 구하기

3.2.1 Original Problem

$$
\begin{aligned}
min \quad & \frac{||\mathsf{w}||_2}{2}^2 + \mathsf{C}\sum_{i=1}^{n}\xi_i \\
subject\ to \quad & \mathsf{y_i}(\mathsf{x_i}\cdot\mathsf{w}+b) \geq 1-\xi_i, \xi_i \geq 0
\end{aligned}
$$

Separable case와 달리 원문제의 목적함수에 slack variable ξ_i를 이용한 term이 추가되었다. C는 ξ_i를 억제하는 penalty로, tuning parameter이다. C가 커지면 error을 많이 허용하지 않게되어 overfit 될 수 있고, C가 작아지면 error을 많이 허용하게되어 underfit 될 수 있다.

3.2.2 Lagrange Primal

목적식에 제약식과 라그랑주 승수를 곱한 항을 더하여 제약이 없는 Lagrange Primal 문제로 변환하여준다. 이때 Separable case와 달리 decision variable로 ξ_i가 추가되어 ξ_i에 대한 term이 존재한다.

$$
\begin{aligned}
\min_{\mathsf{w},b} \quad & \mathcal{L}(\mathsf{w},b,\alpha) = \frac{||\mathsf{w}||_2}{2}^2 - \sum_{i=1}^{n}\alpha_i(\mathsf{y_i}(\mathsf{x_i}\cdot\mathsf{w}+b)-1+\xi_i) - \sum_{i=1}^{n}\gamma_i\xi_i \\
subject\ to \quad & \alpha_i, \gamma_i \geq 0
\end{aligned}
$$

3.2.3 Lagrange Dual

Lagrange primal은 원래 문제의 lower bound가 되므로 원래 문제의 최적해를 찾기 위해서는 lower bound를 최대화 시켜야 한다.

먼저 $\mathcal{L}(\mathsf{w},b,\alpha,\xi,\gamma)$를 최소화하는 w,b,ξ를 찾기 위해 w,b,ξ_i에 대해 미분한다.

$$
\begin{aligned}
\frac{\partial\mathcal{L}}{\partial\mathsf{w}} &= \mathsf{w} - \sum_{i=1}^{n}\alpha_i y_i x_i = 0 \rightarrow \mathsf{w} = \sum_{i=1}^{n}\alpha_i y_i x_i \\
\frac{\partial\mathcal{L}}{\partial b} &= \sum_{i=1}^{n}\alpha_i y_i = 0 \\
\frac{\partial\mathcal{L}}{\partial\xi_i} &= \mathsf{C} - \alpha_i - \gamma_i = 0
\end{aligned}
$$

separable case와 마찬가지로 이 식을 Lagrange Primal의 목적식에 대입하면 다음과 같이 α에 대한 식으로 간단하게 정리된다. 차이점은 $\mathsf{C}-\alpha_i-\gamma_i=0$ 라는 제약이 추가되었다는 점이다.

$$
\begin{aligned}
&\frac{1}{2}(\sum_{i=1}^{n}\alpha_i y_i x_i)(\sum_{j=1}^{n}\alpha_j y_j x_j) - (\sum_{i=1}^{n}\alpha_i y_i x_i)(\sum_{j=1}^{n}\alpha_j y_j x_j) - b\sum_{i=1}^{n}\alpha_i y_i + \sum_{i=1}^{n}\alpha_i \\
&= -\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j x_i x_j y_i y_j + \sum_{i=1}^{n}\alpha_i \\
&where \sum_{i=1}^{n}\alpha_i y_i = 0 \quad and \quad \mathsf{C}-\alpha_i-\gamma_i = 0
\end{aligned}
$$

그런데 여기서 $0 \leq \alpha_i$, $0 \leq \gamma_i$, $C - \alpha_i - \gamma_i = 0$이므로 α의 범위는 0과 C 사이에 놓이게 된다.

