



CLASSIFICATION

- Thuật toán kNN
- Đánh giá giải thuật phân lớp
- Ứng dụng

K-NN ALGORITHM

- Là thuật toán supervised-learning đơn giản nhất.
- Được xếp vào lớp giải thuật *lười học*

K-NN ALGORITHM

- Ý tưởng hoạt động
 - Với mỗi điểm dữ liệu, nhận được suy ra từ nhãn của điểm dữ liệu láng giềng gần nhất trong training dataset
 - Xét điểm gần nhất bằng:
 - độ đo Euclide
 - tính trọng số khoảng cách

K-NN ALGORITHM

- Độ đo khoảng cách Euclide

$$dist(X_1, X_2) = \sqrt{\sum_{i=1}^n (x_{1i} - x_{2i})^2} \quad \text{với} \quad X_1 = (x_{11}, x_{12}, \dots, x_{1n}) \text{ và } X_2 = (x_{21}, x_{22}, \dots, x_{2n})$$

- Chuẩn hóa dữ liệu

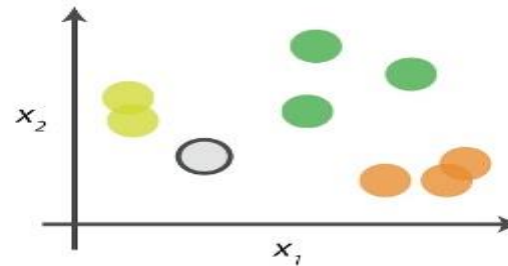
$$v' = \frac{v - \min_A}{\max_A - \min_A}$$

K-NN ALGORITHM

- Minh họa thuật toán

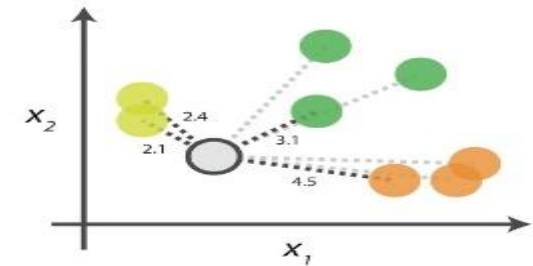
kNN Algorithm

0. Look at the data









Say you want to classify the grey point into a class. Here, there are three potential classes - lime green, green and orange.

1. Calculate distances









Start by calculating the distances between the grey point and all other points.

2. Find neighbours

Point Distance			
	→		2.1 → 1st NN
	→		2.4 → 2nd NN
	→		3.1 → 3rd NN
	→		4.5 → 4th NN

Next, find the nearest neighbours by ranking points by increasing distance. The nearest neighbours (NNs) of the grey point are the ones closest in dataspace.

3. Vote on labels

Class	# of votes	
	2	→ Class  wins the vote! Point  is therefore predicted to be of class  .
	1	
	1	

Vote on the predicted class labels based on the classes of the k nearest neighbours. Here, the labels were predicted based on the k=3 nearest neighbours.

K-NN ALGORITHM

- Ưu điểm:
 - Độ phức tạp tính toán của quá trình training là bằng 0.
 - Không tốn thời gian để *học*.
- Nhược điểm:
 - Tốn thời gian tính toán trong quá trình phân lớp
 - Nhạy cảm với giá trị nhiễu

K-NN ALGORITHM

- Tăng tốc thuật toán
 - Sắp xếp tập dữ liệu đầu vào dưới dạng cây tìm kiếm → $O(\log(n))$
 - Pruning: cắt tỉa những phần tử dữ liệu biết chắc là vô nghĩa

K-NN ALGORITHM

- Code minh họa thuật toán

ĐÁNH GIÁ THUẬT TOÁN

- Một số công thức đánh giá

TP: SL phần tử đoán đúng lớp +1

TN: SL phần tử đoán đúng lớp -1

FP: SL phần tử đoán nhầm +1 sang -1

FN: SL phần tử đoán nhầm -1 sang +1

Bảng 6.2 Ma trận lẫn lộn

<i>Lớp thực tế</i>	<i>Lớp được dự đoán bởi giải thuật phân lớp</i>	
	+1	-1
+1	TP	FN
-1	FP	TN

- Tỉ lệ lỗi tổng thể:

$$Error = \frac{FP + FN}{TP + FP + TN + FN} \times 100\%$$

- Độ chính xác tổng thể:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \times 100\%$$

ĐÁNH GIÁ THUẬT TOÁN

- Chia dữ liệu Theo phương pháp holdout (lấy ngẫu nhiên 2/3 dữ liệu cho training và 1/3 cho test → chất lượng giải thuật rất tốt //rất xấu

ĐÁNH GIÁ THUẬT TOÁN

- Phương pháp: lấy mẫu ngẫu nhiên (random subsampling)

Thực hiện chia dữ liệu k lần

- Mỗi lần thực hiện training và test
- Lấy kết quả trung bình của tất cả các lần test là kết quả của thuật toán

