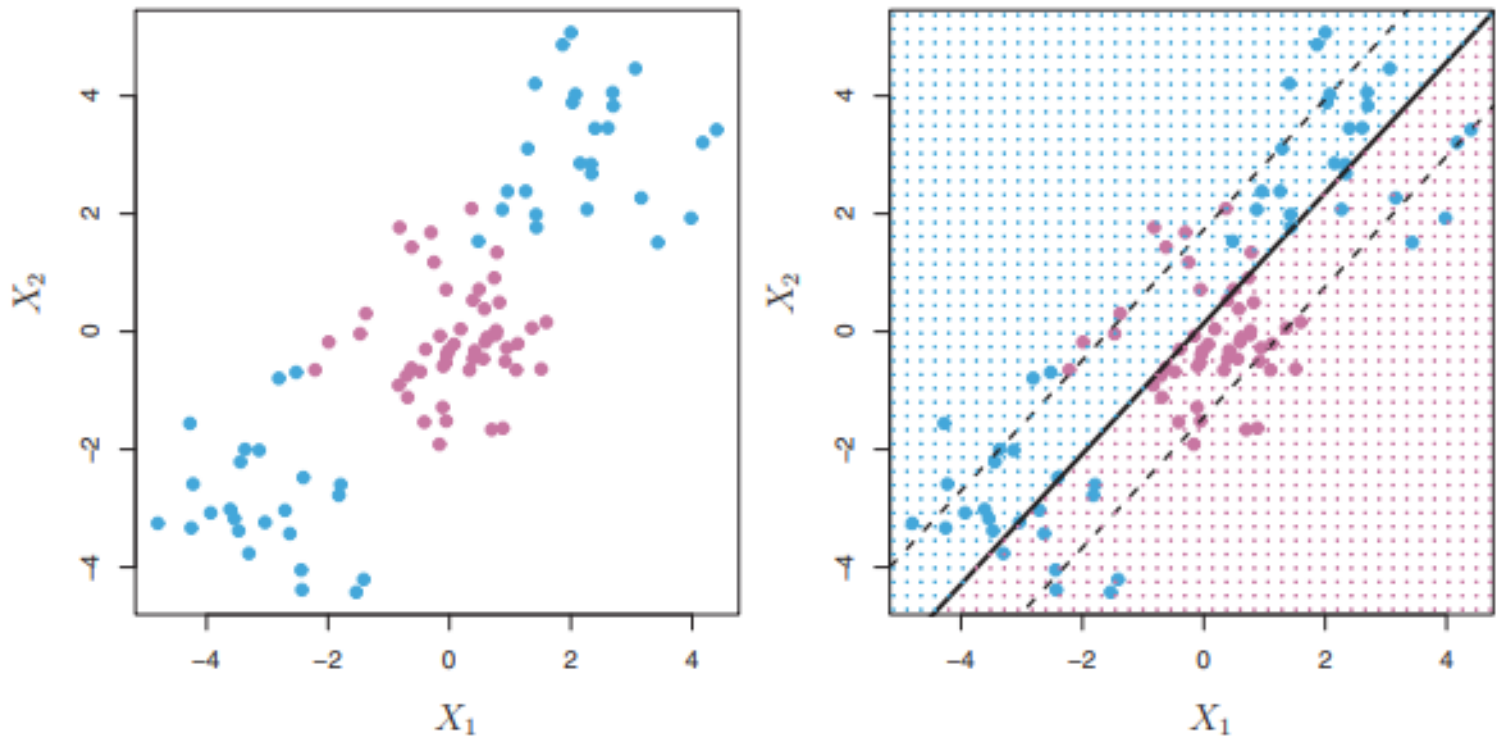
Support Vector Machines

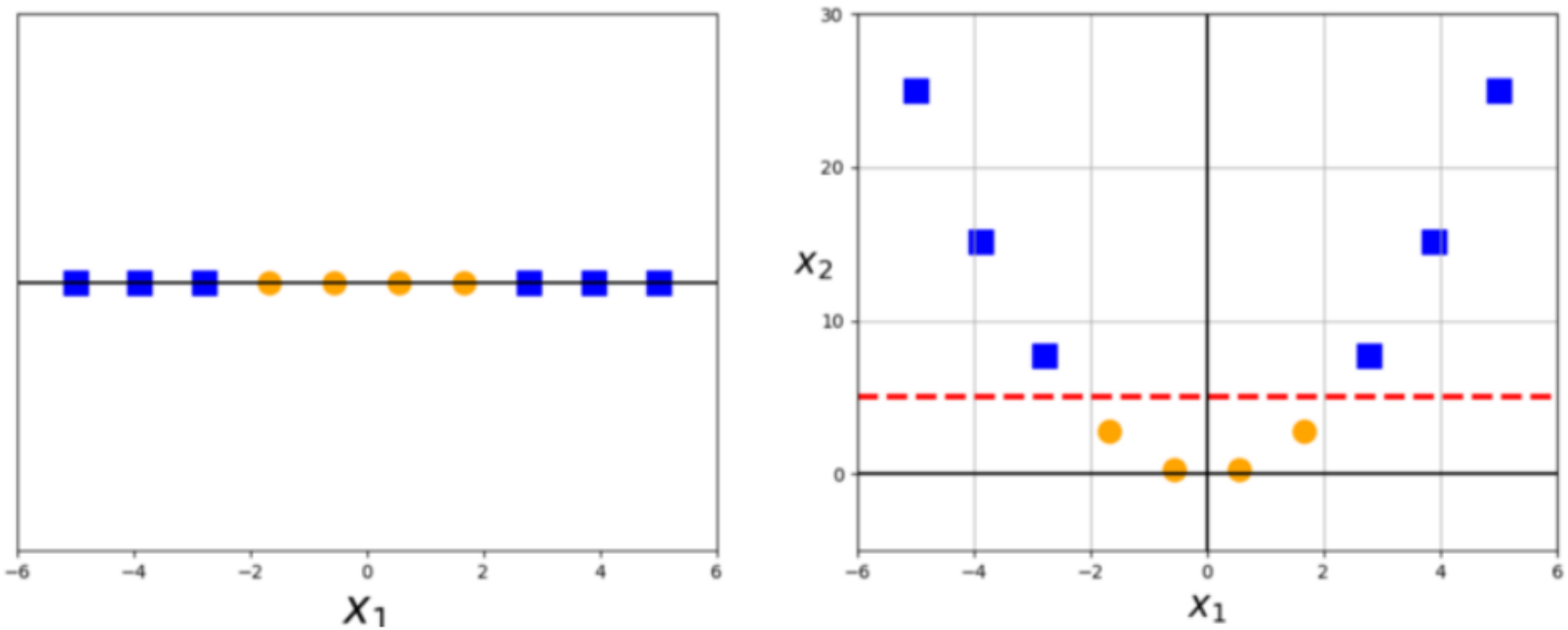
Zofia Dziedzic

Uniwersytet Wrocławski



Rysunek: Działanie klasyfikatora wektorów nośnych w przypadku, gdy obserwacje dzielą się na dwa zbiory rozdzielone nieliniową granicą. Źródło: „Introduction to Statistical Learning"

Transformacja predyktorów – intuicja



Rysunek: Źródło: https://towardsdatascience.com/the-kernel-trick-c98cdbcaeb3f

Można powiększać przestrzeń cech, dodając funkcje obecnych predyktorów, np. ich kwadraty.

$$X_1, X_1^2, X_2, X_2^2, \ldots, X_p, X_p^2$$

Wtedy problem optymalizacyjny wygląda następująco

$$\underset{\beta_0,\beta_{11},\beta_{12},\ldots,\beta_{p1},\beta_{p2},\epsilon_1,\ldots,\epsilon_n}{\text{maksymalizuj}} \quad M$$

