Support Vector Machines

Igor Adamiec

Classification with Non-Linear Decision Boundaries

- Aby znaleźć linię, płaszczyznę itd., które rozdzielą dane nieliniowe należy rozważyć zwiększenie ilości zmiennych (zwiększenie wymiarowości.)

Zamiast $X_1, X_2, \ldots, X_p,$ będzie $X_1, X_1^2, X_2, X_2^2, \ldots, X_p, X_p^2.$

$$\underset{\beta_0,\beta_{11},\beta_{12}\ldots,\beta_{p1},\beta_{p2},\epsilon_1,\ldots,\epsilon_n,M}{\text{maximize}} \quad M \qquad (9$$

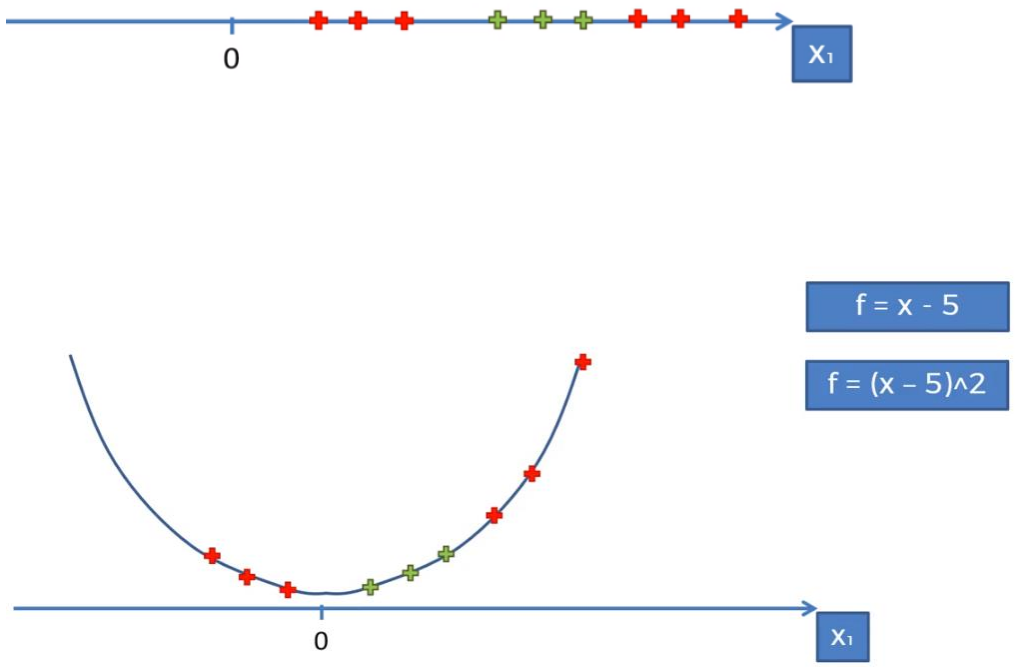
$$\text{subject to } y_i \left(\beta_0 + \sum_{j=1}^{p} \beta_{j1} x_{ij} + \sum_{j=1}^{p} \beta_{j2} x_{ij}^2 \right) \geq M(1 - \epsilon_i),$$

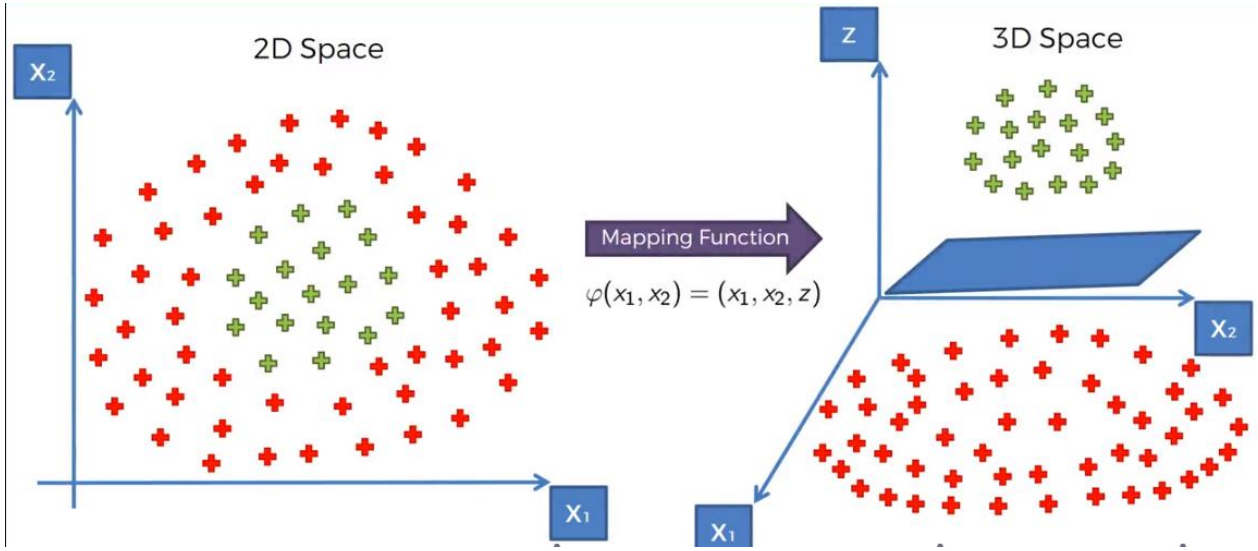
$$\sum_{i=1}^{n} \epsilon_i \leq C, \quad \epsilon_i \geq 0, \quad \sum_{j=1}^{p} \sum_{k=1}^{2} \beta_{jk}^2 = 1.$$

Oznacza to, że granica decyzji jest liniowa, ale w oryginalnej ilości zmiennych przyjmuje postać wielomianową.

