Week4: Lagrangian Dual Problem

강경훈

References

- http://web.stanford.edu/class/ee364a/lectures/duality.pdf
- https://people.eecs.berkeley.edu/~elghaoui/Teaching/EE227A/lecture7.pdf
- https://www.mathworks.com/help/stats/understanding-support-vector-machine-regression.html

Optimization with inequality constraints

다음과 같은 최적화 문제를 생각해보자.

$$\begin{aligned} \textbf{Primal Problem:} \quad & minimize \quad f_0(\mathbf{x}) \quad (\mathbf{x} \in \mathbb{R}^n,\ domain\ \mathcal{D}) \\ & s.t. \quad f_i(\mathbf{x}) \le 0, \quad {}^\forall i \in [m] \\ & \qquad\ \ h_i(\mathbf{x}) = 0 \quad {}^\forall i \in [p] \end{aligned}$$

우리가 익숙한 Lagrange Multiplier에서는 제약식이 등호로만 되어있었지만, 이제는 부등식이 추가되었다. 이러한 경우 부등호 조건식 $f_i(\mathbf{x})$를 **inequality constraints**, 등호 조건식 $h_i(\mathbf{x})$를 **equality constraints**라고 한다.

부등호 조건이 들어간 최적화 문제를 푸는 것은 참 막막한 일이다. 하지만 만일 이 최적화가 어떤 조건을 만족한다면, 우리는 이 문제를 그나마 알고리즘으로 쉽게 풀 수 있는 형태로 바꿀 수 있는데, 그 조건을 **KKT condition**이라고 하고, 원래의 최적화 문제를 살짝 바꾼 형태를 **Lagrangian Dual problem**이라고 한다.

위의 최적화 문제를 **Primal problem**이라고 하자. 이 primal에 해당하는 Lagrangian은 다음과 같이 쓸 수 있다.

$$\textbf{Lagrangian:} \quad L(\mathbf{x}, \lambda, \nu) = f_0(\mathbf{x}) + \sum_{i=1}^{m} \lambda_i f_i(\mathbf{x}) + \sum_{i=1}^{p} \nu_i h_i(\mathbf{x})$$
$$(\lambda_i \ge 0 \quad {}^\forall i \in [m])$$

<u>이제부터 우리는 $\mathbf{x}$의 도메인 전체에서 생각하지 말고, 그 중에서 조건식을 만족하는 $\mathbf{x}$에 대해서만 생각해보자.</u> 이러한 값들을 우리는 feasible한 $\tilde{\mathbf{x}}$라고 부를 것이다 ($\mathbf{x} \in \mathcal{D}^*\ feasible$). 이 feasible한 $\mathbf{x}$ 중에서, 목적함수를 최소화하는 (즉 primal 문제의 해가 되는) $\mathbf{x}_p$에 대한 목적함수의 값을 p^*라고 하자. 이게 조건식을 만족하면서 목적함수를 가장 작게 만들 수 있는 값이며, 우리의 목적은 이 $\mathbf{x}_p$를 구하는 것이다.

만일 부등호 조건이 없는 unconstrained 문제였다면 그냥 $\frac{\partial L}{\partial \mathbf{x}} = 0,\ \frac{\partial L}{\partial \lambda} = 0,\ \frac{\partial L}{\partial \nu} = 0$ 을 풀어서 나오는 연립방정식을 풀면 된다. 그러나 부등호 조건이 있는 constrained 문제에서는 다른 접근이 필요한데, 핵심은 **이 문제를 Convex Optimization 문제로 바꾸는 것**이며, Lagrangian Dual은 그 방법을 제시한다.

Lagrange Dual Function

먼저 이 **라그랑지안의 Lower bound**에 대해 생각해보자. 계수 λ_i, ν_i가 주어졌을 때, feasible한 $\mathbf{x}$를 내맘대로 움직여 가장 낮게 내려간다면 어디까지 갈 수 있을까? 이를 수식으로 보이면 다음과 같다.

$$\begin{aligned}\textbf{Lagrange Dual:}\quad g(\lambda, \nu) &= \inf_{\mathbf{x}\in\mathcal{D}^*} L(\mathbf{x}, \lambda, \nu) \\ &= \inf_{\mathbf{x}\in\mathcal{D}^*} [f_0(\mathbf{x}) + \sum_{i=1}^{m} \lambda_i f_i(\mathbf{x}) + \sum_{i=1}^{p} \nu_i h_i(\mathbf{x})]\end{aligned}$$

(infimum이란 하한선 중 가장 큰 값 (the greatest lower bound)을 의미하는 데, 예컨대 개구간 $(-1, 1)$에서의 infimum은 -1이다.)

