

Buổi học MaSSP số 2

July 2019

Nguyen Dinh Hieu

1 Unified Framework

Với bộ dữ liệu đầu vào X và bộ dữ liệu đầu ra Y , cần xây dựng một phương thức giúp máy tính có thể nhận vào 1 dữ liệu x bất kì và đưa ra kết quả chính xác như ta mong muốn. Đầu tiên, ta dùng phương pháp PCA để giảm số chiều dữ liệu cần phải làm việc (nhưng vẫn giữ lại độ chính xác cao) và thu được các vector coordinate dựa trên những basis functions/features cho sẵn. Sau đó, bằng các thuật toán khác nhau về regression hoặc classification, ta thu được ma trận hệ số W phù hợp giúp thuật toán của ta đạt độ chính xác cao.

2 PCA

Biểu diễn 1 dữ liệu đầu vào X theo dạng

$$X = X_1 y_1 + X_2 y_2 + \dots + X_n y_n$$

với X_1, X_2, \dots là các basis functions cho trước, còn y_1, y_2, \dots là các coordinates tương ứng. Đồng thời, ta có thể giảm số components, giúp giảm độ phức tạp tính toán (do chỉ cần quan tâm đến các chi tiết liên quan tới basis functions)

3 Linear regression

Thuật toán linear regression (tìm hệ số thực cho hàm số tuyến tính) giúp ta gán các hệ số w_0, w_1, \dots cho hàm số cho trước. Với bộ dữ liệu/experience có sẵn, ta dùng linear regression tìm các hệ số để chúng cho ra kết quả y-head gần với kết quả chính xác là y . Công thức:

$$\mathbf{w} = \mathbf{A}^\dagger \mathbf{b} = (\bar{\mathbf{X}}^T \bar{\mathbf{X}})^\dagger \bar{\mathbf{X}}^T \mathbf{y}$$

4 Logistic regression

Logistic regression linh hoạt và dùng được trong nhiều trường hợp. Hàm sigmoid có nhiều tính chất đẹp và thuận lợi để sử dụng như một hàm đánh giá. Ta biểu diễn dữ liệu input và các coordinates dưới dạng xác suất theo cross-entropy. Sau đó, dùng SGD (Stochastic Gradient Descent) để tối ưu hàm số. Phương pháp này hoạt động nhanh với độ chính xác cao bằng việc chọn random một điểm dữ liệu để tính đạo hàm mỗi lần.

Công thức sigmoid:

$$\sigma = \frac{1}{1 + e^{-s}}$$

Công thức cross entropy:

$$J = - \sum_{i=1}^d P_{yi} \log P_{yi}$$

5 Softmax

Lấy dữ liệu Z nhân với ma trận trọng số W ta được W_z . Ta dùng softmax để classify thông qua probability.

Công thức softmax:

$$P_i = softmax_i = \frac{\sigma(y_i)}{\sum_{j=1}^d \sigma(y_j)}$$