1. Nêu khái niệm và ý tưởng thực hiện của Support Vector Machine (SVM) trong classification.

SVM là thuật toán Machine learning phổ biến, tính toán, xử lý hiệu quả. Dùng được trong bài toán classification, regression và outliers detection. Thì trong bài toán classification SVM có các loại như linear SVM, non-linear SVM.

+ Linear SVM: Là sử dụng đường thẳng để phân tách hai nhóm dữ liệu khác biệt riêng giúp nhận dạng mỗi bên dữ liệu là dữ liệu gì. Không chỉ chia đôi, tách biệt hai lớp dữ liệu mà khoảng cách của các instances phải nằm xa đường phân tách đó.

+ Non-linear SVM: Thuật toán linear SVM tốt nhưng chỉ với dữ liệu có thể dùng đường thẳng phân chia mà trên thực tế không phải tập dữ liệu nào cũng như vậy. Bằng cách sử dụng đường cong để phân loại các lớp ra. Thì non-linear SVM có nhiều phương pháp để học ra đường cong cho bài toán đó.

1. So sánh hard margin và soft margin methods trong binary classification.

Hard margin là tất cả những sample của class nằm 1 phía của margin.

Vấn đề: dữ liệu không có đường thẳng nào phân tách ra được thì không bao giờ tìm được hard margin (nonlinear); nhạy cảm theo outlier.

Soft margin methods: cho phép 1 số sample nằm phía còn lại của margin -> large margin -> better generalization

Vấn đề: khả năng sai nhiều hơn so với Hard margin.

1. Trình bày các phương pháp classification cho nonlinear data sử dụng SVM.

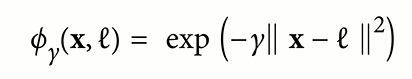
Các phương pháp classification cho nonlinear data sử dụng SVM:

+ Add polinomial features: add features -> fit linear SVM -> non linear boundary (Nhược điểm: thời gian huấn luyện lâu, số lượng feature tăng hoặc không chạy được) -> kernel trick

+ Add similarity features: 1-feature data (1D space) -> 2-feature data (2D space)

Some similarity functions:  
- Euclidean distance

- String kernels

- Gaussian Radial Basis Function 

Nếu gama càng lớn thì nó càng tập trung quanh landmark

Given a 1D data (data with 1 feature):

- 1 landmark -> add 1 new feature -> 1D->2D

- 2 landmark -> add 2 new feature -> 1D->3D

- K landmark -> add K new feature -> 1D->(K+1)D

-> Data are transformed into new feature space (higher dimension)

How to select the landmarks?

The simplest way: each sample -> 1 landmark

-> Higher dimension: more chance to find linear separation

However, with m samples, you add m features -> slow

-> Use kernel trick!

Which kernel to use?

\* Always try the linear kernel first

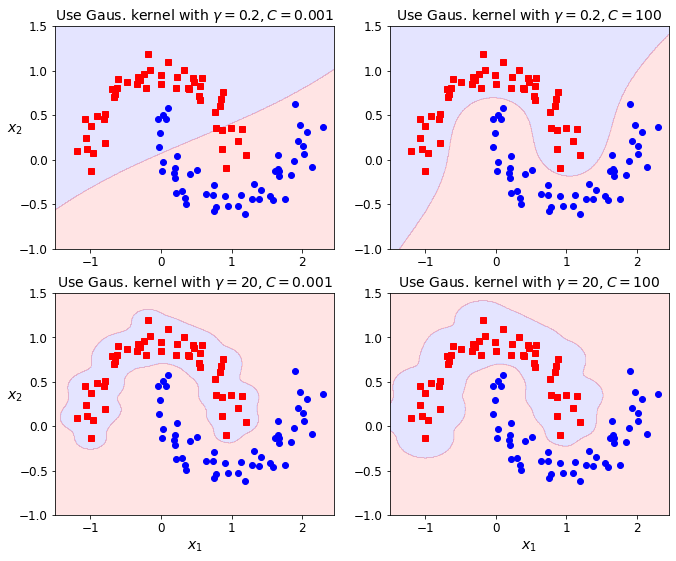
\* Gassian kernel works well in most cases

\* Try a few other kernels using grid-search

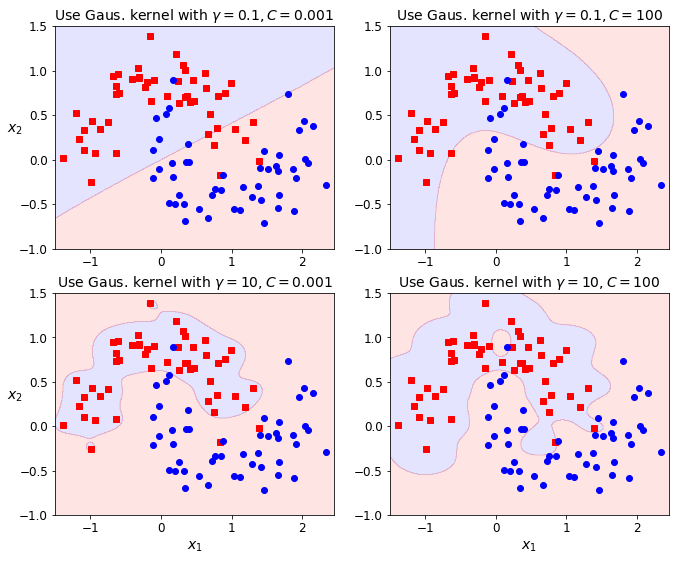
SVM is only directly applicable for two-class tasks.

Bài tập 3: Noise và gamma càng lớn thì ít underfitting và ngược lại

Noise 0.21 gamma 0.2 20



Noise 0.11 gamma 0.1 10



Noise 0.21 gamma 0.2 20

