INTRODUCTION

Briefly, for this binary classification task, we use both the random Fourier transform (RFF) and Polynomial (Poly) transform for original data and Adam update rule for stochastic gradient descent with hinge loss. This method is used to approximate the SVM with kernel and reached 0.9685 for public leaderboard in around 5 mins.

IMPLEMENTATION

For the easy baseline, we simply used online SVM with no optimization and kernel approximation method. For each full batch, we calculate the negative gradient which is –yx. For the Adam or momentum+Nesterov update, we would use the negative gradients above. But this simple method cannot pass the hard baseline. So further, we try to transform the data using random Fourier transform or Poly and get much larger dimension (20000) for data. Also, for training, we use mini-batch for training in each iteration and use momentum + Nestorov or Adam, which are similar in performance. For details, if we use high dimension for random features, but we need to adjust the iteration and mini-batch numbers in order to avoid timeout error.

We found that the learning rate of Adam, the iterations of optimization, the initial of weights are quite important for final result, which have influence on finding the optimal solution for the datasets intuitively. And the most important factor for Poly transform is the number of dimension.