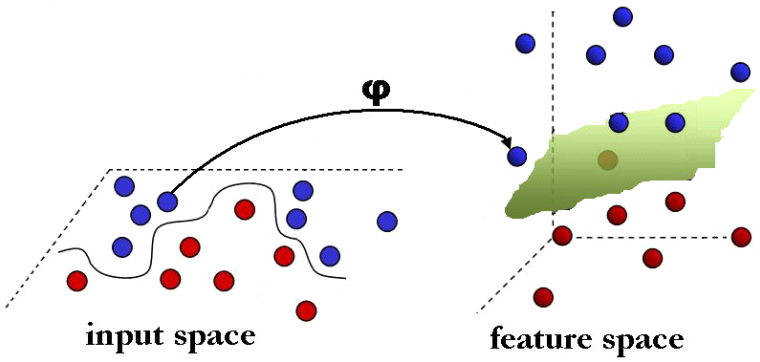
Тема №3: Класификатор с опорни вектори (support vector machine, SVM)

Цел на упражнението

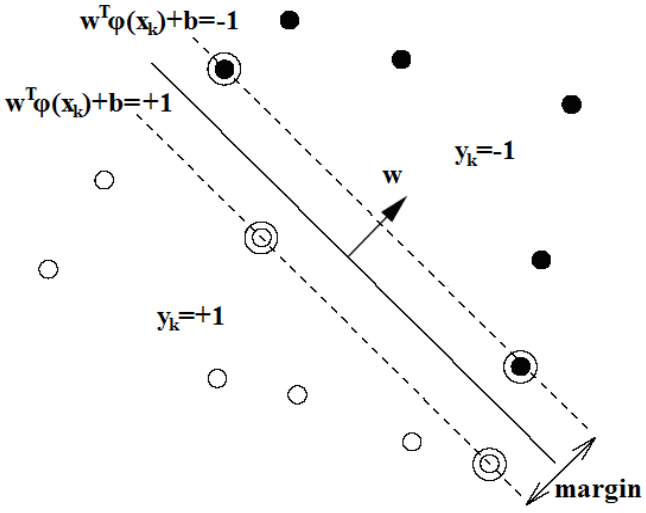
В Тема 3 се запознаваме с обучението и използването на класификатор с опорни вектори, който се явява един от най-съвършените дискриминативни методи за класификация. Създаден е през 70‘те години на XX век в СССР от Владимир Вапник, но дълги години остава засекретен. Едва през 1992 г. методът е публикуван за първи път в англоезични научни списания, след което придобива голяма популярност и става *state-of-art* метод за класификация в два класа.

Понятие за SVM

Концептуално, SVM методът е замислен за класификация в два класа, но по-късно идеята е доразвита за задачи които изискват класификация в много класове. С помощта на специални функции наречени *ядро* (или кърнел), SVM преобразува многомерното пространство на входните вектори в хипер-пространство с повече измерения, като се допуска че това би подпомогнало за по-лесното разделяне на двата класа.



Фиг. 3.1. Класификатор с опорни вектори – трансформация и преминаване от входното пространство с по-малък брой измерения към хиперпространство с повече измерения



Фиг. 3.2. Класификатор с опорни вектори – определяне на разделящата хиперравнина

За разлика от всички други методи, при определянето на хиперравнината, която разделя двата класа, селективно се подбират само част от обучаващите данни, които служат като опорни вектори. Опорните вектори играят ключова роля за получаване на позицията на разделящата хиперравнина в многомерното пространство, и за получаването на компактен модел при класификация.

Пример за класификация с SVM

Нека да разгледаме пример за класификация в три класа {Versicolor, Verginica, Setosa}, така както са дефинирани от Фишер във файл „fisheriris.mat“. Всеки образец е описан с четири вида описатели (дескриптори):

- Petal_Width,
- Petal Length,
- Sepal Width,
- Sepal Length

За целите на визуализация в следния пример избираме да работим с два от тези описатели (двумерни обекти се визуализират интуитивно). За целта последователно ще избираме различни двойки от четирите описатели и ще визуализираме преди да извършим класификация.

Примерна реализация с MATLAB:

Визуализираме данните `X` в двумерна координатна система `X = [meas(:,1), meas(:,2)]`, които определят неговото местоположение в равнината.

```
% зареждаме набора данни на Фишер
load fisheriris

% дефинираме X да ползва два от четирите описателa
X = [meas(:,1), meas(:,2)];

% Избираме един от класовете „Setosa” за да създадем детектор
% за него {versicolor, verginica, setosa}
Y = nominal(ismember(species,'setosa'));

% По случаен начин разделяме оригиналният набор данни на
% обучаващ набор и тестов набор в съотношение ¾ към ¼
% използвайки метода „stratified holdout”
P = cvpartition(Y,'Holdout',0.25);

% Създаваме класификатор “linear support vector machine” използвайки
% обучаващите данни и съответните етикети за класова принадлежност
svmStruct=svmtrain(X(P.training,:),Y(P.training),'showplot',true);

% Използваме класификатора за да обработим тестовия набор от данни
C = svmclassify(svmStruct,X(P.test,:),'showplot',true);
```

```
% Изчисляваме грешката от класификация като броя на грешните решения
% към общия брой тестови вектори (miss-classification rate)
errRate = sum(Y(P.test)~= C)/P.TestSize;
errRate

% Изчисляваме матрицата на решенията на класификатора
% (confusion matrix)
conMat = confusionmat(Y(P.test),C);
conMat
```

Задачи за изпълнение

Задача 1. Създайте SVM класификатор, който да разграничава биологичен вид „Versicolor“ от останалите видове представени в базата данни {Verginica и Setosa}. Изследвайте кой от четирите описатели е най-информативен за целите на класификацията. Сортирайте описателите според критерии „информативност“. Изследвайте кои комбинации от описатели са най-подходящи за разпознаване на „Versicolor“.

Задача 2. Проверете дали степента на информативност на описателите е същата, когато се цели разпознаване на биологичен вид “Verginica“. Направете същото за „Setosa“. Въз основа на резултатите от Задача 1 и Задача 2, сравнете успешността на класификатора, когато се използва като детектор за всеки един от трите класа {Versicolor, Verginica, Setosa} – използвайте критерии „точност на класификация“, която ще изчислите като средна стойност от главния диагонал на confusion matrix.

Задача 3. Създайте SVM класификатор в три класа, който да причислява (т.е. да идентифицира) непознат входен вектор в една от трите известни категории {Versicolor, Verginica, Setosa}, използвайки данните представени във файл „fisheriris.mat“.

Задача 4. Създайте SVM детектор, който да открива състояния означени като „диабет“. За целта създайте класификатор в два класа като използвайте данните от файл „diabetes.mat“. Сравнете получената точност на класификация с тази получена с MLP NN (лаб. упр. 1) и PNN (лаб. упр. 2).

Задача 5. Използвайки SVM класификатора от Задача 4, проверете дали нормализация на данните до диапазона [0, 1] би помогнало за подобряване на точността при класификация. За целта разделете стойностите на всеки от описателите на максималната стойност.