

Machine Learning Lecture Note

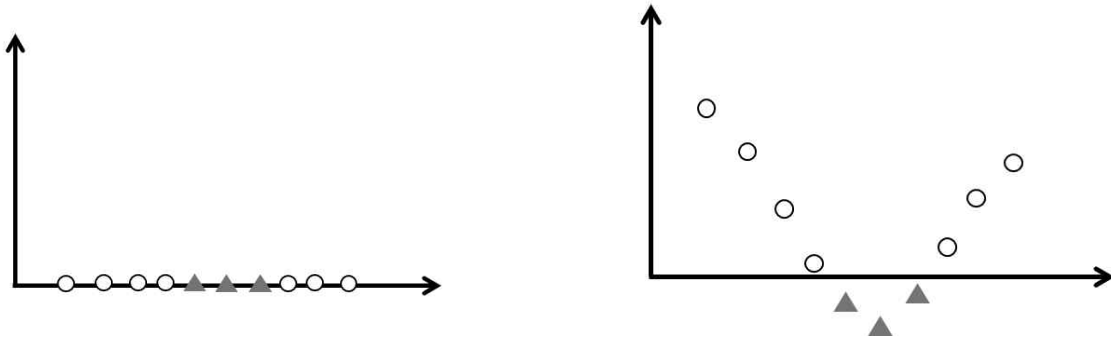
This lecture note is based on the CS229 lecture materials from Stanford University.

Lecture 6. Kernel

내적 (inner product) $\langle x, z \rangle = x^T z$

비선형 특징의 데이터는 선형 분류가 어렵다. 이 경우, 데이터를 고차원 매핑하여 선형 분류가 가능한 feature space를 생성한다. Feature mapping 함수 ϕ 가 주어지면, 커널(Kernel)은 다음과 같이 나타낼 수 있다.

$$K(x, z) = \phi(x)^T \phi(z)$$



알고리즘에서 $\langle x, z \rangle$ 는 $K(x, z)$ 로 대체할 수 있고, 알고리즘은 ϕ 의 feature를 사용하여 학습을 수행한다. $K(x, z)$ 를 연산하기 위한 비용은 많이 들지 않을 수 있지만, ϕ 에 대한 연산은 고차원 데이터를 연산해야 하기 때문에 많은 연산 비용이 발생한다. $K(x, z)$ 의 연산을 효율적으로 줄이기 위한 방법이 커널 트릭 (Kernel trick)이다.

$x, z \in R^n$ 이고 $K(x, z) = (x^T z)^2$ 으로 주어진 예를 살펴보자.

$$K(x, z) = (x^T z)^2 = \left(\sum_{i=1}^n x_i z_i \right) \left(\sum_{j=1}^n x_j z_j \right) = \sum_{i=1}^n \sum_{j=1}^n x_i x_j z_i z_j = \sum_{i,j=1}^n (x_i x_j) (z_i z_j) = \phi(x)^T \phi(z)$$

$$x = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}, x \in R^1, O(n)$$

$$\phi(x) = \begin{bmatrix} x_1 x_1 \\ x_1 x_2 \\ x_1 x_3 \\ x_2 x_1 \\ x_2 x_2 \\ x_2 x_3 \\ x_3 x_1 \\ x_3 x_2 \\ x_3 x_3 \end{bmatrix}, \phi(x) \in R^2, O(n^2)$$

$K(x, z)$ 연산에는 $O(n)$ 이 걸리지만, 고차원으로 feature mapping 한 $\phi(x)$ 는 $O(n^2)$ 이 걸린다.

이처럼, 커널 트릭은 커널 함수의 결과가 고차원 매핑 함수의 결과와 동일하게 맞춰, 연산은 낮은 차원에서 이루어지지만 데이터는 고차원에서 매핑한 효과를 나타내는 것이다.

또 다른 커널 $K(x, z) = (x^T z + c)^2$ 를 살펴보자.

$$K(x, z) = (x^T z + c)^2 = \sum_{i,j=1}^n (x_i x_j) (z_i z_j) + \sum_{i=1}^n (\sqrt{2c} x_i) (\sqrt{2c} z_i) + c^2$$

$$\phi(x) = \begin{bmatrix} x_1x_1 \\ x_1x_2 \\ x_1x_3 \\ x_2x_1 \\ x_2x_2 \\ x_2x_3 \\ x_3x_1 \\ x_3x_2 \\ x_3x_3 \\ \sqrt{2c}x_1 \\ \sqrt{2c}x_2 \\ \sqrt{2c}x_3 \\ c \end{bmatrix}, \phi(x) \in R^2, O(n^2)$$

다항식에 대한 커널로 일반화하면,

커널은 $K(x, z) = (x^T z + c)^d$ 이고, 데이터는 $\binom{n+d}{d}$ 의 feature space에 매핑된다. $\phi(x)$ 는 $O(n^d)$ 의 연산 시간을 가지지만, $K(x, z)$ 는 여전히 $O(n)$ 의 연산 시간을 가진다.

