Extension of Distribution Estimation to Form Classification

Rumuly  
College Station, TX  
masondatminer@tamu.edu

Abstract

Introduction

* Generalize from parameter estimation to find form as well; same reason, good for modeling and analysis of systems
* For this case, look at continuous random variable, not discrete
* Want a way to generally determine which of a set of distributions
* Talk about ‘goodness of fit’
* Frequentist (classical) classification (or Bayesian with equal priors) to strictly compare metrics
* Justify restricted set of characteristic distributions; characterize, best obviously EDF

Objective Functions

* General ideas: EDF and CDF comparison vs. Entropy. Include EDF function, mention that it is non-smooth, hard to work with. Include statistic functions, implementation. All distance-based; lower statistic means higher similarity
* Kullback Leibler, not useful for comparing empirical and theoretic in continuous domain. Entropy-focused; would be important for analysis of discrete distributions
* Kolmogorov-Smirnov, known to converge to 0, basically maximum error between EDF and CDF
* Kuiper, refinement of Kolmogorov-Smirnov; sensitive to tails and median, invariant under cyclic transformations of independent variable
* Anderson-Darling, based on mean-square-error, basic statistic places most of weight on tails, logarithm on cdf makes hard to compute at tails. Critical value dictionary, can reject null hypothesis
* Cramer-von Mises Criterion, refinement of Anderson-Darling, does not include logarithm

Test Method

* Looking at single distribution; samples from multiple distributions mixed together not covered
* Choice of distributions (describe, with PDF functions) and default parameters
  + Normal, Uniform, Exponential: basic, everywhere, proof-of-concept
  + Laplace: Specifically for bamboozling, similar to Normal
* OOO: choose distribution, get sample, estimate each distribution, get statistic for each objective function. Do for sample sizes, do repeatedly and average over values.
* Implementation specifically scalable; easy to add new distributions, including custom ones.

Results

* Accuracy of each objective function under each distribution
* Behavior of each objective function under each distribution

Conclusion

* Success of implementation proves concept; note ease of use in applying to generic input
* Usefulness of each statistic
* Future work ideas
  + Bayesian extension
  + Confidence Proof

Appendix A: Objective Function Implementations

Appendix B: Distribution Implementations

Appendix C: Test Implementation