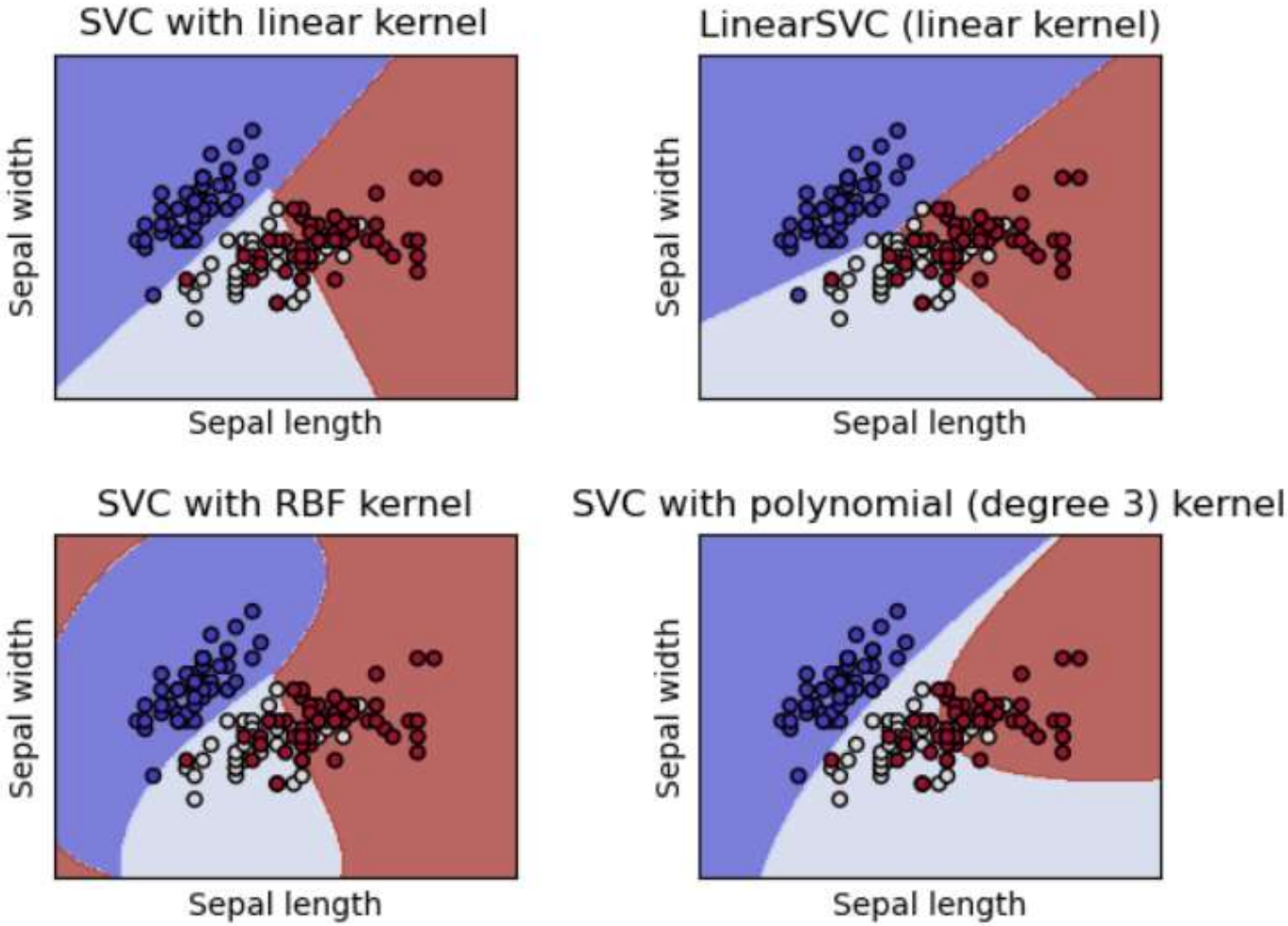
Lecture 9/Cours 9

Support vector Machine(SVM)

Definition

Une machine à vecteurs de support est un algorithme supervisé qui peut classer les cas en trouvant un **séparateur**. SVM fonctionne en mappant d'abord les données à un espace de caractéristiques de grande dimension afin que les données les points peuvent être catégorisés, même lorsque les données ne sont pas autrement linéairement séparables. Ensuite, un séparateur est estimé pour les données. Les données doivent être transformées de manière qu'un séparateur pourrait être dessiné comme un hyperplan.