Lagrange Dual $g(\lambda, \nu)$은 다음과 같은 중요한 특성을 가지고 있다.

1. **모든 feasible한 $\mathbf{x}$에 대해 $g(\lambda, \nu)$는 목적함수 $f_0(\mathbf{x})$의 Lower Bound이다.**

   Lagrange Dual은 λ, ν의 값에 따라 값이 정해지는 λ, ν의 함수인데, 일단 λ, ν가 정해지면 $\mathbf{x}$의 값은 feasible한 $\mathbf{x}$ 중에서 $L(\mathbf{x}, \lambda, \nu)$의 값이 infimum이 되도록 알아서 정해진다. 그런데 이 라그랑지안 $L(\mathbf{x}, \lambda, \nu)$의 값은, $\lambda_i \geq 0$이면 (부등호 조건식이 살아있다면) 항상 목적함수 $f_0(\mathbf{x})$보다 작다. 라그랑지안은 원래 목적함수에 0보다 작은 부등호 조건식을 더한 식이기 때문이다. 그러므로 우리는 모든 feasible한 $\mathbf{x}$에 대하여 Dual이 목적함수보다 작다고 말할 수 있다. 이러한 성질을 **Lower bound property**라고 한다.

$$\begin{aligned}&\textbf{Lower bound property:}\quad g(\lambda, \nu) \leq p^* \quad \text{if} \quad \lambda \geq 0 \\ &\text{proof:}\quad \text{for } \mathbf{x} \in \mathcal{D}^* \text{ and } \lambda \geq 0, \\ &\qquad p^* \geq f_0(\mathbf{x}) \geq L(\mathbf{x}, \lambda, \nu) \geq \inf_{\mathbf{x}\in\mathcal{D}^*} L(\mathbf{x}, \lambda, \nu) = g(\lambda, \nu)\end{aligned}$$

2. **$g(\lambda, \nu)$는 λ, ν에 대해 Concave하다.**

   Lagrange Dual의 식 $g(\lambda, \nu) = \inf_{\mathbf{x}\in\mathcal{D}^*}[f_0(\mathbf{x}) + \sum_{i=1}^{m} \lambda_i f_i(\mathbf{x}) + \sum_{i=1}^{p} \nu_i h_i(\mathbf{x})]$에서 중요한 것은 $f_i(\mathbf{x}) \leq 0$인 조건식이다. 어차피 feasible한 $\mathbf{x}$에 대해서는 등호 조건 $h_i(\mathbf{x})$이 0으로 죽어버리니 ν는 뭐가 되든 상관이 없다. 그러나 만일 어떤 $\mathbf{x}$에 대하여 $f_i(\mathbf{x}) < 0$ 인데, 그 앞에 붙은 계수가 $\lambda_i > 0$이면 (계수가 살아있다면), 라그랑지 듀얼의 값은 $\lambda_i = 0$일 때에 비하여 감소할 것이다. 심지어 어떤 λ_i에 대해서는 라그랑지 듀얼이 $-\infty$가 될 수도 있다. 이런 의미에서 우리는 **Lagrange Dual이 λ, ν에 대한 concave** 함수임을 알 수 있다.

   (정확히 말하면 g가 계수 λ, ν에 대해 affine function 이므로, affine family 에서의 infimum 은 concave 이기 때문이라고 하는데, 뭔 말인지는 나도 모르겠으니 대충 넘어가자. 또한 concave의 대략적인 의미는 그냥 x축에 대해 오목하여, global maximum 이 존재한다 쯤으로 생각하자. 자세한 정의는 https://en.wikipedia.org/wiki/Concave_function)

그렇다면 이제 관점을 바꾸어서, 이렇게 concave한 라그랑지안 듀얼 $g(\lambda, \nu)$의 값을 최대화하는 λ, ν를 구해보자. 즉 primal 라그랑지의 하한선인 Lagrange Dual에 대해, 이를 최대화하는 어차피 feasible한 $\mathbf{x}$에 대해서는 등호 조건식은 다 0이 되버리므로 ν는 중요하지 않으니, 이는 사실상 λ에 대한 최적화 문제와 같다. 이를 써보면 다음과 같다.

$$\begin{aligned}\textbf{Dual Problem:}\quad &maximize \quad g(\lambda, \nu) = \inf_{\mathbf{x}\in\mathcal{D}^*} L(\mathbf{x}, \lambda, \nu) \\ &s.t. \quad \lambda_i \geq 0 \quad {}^\forall i \in [m]\end{aligned}$$

이 문제는 Primal problem보다 풀기가 훨씬 수월하다. 왜냐하면 Primal Problem에서 목적함수 $f_0(\mathbf{x})$는 convex인지, concave인지, 이도저도 아닌지 알 수가 없지만, Dual Problem의 목적함수 $g(\lambda, \nu)$는 concave하므로, (concave 함수의 최대화는 곧 convex 함수의 최소화 문제와 마찬가지니까) convex optimization 알고리즘을 쓸 수 있기 때문이다. 또한 조건식이 $\lambda \geq 0$ 하나로 줄어든 것도 큰 이점이다.

