Estimation of Probability Density Functions and Graphical Models using regularized Sparse Grids.

Karthikeya Sampa Subbarao

Technical University of Munich, Department of Informatics
Boltzmannstr. 3, D-85748 Garching bei Munich, Germany
karthikeya108@gmail.com

Introduction:

Density estimation reconstructs based on observed data, an unobservable underlying probability density function. The Curse of Dimensionality is frequently encountered in the process of density estimation in high-dimensional spaces. In this project we employ model order reduction with sparse grid techniques to deal with the curse of dimensionality. L2 regularised density estimation model has been developed using the Sparse Grid techniques which has proven to be effective in dealing with the curse of dimensionality [2]

$$\widetilde{P} = argmin_{f \in V} \int_{\Omega} (f(x) - p_{\varepsilon}(x))^2 dx + \lambda || \wedge f ||^2_{L^2}$$

In this project we estimate $\widetilde{P}$ defined as the Probability density function of an exponential family. In which, the required function is approximated using sparse grid discretization technique. In other terms MAP (Maximum A Posteriori) approach with sparse grid discretization. Minimization is performed using variational methods: The solution of the underlying

$$\sum_{j=1}^{N} \alpha_j^{i+1} a(\varphi_k, \varphi_j) = \frac{1}{n} \sum_{i=1}^{N} \varphi_k(x_i) - \int \varphi_k(x) \frac{exp(\sum_{j=1}^{N} \alpha_j^i \varphi_j(x))}{\int exp(\sum_{j=1}^{N} \alpha_j^i \varphi_j(z)) dz} dx$$

In the matrix form we have

$$A\alpha^{i+1} = q - \phi(\alpha^i)$$

The challenge lies in solving the following Non-Linear part

$$\phi(\alpha^i) = \int \varphi_k(x) \frac{exp(\sum_{j=1}^{N} \alpha_j^i \varphi_j(x))}{\int exp(\sum_{j=1}^{N} \alpha_j^i \varphi_j(z)) dz} dx$$

We employ Smoothing Spline analysis of variance (ANOVA) to further improve the efficiency of the model by discarding the unwanted components from the Non-linear term. Variable selection and model building using the likelihood basis pursuit in the context of smoothing spline ANOVA has been discussed in [1].

Graphical Models provide a principled approach to deal with uncertainty through the use of probability theory, and an effective approach to cope with the complexity through the use of graph theory. Density estimation and Knowledge discovery are the two main applications of graphical models and in this project we work on deriving a graphical model which will facilitate us to use Messaging passing algorithms (which are variational in nature) to solve the Non-linear term in the equation. Different approaches to solve the Non-linear term, like Monte Carlo Integration has been employed and evaluated in [3]

IDP Objectives

1.The primary objective is to develop a density estimation model which employs Sparse Grid learning methods to reduce the model order along with Smoothing Spline ANOVA – This includes the evaluation of the model in terms of accuracy and runtime efficiency, using several artificial and real world datasets and to compare the results with the output of existing methods.
2. Another objective is to derive a Graphical model to represent the conditional dependence/independence between the variables; based on the model developed in the first step. We use this graphical model to solve the above mentioned problem with the use of Message Passing algorithms.

Implementation

1. The code to implement the above mentioned idea will be written in Python and multiple open source libraries will be used.
2. A specific library for Sparse Grids - SG++ will be used extensively.

Application Area

Successful completion of this project would mean that we have an efficient method to estimate the probability densities for multi-dimensional data. It would also mean that we have achieved another milestone with the application of Sparse Grids in the field of Data Mining. Density estimation can be used in exploration and presentation of data and can be used to solve various types of problems. Classification and Sampling are two such applications.

Relevance of Lecture

Lecture Title: Non-Linear Optimization (Advanced)

Since we are dealing with a minimization problem, we would have to apply different optimization techniques which needs a well understood knowledge of Optimization. Hence the course 'Non-Linear Optimization' is necessary to proceed with the implementation of the concepts described.

References

[1] Yi LIN Meta VOELKER Michael FERRIS Ronald KLEIN Hao Helen ZHANG, Grace WAHBA and Barbara KLEIN. Variable selection and model building via likelihood basis pursuit. 2002.

[2] Benjamin Peherstorfer. Model order reduction of parameterized systems with sparse grid learning techniques. 2013.

[3] Sebastian Soyer. Nonlinear density estimation with applications in astronomy. 2014.