Exemple/Example



Definition

A Support Vector Machine is a supervised algorithm that can classify cases by finding a **separator.** SVM works by first mapping data to a high dimensional feature space so that data points can be categorized, even when the data are not otherwise linearly separable. Then, a separator is estimated for the data. The data should be transformed in such a way that a separator could be drawn as a hyperplane.

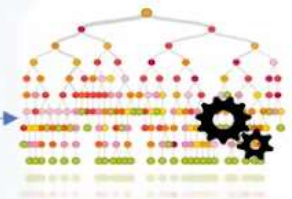
Classification with SVM

ID	Clump	UnifSize	UnifShape	MargAdh	SingEpiSize	BareNuc	BlandChrom	NormNucl	Mit	Class
1000025	5	1	1	1	2	1	3	1	1	benign
1002945	5	4	4	5	7	10	3	2	1	benign
1015425	3	1	1	1	2	2	3	1	1	malignant
1016277	6	8	8	1	3	4	3	7	1	benign
1017023	4	1	1	3	2	1	3	1	1	benign
1017122	8	10	10	8	7	10		7	1	malignant
1018099	1	1	1	1	2	10	3	1	1	benign
1018561	2	1	2	H	2	1	3	1	1	benign
1033078	2	1	1	1	2	1	1	1	5	benign
1033078	4	2	1	1	2	1	2	1	1	benign

Modeling

ID	Clump	UnifSize	UnifShape	MargAdh	SingEpiSize	BareNuc	BlandChrom	NormNucl	Mit	Class
1000015	6	1	1	1	7	1	3	1	1	

Prediction



Imaginez que vous ayez obtenu un ensemble de données contenant des caractéristiques de milliers d'échantillons de cellules humaines extraits de patients que l'on croyait à risque de développer un cancer. L'analyse des données originales a montré que de nombreuses caractéristiques différaient significativement entre les échantillons **bénins** et **malins**. Vous pouvez utiliser les valeurs de ces caractéristiques de cellule dans des échantillons d'autres patients, pour donner une indication précoce si un nouvel échantillon peut être **bénin** ou **malin**.

Imagine that you've obtained a dataset containing characteristics of thousands of human cell samples extracted from patients who were believed to be at risk of developing cancer. Analysis of the original data showed that many of the characteristics differed significantly between **benign** and **malignant** samples. You can use the values of these cell characteristics in samples from other patients, to give an early indication of whether a new sample might be benign or malignant.

What is SVM?

SVM is a supervised algorithm that classifies cases by finding a separator.

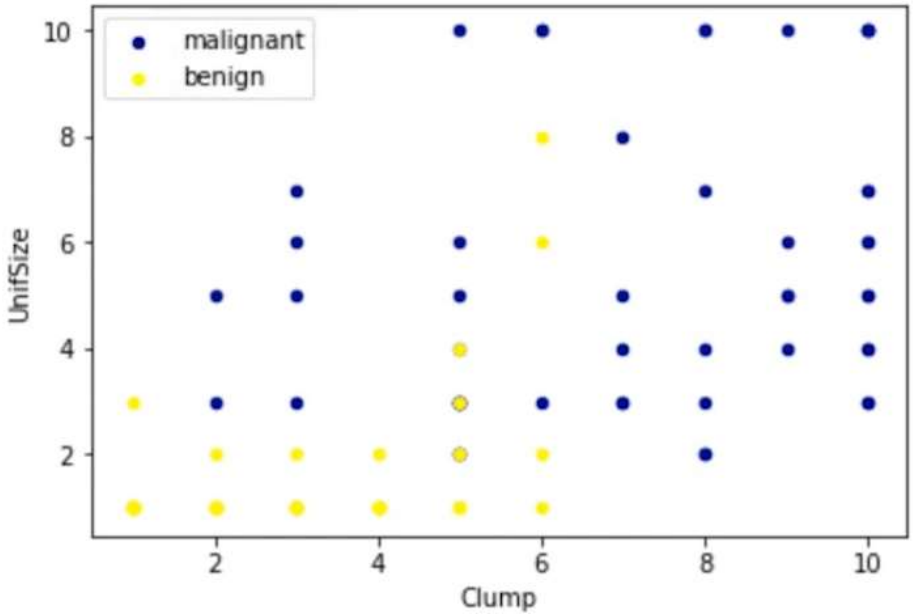
1. Mapping data to a **high-dimensional** feature space
2. Finding a **separator**