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

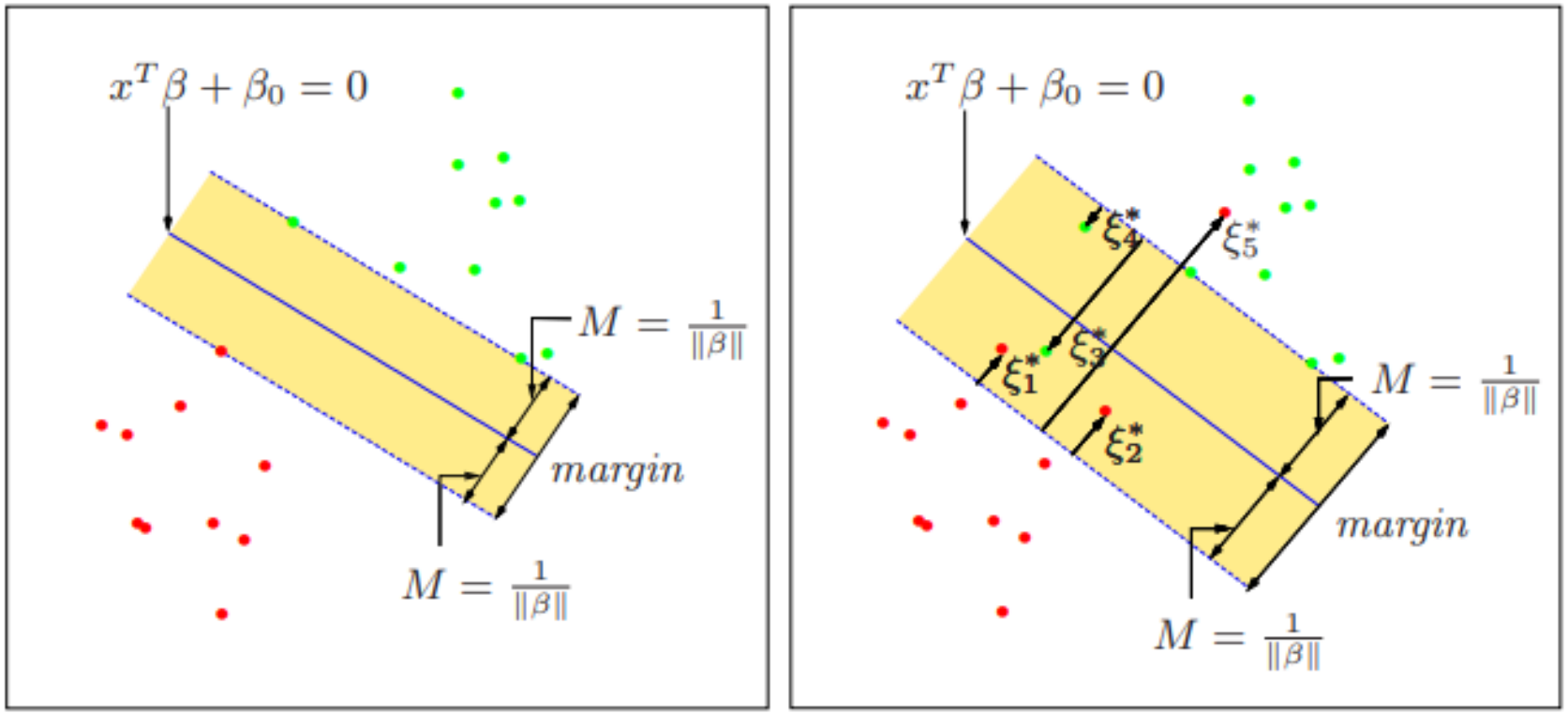
$$\sum_{i=1}^{n} \epsilon_i, \epsilon_i \geq 0, \sum_{j=1}^{p} \sum_{k=1}^{2} \beta_{jk}^2 = 1$$

Takie podejście jest bardzo złożone obliczeniowo, a dla większej liczby cech niemal niewykonalne. Maszyny wektorów nośnych (SVM) pomagają poradzić sobie z tym problemem. Aby pokazać ich działanie, pokażemy najpierw, jak wygląda rozwiązanie problemu optymalizacyjnego klasyfikatora wektorów nośnych.