ĐÁNH GIÁ THUẬT TOÁN

- Phương pháp: thẩm định chéo (cross-validation)

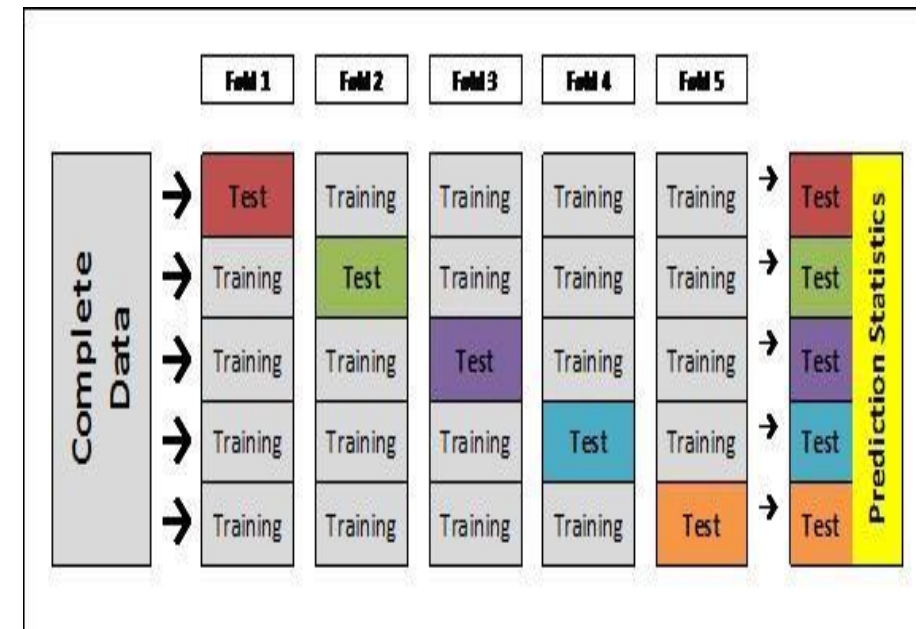
Chia thành k tập với kích thước gần

bằng nhau (D_1, D_2, \dots, D_k)

Lặp lại quá trình đó k lần

- Tại lần lặp thứ i tập D_i sẽ là test set, các tập còn lại là training test.

→ Đảm bảo tính ngẫu nhiên của thuật toán.



ỨNG DỤNG



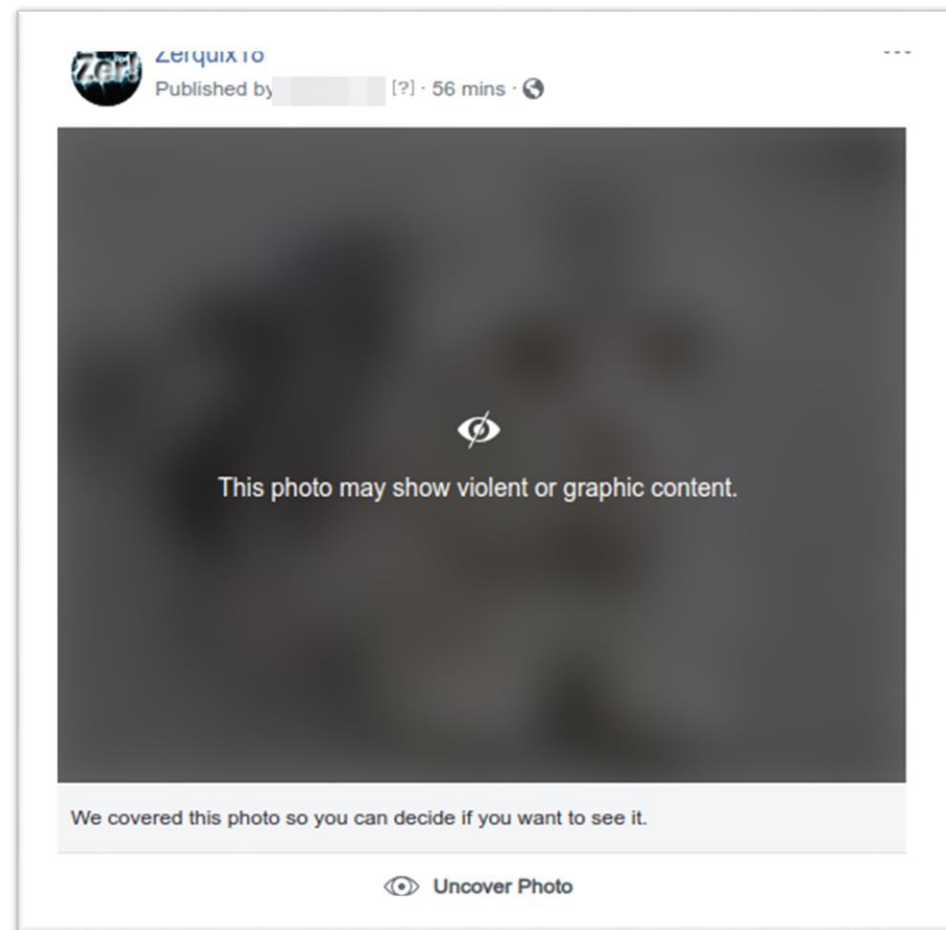
- Trong các mail server (như gmail hay yahoo), chúng ta vẫn thấy các hệ thống lọc thư rác, nó có khả năng phân loại được các thư rác (spam mail) và đưa vào thùng rác. Chức năng này làm cho người dùng thấy rất thuận tiện và tránh được bức mình.