Clump	UnifSize	UnifShape	MargAdh	SingEpiSize	BareNuc	BlandChrom	NormNucl	Mit	Class
5	1	1	1	2	1	3	1	1	benign
5	4	4	5	7	10	3	2	1	benign
3	1	1	1	2	2	3	1	1	malignant
6	8	8	1	3	4	3	7	1	benign
4	1	1	3	2	1	3	1	1	benign
8	10	10	8	7	10		7	1	malignant
1	1	1	1	2	10	3	1	1	benign
2	1	2	H	2	1	3	1	1	benign
2	1	1	1	2	1	1	1	5	benign
4	2	1	1	2	1	2	1	1	benign

Example

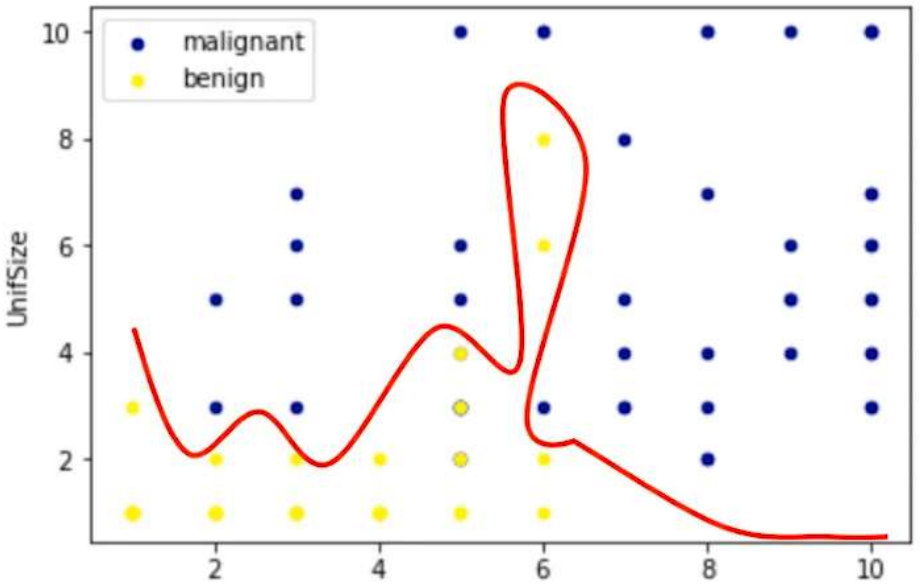
Considérons la figure suivante, qui montre la distribution de petit ensemble de cellules uniquement en fonction de leur taille unitaire et de l'épaisseur de leur touffe. Comme vous pouvez le voir, les points de données se répartissent en deux catégories différentes. Il représente un ensemble de données **linéairement non séparable**. Les deux catégories peuvent être séparées par une courbe mais pas par une ligne. C'est-à-dire qu'il représente un ensemble de données linéairement non séparable, ce qui est le cas pour la plupart des ensembles de données du **monde réel**.

