Maşinile cu vectori suport

Maşinile cu vectori suport (engl. "Support Vector Machines", SVM) reprezintă o metodă de învăţare de nouă generaţie (Shawe-Taylor & Cristianini, 2000), cu o fundamentare matematică riguroasă, bazată pe conceptul de maximizare a „marginii" care separă instanţele din două clase diferite, iar maximizarea este rezolvată analitic, nu empiric. Datorită acestei fundamentări, maşinile cu vectori suport au demonstrat performanţe foarte bune pentru probleme reale cum ar fi clasificarea textelor, recunoaşterea caracterelor scrise de mână, clasificarea imaginilor etc. În general, SVM-urile sunt considerate unele dintre cele mai bune metode de clasificare cunoscute la ora actuală.

Mașinile cu suport vectorial sunt o clasă de algoritmi supervizaţi de învăţare automată care utilizează concepte din teoria învăţării computaţionale. Suportul este o suprafaţă de separare optimală tipic neliniară în spaţiul de intrare şi liniară într-un spaţiu cu mai multe dimensiuni.

Trei idei principale:

- Se definește ce este un hiperplan optimal (într-un mod care poate fi identificat printr-o cale eficientă computațional): maximizează marginea
- Extinde definiția de mai sus pentru probleme liniar neseparabile: are un termen de penalizare pentru clasificări eronate
- Proiectează datele într-un spațiu cu mai multe dimensiuni unde este mai simplă clasificarea cu suprafețe de decizie liniare: reformulează problema astfel încât datele sunt proiectate implicit în acest spați

Istoria SVM

- SVM sunt legate de teoria învăţării statistice
- SVM au fost introduse în 1992
- SVM au devenit populare datorită succeselor în recunoaşterea cifrelor scrise de mână
- SVM sunt privite azi ca un exemplu important de "metode kernel", unul din domeniile cheie ale maşinilor instruibile

Hiperplane de separare. Marginea dintre clase

Să considerăm mai întâi o problemă de clasificare binară cu clase separabile liniar (Figura 1).

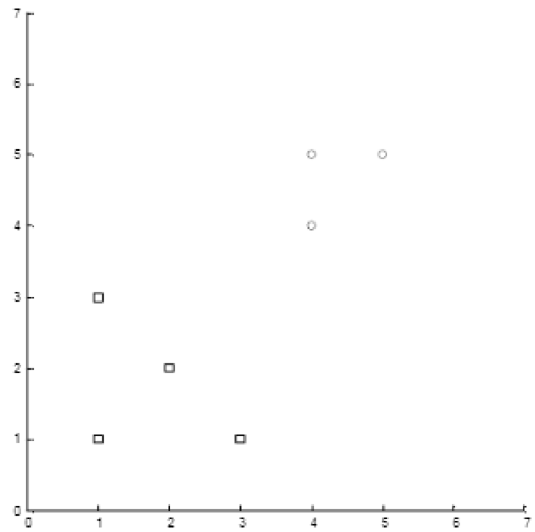

Figura 1. Problema de clasificare cu clase separabile linear.

Un clasificator poate împărţi planul cu o dreaptă în două suprafeţe şi prin urmare poate rezolva corect această problemă.

Dreapta separatoare poate avea parametri diferiţi, după cum se arată în Figura 2. Toate aceste drepte rezolvă problema de clasificare considerată.

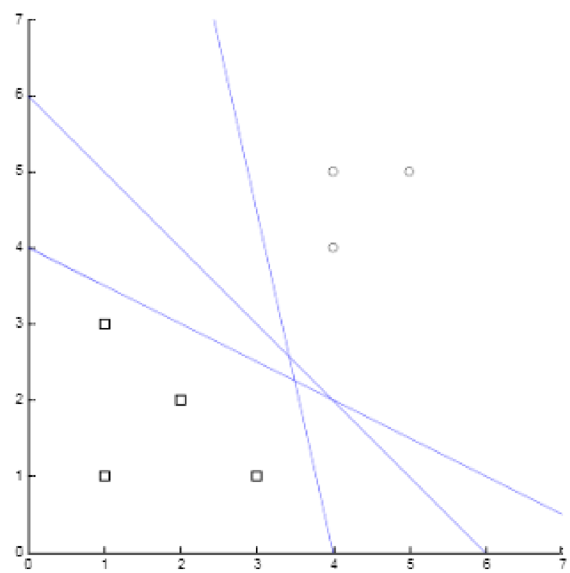

Figura 2. Drepte de separare posibile.