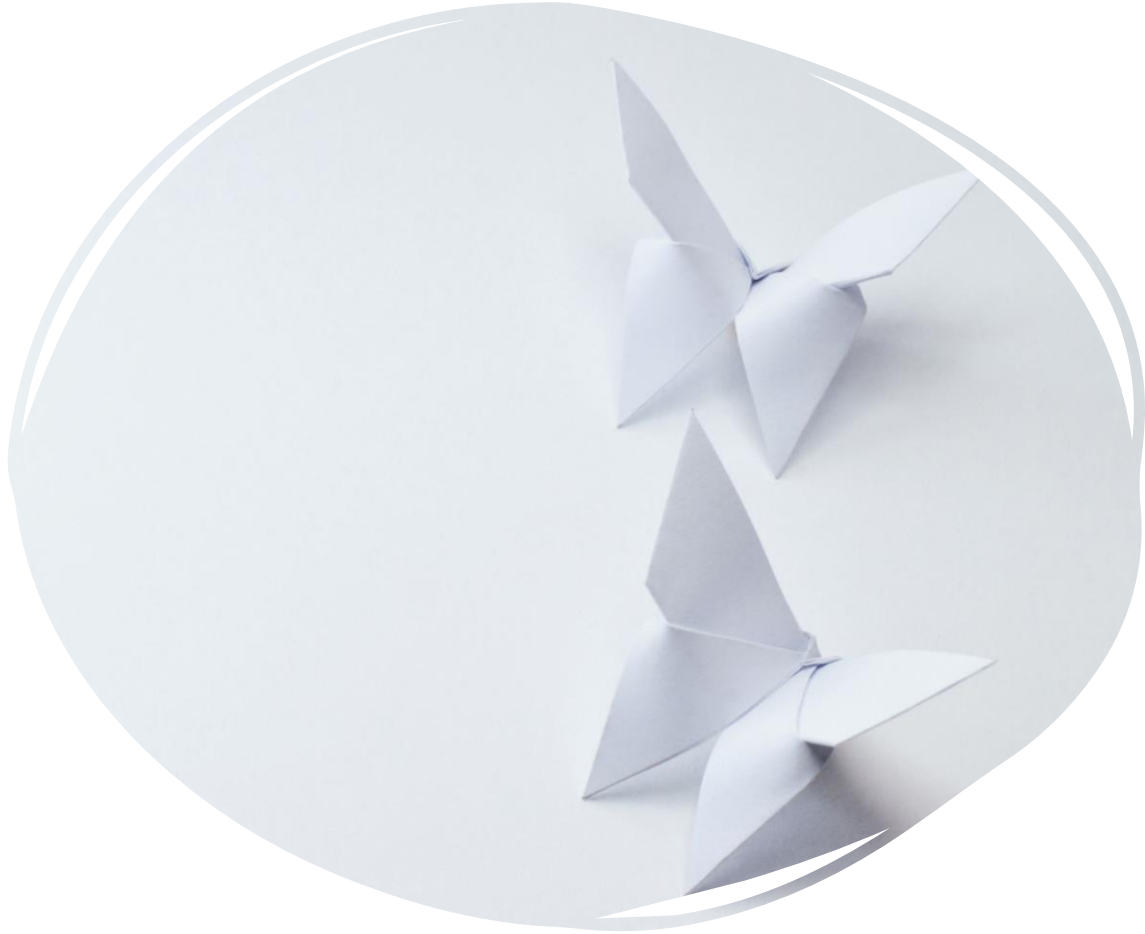
- Trong ngân hàng, khi xem xét hồ sơ của một khách hàng cần vay vốn, nếu ta có thể phân lớp được khách hàng này thuộc lớp “an toàn” hay “mạo hiểm” thì sẽ có ý nghĩa rất quan trọng cho người ra quyết định cho vay vốn.



ỨNG DỤNG

- Các dịch vụ trực tuyến (chia sẻ ảnh, tin hay video) rất cần có một hệ phân lớp có khả năng phát hiện ra các bản tin, các hình ảnh hay video có nội dung không phù hợp như các nội dung dung tục, hay không phù hợp với văn hóa, chính trị, ...





THANK YOU

- Lưu Hoài Linh
- Đinh Xuân Hùng
- Nguyễn Linh