Clump	UnifSize	UnifShape	MargAdh	SingEpiSize	BareNuc	BlandChrom	NormNucl	Mit	Class
5	1	1	1	2	1	3	1	1	benign
5	4	4	5	7	10	3	2	1	benign
3	1	1	1	2	2	3	1	1	malignant
6	8	8	1	3	4	3	7	1	benign
4	1	1	3	2	1	3	1	1	benign
8	10	10	8	7	10		7	1	malignant
1	1	1	1	2	10	3	1	1	benign
2	1	2	H	2	1	3	1	1	benign
2	1	1	1	2	1	1	1	5	benign
4	2	1	1	2	1		1	1	benign



- Les deux catégories peuvent être séparées par une courbe mais pas par une ligne

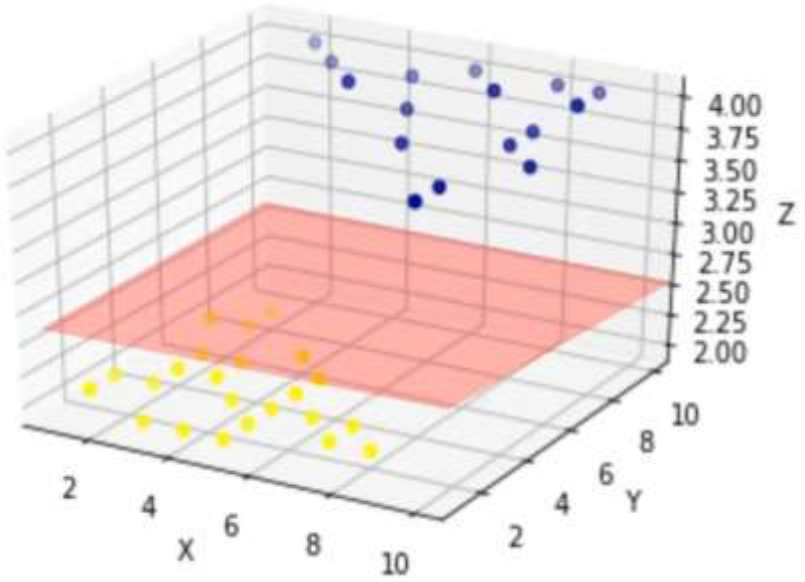
Clump	UnifSize	UnifShape	MargAdh	SingEpiSize	BareNuc	BlandChrom	NormNucl	Mit	Class
5	1	1	1	2	1	3	1	1	benign
5	4	4	5	7	10	3	2	1	benign
3	1	1	1	2	2	3	1	1	malignant
6	8	8	1	3	4	3	7	1	benign
4	1	1	3	2	1	3	1	1	benign
8	10	10	8	7	10		7	1	malignant
1	1	1	1	2	10	3	1	1	benign
2	1	2	H	2	1	3	1	1	benign
2	1	1	1	2	1	1	1	5	benign
4	2	1	1	2	1	2	1	1	benign



For example, consider the following figure, which shows the distribution of a small set of cells only based on their unit size and clump thickness. As you can see, the data points fall into two different categories.It represents a linearly non separable data set. The two categories can be separated with a curve but not a line.That is, it represents a linearly non separable data set, which is the case for most real world data sets.

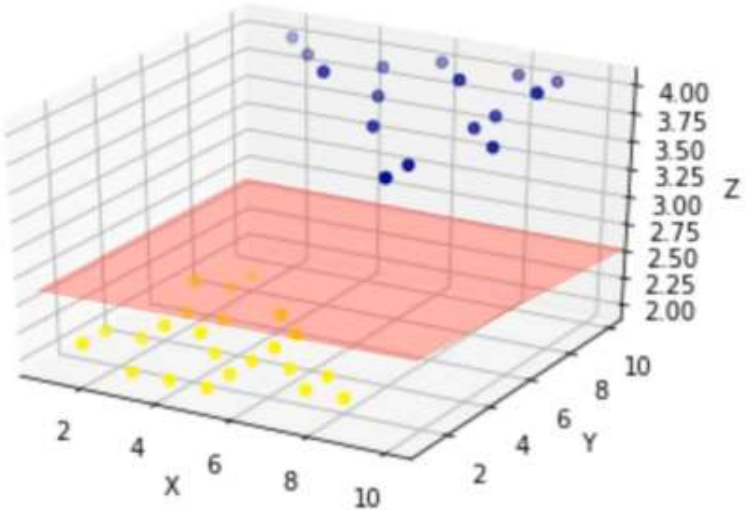
Nous pouvons transférer ces données dans un espace de dimension supérieure, par exemple, en le mappant à un espace tridimensionnel. Après la transformation, la frontière entre les deux catégories peut être définie par un hyperplan.

