Увод у CBM

Ненад Митић

Математички факултет
nenad.mitic@matf.bg.ac.rs

Увод

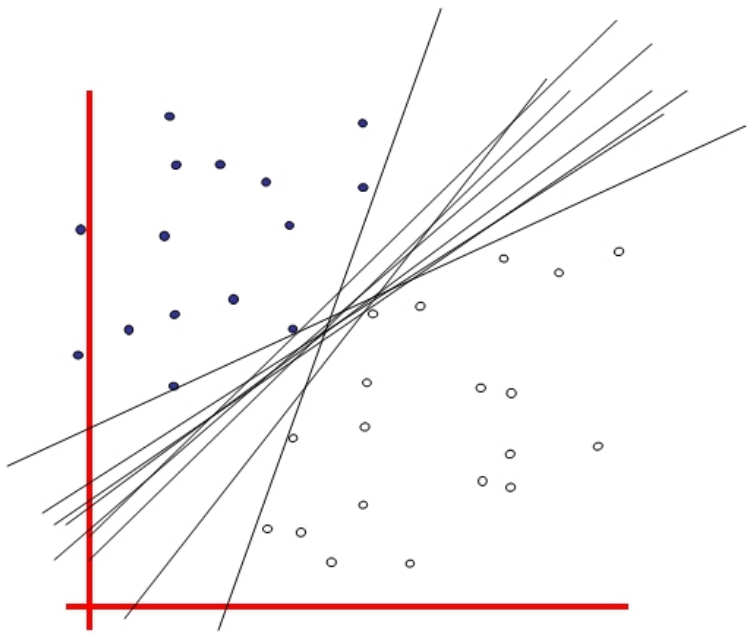
- СВМ (енг. *Support Vector Machine*) метод подржавајућих (потпорних) вектора
- Извор - статистичка теорија учења; Вапник 1979., 1992., 1995. г.
- Техника за класификацију заснована на идеји векторских простора
- Употребљива нарочито у ситуацијама када је број димензија података велики
- Применљива на класификацију (СВК) и регресију (СВР)

Увод

- Модел је *формула* - класа се израчунава
- Основни алгоритам је за бинарну класификацију
- Употребљива за нумеричке податке; категорички подаци се трансформишу увођењем променљиве за сваку вредност категоричког атрибута

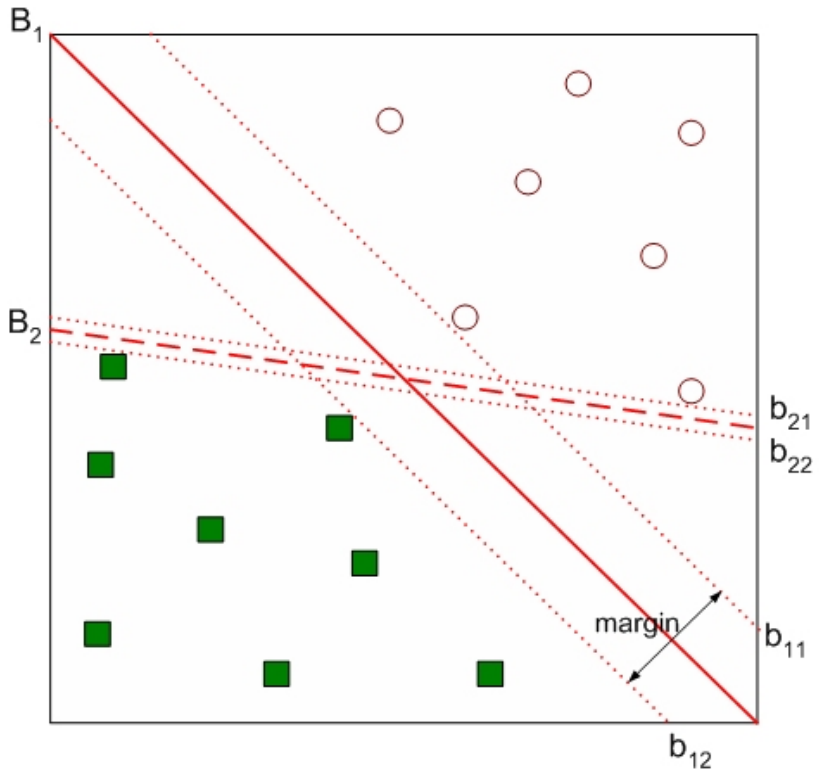
Увод

Идеја: у векторском простору у коме су подаци представљени, наћи раздвајајућу хипер-раван тако да су сви подаци из дате класе са исте стране равни



Како одредити најбољу хипер-раван?

