Comp 9418 Assignment 3 Report

Group member

1. Meng Sun (z5149213)
2. Zechao Li (z5172016)

Abstract

In this assignment the major part of work is pre-processing the data derived from text chunking and utilize Gaussian processing to train model. Using the model to predict probability of the test data. Record the log probabilities of each class on the given test data.

Introduction

The first part of the assignment is to pre-process the training data given. The total amount of files given is 8931. In each file there are about 20-30 chunks of words. There are 211727 chunks of words or data in total. The highest dimension of these data is 2035523. So, the amount of data and the dimension of it is both very high, directly dealing with such huge size of data is impossible. So, the pre-process method is necessary.

As for the model in training, our group has chosen gpflow as the main tool for us to train data. Gpflow is a package which could build Gaussian process models in python, and it uses Tensorflow, a powerful tool for machine learning. Utilizing gpflow is easy to build model and modify parameters.

Our group have 2 group members: Meng Sun and Zechao Li. Meng Sun mainly works on pre-processing the data and write reports. Zechao Li mainly works on building model, training data and Searching files.

Model

In the first part of pre-processing data, the amount and the dimension of data are both problems, if the size of data too big, it leads either break the limit of memory or makes the training process too slow. First thing is to reduce dimension to facilitate process, to reduce dimension we utilize the Truncated SVD in sklearn. The truncated SVD run process the matrix directly, not on the covariance matrix, which not like PCA method. The truncated SVD is good at decrease the dimension of large sparse matrix, which very suitable for this assignment. The data is reduced to 50 dimensions.

Secondly, the amount of data chunks is too large, which makes hard to train the model, so shrinking the data to about 47510 chunks. The choice of data in different label is based on the amount of data this label has. If the number of chunks is less than 3000, in order to preserve information as much as possible, the train set will take all of these data. For labels have more than 3000, it chooses 25% of data inside it. With such amount of data chunk, the model is trainable.

As for the gaussian predicting model, the model our group’s choice is SVGP(Sparse Variational Gaussian approximation) model, This model is suitable for dealing sparse matrix which in this assignment the data is very sparse. The kernel inside chooses RBF(Radial Basis Function), the likelihood is set to multiclass of 23 classes. Pseudo inputs are set to take 1/30 of training data, with size is suitable to train data. The method of maxiter in model optimizer is notebook\_niter, and the iteration time is set to 2000 times and the minibatch size is set to 1000. The reason and the concept of model will be introduced in the following section.

Inference

The kernel used in the model is RBF, a real-valued function which the value depends only on the distance from the origin or other point. The norm is usually Euclidean distance. The model uses this kernel to make approximation process and makes neural network in tensorflow. The RBF in Gaussian process is represented as follows:

(1)

In order to avoid too large data into model to slow down the training speed or reach the memory limit, every time the model only takes small amount of data to make optimization. This helps to reduce model load and avoids the model to reach local optimization. After some test, setting minibatch to 1000 best balances the speed and correctness.

The model also uses matrix of pseudo inputs to lower the load of model. The matrix of pseudo inputs, will labelled Z, is a subset of input data. The space complexity of gaussian process is O(). In order to lower the space complexity, the model could take part of input data to make sparse approximation to lower down the complexity to O(), where M<<N. This significantly reduce the load of data.[1] Because the input data inside the model are randomly chosen, so although the pseudo inputs are chosen by the interval of 30, it could avoid overfitting.

Above all is the set of parameters in model. Gaussian process is decided by the mean m(x) and covariance , In general, assume the mean to be m, a finite collection of inputs {} follow Gaussian distribution , where []=k().

Based on gaussian process the probability could get:

p(f)=N(f;m;) (2)

p(y|f)= (3)

To make predictions, it needs to determine by optimizing the marginal likelihood and then marginalising over the posterior of f.[2]

(4)

p(y\*|y) = (5)

After training the model, to get the predicted data and the predict result, we will put the test data into model to predict, and get a matrix which has 23 dimensions, each row represents the data put in, and each dimension represents the probability of the label that chunk belong to. To get result, just simply choose the label with highest probability.

Parameter Estimation

The parameter estimation algorithm is based on MLE(maximum likelihood estimation), as (4) shows, In this model, also we select the data to train by the labels, this is only for avoid overfit, and preserve the information, to predict the label, the result only decided by the maximum probability of labels, so the parameter estimation belongs to MLE.

Results

To verify the result, our group will choose subset of data as validation set. The validation set will be chosen randomly from the whole data set. After, first our group tries to make sure every kinds of label appear in the validation set, likes the way of choosing training set. But the result will become lower accuracy, because the amount of data each label contain varies much. The label with less data has lower accuracy of prediction, for the information provided is limited. The best way to show the accuracy of model is to choose data of validation set randomly, ignore the label of data.

This assignment uses error rate(ER) and mean negative log probability(MNLP) to examine the model, which these 2 methods are presented as follows:

(6)

(7)

Except using ER and MNLP to justify the model trained, our group also utilizes logistic regression in sklearn to train data and predict result, and also uses ER and MNLP to verify results.

In the validation set, the amount data used to validate is 10000. The gaussian process model and logistic regression model use the same set of data to train and use same validation set. The ER of gaussian process model is about 0.127, which means the probability of making correct prediction of label is 0.873. The MNLP is 179. For logistic regression, the ER is about 0.146, which means the probability of making correct prediction of label is 0.854. And the MNLP is 214.

Above results prove the model trained by the SVGP module in gpflow has high accuracy and the model is convincible. Except for the models its selves, the data provided in this assignment is very huge in both amount and dimensions. It will consume too much resource and time to train the model. So, our group uses some methods and trick to lower the load of training model, which has mentioned above. The result proves the pre-processing step has reaches great balance between efficiency and correctness.

References

[1] Edward Snelson, Zoubin Ghahramani, Sparse Gaussian Processes using Pseudo-inputs. Advances in Neural Information Processing Systems 18(NIPS2005), 2005

[2] Matthias Bauer, Mark van der Wilk, Carl Edward Rasmussen. Understanding Probabilistic Sparse Gaussian Process Approximations. arXiv.org, 2017.