Użyjemy dla wygody obliczeń równoważnego zapisu tego zagadnienia

$$\min_{\beta_0,\beta} \frac{1}{2}||\beta|| + C\sum_{i=1}^{N} \xi_i$$

$$\text{pod warunkiem } y_i(x_i^T\beta + \beta_0) \geq 1 - \xi_i \ \ \forall i$$



Rysunek: Źródło: „The Elements of Statistical Learning"

$$L_P = \frac{1}{2}\|\beta\|^2 + C\sum_{i=1}^{N}\xi_i - \sum_{i=1}^{N}\alpha_i[y_i(x_i^T\beta + \beta_0) - (1-\xi_i)] - \sum_{i=1}^{N}\mu_i\xi_i$$

$$\begin{aligned}
\beta &= \sum_{i=1}^{N}\alpha_i y_i x_i, \\
0 &= \sum_{i=1}^{N}\alpha_i y_i, \\
\alpha_i &= C - \mu_i,\ \forall i,
\end{aligned}$$

$$L_D = \sum_{i=1}^{N}\alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{i'=1}^{N}\alpha_i\alpha_{i'}y_i y_{i'}x_i^T x_{i'},$$

$$\hat{\beta} = \sum_{i=1}^{N}\hat{\alpha}_i y_i x_i$$

Klasyfikator SVC

Ostatecznie klasyfikator przyjmuje więc postać

$$\hat{f}(\mathrm{x}) = \mathrm{x}^T \hat{\beta} + \hat{\beta}_0 = \sum_{i=1}^{N} \hat{\alpha}_i y_i \langle \mathrm{x}, \mathrm{x}_i \rangle + \hat{\beta}_0.$$

Podobnie dla przypadku predyktorów przekształconych przez funkcję h będziemy mieli

$$\hat{f}(\mathrm{x}) = \sum_{i=1}^{N} \hat{\alpha}_i y_i \langle h(\mathrm{x}), h(\mathrm{x}_i) \rangle + \hat{\beta}_0.$$

Jak widzimy, w tym rozwiązaniu istotne są więc jedynie iloczyny skalarne obserwacji testowej i obserwacji treningowych (a nawet więcej – tylko wektorów nośnych).

Funkcja jądra

Możemy zastąpić ten iloczyn skalarny pewnym jego uogólnieniem. Taką funkcję nazwiemy funkcją jądra (kernel function). Określa ona podobieństwo między dwiema obserwacjami.

Kernel function

Dla $x, x' \in X$ i funkcji rzutującej $h : X \longrightarrow \mathcal{R}^n$ funkcją jądra nazywamy funkcję K, taką że

$$K(x, x') = \langle h(x), h(x') \rangle$$

Jest to więc funkcja, która przyjmuje za argumenty wektory z oryginalnej przestrzeni, a zwraca iloczyn skalarny wektorów z przestrzeni cech (feature space).

$$\mathbf{x} = (x_1, x_2, x_3)^T$$
$$\mathbf{y} = (y_1, y_2, y_3)^T$$

$$\phi(\mathbf{x}) = (x_1^2, x_1x_2, x_1x_3, x_2x_1, x_2^2, x_2x_3, x_3x_1, x_3x_2, x_3^2)^T$$
$$\phi(\mathbf{y}) = (y_1^2, y_1y_2, y_1y_3, y_2y_1, y_2^2, y_2y_3, y_3y_1, y_3y_2, y_3^2)^T$$

