BIG DATA CHAPTER 9 SUMMARY

Author
Kevin Loun

Introduction

The exponential family of distributions is crucial in both statistics and machine learning, encompassing common distributions like Gaussian, Bernoulli, and multinoulli. It forms the foundation for constructing both generative and discriminative models, including generalized linear models (GLMs).

Exponential Family

Definition

A distribution belongs to the exponential family if it can be expressed in the form:

$$p(x|\theta) = h(x) \exp(\theta^T \phi(x) - A(\theta)) \tag{1}$$

where θ are the natural parameters, $\phi(x)$ the sufficient statistics, and $A(\theta)$ the log-partition function. This formulation allows for efficient computation of statistics and simplifies the creation of conjugate priors.

Importance

The exponential family is significant due to:

- The existence of finite-sized sufficient statistics, aiding in data compression and online learning.
- The availability of conjugate priors, simplifying Bayesian inference.
- Formulation of generalized linear models, where the response variable's distribution follows the exponential family.

Generalized Linear Models (GLMs)

GLMs extend linear models by allowing the mean of the distribution to be a nonlinear function of the linear predictors. This is achieved through a link function connecting the mean of the distribution to the linear combination of input variables.

Applications and Estimation

- GLMs are used for various regression models including logistic regression and Poisson regression.
- The parameters of GLMs can be estimated using maximum likelihood estimation (MLE), ensuring the model captures the essential data patterns without overfitting.

Bayesian Approach

In Bayesian statistics, GLMs can be combined with conjugate priors to form a posterior that is easy to update with new data, adhering to the Bayesian updating rules.

Conclusion

Understanding the exponential family and GLMs equips data scientists and statisticians to build more robust, efficient, and interpretable models that are foundational in machine learning and statistical analysis.