Având mai multe opţiuni la dispoziţie, se pune problema care dintre acestea este cea mai bună, criteriul luat în calcul fiind capacitatea de generalizare a modelului, adică posibilitatea acestuia de a clasifica în mod corect instanţe noi.

Odată găsită o dreaptă de separare între instanţele de antrenare, putem spune că este destul de probabil ca o nouă instanţă (1, 2) să aparţină clasei pătratelor iar una (7, 7) să aparţină clasei cercurilor. Probleme vor apărea în zona de demarcaţie dintre cele două clase. Pentru o instanţă (3, 4), folosind două modele diferite, vom avea rezultate diferite, după cum se poate vedea în figura 3: o dreaptă o clasifică drept cerc iar cealaltă dreaptă – drept pătrat.

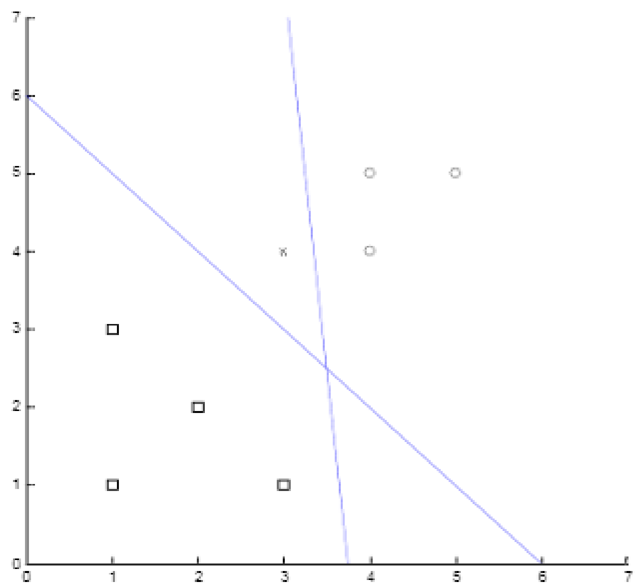

Figure 3. Drepte de separare diferite pot determina rezultate diferite.

Una din ideile de bază ale maşinilor cu vectori suport este că modelul care generalizează cel mai bine este acela care desparte cel mai mult clasele, adică cel care asigură marginea cea mai mare de demarcaţie între clase. Marginea este distanţa dintre dreptele paralele cu dreapta de separare care ating cel puţin una din instanţele fiecărei clase. Pentru cele două modele din figura 3, marginea dată de dreapta 1 este cea din figura 4a iar marginea dată de dreapta 2 este cea din figura 4b.

1.