Clump	UnifSize	UnifShape	MargAdh	SingEpiSize	BareNuc	BlandChrom	NormNucl	Mit	Class
5	1	1	1	2	1	3	1	1	benign
5	4	4	5	7	10	3	2	1	benign
3	1	1	1	2	2	3	1	1	malignant
6	8	8	1	3	4	3	7	1	benign
4	1	1	3	2	1	3	1	1	benign
8	10	10	8	7	10		7	1	malignant
1	1	1	1	2	10	3	1	1	benign
2	1	2	H	2	1	3	1	1	benign
2	1	1	1	2	1	1	1	5	benign
4	2	1	1	2	1	2	1	1	benign



We can transfer this data to a higher-dimensional space, for example, mapping it to a three-dimensional space. After the transformation, the boundary between the two categories can be defined by a hyperplane.

Clump	UnifSize	UnifShape	MargAdh	SingEpiSize	BareNuc	BlandChrom	NormNucl	Mit	Class
5	1	1	1	2	1	3	1	1	benign
5	4	4	5	7	10	3	2	1	benign
3	1	1	1	2	2	3	1	1	malignant
6	8	8	1	3	4	3	7	1	benign
4	1	1	3	2	1	3	1	1	benign
8	10	10	8	7	10		7	1	malignant
1	1	1	1	2	10	3	1	1	benign
2	1	2	H	2	1	3	1	1	benign
2	1	1	1	2	1	1	1	5	benign
4	2	1	1	2	1	2	1	1	benign



Remarque/Remark

Ce hyperplan peut être utilisé pour classer les cas nouveaux ou inconnus.Par conséquent, l'algorithme SVM génère un hyperplan optimal qui catégorise les nouveaux exemples.

Maintenant, il y a deux questions difficiles à considérer.

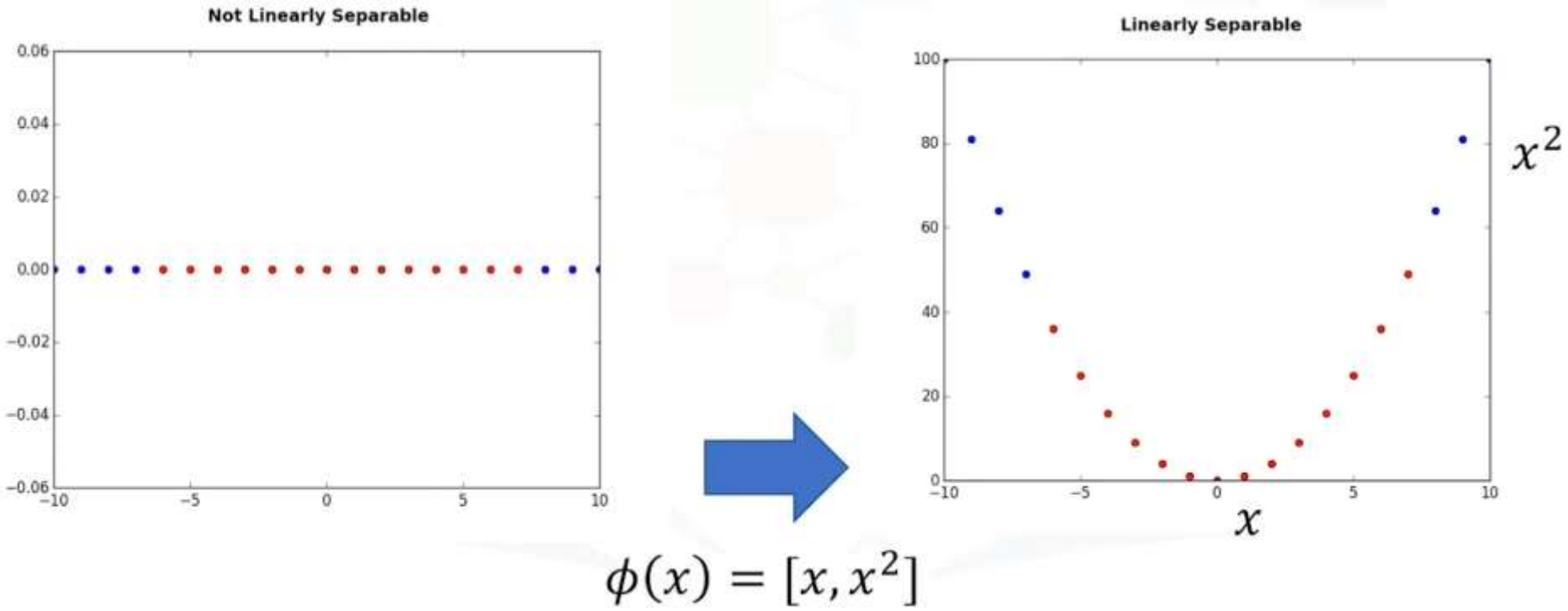
1. Premièrement, comment transférons-nous des données de telle manière qu'un séparateur pourrait être dessiné comme un **hyperplan** ?