이러한 Dual Problem를 최적화하는 해답을 λ_d, ν_d라고 하고, 이 때의 라그랑지안 듀얼 $g(\lambda_\mathbf{d}, \nu_\mathbf{d}) = \inf_{\mathbf{x}\in\mathcal{D}^*} L(\mathbf{x}, \lambda_\mathbf{d}, \nu_\mathbf{d})$의 값을 d^*라고 하자. 지금까지의 논의를 종합하면 d^*는 다음과 같이 쓸 수 있다.

$$d^* = \max_{\lambda,\nu} \inf_{\mathbf{x}\in\mathcal{D}^*} L(\mathbf{x}, \lambda, \nu)$$

라그랑지 듀얼이 항상 목적함수 $f_0(\mathbf{x})$의 lower bound이므려, 라그랑지 듀얼의 최댓값인 d^*도 p^* 보다 작거나 같을 것인데, 이런 당연한 성질을 **weak duality**라고 한다. 이건 우리의 관심사가 아니다. 그러나 특정한 조건을 만족하거나, 혹은 원래 Primal의 목적함수가 convex할 경우에는 대부분 $d^* = p^*$가 되는데, 이를 **strong duality**라고 하며, 이렇게 되는 조건을 **constraint qualifications**라고 한다. **이런 조건을 만족하면, 우리는 원래의 Primal 문제를 우회하여 더 수월한 Dual 문제를 풀 수 있는 것이다.**

$$\begin{aligned} \textbf{Weak Daulity:} \quad & d^* \le p^* \quad \text{(always holds)} \\ \textbf{Strong Daulity:} \quad & d^* = p^* \quad \text{(iff constraint qualifications hold)} \end{aligned}$$

Lagrange Dual: Geometric Intuition

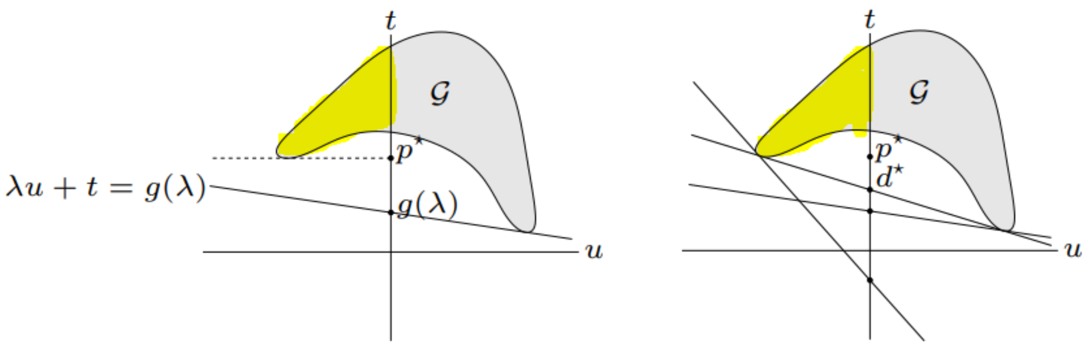
d^*, p^*의 관계를 그림으로 나타내면 다음과 같다. 여기에서 t는 목적함수 $f_0(\mathbf{x})$, u는 부등호 조건식 $f_1(\mathbf{x})$를 의미하며, 이 두 값을 결정하는 $\mathbf{x}$는 그래프에 나오지 않은 매개변수로 생각할 수 있다. $\mathcal{G}$는 $\mathbf{x}$의 도메인에서 목적함수와 조건식이 가질 수 있는 모든 값의 영역을 나타내며, feasible한 $\mathbf{x}$에 대해서는 $u \le 0$이 되어야 하므로 feasible한 영역은 노란색으로 표시한 부분이다.

Geometric interpretation

for simplicity, consider problem with one constraint $f_1(x) \le 0$

interpretation of dual function:

$$g(\lambda) = \inf_{(u,t)\in\mathcal{G}} (t + \lambda u), \qquad \text{where} \quad \mathcal{G} = \{(f_1(x), f_0(x)) \mid x \in \mathcal{D}\}$$



- $\lambda u + t = g(\lambda)$ is (non-vertical) supporting hyperplane to $\mathcal{G}$
- hyperplane intersects t-axis at $t = g(\lambda)$

(http://web.stanford.edu/class/ee364a/lectures/duality.pdf)

조건식을 만족하면서 (feasible하면서) 목적함수의 값 t를 최소화하는 지점은 노란색 영역의 가장 왼쪽 아래 끄트머리 지점이며, 이 지점에서의 매개변수가 $\mathbf{x}^*$일 것이고, 이 때의 목적함수의 값이 p^*가 될 것이다.

