# Abstract

Each word or token is treated as one data, obtain the dense vector of word or token are representing as a sparse vector (the dimensional is 2,035,523) and use classification algorithms to do it such as Gaussian Process.

Because the dimension is too high, PCA is used to reduce dimension (the dimensional is 3000).

The value of ER is: 0.43

The value of MNLP is: 42

# Introduction

This is a problem of sequence annotation, and it is also a classification problem. Here we consider it as a classification problem, each word corresponds to a category, regardless of the context. Of course, this accuracy rate will be very low. We use Gaussian Process to classification.

# Model

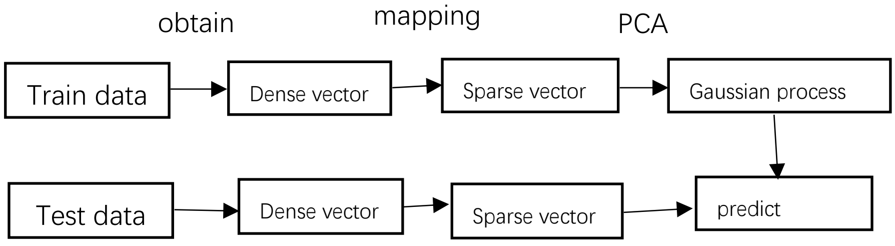
$$\mathrm{Data{:}\ D} = \{(X_i, Y_i)\}_i^N, \mathrm{X} \in R^D, y \in \{0,1,\cdots 22\}$$ -> there are 23 categories.

$$\mathrm{Input{:}}\ (\mathrm{X})_{D*N},\ \ Target{:}\ (y)_{N*1}$$

$$\mathrm{Prior{:}\ f{\sim}GP}\big(0, \mathrm{K}(\mathrm{x}, \acute{x}; \theta)\big), \mathrm{p(f|X)} = N(f|0, K)$$

$$\mathrm{Likelihood{:}\ p(y|f)} = \sigma\big(\mathrm{f(x)}\big) = \prod_{i=1}^{N} p(y_i|f(x_i))$$

$$\sigma(\mathrm{z}) = \frac{1}{1 + \exp(z)}$$



# Inference

Compute predictive distribution of latent functions:

$$\mathrm{p}(f_*|X,y,x_*)=\int p(f_*|X,x_*,f)\cdot p(f|X,y)df$$

$$\mathrm{p(f|X,y) \propto p(f|X) \cdot p(y|f)}$$

Compute probabilistic predictions:

$$\mathrm{p}(y_*=+1|X,y,x_*)=\int \sigma(f_*)\cdot p(f_*|X,y,x_*)df_*$$

# Parameter Estimation

Use Laplace approximation

$$\mathrm{p(f|X,y,\theta) \approx N(f|}\hat{f},A^{-1})$$

where: $\hat{f}=\mathrm{argmax\ p(y|f,\theta)\cdot p(f|X,\theta)}$ and A is the Hessian of the negative log-posterior

evaluated at $\hat{f}$.

# Results

There are two indicators for evaluating the performance of models. ER and MNLP

$$\mathrm{ER}=1-\frac{1}{N}\sum_{i=1}^{N}\frac{1}{T_i}\sum_{k=1}^{T_i}\llbracket[\hat{y}_k^{(i)}=y_i^{(i)}]$$

$$\mathrm{MNLP}=-\frac{1}{N}\sum_{i=1}^{N}\frac{1}{T_i}\sum_{k=1}^{T_i}\sum_{j=1}^{C}\llbracket[y_k^{(i)}=y]logP_{model}(y_k^{(i)}|x^{(i)})$$

The value of ER is: 0.43

The value of MNLP is: 42

If we use the method of cross-validation, the performance will be better.

# References

Gaussian Process Classification, Approximations and Other GP Models - UNSW Sydney

http://scikit-learn.org

Variational Learning of GP Models - UNSW Sydney

# Appendix

The dimension of the feature space is not equal 2035523, only have 90000 dimension, so we using the value of 90000 to mapping sparse vector, will be faster and use less memory.