2. Et deux, comment pouvons-nous trouver le meilleur ou

Séparateur hyperplan optimisé après transformation ?

Imaginez que notre ensemble de données est des données unidimensionnelles. Cela signifie que nous n'avons qu'une seule caractéristique x.Comme vous pouvez le voir, il n'est pas linéairement séparable. Alors que pouvons-nous faire ici ?

Eh bien, nous pouvons le transférer dans un espace à deux dimensions. Par exemple, vous pouvez augmenter la dimension des données en mappant x dans un nouvel espace utilisant une fonction avec les sorties x et x au carré.



Data Transformation

For the sake of simplicity, imagine that our dataset is one-dimensional data. This means we have only one feature x.

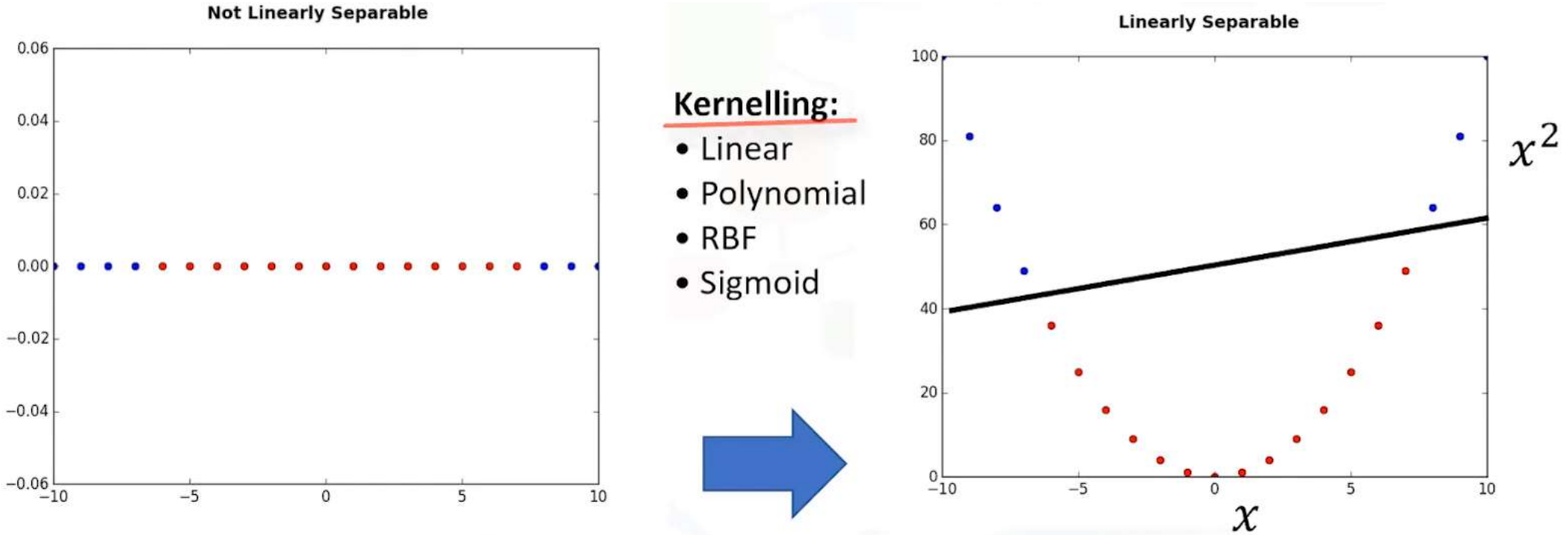
As you can see, it is not linearly separable. So what can we do here?