λ의 값이 주어졌다면 라그랑지안 듀얼

$$g(\lambda) = \inf_{\mathbf{x}\in\mathcal{D}^*} L(\mathbf{x}, \lambda) = \inf_{\mathbf{x}\in\mathcal{D}^*} [f_0(\mathbf{x}) + \lambda f_1(\mathbf{x})] = \inf_{\mathbf{x}\in\mathcal{D}^*} [t + \lambda u]$$

은, (t, u) 좌표평면에서 $t + \lambda u = g(\lambda)$를 만족하는 모든 점들의 집합으로, 직선으로 나타날 것이다. 이 직선을 **Supporting Hyperplane**이라고 한다.

Supporting Hyperplane이 가지는 중요한 특성은 다음과 같다.

1. Supporting Hyperplane은 직선이다. 이건 당연하지.
2. Supporting Hyperplane은 항상 $\mathcal{G}$보다 아래에 있어야한다. u축의 값을 하나 고정하여 그린 수직선과 $\mathcal{G}$가 만나는 지점들의 t값들은, 조건식 $u = f_1(\mathbf{x})$의 값이 주어졌을때의 목적함수 $t = f_0(\mathbf{x})$가 가질 수 있는 값들의 범위와 같다. 이때 라그랑지 듀얼은 첫 번째 성질 **"모든 feasible한 $\mathbf{x}$에 대해 $g(\lambda, \nu)$는 목적함수 $f_0(\mathbf{x})$의 Lower Bound이다."**에 의하여, u축에서 그은 수직선 위에서 항상 $\mathcal{G}$보다 아래에 있어야한다.
3. Supporting Hyperplane의 위치는 λ가 결정한다. 우선 기울기가 λ에 따라 결정되며, λ가 바뀌면 $g(\lambda) = \inf_{\mathbf{x} \in \mathcal{D}^*} L(\mathbf{x}, \lambda)$를 만족하면 $\mathbf{x}^*$의 값도 다를 것이므로 직선의 위치도 바뀐다.
4. Supporting Hyperplane은 항상 feasible한 도메인에 접해야 한다. 때문에 왼쪽 그림에서 그려진 직선은 엄밀히 말하면 $\mathbf{x}$의 feasible한 도메인 밖에 있기 때문에 그릴 수 없는 직선이다.

이를 모두 종합해 고려하면, **$g(\lambda)$를 최대화하는 λ를 고르는 Dual 문제는 $\mathcal{G}$의 feasible한 영역 중 한 점을 골라 $\mathcal{G}$를 떠받드는 supporting hyperplane을 그리는 문제**로 볼 수 있다. 또한 오른쪽 그림을 보면, λ의 값에 따라 그릴 수 있는 수많은 supporting hyperplane중에서, 위 네 가지 성질을 만족하면서 t축에서의 값 $g(\lambda)$가 최대가 되는 직선은 하나이며, 거기에서의 값이 d^*임을 알 수 있다.

위 사례는 목적함수 $f_0(\mathbf{x})$이 concave한 경우에 해당한다. 만일 $f_0(\mathbf{x})$이 convex하다면 그려지는 $\mathcal{G}$도 그 모양이 아래로 볼록하며, 그 결과 $p^* = d^*$가 될 것임을 어렵지 않게 짐작할 수 있다. 즉 **Convex optimization 문제에서는 Primal Problem을 Dual Problem으로 바꿔서 풀 수 있다는 것이다.**

그러나 이 조건만으로는 Dual Problem을 풀기에 부족하다. Convex하지만 $d^* < p^*$일 수도 있고, convex하지 않은데 $d^* = p^*$일 수도 있다. 때문에 좀 더 일반적인 조건이 필요하다.

Complementary Slackness

문제의 방향을 거꾸로 틀어, 일단 **Strong Duality가 성립한다**고 가정해보자. 이때 $\mathbf{x}^*$를 Primal 문제의 해라고 해서 Primal Optimal이라고, (λ^*, ν^*)를 Dual 문제의 해라고 해서 Dual Optimal이라고 부른다. $d^* = p^*$이므로 다음과 같이 쓸 수 있다.