Увод



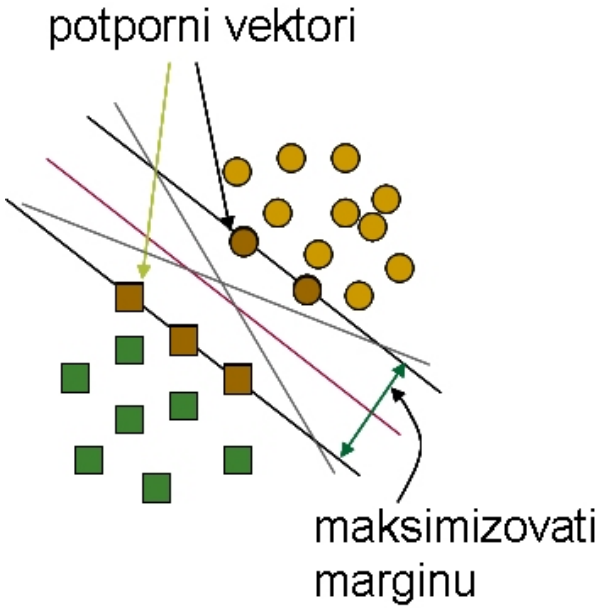
Како одредити најбољу хипер-раван?

Како изабрати најбољу хиперраван

- Постоје једноставни алгоритми (нпр. перцептрон) који могу да одреде раздвајајућу раван, али не и оптималну
- СВМ одређује оптимално решење које максимизује раздаљину између хипер-равни и тачака које су близу потенцијалне линије раздвајања
- Интуитивно: ако нема тачака близу линије раздвајања, онда ће класификација бити релативно лака

Како изабрати најбољу хиперраван

- СВМ максимизује маргину око раздвајајуће хипер-равни
- Резултат: раздвајајућа хипер-раван је потпуно одређена специфичним подскупом тренирајућих података, који се зову подржавајући (потпорни) вектори