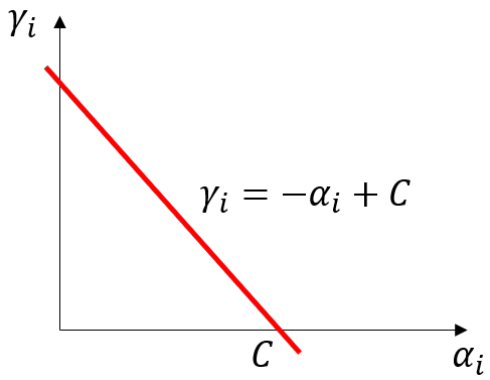


Figure 7: range of alpha

따라서 dual 식의 제약조건에 α_i가 C 이하라는 조건이 separable case와 달리 추가되었다.

$$\max_{\alpha} \quad -\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j x_i x_j y_i y_j + \sum_{i=1}^{n}\alpha_i$$
$$subject\ to \quad 0 \leq \alpha_i \leq C$$

3.3 Support Vector

separable case와 마찬가지로 $w^* = \sum_{i=1}^{n}\alpha_i y_i x_i$
로 구해져 $\alpha_i > 0$인 관측치들만 hyperplane을 구하는데 영향을 준다는 것을 알 수 있다. 또, KKT condition의 compete slackness에 의해 $\sum_{i=1}^{n}\alpha_i(y_i(x_i \cdot w + b) - 1 + \xi_i) = 0$, $\gamma_i\xi_i = 0$조건이 만족되는데 이를 α의 case별로 살펴보자.

1) hyperplane 결정에 영향을 미치치 않는 $\alpha_i = 0$인 경우, x_i는 각 plus/minus plane 안에 있는 점들이다.

$$\alpha_i = 0 \quad \rightarrow \quad \gamma_i = C,\ \xi_i = 0 \quad \Rightarrow \quad y_i(x_i \cdot w + b) \neq 1$$

2) hyperplane 결정에 영향을 미치는 $0 < \alpha_i < C$인 경우, x_i는 plus/minus plane위에 있다. 이 x_i들은 support vector이다.

$$0 < \alpha_i < C \quad \rightarrow \quad \gamma_i > 0,\ \xi_i = 0 \quad \Rightarrow \quad y_i(x_i \cdot w + b) = 1$$

3) hyperplane 결정에 영향을 미치는 $\alpha_i = C$인 경우, x_i는 각 plus/minus plane 밖에 있다. 이 x_i들도 support vector이다.

$$\alpha_i = C \quad \rightarrow \quad \gamma_i = 0,\ \xi_i > 0 \quad \Rightarrow \quad y_i(x_i \cdot w + b) - 1 = 1 - \xi_i$$

4 Non-linear & Non-seperable

4.1 고차원 매핑

지금까지 Linear & Seperable, Linear & Non-seperable의 경우들에 대해 어떻게 svm의 목적식과 제약식에서 최적화를 할 수 있는지를 보았다. 하지만 맨 처음 언급했듯이, 선형 초평면으로는 $\mathbb{R}^d$ 상에서 $d+1$ 개의 점만 shatter 할 수 있다. 가장 간단한 예로 2차원에서 선형분류기는 XOR 문제를 풀 수 없다. 우리가 다뤄야 할 점(관측값)은 그것보다 훨씬 더 많이 때문에, 비선형 분리의 필요성이 생긴다. 그래서 우리는 현재 차원 $\mathbb{R}^d$을 고차원 $\mathbb{R}^D$, $D >> d$ 으로 대응시켜 고차원 $\mathbb{R}^D$ 상에서 선형으로 분리하려 한다. 2차원에서 이차항까지 고려하는 고차원으로 매핑하는 예시를 들어보자.

$$\Phi : (x_1, x_2) \rightarrow ({x_1}^2, {x_2}^2, \sqrt{2}x_1, \sqrt{2}x_2, \sqrt{2}x_1x_2, 1)$$