$$\begin{aligned} p^* = f_0(\mathbf{x}^*) = d^* &= \inf_{\mathbf{x} \in \mathcal{D}^*} L(\mathbf{x}, \lambda^*, \nu^*) \\ &= \inf_{\mathbf{x} \in \mathcal{D}^*} [f_0(\mathbf{x}) + \sum_{i=1}^m \lambda_i^* f_i(\mathbf{x}) + \sum_{i=1}^p \nu_i^* h_i(\mathbf{x})] \\ &\le f_0(\mathbf{x}^*) + \sum_{i=1}^m \lambda_i^* f_i(\mathbf{x}^*) + \sum_{i=1}^p \nu_i^* h_i(\mathbf{x}^*) \\ &\le f_0(\mathbf{x}^*) = p^* \end{aligned}$$

첫 번째 줄과 두 번째 줄은 정의를 그대로 적었을 뿐이다. 핵심은 (1) 두 번째에서 세 번째로 내려가는 부분과, (2) 세 번째에서 막줄로 가는 부분이다. 만일 $d^* = p^*$가 성립한다고 하면 맨 처음과 맨 마지막이 다 똑같으니 중간에 있는 모든 부등호도 등호로 바뀌어야 한다. 때문에 다음이 성립한다.

1. Dual 문제를 풀어서 (λ^*, ν^*)를 구했다고 치자. 이걸 라그랑지에 넣어서 $L(\mathbf{x}, \lambda^*, \nu^*)$를 구했다고 하자. 이때 이 식을 최소화하는 $\mathbf{x}$는 바로 Primal의 해 $\mathbf{x}^*$이다. 때문에 Primal Problem을 풀 필요 없이 Dual Problem을 풀어 라그랑지 식을 $\mathbf{x}$에 대해 최소화하면 된다.
2. $\lambda_i \ge 0$이고 , $f_i(\mathbf{x}) \le 0$임을 기억하자. (2)이 의미하는 바는 다음과 같다.

$$f_0(\mathbf{x}^*) + \sum_{i=1}^m \lambda_i^* f_i(\mathbf{x}^*) + \sum_{i=1}^p \nu_i^* h_i(\mathbf{x}^*) = f_0(\mathbf{x}^*)$$

$h_i(\mathbf{x})$는 어차피 feasible한 $\mathbf{x}$에 대해서는 0이 되니 무시하자. 문제는 $f_i(\mathbf{x})$인데, 위 식이 성립하려면 모든 조건식에 대해 $\lambda_i f_i(\mathbf{x}) \leq 0$이어야만 한다. 즉 $\lambda_i^* = 0$이면 조건식이 음수든 뭐든 상관이 없지만, $\lambda_i^* \neq 0$으로 살아있다면 조건식은 무조건 0이 되어야 한다는 것이다. 이 조건을 **Complementary Slackness**라고 부른다.

이를 벡터로 생각해본다면, Primal optimal $\mathbf{x}^*$가 주어졌을 때 두 벡터 λ와 $f_i(\mathbf{x}^*)$는 complementary sparse해야한다는 것. 즉 어떤 행에서 한 놈이 0이 아니면 다른 놈이 0이 되어야 한다는 것이다. 쉽게 말하면 **람다가 살아있는 부등호 조건식은 프라이멀 옵티멀에서 0이 되어야 한다**로 생각할 수 있겠다.

Karush-Kuhn-Tucker conditions

지금까지의 논의를 종합해보자. 만일 어떤 Primal Problem에 대해 Strong duality가 성립한다면, $\mathbf{x}^*$ Primal Optimal, (λ^*, ν^*) Dual Optimal은 다음의 조건을 만족해야 하는데, 이를 통틀어 **KKT conditions**이라고 한다.

1. primal constraints: $f_i(x) \leq 0,\ i = 1, \ldots, m,\ h_i(x) = 0,\ i = 1, \ldots, p$
2. dual constraints: $\lambda \succeq 0$
3. complementary slackness: $\lambda_i f_i(x) = 0,\ i = 1, \ldots, m$
4. gradient of Lagrangian with respect to x vanishes:

$$\nabla f_0(x) + \sum_{i=1}^{m} \lambda_i \nabla f_i(x) + \sum_{i=1}^{p} \nu_i \nabla h_i(x) = 0$$

(http://web.stanford.edu/class/ee364a/lectures/duality.pdf)

1번은 primal feasibility, 2번은 dual feasibility라고도 부르는데, 이는 $\mathbf{x}^*$가 Primal Optimal, (λ^*, ν^*)가 Dual Optimal이 되기 위한 조건이므로 Strong duality와는 상관이 없다. 여기서 중요한 것은 앞서 살펴본 3번 complementary slackness와 4번 $\nabla \mathbf{x} L = 0$이다. 4번이 " Dual을 풀어서 나온 $L(\mathbf{x}, \lambda^*, \nu^*)$을 최소화하는 $\mathbf{x}$는 바로 Primal Optimal $\mathbf{x}^*$"라는 말이다.