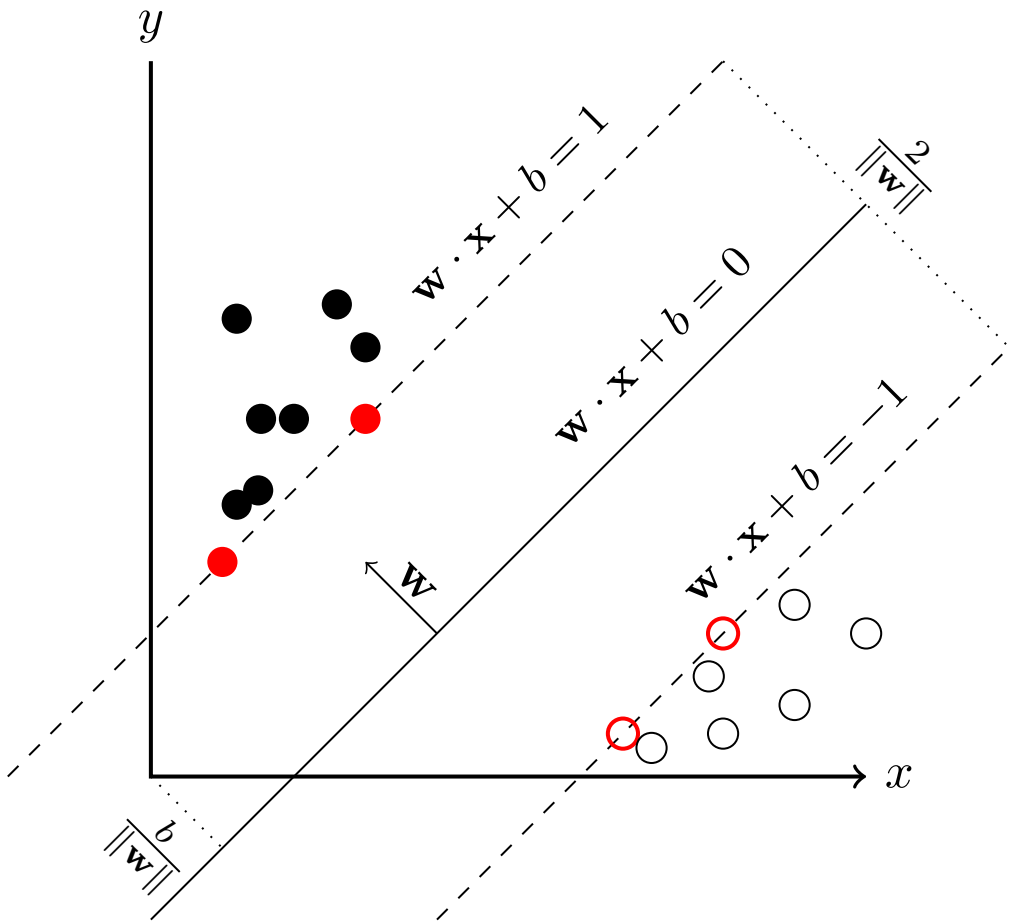
Како изабрати најбољу хиперраван

- Претпоставља се да су подаци линеарно раздвојиви
- Тренирање: наћи оптималну раздвајајућу хипер-раван, тј. раван са максималном маргином односно растојањем од података за тренинг
- Једначина те хипер-равни је модел
- Примена модела: израчунати растојање од хипер-равни и на основу тога одредити класу (изнад/испод равни)

Линеарно раздвојиви подаци

- Бинарни класификациони проблем са N примерака за тренинг
- Сваки примерак представљен торком (x_i, y_i),где су атрибути $x_i = (x_{i1}, x_{i2}, ..., x_{id})^T$, а $y \in \{-1, 1\}$ представља ознаку класе
- Једначина хипер-равни: $w \cdot x + b = 0$
- За тачке x_a и x_b које припадају хипер-равни важи
  $w \cdot x_a + b = 0$
  $w \cdot x_b + b = 0$
  $\longrightarrow W \cdot (x_b - x_a) = 0$

Линеарно раздвојиви подаци



Линеарно раздвојиви подаци

- За било коју тачку x_i изнад граничне линије важи $w \cdot x_i + b = k, k > 0$
- За било коју тачку x_i испод граничне линије важи $w \cdot x_i + b = k', k' < 0$
- Ознака класе произвољне тачке z се одређује

$$y = \begin{cases} 1, & ako\ w \cdot z + b > 0 \\ -1, & ako\ w \cdot z + b < 0 \end{cases}$$

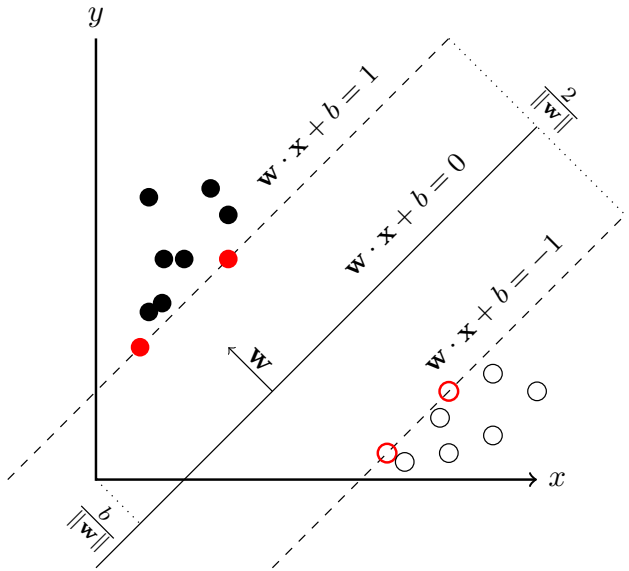
Одређивање маргине

- Рескалирањем параметара w и b из једначине хипер-равни добијају се две паралелне хипер-равни h_{i1} и h_{i2} које представљају границе маргина
- њихове једначине су
  $h_{i1} : w \cdot x + b = 1$ i
  $h_{i2} : w \cdot x + b = -1$
- Ако $x_1 \in h_{i1}$ и $x_2 \in h_{i2}$ тада се величина маргине d добија као $w \cdot (x_1 - x_2) = 2$, односно

$$||w|| \times d = 2$$

$$d = \frac{2}{||w||}$$

Одређивање маргине



- Параметри w и b морају да задовоље услове

$$w \cdot x_i + b \geq 1 \quad \text{ako } y_i = 1$$
$$w \cdot x_i + b \leq -1 \quad \text{ako } y_i = -1$$

