Data Leakage Discussion
ATOM

Nov 28th, 2017

Matt Drury

Leaky faucets move a substance from the correct place to somewhere you do not want it.

Opening Example

What’s wrong here?

```
standardizer = StandardScaler()
standardizer.fit(X_raw, y_raw)
X_train_std = standardizer.transform(X_train)
X_test_std = standardizer.transform(X_test)
```

In what situations could this go wrong? Is it ever safe?

Leakage Definitions

“The unintentional introduction of predictive information from the target by the data collection, aggregation, and preparation process.”

“Introduction of information about the target of a data mining problem, which should not be legitimately available to mine from.”

Any others?

Leakage From Features

```
Logistic Model: Probability of Churn
------------------------------------


Variable Name              Parameter Estimate
...
has_full_pay_discount      -0.2
has_good_driver_discount   -0.3
has_good_payer_discount    +1.5
has_good_credit_score      -0.2
percent_change_premium     +0.2
...
```

Leakage From Features

```
Logistic Model: Probability of Churn
------------------------------------

Variable Name              Parameter Estimate
...
has_full_pay_discount      -0.2
has_good_driver_discount   -0.3
has_good_payer_discount    +1.5 <- Under reasonable assumptions, this TRIPLES the
                                   probability of churn!
has_good_credit_score      -0.2
percent_change_premium     +0.2
...
```

Ok… So What Happened...

Leakage From ???

```
Logistic Model: Probability of Purchase
---------------------------------------

Variable Name              Parameter Estimate
...
is_male                   -0.05
has_good_credit           +0.2
accident_free             +0.1
has_dui                   +1.0  # <- Oh geez, that's bad…
                                   Should I believe it?
owns_car                  +0.1
...
```

Leakage From ???

```
Logistic Model: Probability of Purchase
---------------------------------------


Variable Name              Parameter Estimate
...
is_male                 -0.05
has_good_credit          0.2
accident_free            0.1  # <- Oh geez, that's bad…  Should I believe it?
has_dui                  1.0
owns_car                 0.1
...
```

What could have happened here?

Other Examples?

Structural Example: Model Stacking

```
model_1 = model_1.fit(X_train, y_train)
model_2 = model_2.fit(X_train, y_train)

stacked_train = DataFrame({
    'm1': model_1.predict(X_train),
    'm2': model_2.predict(X_train)
})

stacked_model = stacked_model.train(stacked_train, y_train)
#Profit!
```

Structural Example: Model Stacking

```
model_1 = model_1.fit(X_train, y_train)
model_2 = model_2.fit(X_train, y_train)

stacked_train = DataFrame({
    'm1': model_1.predict(X_train),
    'm2': model_2.predict(X_train)
})

stacked_model = stacked_model.train(stacked_train, y_train)
#Profit!
```

How can we correct this?

Other Examples?

Leakage From the Future

Again, predicting churn.

- We expect that agents (sales representatives) have some influence on the churn of their customers.
- Too many agents to estimate a parameter for each one…

Solution: Compute the average churn rate for each agent in the training data, use this as a predictor. Now we are only estimating one parameter (or a small number, to account for non-linearity).

Yes?

Leakage From the Future

No.

We have implicitly leaked information about the agent’s performance in the future to make predictions about their past performance!

Suggested Fixes?

Moar Examples?

Non-Explicit Leakage

Exploratory Data Analysis: GOOD.

Using your entire data set: BAD.

“Every decision you make is an opportunity to be wrong.
Your only protection is a test set.”

Combating Leakage

- **Understand Your Data!**
    - **Talk to People! Develop Intuition and priors!**

- Exploratory Data Analysis
    - Predictor vs. Response Plots

- Model Diagnostics
    - Parameter Estimates
    - Variable Importance Metrics
    - Partial Dependency Plots

Combating Leakage

“This is completely anecdotal, but I've found variable importance useful in identifying mistakes or weaknesses in GBMs.

Variable importance gives you a kind of huge cross-sectional overview of the model that would be hard to get otherwise. Variables higher in the list are seeing more activity (whether or not they are more 'important' is another question). Often a poorly behaving predictor (for instance something forward-looking, or a high-cardinality factor) will shoot to the top.

If there's a big disagreement between intuition variable importance and GBM variable importance, there's usually some valuable knowledge to be gained or a mistake to be found.”

Declan Groves (https://stats.stackexchange.com/a/202827/74500)

Leakage In Competitions...

Moral

or

Immoral?

Free Discussion!

Image Sources

https://static.comicvine.com/uploads/scale_small/0/40/711130-112808_86953_homer_simpson_super.jpg

http://www.popularmechanics.com/home/interior-projects/how-to/a3095/5-steps-to-fix-a-leaky-faucet-15470175/

https://www.faucetmag.com/fix-leaking-kitchen-faucet/

http://indianaontap.com/news/wasted-how-the-craft-beer-movement-abandoned-jim-koch-and-his-beloved-sam-adams/