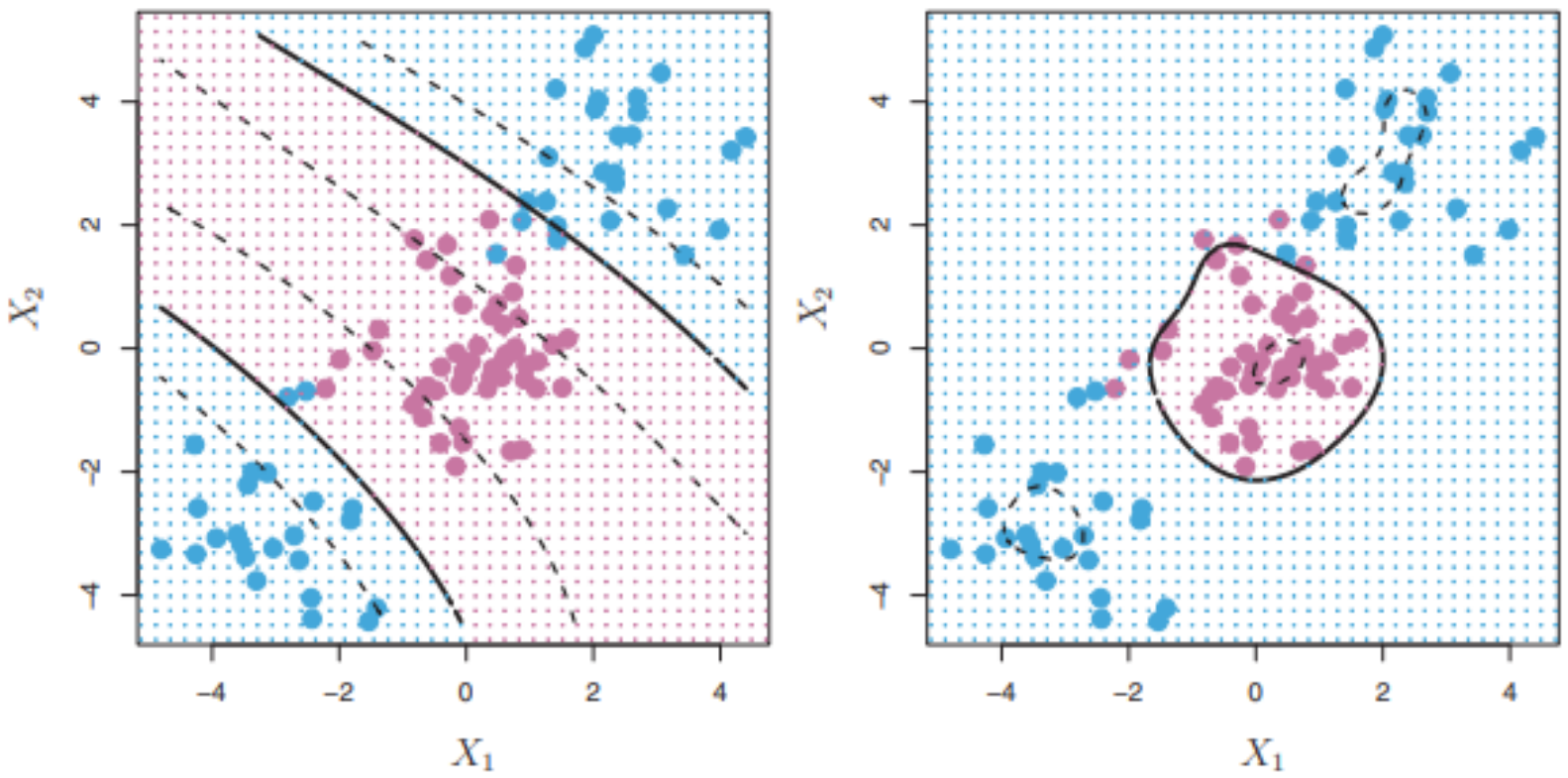
$$\phi(\mathbf{x})^T\phi(\mathbf{y}) = \sum_{i,j=1}^{3} x_ix_jy_iy_j$$

$$\begin{aligned} k(\mathbf{x}, \mathbf{y}) &= (\mathbf{x}^T\mathbf{y})^2 \\ &= (x_1y_1 + x_2y_2 + x_3y_3)^2 \\ &= \sum_{i,j=1}^{3} x_ix_jy_iy_j \end{aligned}$$