지금까지 본 것은 **Strong Duality** $\rightarrow$ **KKT Conditions**이다. 이는 Primal의 목적함수가 convex이든 concave이든 항상 성립한다. 그러나 만일 **Primal이 convex이면 Strong Duality** $\leftarrow$ **KKT Conditions**가 성립한다. 왜 그런지는슬라이드에서 안 알랴줘서 나도 모르겠는데, 위의 geometric interpretation을 보면 대충 감이 오지 않을까? 때문에 만일 내가 푸는 최적화 문제가 부등호 조건이 들어갔는데, 이 놈의 Dual을 풀어보니 KKT condition을 만족하는 (λ^*, ν^*)이 존재하면, 그냥 Dual 문제를 풀어 나온 (λ^*, ν^*)를 대입한 라그랑지를 최소화하는 $\mathbf{x}^*$를 구하면 된다는 것.

이제 이걸 SVM에 적용해보자.

Applications: Support Vector Machines

SVM에 대한 자세한 내용은 04/23 학회 세션과 유투브 강의를 참고하도록 하자.

- https://github.com/YonseiESC/ESC20-WINTER/blob/master/ISL/lectureNotes/ISL09.pdf
- https://www.youtube.com/watch?v=DIpC35L9Ons&list=PLTGzWF3DajHQZ7zXesjid0zxmGdaNS4-K&index=36

- https://www.youtube.com/watch?v=O6Ha_XyA9ys&list=PLTGzWF3DajHQZ7zXesjid0zxmGdaNS4-K&index=37
- https://www.youtube.com/watch?v=OfykM7rnrts&list=PLTGzWF3DajHQZ7zXesjid0zxmGdaNS4-K&index=38
- https://www.youtube.com/watch?v=QZtcXkaF0m8&list=PLTGzWF3DajHQZ7zXesjid0zxmGdaNS4-K&index=39
- https://www.youtube.com/watch?v=dKcNWAWTML4&list=PLTGzWF3DajHQZ7zXesjid0zxmGdaNS4-K&index=40

여기서는 Lagrange Dual이 어떻게 SVM에 적용되는지에 대해서만 살펴보겠다.

Classification에서 SVM이란 서로 다른 클래스의 데이터 산점도 사이에 어떤 중앙선을 그리고, 그 중앙선 양 옆으로 2차선 도로를 그리는데, 그 도로의 폭이 최대한 넓도록 중앙선을 그리는 것이다. 그 중앙선을 Hyperplane으로 볼 수 있는데, 식으로 쓰면 다음과 같다.

$$\textbf{Hyperplane} \quad y(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + b = 0$$

가장 넓은 도로 폭을 그리는 문제를 최적화 문제로 나타내면 다음과 같다.

$$\begin{aligned} \textbf{Primal Problem} \quad & \min_{\mathbf{w},b} \quad \frac{1}{2}\|\mathbf{w}\|^2 \\ & s.t. \quad 1 - t_i(\mathbf{x}^\mathbf{T}\mathbf{x}_i + b) \le 0 \;\; {}^\forall i \in [n] \end{aligned}$$

이 최적화 문제의 라그랑지안과 KKT 조건은 다음과 같다.

$$\textbf{Lagrangian:} \quad L(\mathbf{w}, b, \lambda) = \frac{1}{2}\|\mathbf{w}\|^2 + \sum_{i=1}^{n} \lambda_i[1 - t_i(\mathbf{w}^\mathbf{T}\mathbf{x}_i + b)]$$

$$\textbf{KKT Conditions} \quad \begin{cases} \textbf{Primal feasibility:} & 1 - y_i(\mathbf{w}^\mathbf{T}\mathbf{x}_i + b) \le 0 \;\; {}^\forall i \in [n] \\ \textbf{Dual feasibility:} & \lambda_i \ge 0 \;\; {}^\forall i \in [n] \\ \textbf{Complementary slackness:} & \lambda_i[1 - t_i(\mathbf{w}^\mathbf{T}\mathbf{x}_i + b)] = 0 \;\; {}^\forall i \in [n] \\ \textbf{Stationarity:} & \dfrac{\partial L}{\partial w_i} = \dfrac{\partial L}{\partial b} = 0 \end{cases}$$

Primal Problem의 목적함수 $\frac{1}{2}\|\mathbf{w}\|^2$은 convex 함수이다. 때문에 Primal과 Dual의 해가 이 조건을 만족한다고 가정하면, 우리는 Dual 문제를 풀어 Primal의 해인 Hyperplane $(\mathbf{w}, b)$ 를 구할 수 있다. 듀얼을 푸는 방법은 다음과 같다. 프라이멀과 듀얼의 옵티멀은 KKT 조건을 만족할 것이다. 때문에 **KKT 조건을 $(\mathbf{w}, b, \lambda)$에 대한 연립방정식으로 보고 문제를 풀면 된다.**