- Обе неједнакости могу да се запишу као $y_i(w \cdot x_i + b) \geq 1, i = 1, 2, \ldots N$

Одређивање маргине

Захтев да маргина буде максимална је еквивалентан одређивању

$$\max_w \frac{2}{||w||}$$
$$\text{уз услов} \quad y_i(w \cdot x_i + b) \geq 1, i = 1, 2, ...N$$

Или еквивалентно, одређивању

$$\min_w \frac{||w||^2}{2}$$
$$\text{уз услов} \quad y_i(w \cdot x_i + b) \geq 1, i = 1, 2, ...N$$

У питању је квадратни оптимизациони проблем са линеарним ограничењима који има јединствени минимум

Лагранжови множиоци

Претходни запис може да се трансформише у Лагранжову формулацију проблема. Разлози:

- Ограничење $y_i(w \cdot x_i + b) \geq 1, \forall i$
  биће замењено ограничењима самих Лагранжових множилаца, што је лакше за рад
- У новој форми тренинг подаци се јављају једино у облику скаларног производа, што омогућава генерализацију решења на случај нелинеарно раздвојивих података

Лагранжови множиоци

- Нека је $f(x_1, x_2, ..., x_n) : R^n \longrightarrow R$ диференцијабилна функција чији се минимум (максимум) тражи, уз ограничење $g(x_1, x_2, ..., x_n) = 0$
- Како се градијент функције мења на исти начин као и градијент ограничења, важи једначина $\nabla f(x) = \lambda \nabla g(x)$, при чему је λ коефицијент промене (Лагранжов множилац)
- Комбинацијом једначине и услова формира се *Лагранжијан* $L(x, \lambda) = f(x) - \lambda g(x)$
- Вредност λ се одређује тако да је $\nabla L(x, \lambda) = 0$
- Заменом вредности λ, x одређује се минимум ф-је.

Лагранжови множиоци и СВМ

- Ограничења су неједнакости - нису згодне за израчунавање
- Функција која треба да се минимизује записује се у облику Лагранжијана

$$L_p = \frac{1}{2}||w||^2 - \sum_{i=1}^{N} \lambda_i (y_i(w \cdot x_i + b) - 1)$$

odnosno

$$L_p = \frac{1}{2}||w||^2 - \sum_{i=1}^{N} \lambda_i y_i (w \cdot x_i + b) + \sum_{i=1}^{N} \lambda_i$$

- Сви λ_i су позитивни јер су сва ограничења увек ≥ 0

Одређивање Лагранжових множилаца

За минимизацију Лагранжиана потребно је да се одреди

- $\frac{\partial L_p}{\partial w} = 0 \Longrightarrow w = \sum_{i=1}^{N} \lambda_i y_i x_i$
- $\frac{\partial L_p}{\partial b} = 0 \Longrightarrow 0 = \sum_{i=1}^{N} \lambda_i y_i$
- Ако су ограничења једнакости, из скупа једначина могу да се добију w, b и λ

Одређивање Лагранжових множилаца

- Како су ограничења неједнакости, а важи $\lambda_i \geq 0$ могућа је трансформација у ограничења са једнакостима
- *Karush-Kuhn-Tucker*-ови услови (w и x су вектори)
  1. $\frac{\partial L_p}{\partial w} = w - \sum_i \lambda_i y_i x_i$
  2. $\frac{\partial L_p}{\partial b} = -\sum_i \lambda_i y_i$
  3. $y_i(w \cdot x + b) - 1 \geq 0$
  4. $\lambda_i \geq 0$
  5. $\lambda_i(y_i(w \cdot x + b) - 1) = 0$
- Важи $\lambda_i = 0$ осим ако је $y_i(w \cdot x + b) = 1$
- Инстанце код којих је $\lambda_i > 0$ јесу *потпорни вектори*
- Ограничења функције се замењују ограничењима Лагранжових множилаца

