Comparing the number of real roots in "real-world" polynomials and randomly-generated polynomials

Tereso del Río, Matthew England

Motivation

- Growing interest in using Machine Learning in symbolic computation.
- Huge amounts of data are needed and "real-world" objects are limited.
- Some papers have been criticized for using random data because it is believed that random and "real-world" objects behave in a different way.
- We wanted to study how to generate synthetic data that behaves similarly to "real-world" data.

What we did

- Create a tool to study how similar random and synthetic problems were to "real-world" problems.

What we did

- Create a tool to study how similar random and synthetic problems were to "real-world" problems.
  - Extract univariate polynomials from the problems using CAD projection.
  - Compute the number of real roots of these polynomials counting multiplicity.
  - Use the bootstrap method to determine if the number of real roots in two different families follow a similar distribution.

What we did

- Create a tool to study how similar random and synthetic problems were to "real-world" problems.
  - Extract univariate polynomials from the problems using CAD projection.
  - Compute the number of real roots of these polynomials counting multiplicity.
  - Use the bootstrap method to determine if the number of real roots in two different families follow a similar distribution.
- We wanted to compare families from different origins but in the process we observed that families of real-world problems are very different to each other.

Obtaining families of problems

- **"Real-world" problems:** QF_NRA category of the SMT-LIB (ex: *Geogebra* and *meti-tarski*).
- **Synthetic problems:** Using `randpoly()` conserving some features of the "real-world" problems (ex: *random_Geogebra*).
- **Random problems:** Using `randpoly()` (ex: *random* and *meti-tarski*).

Comparing two samples of the number of real roots

Basically there are two list of numbers and we want to determine if they were generated using the same distribution.

Hypothesis: *Sample 1* and *Sample 2* were generated by the same distribution *Distribution 1*.

Comparing two samples of the number of real roots

Basically there are two list of numbers and we want to determine if they were generated using the same distribution.

Hypothesis: *Sample 1* and *Sample 2* were generated by the same distribution *Distribution 1*.

- *Auxiliary Sample* is extracted from *Distribution 1*.

Comparing two samples of the number of real roots

Basically there are two list of numbers and we want to determine if they were generated using the same distribution.

Hypothesis: *Sample 1* and *Sample 2* were generated by the same distribution *Distribution 1*.

- *Auxiliary Sample* is extracted from *Distribution 1*.
- The probability of both *Sample 2* and *Auxiliary Sample* are compared.
    - If *Sample 2* is more probable than *Auxiliary Sample* that indicates is likely that the hypothesis is true.
    - Else that indicates the hypothesis might not be true.

Comparing two samples of the number of real roots

Basically there are two list of numbers and we want to determine if they were generated using the same distribution.

Hypothesis: *Sample 1* and *Sample 2* were generated by the same distribution *Distribution 1*.

- *Auxiliary Sample* is extracted from *Distribution 1*.
- The probability of both *Sample 2* and *Auxiliary Sample* are compared.
  - If *Sample 2* is more probable than *Auxiliary Sample* that indicates is likely that the hypothesis is true.
  - Else that indicates the hypothesis might not be true.
- This is repeated many times replicating the idea of the bootstrap method *Freund et al. 1995* to get a closer idea of how likely is that the hypothesis is true.

Results

The numbers in this table represent the certainty to discard that the named family is the same as "Geogebra".

Degree	Geogebra	random-Geogebra	random	meti-tarski
2	0.45272	0.80008	0.91754	1.00000
3	0.45784	0.78931	0.67135	0.99998
4	0.45102	0.93661	0.85620	1.00000
5	0.45556	0.81697	0.99783	1.00000
6	0.42942	0.82243	0.99983	1.00000
7	0.43385	0.93313	0.97233	0.99888
8	0.42578	0.99090	0.99971	1.00000

The bigger the number the more evidence that the samples come from different distributions. It is standard to discard the hypothesis if higher than 0.95.

Main conclusions

- Conserving features from the original family results in similarities on the distribution of the number of real roots.
- There is no such a thing as the properties of the "real-world" polynomials. The QF_NRA collection is quite heterogeneous.
- This, together with the imbalance of this collection implies that one should be **very careful** when training a Machine Learning model on it.

If there is a paper... and future directions

- A more exhaustive analysis of how heterogeneous the QF_NRA collection is.
- How much the distribution of the number of real roots changes when different features are conserved.
- Possible solutions to the imbalance and heterogeneity of the QF_NRA and other collections.
- Comparing the performance on "real-world" data of models trained with "real-world", synthetic and random data.