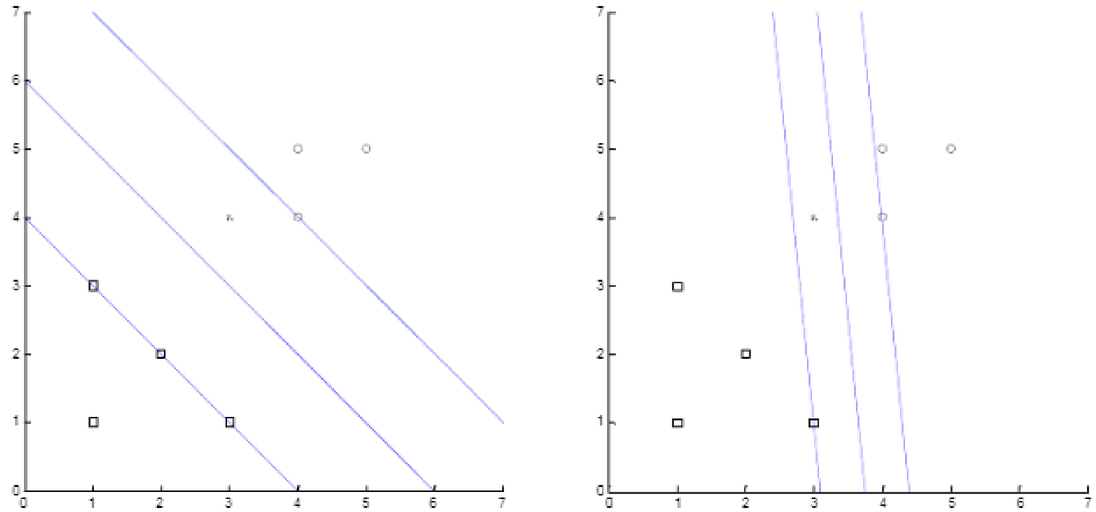

Figura 4. Marginile generate de doua drepte de separare diferite.

Se vede că marginea determinată de primul model este mai mare, prin urmare acesta este mai bun decât celălalt.

În cazul de faţă, problema de clasificare are 2 atribute şi suprafaţa de separare este o dreaptă care împarte planul în 2 semiplane. Dacă problema ar fi avut 3 atribute, suprafaţa de separare ar fi fost un plan care împarte în două spaţiul tridimensional. În general, hiperspaţiul nD corespunzător unei probleme cu n atribute este împărţit în două de către un hiperplan (n – 1)D.

Reprezentarea matematica

Pentru a formaliza această descriere, să considerăm că o suprafaţă de separare este descrisă de următoarea ecuaţie:

$$(\mathbf{w} \cdot \mathbf{x}) + b = 0,$$

Unde w si x sunt vectori iar b este un scalar.

$(\mathbf{w} \cdot \mathbf{x})$ este produsul scalar al celor doi vectori, adică:

$$(\mathbf{w} \cdot \mathbf{x}) = w_1 x_1 + \cdots + w_n x_n,$$

unde n este dimensiunea lor.

Exemplu:

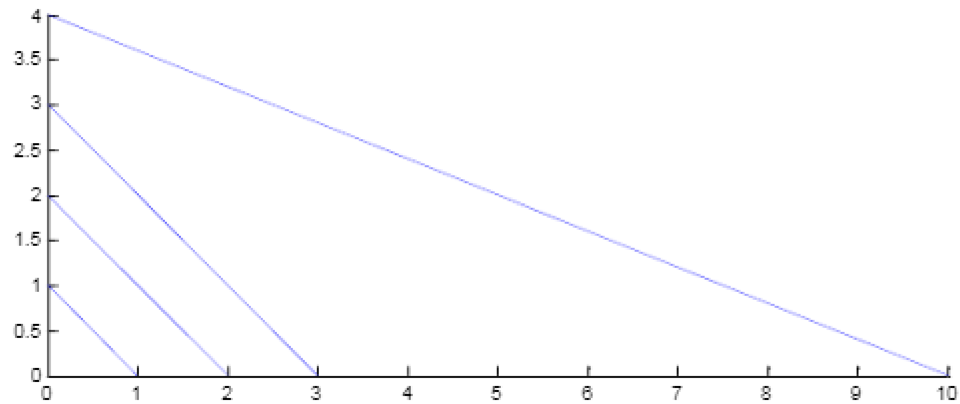

Figure 5. Drepte cu diferite ecuatii.

În figura 5, ecuaţia primei drepte este:

$$x_2 = 1 - x_1 \Rightarrow x_1 + x_2 - 1 = 0.$$

Considerăm aici că numele axelor sunt x1, x2 etc., întrucât y are o semnificaţie deosebită, indicând clasa unei instanţe şi nu un atribut (corespunzător unei axe de coordonate).