Одређивање Лагранжових множилаца - дуални проблем

- Свођење на дуални проблем трансформацијом Лагранжијана у чисте Лагранжове множиоце
- Заменом $w = \sum_{i=1}^{N} \lambda_i y_i x_i$ и $\sum_{i=1}^{N} \lambda_i y_i = 0$ уместо b у

$$L_p = \frac{1}{2}||w||^2 - \sum_{i=1}^{N} \lambda_i y_i (w \cdot x_i + b) + \sum_{i=1}^{N} \lambda_i$$

Добија се дуални проблем

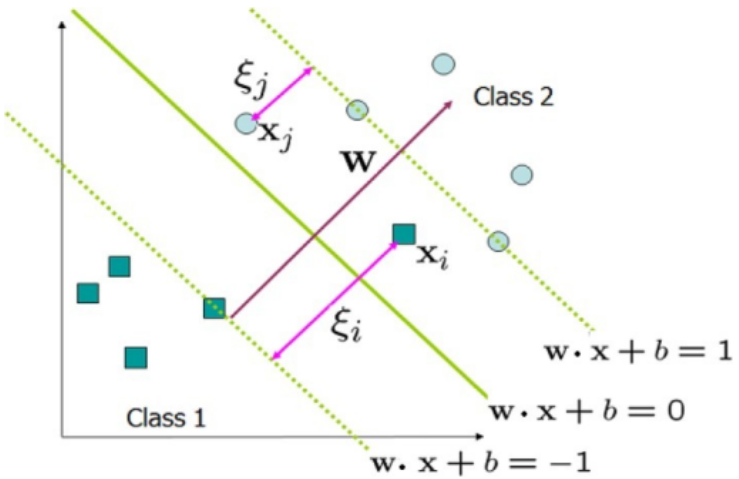
$$L_D(\lambda_i) = \sum_{i=1}^{N} \lambda_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \lambda_i \lambda_j y_i y_j x_i \cdot x_j$$

Одређивање Лагранжових множилаца - дуални проблем

- Вредност λ_i се одређују узимањем извода по λ и изједначавањем са нулом: $\sum_{i=1}^{N} \lambda_i y_i = 0$
- w који одређује максималну маргину хиперравни се добија као $w = \sum_{i=1}^{N} \lambda_i y_i x_i$
- отклон b се добија као $b_i = \frac{1}{y_i} - w^T x_i = y_i - w^T x_i$, $b = \underset{i,\lambda_i>0}{avg}\ b_i$
- за непознату тачку z класа се одређује на основу знака $f(z) = sign(w \cdot z + b) = sign((\sum_{i=1}^{N} \lambda_i y_i x_i \cdot z) + b)$

Линеарни СВМ - нераздвојиви подаци

Ако скуп за тренирање није линеарно раздвојив увести променљиве $\xi_i > 0$ - величина за коју може да се олабави граница тако да се толеришу (мале) грешке при класификацији



$$w \cdot x_i + b \geq 1 - \xi_i \quad \text{ako } y_i = 1$$
$$w \cdot x_i + b \leq -1 + \xi_i \quad \text{ako } y_i = -1$$

Линеарни СВМ - нераздвојиви подаци

- Може да се примени исти поступак уз претходне услове да би се одредиле границе хиперравни
- Не постоји ограничење на број грешака (у класификацији), могуће је да се одреди широка маргина са великим бројем лоше класификованих података
- Поставља се ограничење на величину ξ, односно ограничава 'пропусност' границе

Линеарни СВМ - нераздвојиви подаци

Функција која треба да се минимизује (максимизује) је

$$f(w) = \frac{||w||^2}{2} + C \sum_{i=1}^{N} \xi_i^k$$

где су C и k константе које одређују цену погрешне класификације

- Терм $\sum_{i=1}^{N} \xi_i^k$ означава губитак, одн. процену девијације од случаја са раздвојивим подацима

- C се бира емпиријски и представља *константу регуларизације* која контролише однос између максимизације маргине и минимизације губитка

Линеарни СВМ - нераздвојиви подаци

$$f(w) = \frac{||w||^2}{2} + C\sum_{i=1}^{N}\xi_i^k$$

- Када $\lim_{C\to 0}$ губитак ефективно нестаје и маргина се максимизује
- Када $\lim_{C\to\infty}$ маргина практично нестаје и потребно је минимизовати губитак
- k покрива облик губитка - обично се поставља на 1 или 2
- Ako к=1 (енг. *hinge loss*- шарка) - минимизује се збир ξ_i
- Ako к=2 (енг. *quadratic loss*) - минимизује се збир квадрата ξ_i