Well, we can transfer it into a two-dimensional space. For example, you can increase the dimension of data by mapping x into a new space using a function with outputs x and x squared.

NB

Notez que comme nous sommes dans un espace à deux dimensions, l'hyperplan est une ligne diviser un hyperplan en deux parties où chaque classe se trouve de chaque côté.

Nous pouvons maintenant utiliser cette ligne pour classer les nouveaux cas. Fondamentalement, le mappage des données dans un espace de dimension supérieure est appelé, noyautage(kernelling).



$$\phi(x) = [x, x^2]$$

La fonction mathématique utilisée pour la transformation est connue sous le nom de noyau fonction, et peut être de différents types, tels que linéaire, polynôme, fonction de base radiale ou RBF et sigmoïde.

Chacune de ces fonctions a ses propres caractéristiques, ses avantages et ses inconvénients, et son équation.

Mais la bonne nouvelle est que **vous n'avez pas besoin de les connaître** car la plupart d'entre eux sont déjà implémentés dans des bibliothèques de langages de programmation de science des données(Machine Learning Language)

The mathematical function used for the transformation is known as the kernel function, and can be of different types, such as linear, polynomial, Radial Basis Function,or RBF, and sigmoid.

Each of these functions has its own characteristics, its pros and cons, and its equation.But the good news is that you don't need to know them as most of them are already implemented in libraries of data science programming languages.

NB:

Notice that as we are in a two-dimensional space, the hyperplane is a line dividing a plane into two parts where each class lays on either side. Now we can use this line to classify new cases. Basically, mapping data into a higher-dimensional space is called, kernelling.

Pros and Cons of SVM

Advantages:

The two main advantages of support vector machines are that they're **accurate in** high-dimensional spaces**.** And they use a subset of training points in the decision function called, support vectors, so it's also **memory efficient.**

Disadvantages:

The disadvantages of Support Vector Machines include the fact that the algorithm is **prone for over-fitting** if the number of features is much greater than the number of samples. Also, SVMs do not directly **provide probability estimates**, which are desirable in most classification problems. And finally, SVMs are not very efficient computationally if your **dataset is very big**, such as when you have more than 1,000 rows.

Pos & Cons

Avantages:

Les deux principaux avantages des machines à vecteurs de support sont qu'elles sont **précises** dans les espaces de grande dimension. Et ils utilisent un sous-ensemble de points d'apprentissage dans la fonction de décision appelée vecteurs de support, donc c'est aussi **efficace en mémoire**.

Désavantages:

Les inconvénients des machines à vecteurs de support incluent le fait que l'algorithme est sujet au **surajustement** si le nombre de caractéristiques est bien supérieur au nombre d'échantillons. De plus, les SVM ne fournissent pas directement **d'estimations de probabilité**, ce qui est souhaitable dans la plupart des problèmes de classification. Et enfin, les SVM ne sont pas très efficaces en termes de calcul si votre ensemble de **données est très volumineux(big dataset)**, par exemple lorsque vous avez plus de 1 000 lignes

SVM Applications

In which situation should we use SVM?

1. SVM is good for **image analysis tasks**, such as image **classification and handwritten digit recognition.**
2. Also, SVM is very effective **in text mining tasks**, particularly due to its effectiveness in dealing with high-dimensional data.
3. For example, it is used for **detecting spam**, **text category assignment** and **sentiment analysis**.
4. Another application of SVM is in **gene expression data classification**,

Questions

Tutorial