SVM에서의 KKT 조건 중에 가장 중요한 것은 **Complementary Slackness**이다. 제대로 분류된 데이터는 모두 중앙선 도로 밖에 있으니 $t_i(\mathbf{w}^\mathbf{T}\mathbf{x}_i + b) > 1$이며, 때문에 이에 해당하는 부등호 조건식 $1 - t_i(\mathbf{x}^\mathbf{T}\mathbf{x}_i + b)$은 0보다 작은 음수이다. 이때 KKT 조건에 의해 이 조건식에 해당하는 λ_i는 0이 되어야 한다. **이는 즉 SVM에서 도로 가에 위치한 Support vector 외에 다른 모든 관측치들은 hyperplane의 결정에 아무런 영향이 없다는 것이다.**

먼저 네 번째 조건에 따라 라그랑지안을 $\mathbf{w}, b$에 대해 미분한다.

$$\nabla_\mathbf{w} L|_{(\mathbf{w}=\mathbf{w}^*)} = \mathbf{w}^* - \sum_{i=1}^{n} \lambda_i t_i \mathbf{x}_i = 0$$

$$\frac{\partial L}{\partial b}|_{b=b^*} = \sum_{i=1}^{n} \lambda_i t_i = 0$$

이를 라그랑지안 $L(\mathbf{w}, b, \lambda)$에 대입하면 Dual의 목적함수 $L(\mathbf{w}^*, \mathbf{b}^*, \lambda)$를 구할 수 있다.

$$L(\mathbf{w^*}, \mathbf{b^*}, \lambda) = \sum^n \lambda_i - \frac{1}{2}\sum^n_{i,j} t_i t_j \lambda_i \lambda_j \langle \mathbf{x}_i, \mathbf{x}_j \rangle$$

이렇게 해서 Dual Problem을 써보면

$$\begin{aligned} \textbf{Dual Problem} \quad & \max_{\lambda} L(\mathbf{w^*}, \mathbf{b^*}, \lambda) = \sum^n \lambda_i - \frac{1}{2}\sum^n_{i,j} t_i t_j \lambda_i \lambda_j \langle \mathbf{x}_i, \mathbf{x}_j \rangle \\ s.t. \quad & \lambda_i \geq 0 \ \ ^\forall i \in [n] \\ & \sum_{i=1}^n \lambda_i t_i = 0 \end{aligned}$$

이렇게 α^*를 구하고 나면 $\mathbf{w}^*$는 $\frac{\partial L}{\partial w_i} = 0$ 인 조건을 이용해 구할 수 있으며, b^*는 Hyperplane의 위치를 생각해보면 아래처럼 쉽게 구할 수 있다.

$$\begin{aligned} \mathbf{w}^* &= \sum^n \lambda_i t_i \mathbf{x}_i \\ b^* &= -\frac{1}{2}(\max_{i|t_i=-1} \mathbf{w}^{*T}\mathbf{x}_i + \min_{i|t_i=1} \mathbf{w}^{*T}\mathbf{x}_i) \end{aligned}$$

Kernel SVM for Non-linear Decision Boundary

이 식을 보면, 새로운 관측치 벡터 $\mathbf{x}_h$가 주어졌을 때 이는 어디로 분류될 것인지는 전적으로 새로운 벡터와 기존 training set의 모든 벡터와의 내적에 의해 결정되는 것을 알 수 있다.

$$\begin{aligned} \mathbf{w^T}\mathbf{x}_h + b &= (\sum_{i=1}^n \lambda_i t_i \mathbf{x}_i)^T \mathbf{x}_h + b \\ &= \sum_{i=1}^n \lambda_i t_i \langle \mathbf{x}_i, \mathbf{x}_h \rangle + b \end{aligned}$$

때문에 만일 우리가 관측치 벡터에 대해 feature extraction을 했다면, 내적의 자리에 $\langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_h) \rangle$를 집어 넣으면 되는 것이고, 만일 feature extraction 없이 Mercer Theorem을 이용해 곧바로 kernel를 구했다면, 그 함수를 저 자리에다가 넣기만 하면 되는 것이다.

중요한 점은 앞서 살펴봤듯이 complementary slackness에 의해 support vector 외의 점에서는 $\lambda_i = 0$이므로, 결국은 새로운 데이터의 예측을 위해서 support vector와의 내적만 계산하면 된다는 것. PRML 책 6장의 6.1절을 보면 알겠지만 Linear Regression의 경우 새로운 데이터의 예측을 위해 모든 점과의 내적을 구해야 했다. 그러나 SVM은 몇몇 벡터와의 내적만 계산하면 된다. 때문에 SVM을 **Sparse Kernel** machine이라고도 부른다.