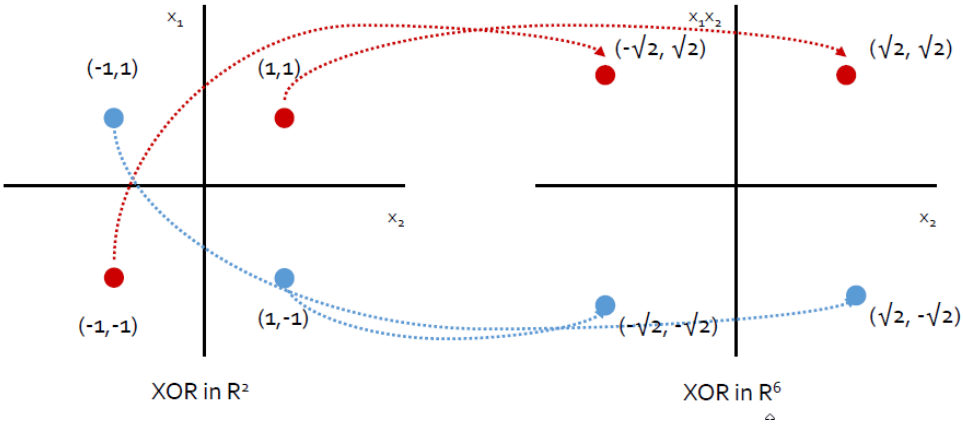


Figure 8: High-Dimensional Mapping

위에서 $(x_1x_2,\ x_2)$축에서 선형으로 나타나듯이, 변수들이 고차원에서는 선형으로 분리되는 형태가 존재할 수 있다. 정확히 언제 선형으로 분리되는지는 모르지만, 매핑할 뿐이다. 그래서 우리의 목표는 현재 데이터를 잘 분리할 수 있는 유연한(Flexible)한 분류기를 만듬과 동시에, Margin을 최대화 함으로써 분류의 일반화 성능을 높여야 한다. 이를 목적함수와 제약식으로 나타내면, 이전과 매우 유사한 형태이다.

$$\begin{aligned} min \quad & \frac{1}{2}||w||^2 + C\sum_{i=1}^{n}\xi_i \\ s.t. \quad & y_i(w^T\Phi(x_i) + b) \geq 1 - \xi_i,\ \ \xi_i \geq 0 \end{aligned}$$

이전에는 입력공간 x 자체에서 제약식을 다뤘다면, 이제는 매핑된 입력공간 $\Phi(x)$에서 제약식을 다루는 것이 다르다. 하지만 이에따라 달라지는 것은 없다! 똑같이 라그랑지안을 통해 듀얼문제로 변형하게 되면 다음과 같다.

$$\begin{aligned} max \quad & L_D = \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j \Phi(x_i)\Phi(x_j) \\ s.t. \quad & \sum_{i=1}^{n}\alpha_i y_i = 0,\ \ 0 \leq \alpha_i \leq C \end{aligned}$$

4.2 Kernel Trick

그런데 우리가 직접 $\Phi(x)$를 매핑하고, 다시 내적 $\Phi(x_i)\Phi(x_j)$를 계산하는 일은 쉬운일이 아니다. 그렇다면 직접 내적을 계산하지 않고, 내적이 계산된 내적공간의 함수로서 $\Phi(x_i)\Phi(x_j) = K(x_i, x_j)$로서 문제를 바꾼다면 엄청나게 큰 계산비용을 절감할 수 있다. 이런 방법을 Kernel Trick이라 한다.

이런 커널 트릭의 장점은

1) 엄청나게 계산비용을 절감할 수 있다는 점
2) 특정 조건(Mercer's Condition)만 만족하면 여전히 수학적으로 문제가 없이 linear svm처럼 다룰 수 있다는 점이다.

