SVM PROJET

Elaboré par :
Khawla Chouchene Bouhmila
MASTER SINT 1

Table des matières

INTRODUCTION

1. Définition de Machine Learning

Le Machine Learning est une technologie d'intelligence artificielle permettant aux ordinateurs d'apprendre sans avoir été programmés explicitement à cet effet. Pour apprendre et se développer, les ordinateurs ont toutefois besoin de données à analyser et sur lesquelles s'entraîner.

2. Définition de Big Data

Le Big data bouscule de fond en comble nos manières de faire du "business". Le concept, tel qu'il est défini actuellement, englobe un ensemble de technologies et de pratiques destinées à stocker de très grandes masses de données et à les analyser très rapidement.

De fait, le Big Data est l'essence du Machine Learning, et c'est la technologie qui permet d'exploiter pleinement le potentiel du Big Data.

3. Définition de Support Vecteur Machines (SVM)

Les Support Vecteur Machines souvent traduit par l'appellation de Séparateur à Vaste Marge (SVM) sont une classe d'algorithmes d'apprentissage initialement définis pour la discrimination c'est-à-dire la prévision d'une variable qualitative binaire.

Problématique

1. Présentation de Corpus

Le cancer de sein est le cancer le plus répandu chez les femmes dans le monde.

La détection précoce est le moyen le plus efficace de réduire le nombre de décès par cancer du sein.

Le diagnostic précoce nécessite une procédure précise et fiable permettant de distinguer les tumeurs bénignes du sein des tumeurs malignes.

Il y trois types de tumeurs du sein : tumeurs bénignes du sein, cancers in situ et cancers invasifs.

La majorité des tumeurs du sein détectées par mammographie sont bénignes. Ce sont des excroissances non cancéreuses et ne peuvent pas se propager à l'extérieur du sein vers d'autres organes.

Dans certains cas, il est difficile de distinguer certaines masses bénignes des lésions malignes avec mammographie. Si les cellules malignes ne sont pas passées à travers la membrane basale mais sont complètement contenues dans le lobule ou les canaux, le cancer est dit in situ ou non invasif. Si le cancer a traversé la membrane basale et s'est propagé dans les tissus environnants, il est appelé invasif. Cette analyse aide à différencier les tumeurs bénignes des tumeurs malignes.

2. Base de données

Breast_cancer Database (breast.cancer.csv).

3. Nombres d'instances

286 instances.

4. Attributs

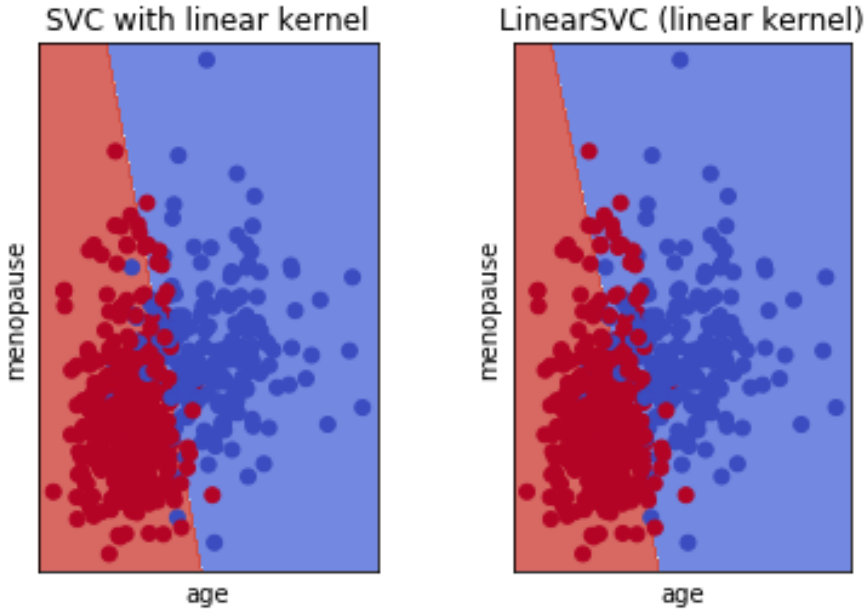
@attribute age {'10-19','20-29','30-39','40-49','50-59','60-69','70-79','80-89','90-99'}
@attribute menopause {'lt40','ge40','premeno'}
@attribute tumor-size {'0-4','5-9','10-14','15-19','20-24','25-29','30-34','35-39','40-44','45-49','50-54','55-59'}
@attribute inv-nodes {'0-2','3-5','6-8','9-11','12-14','15-17','18-20','21-23','24-26','27-29','30-32','33-35','36-39'}
@attribute node-caps {'yes','no'}
@attribute deg-malig {'1','2','3'}
@attribute breast {'left','right'}
@attribute breast-quad {'left_up','left_low','right_up','right_low','central'}
@attribute 'irradiat' {'yes','no'}
@attribute 'Class' {'no-recurrence-events','recurrence-events'}