Лагранжови множиоци за меку маргину

Функција која треба да се минимизује записује се у облику Лагранжијана

$$L_p = \frac{1}{2}||w||^2 - C\sum_{i=1}^{N}\xi_i \sum_{i=1}^{N}\lambda_i(y_i(w\cdot x_i + b) - 1 + \xi_i) - \sum_{i=1}^{N}\beta_i\xi_i$$

(последњи сабирак је резултат услова $\xi \geq 0, \forall i$)

Ограничена са неједнакостима могу да се трансформишу у ограничења са једнакостима користећи KKT услове

$$\xi \geq 0, \lambda_i \geq 0, \beta_i \geq 0, \forall i$$
$$\lambda_i(y_i(w\cdot x_i + b) - 1 + \xi_i) = 0$$
$$\beta_i\xi_i = 0$$

Мека маргина - дуални Лагранжијан

Дуални Лагранжијан се добија изједначавањем првог извода по w, b и ξ са 0

$$\frac{\partial L_p}{\partial w} = w - \sum_{i=1}^{N} \lambda_i y_i x_i = 0 \implies w = \sum_{i=1}^{N} \lambda_i y_i x_i$$

$$\frac{\partial L_p}{\partial b} = -\sum_{i=1}^{N} \lambda_i y_i = 0 \implies \sum_{i=1}^{N} \lambda_i y_i = 0$$

$$\frac{\partial L_p}{\partial \xi_i} = C - \lambda_i - \beta_i = 0 \implies \lambda_i + \beta_i = C$$

Мека маргина - дуални Лагранжијан

И заменом добијених вредности у L_p

$$L_D = \frac{1}{2} w^T w - w^T \underbrace{(\sum_{i=1}^{N} \lambda_i y_i x_i)}_{w} - b \underbrace{(\sum_{i=1}^{N} \lambda_i y_i)}_{0} + \sum_{i=1}^{N} \lambda_i + C \sum_{i=1}^{N} \xi_i - \sum_{i=1}^{N} (\lambda_i + \beta_i) \xi_i$$

$$L_D = -\frac{1}{2} w^T w + \sum_{i=1}^{N} \lambda_i + C \sum_{i=1}^{N} \xi_i - \sum_{i=1}^{N} (\lambda_i + C - \lambda_i) \xi_i$$

$$L_D = \sum_{i=1}^{N} \lambda_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \lambda_i \lambda_j y_i y_j x_i \cdot x_j$$

Мека маргина - дуални Лагранжијан

Добијена циљна функција чија се екстремна вредност тражи је

$$\max_{\lambda_i} L_D = \sum_{i=1}^{N} \lambda_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \lambda_i \lambda_j y_i y_j x_i \cdot x_j$$