Przykłady funkcji jądra

Najczęściej stosowane funkcje jądra:

$$\begin{aligned} d\text{th-Degree polynomial: } & K(x, x') = (1 + \langle x, x' \rangle)^d, \\ \text{Radial basis: } & K(x, x') = \exp(-\gamma \|x - x'\|^2), \\ \text{Neural network: } & K(x, x') = \tanh(\kappa_1 \langle x, x' \rangle + \kappa_2) \end{aligned}$$



Rysunek: Po lewej: przykład zastosowania wielomianowego jądra stopnia 3 do danych z Rys. 1. Po prawej: jądro radialne. Źródło: „Introduction...”

$$\exp\left(-\frac{1}{2}\|\mathbf{x}-\mathbf{x}'\|^2\right) = \exp(\frac{2}{2}\mathbf{x}^\top\mathbf{x}' - \frac{1}{2}\|\mathbf{x}\|^2 - \frac{1}{2}\|\mathbf{x}'\|^2)$$

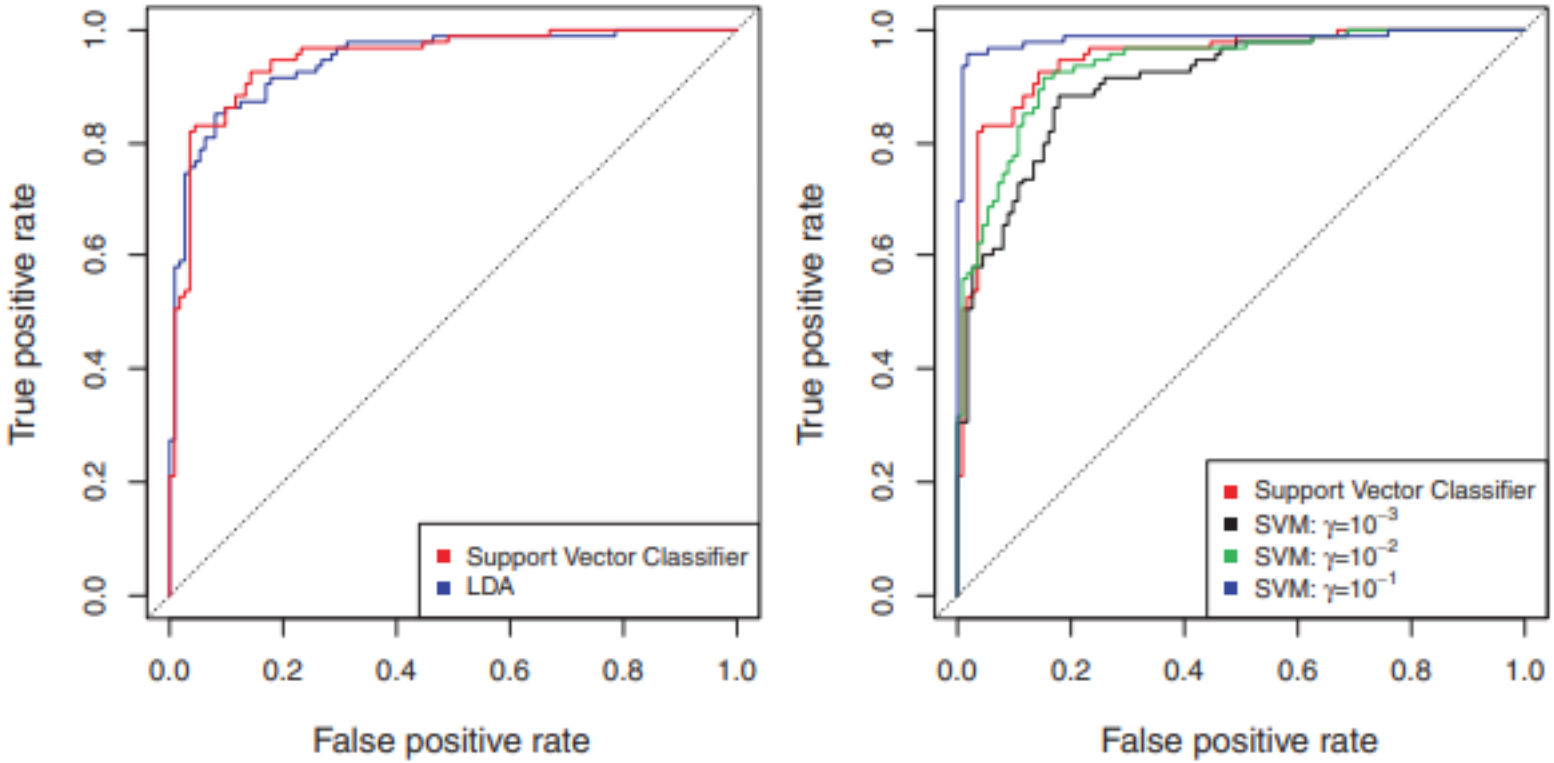
$$= \exp(\mathbf{x}^\top\mathbf{x}')\exp(-\frac{1}{2}\|\mathbf{x}\|^2)\exp(-\frac{1}{2}\|\mathbf{x}'\|^2)$$

