Research Interests

of *Jiafeng Liu* (Applicant for Fall—2017)

I am a senior student and am planning to pursue a PhD degree after graduating next summer. Below I would like to talk about my research interest. My major research interest is in Bayesian methodology, machine learning and their applications, especially kernel methods.

Research Interests:

(Based on my past research experiences)

Bayesian Statistics#1: I approached Bayesian theory in the course Mathematics Statistics where I learnt about the priori probability for the first time. Bayesian probabilities are based on "degree of belief" and it plays a significant role in making predictions. One of the highlight is that it predicts the probability by previous knowledge and observation. This method keeps updating the priori distribution with the latest observations. It tells us that when we predict an event, we should firstly base on existing experience and knowledge to infer a priori probability, and then we should also update our knowledge by accumulating more evidence.

For instance, in bayesian statistics, the parameter θ is regarded as a random variable, we have following procedure to do bayesian estimation. Firstly, we have observations (training data) $D = \{(x_1, y_1), \ldots, (x_n, y_n)\}$, the parameter of the model is θ which have cdf $\pi(\theta)$. We can regard the observations D is generated under one exact parameter θ_0. Thus we have observations conditional distribution: $p(D|\theta_0)$. While the parameter is not one exact number but follows $\pi(\theta)$,therefore we have **joint distribution:** $h(D, \theta) = p(D|\theta)\pi(\theta)$, **posteriori distribution** is:

$$\pi(D|\theta) = \frac{p(D|\theta)\pi(\theta)}{\int_\Theta p(D|\theta)\pi(\theta)\,d\theta}$$

However the complex integral is difficult to calculate, so we use the Monte Carlo method to get approximate the solution. Monte Carlo Methods provide a new way to approach and also provide a new idea to think of certain event. We can simulate or approach a certain event by doing random event. Bayesian statistics is limited by the computing power. MCMC and other sampling methods improve the range of application. While nowdays, statisticians must face the challenge of big data. New methods or improved methods are needed to increase the efficiency of bayesian formula. That's the area that I am interested in, **a. use less data to approach target function** and **b. overcome the difficulty in computing high dimensional data.**

Kernel Methods#2: [1] I got to know kernel function when I did a project with SVM methods. At the beginning, I thought the kernel trick is just a point-to-point mapping from one dimensional space to high dimensional space. This feature leads me to the study of kernel methods.

After going through literatures about kernel methods and Hilbert space[2], I got a deeper understanding of kernel, that it is an independent area. Not only does kernel map establishes point-to-point mapping, it can also provide methods to represent a probability distribution as an element of reproducing kernel Hilbert space[3]. To my understanding, this feature works in this way: given two independent samples, they can be mapped into an infinite-dimensional space as two points. And then we could figure out whether they obey the same distribution by measuring the distance of two points. One commonly used method to measure the distance is the Maximum Mean Discrepancy (MMD)[4].

However, one drawback of MMD is that it depends on the choice of kernel. Different kernels may result in different test statistics. So the setting of the kernel is of great importance. I would like to propose a new choice of kernel for hypothesis testing, which will markedly optimize the performance of the hypothesis test. My research interest in kernel methods is mainly about enhancing the efficiency of MMD regardless of the structure and size of the data[5].

In a research project of lecture Statistical Computing. I generate a bivariate random sample from a joint distribution based on the Metropolis-Hastings sampling algorithm. Then I used kernel function kde2d() in R from MASS package to estimate the distribution the bivariate random sample would follow. But comparing the kernel estimation distribution and real joint mixture distribution, I found estimation was not stable, especially when I tried kernel estimation several times using different bivariate random sample. And this estimation method also cannot perform well in a large sample size. While I changed the given real joint distribution, according to the contour plot, I found that kernel estimation performed well when there are few peak in the distribution. I realized that the performance may relate with the distribution of the observations .If there is kernel function could perform stable or more efficiency in certain circumstance, we can infer the original distribution from the observations accurately. I have interests in this field, I would like to explore new rules and methods to improve the estimation performance.

References

[1] T. Hofmann, B. Schölkopf, and A. Smola. Kernel methods in machine learning. *The Annals of Statistics*, 36(3)(1171-1220), 2008.

[2] A. Smola, A. Gretton, L. Song, and B. Schölkopf. A hilbert space embedding for distributions. *Algorithmic Learning Theory: 18th International Conference.*, 2007.

[3] L. Song. Learning via hilbert space embedding of distributions. *PhD Thesis, University of Sydney*, 2008.

[4] A. Gretton, K. Borgwardt, M. Rasch, and A. Smola B. Schölkopf. A kernel two-sample test. *Journal of Machine Learning Research*, 13(723-773), 2012.

[5] Qinyi Zhang, Sarah Filippi, Arthur Gretton, and Dino Sejdinovic. Large-scale kernel methods for independence testing. 2016.