Implèmentation Python

```
import numpy as np
import matplotlib.pyplot as plt
from sklearn import svm, datasets
# Chargement des données
cancer = datasets.load_breast_cancer()
# Garder juste les deux premiers attributs
X = cancer.data[:, :2]
y = cancer.target
# Pour afficher la surface de décision on va discrétiser l'espace avec un pas h
h = .02
C = 1.0 # paramètre de régularisation
svc = svm.SVC(kernel='linear', C=C).fit(X, y)
lin_svc = svm.LinearSVC(C=C).fit(X, y)

# Créer la surface de décision discretisée
x_min, x_max = X[:, 0].min() - 1, X[:, 0].max() + 1
y_min, y_max = X[:, 1].min() - 1, X[:, 1].max() + 1
xx, yy = np.meshgrid(np.arange(x_min, x_max, h), np.arange(y_min, y_max, h))
titles = ['SVC with linear kernel', 'LinearSVC (linear kernel)']

for i, clf in enumerate((svc, lin_svc)):
    plt.subplot(1, 2, i + 1)
```

```
    plt.subplots_adjust(wspace=0.4, hspace=0.4)

    Z = clf.predict(np.c_[xx.ravel(), yy.ravel()])
    # Utiliser une palette de couleurs
    Z = Z.reshape(xx.shape)
    plt.contourf(xx, yy, Z, cmap=plt.cm.coolwarm, alpha=0.8)
    # Afficher aussi les points d'apprentissage
    plt.scatter(X[:, 0], X[:, 1], c=y, cmap=plt.cm.coolwarm)
    plt.xlabel('age')
    plt.ylabel('menopause')
    plt.xlim(xx.min(), xx.max())
    plt.ylim(yy.min(), yy.max())
    plt.xticks(())
    plt.yticks(())
    plt.title(titles[i])
plt.show()
```

Exécution


SVC with linear kernel
menopause
age
LinearSVC (linear kernel)
menopause
age

Conclusion

Le machine learning est un outil très puissant qui permet d'effectuer de multiples actions comme classifier des données, faire apprendre à un programme à partir d'expèrimentations ou encore de crèer un programme èvolutionnaire qui s'amèliore sans cesse. Ainsi, même avec un èchantillon peu fourni (le machine learning nècessite habituellement des èchantillons avec 50 spècimens) et des donnèes influencèes par la subjectivitè de celui qui les mesure, le machine learning reste relativement prècis malgrè quelques lacunes.

Nèanmoins, le machine learning n'a pas que des qualitès, il doit être constamment adaptè au problème qu'il tente de rèsoudre. En effet, le programmeur doit tout d'abord se procurer l'èchantillon le plus reprèsentatif possible. Ensuite il faudra qu'il choisisse la fonction la plus fidèle à l'èchantillon, ce qui n'est pas nècessaire dans notre cas car les classes sont suffisamment distinctes pour que changer la forme de la fonction n'influe pas les rèsultats obtenus. Enfin, le machine learning doit être utilisè comme un outil car tous les problèmes ne nècessitent pas un programme complexe en machine learning.