$$= \sum_{j=0}^{\infty}\frac{(\mathbf{x}^\top\mathbf{x}')^j}{j!}\exp\left(-\frac{1}{2}\|\mathbf{x}\|^2\right)\exp\left(-\frac{1}{2}\|\mathbf{x}'\|^2\right)$$

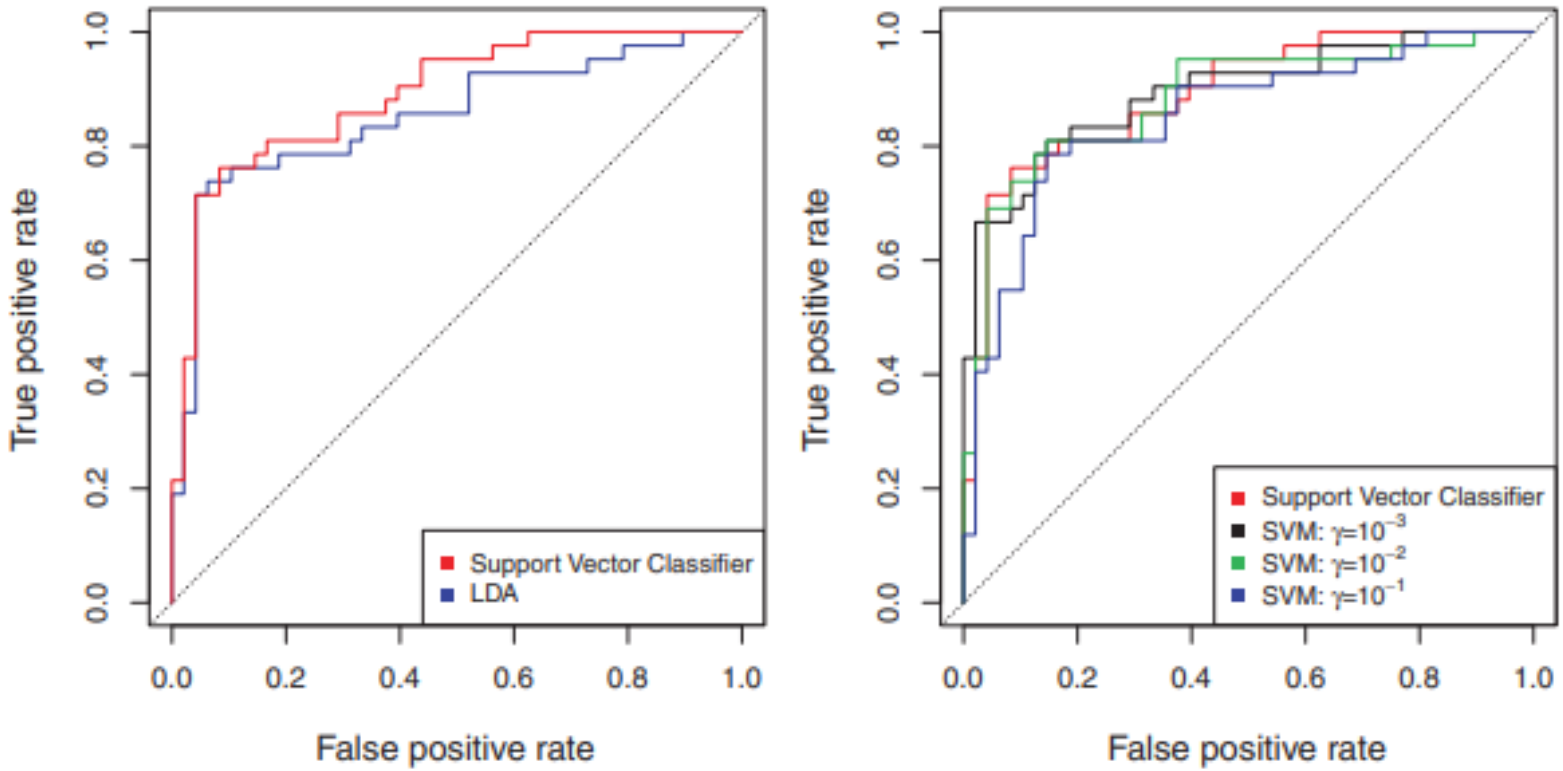
$$= \sum_{j=0}^{\infty}\sum_{\sum n_i=j}\exp\left(-\frac{1}{2}\|\mathbf{x}\|^2\right)\frac{x_1^{n_1}\cdots x_k^{n_k}}{\sqrt{n_1!\cdots n_k!}}\exp\left(-\frac{1}{2}\|\mathbf{x}'\|^2\right)\frac{{x'}_1^{n_1}\cdots {x'}_k^{n_k}}{\sqrt{n_1!\cdots n_k!}}$$

Rysunek: Źródło: https://en.wikipedia.org

Zastosowanie: choroby serca



Rysunek: Po lewej: porównanie krzywych ROC dla SVC i LDA. Po prawej: porównanie krzywych ROC dla różnych paramatrów γ w radialnej funkcji bazowej.Źródło: „Introduction...”



Rysunek: Krzywe ROC jak w Rysunku powyżej, dla zbioru testowego.Źródło: „Introduction...”

One-Versus-One Classification

Rozważmy teraz, jak zastosować SVM do więcej niż dwóch różnych klas. Pokażemy dwa najczęściej stosowane podejścia.

One-Versus-One Classification polega na dopasowaniu $\binom{K}{2}$ modeli SVM, parami dla każdych dwóch z K klas, a następnie zaklasyfikowaniu obserwacji testowej do jednego zbioru z każdej pary, korzystając ze stworzonych modeli. Na koniec należy zliczyć częstotliwość klasyfikacji do każdej z klas i wybrać te klasę, do której obserwacja testowa „wpadała" najczęściej.

One-Versus-All Classification

W One-Versus-All Classification tworzymy K modeli, które pozwalają zbadać, czy obserwacja zalicza się do k-tej klasy, czy do którejś z $K-1$ pozostałych. Następnie dla testowej obserwacji x^* obliczamy $\beta_{0k} + \beta_{1k}x^* + \ldots + \beta_{pk}x^*$ i wybieramy tę klasę, dla której ta wartość jest największa.