Machine learning note

Khi phân cụm K means làm việc không hiệu quả do 2 cụm bị đè lên nhau có thể sử dụng phân cụm spectral

Cách thay một điểm dữ liệu bằng tâm cụm tương ứng là một trong số các kỹ

thuật có tên chung là vector quantization – VQ

Cơ sở dữ liệu Ling-Spam dataset

Kỹ thuật sử dụng để giảm tốc độ học trong gradient descent - learning rate decay

Nesterov accelerated gradient giúp giảm tốc độ hội tụ cho gradient descent with momentum

stochastic gradient descent (SGD) được sử dụng khi cần tính toán nhanh, cho các hệ thống online learning, thuật toán này có thể gây ra sai số nhưng mang lại lợi ích về mặt tính toán. Mỗi lượt duyệt qua hết N điểm dữ liệu gọi là 1 epoch, batch GD thì 1 epoch tương đương với t lần cập nhật theta, SGD thì 1 epoch tương đương với N lần cập nhất theta, SGD có thể hội tự chỉ sau vài epoch

perceptron learning algorithm (PLA) là thuật toán đi tìm siêu phẳng trong bài toán phân loại nhị phân nếu ranh giới đó tồn tại (linearly seperable)

Khắc phục nhược điểm cho tập dữ liệu gần tuyến tính của thuật toán PLA là thuật toán bỏ túi (pocket algorithm)

solver = ‘lbfgs’ là phương pháp tối ưu rất mạnh dùng trong gradient descent

Hàm ReLU đang được sử dụng nhiều trong mạng neuron, có các biến thể như leaky rectified linear unit, parametric rectified linear unit, randomized leaky rectified linear units

||X||^2\_F = trace(X \* X.T)

Khi ma trận A lớn, việc tính toán SVD tốn nhiều thời gian. Cách tính SVD

cắt ngọn với k như được trình bày trở nên không khả thi. Có một phương

pháp lặp giúp tính các trị riêng và vector riêng của một ma trận lớn một cách

hiệu quả. Trong phương pháp này, ta chỉ cần tìm k trị riêng lớn nhất của A \* A.T và các vector riêng tương ứng. có thể tìm thấy tại Power method for approximating eigenvalues

Tổng hợp các thuật toán

* K lân cận ( KNN) – học có giám sát,
* K means

Hạn chế: số cụm K cần được xác định trước. Trong trường hợp chưa biết trước thì có thể dùng phương pháp elbow để xác định

Nghiệm cuối cùng phụ thuộc vào các tâm cụm ban đầu

Các cụm có số lượng điểm gần bắng nhau

Khi một cụm nằm trên cụm khác K means ko cho kết quả tốt

* Naive Bayes ( thường được sử dụng trong các bài toán phân loại email )
* Thuật toán perceptron

PLA là một thuật toán đơn giản giúp tìm ranh giới siêu phẳng cho bài toán phân loại nhị phân trong trường hợp tồn tại siêu phẳng đó. Nếu hai dữ liệu là tách biệt tuyến tính

* Hồi quy tuyến tính (dùng cho bài toán phân loại nhị phân)
* Lasso Regression
* XGBoost regressor
* Random Forest
* Single Decision Tree
* Hồi quy logistic (dùng cho bài toán phân loại nhị phân là chủ yếu)
* Hồi quy softmax (mở rộng của hồi quy logistic, dành cho bài toán phân loại đa lớp mà logistic có nhiều hạn chế)
* Các thuật toán giảm chiều dữ liệu

+ Phân tích giá trị suy biến (SVD)

+ Phân tích thành phần chính (PCA)

+ Phân tích biệt thức tuyến tính (LDA)

* Máy vector hỗ trợ (Support vector machine)
* **Under sampling là** việc ta giảm số lượng các quan sát của nhóm đa số để nó trở nên cân bằng với số quan sát của nhóm thiểu số.
* **WCSS value**

Sigmoid and Tanh function has vanishing gradient problem

Activation function for hidden layers in neural networks maybe ReLU, Leaky ReLU