Mercer's Condition은 SVM의 커널 함수가 가져야하는 조건이다. 이는 커널함수 $K(x_i, x_j)$를 기존 x의 매핑함수인 $\Phi(x)$의 내적공간 $\Phi(x_i)\Phi(x_j)$으로 다룰 수 있어야 함과 연관된다. 우리가 내적을 했을 때, 내적 순서를 바꾼다고 해서 값이 변하지 않고, 내적값은 언제나 0보다 크거나 같다. 이 조건을 해당 함수가 만족하면 된다. 따라서, 커널함수는 다음의 조건을 만족하면 된다.

$$1)\ K(x_i, x_j) = K(x_j, x_i) \qquad \text{(symmetic)}$$
$$2)\ K(x_i, x_j) \geq 0, \qquad \text{(positive semi-definite)}$$

4.3 Canonical Kernel

그래서 많이 쓰이는 커널을 살펴보게 되면 보통 다음의 커널을 많이 사용한다.

$$Polynomial: \quad K(x, y) = (x \cdot y + c)^p, \ \ c > 0$$
$$Gaussian(RBF): \quad K(x, y) = exp(-\frac{||x - y||^2}{2\sigma^2}), \ \ \sigma \neq 0$$

$x \in \mathbb{R}^d$를 $\mathbb{R}^D$로 보내는 polynomial mapping $\Phi(x) \in \mathbb{R}^D$를 생각하자. 이때 Polynomial Kernel의 차원은 $\binom{d+p-1}{p}$가 된다. 즉 변수가 10개 있고 5차항까지 확장시킨다면, $\mathbb{R}^{10} \rightarrow \mathbb{R}^{2002}$가 된다. 이 과정에서 각각 변수의 교차항과 상호작용항들이 자연스럽게 고려되기 때문에, 굳이 svm에서는 이런 고차항이나 상호작용항을 직접 넣어줄 필요가 없다. 다음의 예시를 확인하면, 차수를 높임에 따라 더 유연(flexible)한 분류를 할 수 있지만, 동시에 margin이 늘어나서 일반화 정도가 높아짐을 확인할 수 있다.

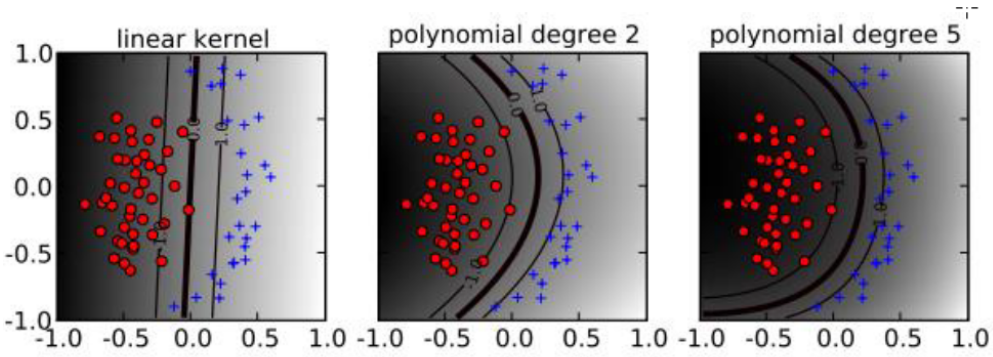


Figure 9: Example of Polynomial Kernel

Gaussian 커널은 이런 계산조차 불가능하게 무한차원으로 매핑한다. 왜냐하면 exp 함수는 테일러 근사를 하면 유한차수 다항함수의 합으로 무한하게 표현되기 때문이다. 다음의 예시를 보면 가우시안 커널의 $\gamma = \frac{1}{\sigma^2}$ 값을

조정함에 따라 결정되는 hyperplane초평면을 보여준다. γ 값을 키울수록, 더 복잡한 초평면이 나오고, 이는 곧 더 초고차원상에서 선형분리함을 의미한다. 이 복잡도를 높이고 높이면, 우리의 모든 n개의 관측치를 shatter할수도 있다.(zero-training error)