Soft-Penailizing Errors

만일 도로 안에 몇 개 데이터가 있어도 되고, 심지어 중앙선을 넘어도 허용하는 soft-penalizing의 경우 Primal의 조건식이 이에 맞게 살짝 바뀌는데, 핵심은 $t_i(\mathbf{x^T}\mathbf{x}_i + b) \leq 1 - \epsilon_i$으로 조금의 오차가 허용되는 대신, 그 오차의 총량 $C\sum_{i=1}^n \epsilon_i$을 목적함수에 갖다넣어서 오차의 총량을 규제하는 것이다. 이 때도 마찬가지로 KKT 조건을 이용해 Dual 문제를 풀면 위와 동일한 결과가 나온다. 자세한 내용은 세션을 참고하자.

Support Vector Regression

Regression에서 SVM은 Classification과 반대로 어떤 Hyperplane을 그리는데, 중앙선 양 옆 2차선 도로에 최대한 많은 데이터를 집어넣을 수 있는 중앙선을 그리는 방법이다. 이때 도로의 폭은 **개별 관측치가 도로 밖에 얼마나 멀리 떨어져 있어도 되냐**에 따라 결정되는데, 만일 도로에서 탈선할 정도를 최대한 적게 하려면 데이터의 분포에 꼭 잘 들어맞는 도로가 만들어질 것이며, 반대로 탈선할 정도를 많이 눈감아주면 데이터의 모양에 대략적으로 들어맞는 도로가 나올 것이다.

SV regression에 해당하는 Primal Problem은 다음과 같이 쓸 수 있다.

$$J(\beta) = \frac{1}{2}\beta'\beta + C\sum_{n=1}^{N}\left(\xi_n + \xi_n^*\right).$$

subject to:

$$\forall n : y_n - (x_n{}'\beta + b) \le \varepsilon + \xi_n$$
$$\forall n : (x_n{}'\beta + b) - y_n \le \varepsilon + \xi_n^*$$
$$\forall n : \xi_n^* \ge 0$$
$$\forall n : \xi_n \ge 0\,.$$

(https://www.mathworks.com/help/stats/understanding-support-vector-machine-regression.html)

여기서 ϵ은 도로의 폭인데, 뭐가 되는 크게 상관이 없다. 중요한 것은 탈선한 정도인 ξ 와, 탈선한 정도의 총량을 규제하는 상수 C이다. C의 값이 크면 탈선을 많이 규제하는 것이다. C의 값에 따른 SVR의 결과는 다음과 같다.

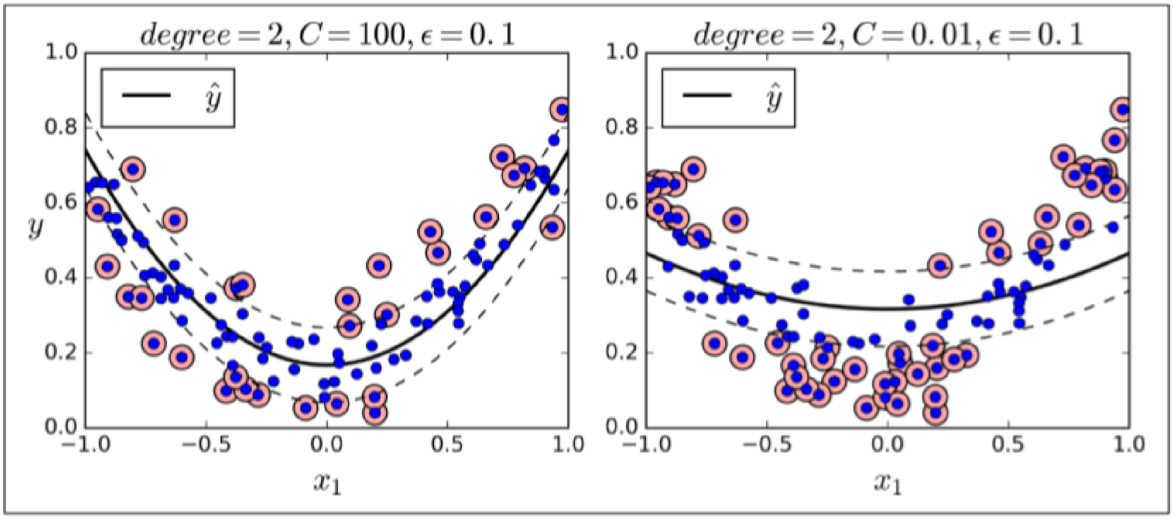


Figure 5-11. SVM regression using a 2nd-degree polynomial kernel

(Hands on Machine Learning)