Analog, ecuaţiile următoarelor două drepte sunt:

$$x_1 + x_2 - 2 = 0,$$

respectiv:

$$x_1 + x_2 - 3 = 0.$$

În aceste cazuri, w = (1, 1) iar b diferă. Dreptele cu aceeaşi orientare w sunt paralele iar translaţia este dată de termenul b.

A patra dreaptă nu este paralelă cu celelalte trei; aici w = (2, 5) iar b = –20. w este un vector perpendicular pe dreapta corespunzătoare sau, în general, pe suprafaţa de separare considerată.

Reprezentarea problemei pentru SVM presupune ca ecuaţiile suprafeţelor de separare să fie normalizate, astfel încât în limitele care trec prin vectorii suport ai celor două clase să fie

$(\mathbf{w} \cdot \mathbf{x}) + b = -1$

respectiv

$(\mathbf{w} \cdot \mathbf{x}) + b = 1$

(figura 6).
Instanţele de pe aceste două suprafeţe se numesc vectori suport, care dau şi numele metodei.

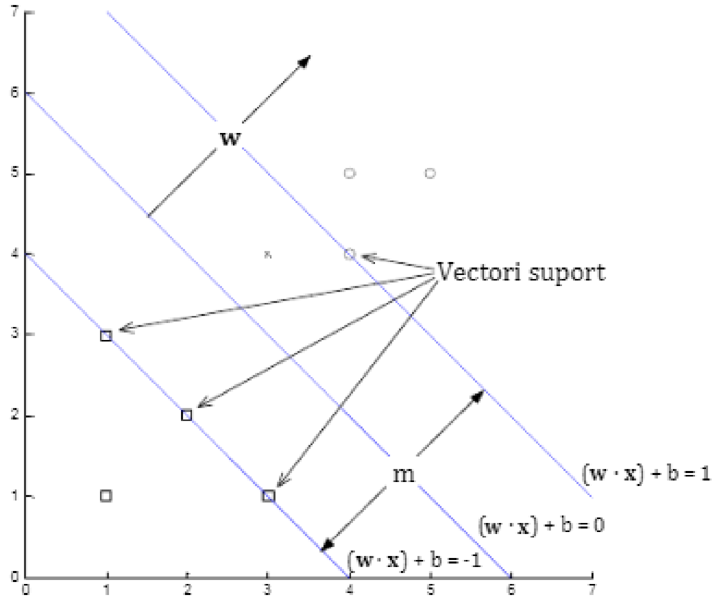

Figura 6. Vectorii support si ilustrarea botiunilor caracteristice problemei.

Clasificarea propriu-zisă presupune determinarea parametrilor w (en. weights) şi b (en. bias) care maximizează marginea de separare dintre clase. Apoi, o instanţă nouă z va fi clasificată în una din cele două clase (–1 sau 1) cu ajutorul funcţiei discriminant:

$$f_d(\mathbf{z}) = \operatorname{sgn}\left((\mathbf{w} \cdot \mathbf{z}) + b\right),$$

unde

$$\operatorname{sgn}(a) = \begin{cases} -1, & \textit{dacă} \;\; a < 0 \\ 1, & \textit{dacă} \;\; a \geq 0 \end{cases}.$$

Fie yi clasa instanţei i. Odată determinaţi parametrii wwww şi b, următoarele relaţii trebuie respectate:

$$(\mathbf{w} \cdot \mathbf{x}_i) + b \leq -1 \;\; \textit{dacă} \;\; y_i = -1\,,$$

$$(\mathbf{w} \cdot \mathbf{x}_i) + b \geq 1 \;\; \textit{dacă} \;\; y_i = 1\,.$$

De exemplu, pentru clasa 1, instanţele de pe dreapta-graniţă a marginii vor avea cantitatea

$(\mathbf{w} \cdot \mathbf{x}_i) + b$

egală strict cu 1, fiind vectori suport, iar celelalte instanţe din aceeaşi clasă, mai depărtate de margine, vor avea cantitatea respectivă mai mare strict decât 1.