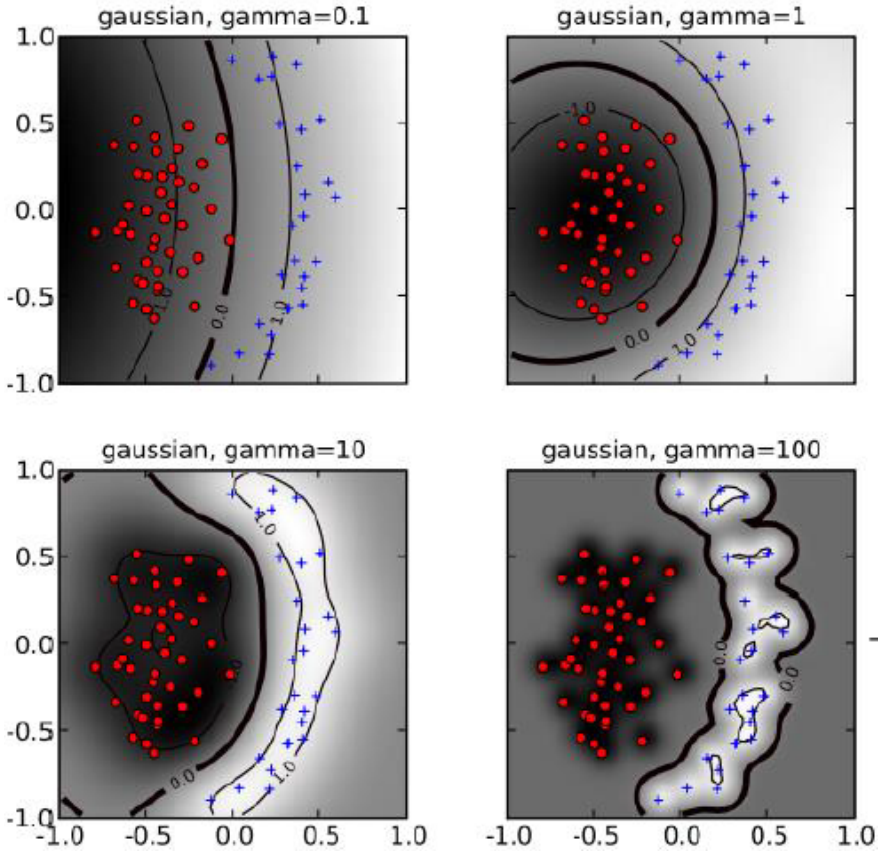


Figure 10: Example of Gaussian Kernel

그래서 우리의 최종 분류기는 $f(\mathrm{x}) = sign(\sum \alpha_i y_i K(\mathrm{x_i}, \mathrm{x_j}) + b)$를 통해 결정된다.

5 SVM의 특징

5.1 VC dimension of SVM

커널트릭을 이용해서 비선형적인 분류를 할 수 있게됨에 따라 더 좋은 분류기가 만들어지지만, 동시에 모델의 복잡도가 높아진다. 만약 극단적으로 Gaussian Kernel에서 σ^2 값을 매우매우 작게 만든다면, 엄청나게 유연한 분류기가 만들어질 수 있고, 그에 따라 모든 관측치 n개를 다 shatter할 수 있게 된다. 이는 우리가 목표인 'Structural Risk Minimization'를 달성하기 어려워 질 수 있다. 저런 극단적인 예시가 아니라도, Gaussian Kernel 자체는 무한차원으로 문제를 가져가기 때문에, 복잡도가 매우 높다. 하지만 우리는 실제 데이터 분석에서 SVM이 가지는 높은 성능을 보아왔다. 'Structural Risk Minimization'의 관점에서 SVM의 높은 성능을 어떻게 해석할 수 있을까?