Oczywiście można dodać wyższe rzędy wielomianów.

Podnoszenie wymiarowości obrazowo





SVM to rozszerzenie suport vector classifiera, które skutkuje zwiększeniem wymiarowości poprzez użycie kerneli

Dot product $\langle a, b \rangle = \sum_{i=1}^{r} a_i b_i.$

$$\langle x_i, x_{i'} \rangle = \sum_{j=1}^{p} x_{ij} x_{i'j}.$$

Kernel liniowy

$$f(x) = \beta_0 + \sum_{i=1}^{n} \alpha_i \langle x, x_i \rangle,$$

Mamy n parametrów alfa (tyle ile obserwacji.
Aby obliczyć parametry alfa potrzebujemy

$$\binom{n}{2} = \frac{n!}{2 * (n-2)!} = \frac{n * (n-1) * (n-2)!}{2 * (n-2)!} = \frac{n(n-1)}{2}$$

Okazuje się, że parametry alfa są zerowe dla treningowych obserwacji, które nie są wektorami wspornymi.
Możemy więc uprościć równanie do (s to zbiór wektorów wspornych):

$$f(x) = \beta_0 + \sum_{i \in \mathcal{S}} \alpha_i \langle x, x_i \rangle,$$

Załóżmy, że każdy inner product, który pojawia się w rozwiązaniu z poprzedniego slajdu zastępujemy uogólnieniem.

$$K(x_i, x_{i'}),$$

K to funkcja nazwana dalej kernelem. Określa ona podobieństwo dwóch obserwacji. Najprościej pokazać ją w poniższy sposób, gdzie jest ona zwykłym liniowym svc.

$$K(x_i, x_{i'}) = \sum_{j=1}^{p} x_{ij} x_{i'j},$$

Polynomial kernel

$$K(x_i, x_{i'}) = (1 + \sum_{j=1}^{p} x_{ij} x_{i'j})^d.$$

$$f(x) = \beta_0 + \sum_{i \in \mathcal{S}} \alpha_i K(x, x_i).$$

Radial kernel

$$K(x_i, x_{i'}) = \exp(-\gamma \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2).$$

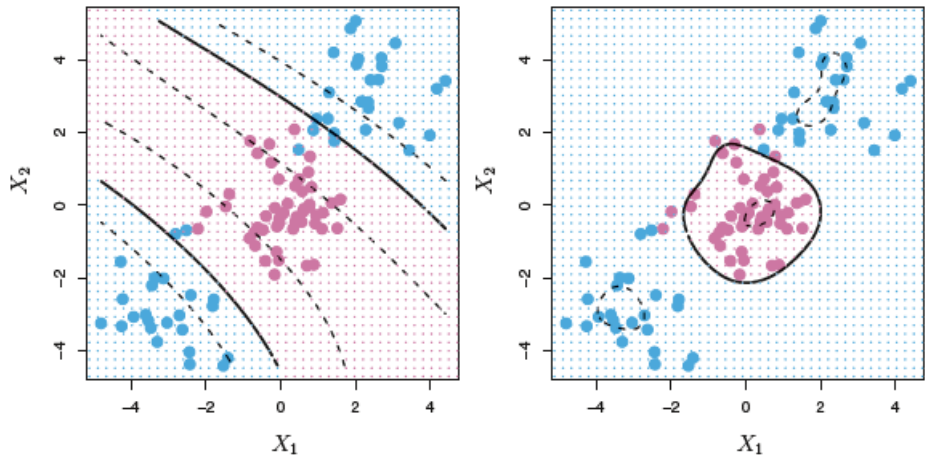


FIGURE 9.9. Left: *An SVM with a polynomial kernel of degree 3 is applied to the non-linear data from Figure 9.8, resulting in a far more appropriate decision rule.* Right: *An SVM with a radial kernel is applied. In this example, either kernel is capable of capturing the decision boundary.*

Jak działa radial kernel

γ jest dodatnią stałą.

Jeżeli obserwacja testowa

$x^* = (x_1^* \dots x_p^*)^T$

Jest daleko od danej obserwacji x, to

$\sum_{j=1}^{p}(x_j^* - x_{ij})^2$ Jest duże

A wartość funkcji kernelowej jest mała. Oznacza to, że treningowe obserwacje, które są daleko od danej obserwacji testowej, nie mają wpływu na przewidywanie wyniku klasyfikacji.

Używanie kerneli jest dużo mniej obciążające dla maszyny niż zwiększanie wymiarowości

Dla więcej niż dwóch klas

- One-Versus-One

Tworzymy kombinację $\binom{K}{2}$ porównań klas (każda z każdą). Testową obserwację testujemy na każdym modelu i wybieramy tę klasę, która pojawiała się najczęściej

- One-Versus-All

Dopasowujemy K modeli, które porównują K-tą klasę ze wszystkimi pozostałymi (K-1). Przypasowanej klasie przyporządkowujemy wartość 1, a nieprzyporządkowanej -1. Przypisujemy obserwację do klasy, dla której $\beta_{0k}+\beta_{1k}x_1^*+\beta_{2k}x_2^*+\ldots+$ jest największe.