Optimal margin classifier + Kernel trick = Support Vector Machine (SVM)

만약 $\phi(x)$ 와 $\phi(z)$ 가 서로 가까이 있어 유사하다면 $K(x, z) = \phi(x)^T \phi(z)$ 는 커지고, 만약 $\phi(x)$ 와 $\phi(z)$ 가 서로 멀리 떨어져 있어 다르다면 (즉 거의 직교 한다면) $K(x, z) = \phi(x)^T \phi(z)$ 는 작아진다.

이러한 직관을 가지고 Gaussian kernel을 살펴보자.

$$K(x, z) = \exp\left(-\frac{\|x - z\|^2}{2\sigma^2}\right) \quad \begin{array}{l} x \text{와 } z \text{가 유사하면 1에 가깝고, 서로 멀리 떨어져} \\ \text{있으면 0에 가깝다. 이 특징을 SVM에 활용한다.} \end{array}$$

커널은 어떻게 정의할 수 있나?

어떤 feature mapping ϕ 에 해당하는 유효한 커널 K 를 가정하면,

$$K(x, x) = \phi(x)^T \phi(x) \geq 0$$

Let $\{x^1, \dots, x^n\}$ be n points; Let $K \in R^{n \times n}$; $K_{ij} = K(x^i, x^j) = K(x^j, x^i) = K_{ji}$ (symmetric)

Given any vector z ,

$$\begin{aligned} z^T K z &= \sum_i \sum_j z_i K_{ij} z_j = \sum_i \sum_j z_i \phi(x^i)^T \phi(x^j) z_j \\ &= \sum_i \sum_j z_i \sum_k (\phi(x^i))_k (\phi(x^j))_k z_j \\ &= \sum_k \sum_i \sum_j z_i (\phi(x^i))_k (\phi(x^j))_k z_j \\ &= \sum_k \left(\sum_i z_i (\phi(x^i))_k \right)^2 \geq 0 \end{aligned}$$

So, $K \geq 0$

Theorem (Mercer)

커널 $K: R^n \times R^n \rightarrow R$ 가 주어졌을 때, K 가 유효한 커널 함수가 되기 위해서는 n 개의 데이터 포인트들에 대해 해당 kernel matrix 가 0보다 큰 경우 ($K \geq 0$)에만 가능하다.

커널의 종류

Linear kernel	$K(x, z) = x^T z$
Polynomial kernel	$K(x, z) = (x^T z + c)^d$
Gaussian kernel	$K(x, z) = \exp\left(-\frac{\ x - z\ ^2}{2\sigma^2}\right)$

Regularization

비선형적으로 분리가능한 데이터 세트에 대해 알고리즘을 작동시키고 이상치에 대해 덜 민감하게 만들기 위해 l_1 -regularization을 사용하여 최적화를 재구성한다.

L1 norm soft margin SVM

$$\begin{aligned} \min_{w, b} \quad & \frac{1}{2} \|w\|^2 + C \sum_{i=1}^m \xi_i \\ \text{s.t.} \quad & y^i (w^T x^i + b) \geq 1 - \xi_i, \quad i = 1, \dots, m \\ & \xi_i \geq 0, \quad i = 1, \dots, m \end{aligned}$$

Lagrangian:

$$L(w, b, \xi, \alpha, \beta) = \frac{1}{2} w^T w + C \sum_{i=1}^m \xi_i - \sum_{i=1}^m \alpha_i [y^i (w^T x^i + b) - 1 + \xi_i] - \sum_{i=1}^m \beta_i \xi_i$$

Dual problem

$$\begin{aligned} \max_{\alpha} \quad & W(\alpha) = \sum_{i=1}^m \alpha_i - \frac{1}{2} \sum_{i,j=1}^m \alpha_i \alpha_j y^i y^j \langle x^i, x^j \rangle \\ \text{s.t.} \quad & 0 \leq \alpha_i \leq 1 \quad i = 1, \dots, m \\ & \sum_{i=1}^m \alpha_i y^i = 0 \end{aligned}$$