1) SVM은 현재 데이터들이 존재하는 $\mathbb{R}^d$ 상에서 분류를 하는 것이 아니라, 훨씬 더 초고차원인 $\mathbb{R}^D$ 상에서 선형분류를 하게 된다. 만약 이런 초고차원 상에서 Margin을 충분히 극대화할 수 있다면, 모델의 VC Confidence는

적절하게 조절될 수 있다 (극단적으로 커지지 않는다). 이를 수식으로 확인하면 다음과 같다.

$$h \leq min(\frac{R^2}{\triangle^2}, d) + 1$$

$$R(\alpha) \leq R_{emp}(\alpha) + \sqrt{\frac{h(log(2l/h) + 1) - log(\eta/4)}{l}}$$

여기서 R은 입력데이터를 입력공간에서 감싸는 초구 (hypersphere) 의 반지름이고, $\triangle$은 margin이고, d는 입력공간의 차원이다. 여기서 R과 d는 고정이기 때문에, 모델의 capacity는 $\triangle$에 의존한다. 결국 마진을 최대화할 수 있다면, 모델의 generalizability가 높아져서 capacity를 통제할 수 있게 된다.

2) 또한 비선형분리를 하더라도, 적절한 파라미터 튜닝을 하게 되면 모델의 복잡도를 적정 수준으로 조정할 수 있다.

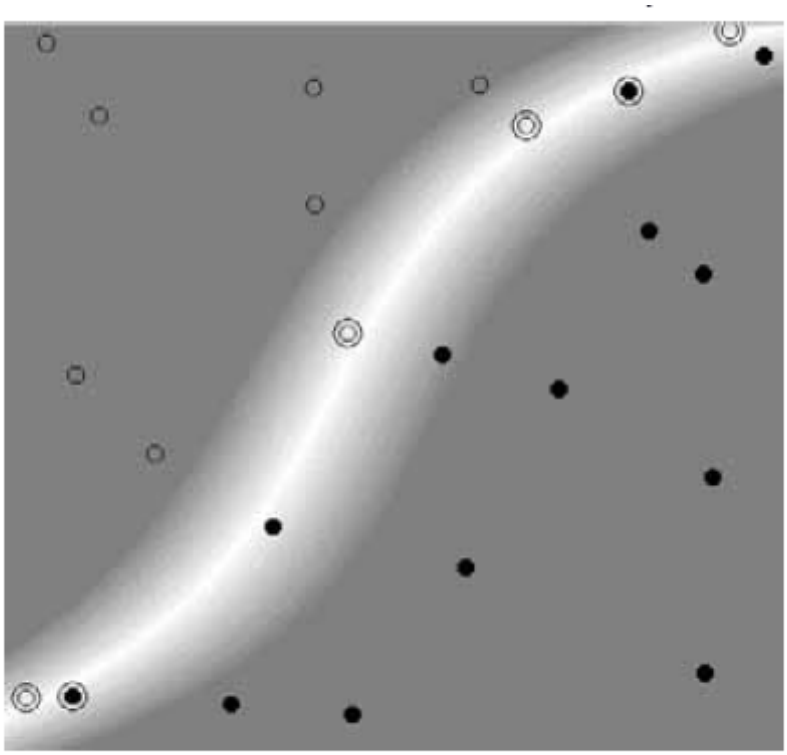


Figure 11: High-Dimensional Mapping

5.2 SVM의 장단점

단점 먼저…!

1) 커널과 파라미터 선택에 따른 성능차이가 극명하다.
2) 학습 속도가 느리고, 데이터의 개수가 매우 많을 경우 최적화를 다루는 것이 어렵다.
3) 이진분류에 비해 다항분류는 상대적으로 다루기 어렵다.

하지만 SVM의 가장 큰 장점은 Convex optimization 문제이기 때문에 언제나 global optimum을 반환한다는 점이다.