Fullname: Nguyễn Ngọc Thiện

Student ID: 19110148

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

Khái niệm: SVM là một thuật toán giám sát có độ chính xác cao, được sử dụng rộng rãi và linh hoạt, có thể thực hiện Classification, Regression, Clustering, ... Tuy nhiên nó thường được sử dụng chủ yếu cho việc phân loại

Ý tưởng: tìm một đường thẳng để ngăn cách các điểm dữ liệu với nhau thành các miền giá trị riêng biệt, để tìm ra đường thẳng hoàn hảo nhất cách xa các miền giá trị riêng biệt nhất. Bằng cách dựng 1 trục chính giữa và 2 đường thẳng 2 bên, sau đó đẩy nó ra sao cho chạm điểm đầu tiên của miền giá trị. Xoay trục sao cho khoảng cách của 2 đường thẳng 2 bên là xa nhất => tốt nhất.

1. *[Không bắt buộc]* So sánh hard margin và soft margin methods trong binary classification.

Giống nhau: Đều ngăn chia feature space thành các miền khác nhau

|  |  |  |
| --- | --- | --- |
|  | Hard margin | Soft margin |
| Khác nhau | Nó không có phép các miền giá trị có các loại dữ liệu khác nhau -> cứng ngắt, khó tính. Một class chỉ thuộc một side.  Có thể margin sẻ nhỏ | Ngược lại với hard margin thì nó cho phép một class chứ 2 hoặc nhiều hơn dữ liệu khác nhau thuộc 1 side -> linh hoạt, dễ tính  Thường magrin sẽ lớn (giữ nguyên cực đại) vì nó chấp nhận sample khác nhau |
| Vấn đề | Có vài sample nằm khác phía (dữ liệu lạ) -> kết quả error  Không làm được nếu dữ liệu không tách rời nhau | Chấp nhận sample nằm ở side khác, vì vậy kết quả có thể kém chính xác.  Tuy nhiên với tập dữ liệu thực tế nó sẽ chạy rất tốt. |

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

Có 2 phương pháp:

* Add polinomial features: thêm các feature bậc cao để cho model học và tạo ra đường phi tuyến để chia các dữ liệu. Để làm được điều này ta sử dụng thêm kernel trick để tạo ra decision boundary phức tạp xấp xĩ polinomial bậc cao nhưng không cần thêm features bậc cao
* Add similarity features : tạo các features tương tự trong không gian 1, 2, ... chiều thành nhiều chiều. Không gian càn nhiều chiều sẽ mô tả tốt các dữ liệu hơn.

1. *[Không bắt buộc]* Đọc hiểu và trình bày tóm tắt các phần sau trong sách (Géron, 2019):
   * Decision Function and Predictions (page 166)
   * Training Objective (page 167)
   * The Dual Problem (page 170)
   * Kernelized SVM (page 171)

Phần lập trình

1. Chọn 2 classes trong Iris data (**khác với 2 class** demo trong bài học), thực hiện chạy từng bước các thuật toán được học và giải thích kết quả.
2. Chọn một số giá trị khác nhau cho 2 tham số **degree** (trong PolynomialFeatures) và **C** (trong LinearSVC) ở **mục 4.2 (Polynomial kernel)**. Huấn luyện SVM và giải thích kết quả.   
   *Yêu cầu:*   
   Chạy thử ít nhất 4 trường hợp: 2 giá trị khác nhau (một giá trị nhỏ hơn, một giá trị lớn hơn trong code demo) cho mỗi tham số **degree** và **C**.
3. Thử ít nhất 2 giá trị khác nhau cho tham số **noise** (hàm **make moons**()) và các hyperparameters **gamma** và **C** của Gaussian kernels trong **mục 5.6 (Gaussian kernel)**.   
   So sánh và giải thích kết quả classification của từng bộ giá trị (*ghi comments trong code*).