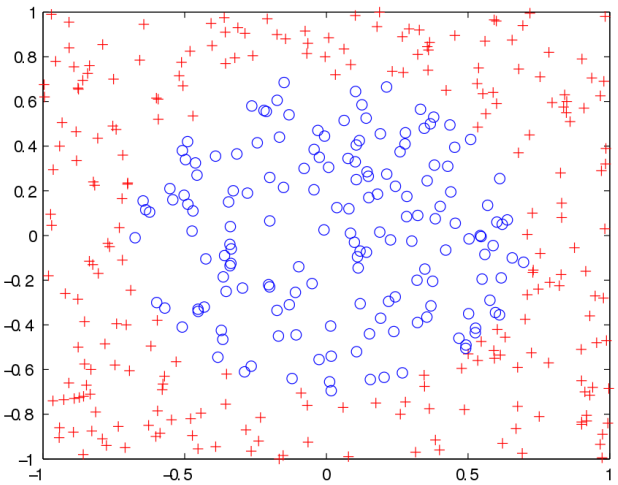
уз ограничења

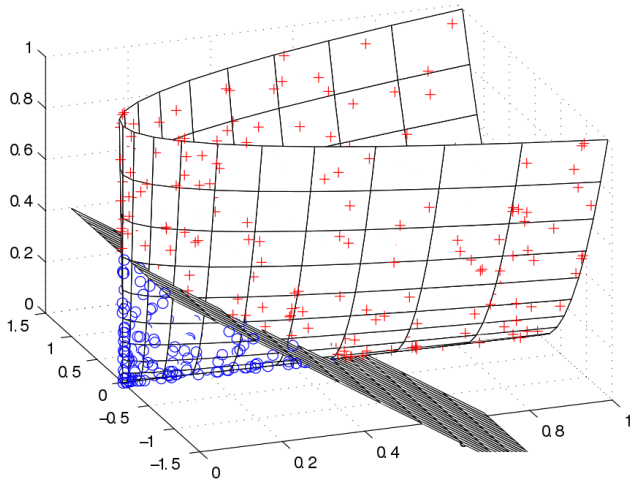
$$0 \leq \lambda_i \leq C, \forall i$$

и

$$\sum_{j=1}^{N} \lambda_i y_i = 0$$

Нелинеарни СВМ - увод





Идеја: одредити функцију $\Phi : \mathrm{R}^m \longrightarrow \mathrm{R}^n$ тако да хиперраван $\mathrm{w} \cdot \Phi(\mathrm{x}) + b = 0$ раздваја трансформисане податке

Нелинеарни СВМ - увод

Сличним поступком као код линеарно раздвојивих СВМ максимална маргина се добија минимизацијом

$$\min_{w,b,\xi_i} \left\{ \frac{||\mathrm{w}||^2}{2} + C \sum_{i=1}^{N} \xi_i^k \right\}$$

уз услов да $\forall \mathrm{x_i} \wedge i = 1, 2, \ldots N$ важи

$$y_i(\mathrm{w} \cdot \Phi(\mathrm{x}_i) + b) \geq 1 - \xi_i \ \wedge \ \xi_i \geq 0$$

Нелинеарни СВМ - увод

Аналогно, дуални Лагранжијан је

$$L_D(\lambda_i) = \sum_{i=1}^{N} \lambda_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \lambda_i \lambda_j y_i y_j \Phi(\mathrm{x}_i) \cdot \Phi(\mathrm{x}_j)$$

уз услове

$$\sum_{i=1}^{N} \lambda_i y_i = 0$$

и

$$0 \leq \lambda_i \leq C, i = 1, 2, \ldots N$$

Нелинеарни СВМ - увод

Параметри

$$\mathrm{w} = \sum_{i=1}^{N} \lambda_i y_i \Phi(\mathrm{x_i}) \qquad b = \underset{i,\lambda_i>0}{avg}\ b_i$$

$$\text{gde} \quad \lambda_i \{ y_i (\sum_{j=1}^{N} \lambda_j y_j \Phi(\mathrm{x}_j) \cdot \Phi(\mathrm{x}_i) + b) - 1 + \xi_i \} = 0$$

За непознату тачку z класа се одређује на основу знака

$$f(z) = sign(w \cdot z + b) = sign(\sum_{\lambda_i \geq 0, i=1}^{N} \lambda_i y_i \Phi(\mathrm{x}_i) \cdot \Phi(\mathrm{y}) + b)$$

Кернели

- Нејасно како одредити Φ
- Израчунавање може да буде сложено
- Могућа мера сличности у трансформисаном простору $\Phi(x_i) \cdot \Phi(x_j)$
- *Кернел трик* - рачунање сличности у трансформисаном простору користећи оригинални скуп атрибута
- Предност - не третира се експлицитно n-димензиони простор
- *Кернел* функција - ф-ја сличности K која се рачуна помоћу оригиналног скупа атрибута

Мерцерова теорема

Кернел функција K може да се представи као $K(x, y) = \Phi(x) \cdot \Phi(y)$ акко важи

$$\text{ako } \forall g(x) : \int g(x)^2 dx \text{ je konačno}$$

$$\Longrightarrow \int K(x, y) g(x) g(y) dx dy \geq 0$$

Мерцерова теорема (алтернативна дефиниција)

Ако Симетрична функција $K(\mathrm{x},\mathrm{x}')$ задовољава услов

$$\sum_{i,j=1}^{M} h_i h_j K(\mathrm{x},\mathrm{x}') \geq 0$$

$\forall M \in \mathbb{N}, \forall h_i \in \mathbb{R}, \forall x_i$, тада постоји функција пресликавања $\phi(\mathrm{x})$ која пресликава x у простор са особином скаларног производа, где $\phi(\mathrm{x})$ задовољава услов

$$K(\mathrm{x},\mathrm{x}') = \phi^T(\mathrm{x})\phi(\mathrm{x}')$$

Мерцерова теорема (алтернативна дефиниција)

