COMS W6998

Scorpion

Presenter: Richard Zhang

Scribe: Zachary Huang

How would you design an experiment?

Possible.

Similar to wrangler. Let users get dataset have domain meaning.

Select points outliers.

Whether not predicate drive explanation.

Scale ranking.

How well predications fit into what people expect?

User study? Possible for users to assess?

Hard to understand the results if not familiar with the study itself.

Given predicates, can make sense to some dataset.

Define data points are outlier/hold-out is difficult for users.

Hard for users to understand selecting tuples as outliers/hold-out influence the final predicate results.

After seeing the demo:

What happen if remove the data.

Select: not sure what it means?

Should everything changes? Should subset changes? Should others not change?

How to tune? What is the best one?

Trade-off is not something system can do for you.

Interactive part feels like more useful.

In the paper can't see which tuples are removed.

Better build a demo than to explain.

From users side, if system generates some results, how can users assess whether it's the best answer?

What's metrices of metrics?

Causal inference?

RSexplain use causal inference view of the world.

Saw system anomalies. Assume you have infinite time, what would be a satisfying answers?

All the explanation systems help your manual efforts.

No true answers. Hard to know.

Scorpion injects errors.

Real datasets manually look into it and hope for the best.

One option is to try things and hope it works.

Given a result can assess whether sensible, but don't know whether it's better.

Are explanations related to time series?

Time series are to find outliers. Maybe related.

It's common place where things change all the time.

Can do the same analysis of population between new york and California.

All tools group by state. One result per state. Ask why they are different. Explanations are predicates not on the attribute. Meaningless to say new york is different because new york is different. Need to se subgroup.

Working women and children are related to first immigration. Take a lot of time to do manually. One things make something in common.

Difference between explanations of databases and machine learning?

Last week focuses machine learning interpretability.

DIFF/scorpion are about data, schema, columns opposed to breakdown of machine learning which focuses on process.

Part of this week's systems can be used to last week's system if you need to know attribution and find predicates. If use pixels and cells, able to do aggregation and find some formal predicates.

Form: data -> f() -> output.

ML: records = image, focus on f(). If f = g&h, want to know g,h's input and output. Given change of predication, what changes the input? Some people care about subset of output query. More aggregate functions to results. Similar to database framework.

DB: f well studied. f doesn't do anything changes the data. Do join, filter aggregate. Have outputs records. To understand records is easy. If I perturb set of output records. Set has relations have relationships of input records.

Model: precision, recall. Know whether the records are classified right.

Can recast it to DIFF/Scorpion.

Train give you model parameters. A lot of similar process.