Mai compact, constrângerile din ecuatiile de mai sus pot fi scrise astfel:

$$y_i\left((\mathbf{w}\cdot\mathbf{x}_i)+b\right)\geq 1.$$

Din figura 7, putem observa că marginea m este de fapt proiecţia distanţei dintre oricare doi vectori suport x1 si x2 pe directia vectorului w.

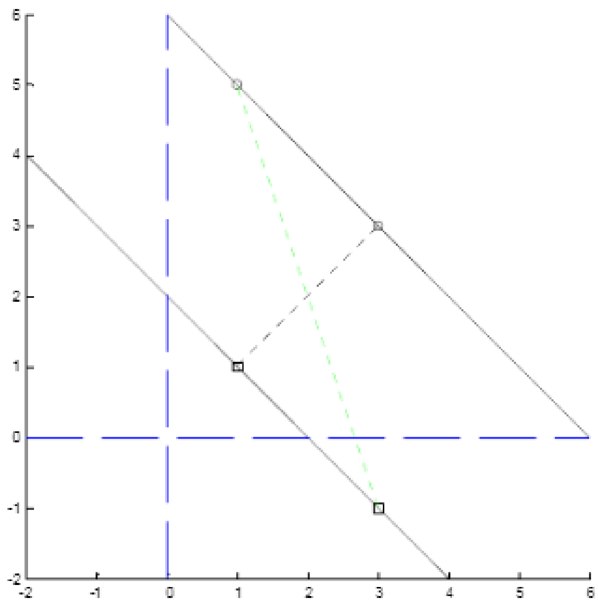

Figura 7. Ilustrarea modalitatii de calcul a marginii.

Prin urmare, valoarea lui m se poate obţine de exemplu scăzând ecuaţiile vectorilor x1 şi x2:

$$(\mathbf{w}\cdot\mathbf{x}_1)+b=-1$$
$$(\mathbf{w}\cdot\mathbf{x}_2)+b=1$$
$$\Rightarrow(\mathbf{w}\cdot(\mathbf{x}_2-\mathbf{x}_1))=2$$
$$\Rightarrow m=\frac{\mathbf{w}}{\|\mathbf{w}\|}\cdot(\mathbf{x}_2-\mathbf{x}_1)=\frac{2}{\|\mathbf{w}\|},$$

unde ||·|| este norma euclidiană, adică lungimea vectorului considerat:

$$\|\mathbf{w}\| = \sqrt{w_1^2 + \cdots + w_n^2}\ .$$

Exemplu:

În situaţia din figura 7, marginea m este în mod evident distanţa dintre punctele (1, 1) şi (3, 3):

$$m = \sqrt{(3-1)^2 + (3-1)^2} = 2\sqrt{2} = 2,83.$$

Însă aceeaşi valoare se obţine proiectând distanţa dintre x1(3,-1) si x2(1,5) pe direcţia w(1,1).

De fapt, distanţa dintre oricare 2 puncte situate pe cele 2 drepte paralele proiectată pe direcţia lui w poate fi utilizată pentru a ajunge la acelaşi rezultat.

Exemplu de aplicare a algoritmului SVM pe un set de date

```
from sklearn.svm import SVC
model=SVC(kernel='linear',gamma=1)
url = "https://raw.githubusercontent.com/jbrownlee/Datasets/master/pima-
indians-diabetes.data.csv"
names = ['preg','plas','pres','skin','test','mass','pedi','age','class']
import pandas
dataframe = pandas.read_csv(url, names=names)
array = dataframe.values
X = array[:,0:8]
Y = array[:,8]
test_size = 0.33
seed = 7
from sklearn import model_selection
X_train, X_test, Y_train, Y_test = model_selection.train_test_split(X,
Y,test_size=test_size, random_state=seed)
model.fit(X_train,Y_train)
model.score(X_test,Y_test)
```