Тада важи Мерсеров услов

$$\sum_{i,j=1}^{M} h_i h_j K(x_i, x_j) = \left(\sum_{i=1}^{M} h_i \phi^T(x_i) \right) \left(\sum_{i=1}^{M} h_i \phi(x_i) \right) \geq 0$$

Функција која задовољава Мерсеров услов се назива позитиван семидефинитни кернел или Мерсеров Кернел

Примери кернела

linearni kernel: $K(x,y) = x^T \cdot y$

polinomijalni kernel: $K(x,y) = (x^T \cdot y + r)^d$

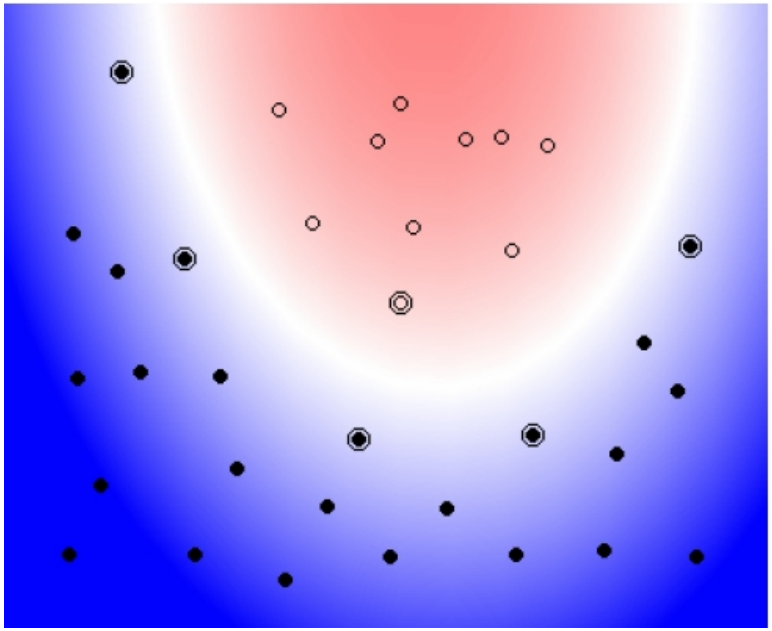
Gausov RBF kernel: $K(x,y) = e^{-\gamma||x-y||^2}$, često $\gamma = \frac{1}{2\sigma^2}$

eksponencijalni RBF kernel: $K(x,Y) = e^{-\frac{||x-y||}{2\sigma^2}}$

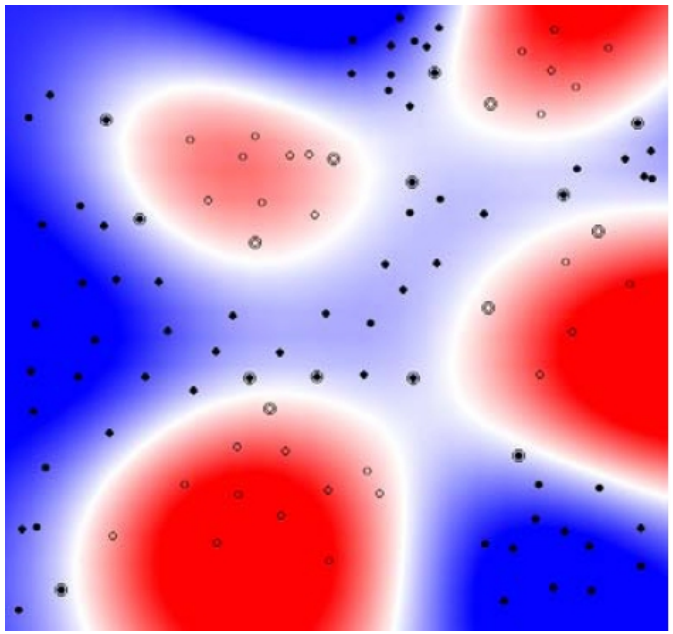
sigmoid (tangens hiperbolički): $K(x,y) = tanh(\kappa x \cdot y - \delta), \kappa > 0, \delta > 0$

... ...

Примери кернела



СВМ са полиномијалним кернелом 2. степена



SVM sa RBP kernelom