Measuring Average Treatment Effect

In Reddit’s “Change My View” forum

Reut Apel 201316346

Shimon Sheiba 201110376

**INTRO**

**C**hange **M**y **V**iew

Reddit community formed an arena for inviting others to explain why they disagree with you in the pursuit for greater understanding on the matter. By submitting a well explained opinion that is opened to change, filled according to a strict set of rules (which apply on the responders as well), one is actually starting a civil discourse that has a great potential of being content rich and very much diverse. This unique DNA is becoming especially vital in an era where all other social platforms and intelligent agents are oriented to extend one’s stay by exposing only similar opinions and content.

“In order to resolve our differences, we must first understand them. We believe that productive conversation requires respect and openness, and that certitude is the enemy of understanding.”[CMV WIKI](https://www.reddit.com/r/changemyview/wiki/index)

**Data structure**

An opinion is submitted as a new topic in the subreddit, and by that a new hierarchical tree of comments is started for that topic, while everyone is free to respond to anyone, meaning directly to the topic submission or to any comment inside the tree.

CMV evolved throughout the years of its existence, adding new mechanisms that contribute to the discussions and provide important social insights on civil discourse, persuasiveness and opinions. An evolution that created a breeding ground for many researches from different domains including social science, artificial intelligence etc.

The most familiar mechanism is “Delta”, a sign taken from mathematics as a symbol for change, the delta is an incentive that anyone can grant anyone else, again if aligned with the rules, for acknowledging the influence that latter had on the former’s opinion.

**PROJECT OUTLINE**

In the framework of our project, the delta is taken as the outcome, meaning the information we wish to predict and measure the effect of other characteristics in the discourse on it, in a causal manner.

**Definitions**

Data:

- Submissions and comments from Change my view subreddit in a feature vector representation.

- Initial raw data ~35K submissions and their ~2.5M comments while ~15K of them got delta by the submitter.

* Filters performed on raw data, including:
* Length of Comment text > 200 characters.
* Comments creation date is at most one week after the submission creation date.
* Only discourses which at least one delta was given in them by the submitter were taken.
* After filters: ~ 300K comments, ~12.5K of them got delta – increasing the minority class from 0.6% to 4.3%.
* On this data the features of the predicting model, including the treatment will be built.

Outcome **Y**: A binary label for each comment in the data, whether the submitter gave a delta to the comment or not.

Treatment **T**: A binary value indicating quotation, whether the comment quoted the submitter: either the submission itself or the comments’ parent if the latter was written by the submitter.

Causal question: What is the causal effect, measured by ATE, of quoting the submitter in a comment on the decision of the submitter rather to give that comment a delta or not?

Units: Each unit will be composed of a comment and its submission, excluding comments that were written by the submitter.

* On this data the ATE will be measured.

Features **X**:

Upon building the feature set, a number of principals were tried to be reflected:

* Level of involvement by the submitter in the discourse.
* Environmental and quantitative characteristics of the discourse.
* Behavior of the submitter/commenter.
* Different textual signals.

All features were built on data prior to comment creation time, in order not to include post-treatment data.

1. Comment’s depth.
2. Length of the submission.
3. Length of the submissions’ title.
4. Length of the comment.
5. Sentiment of the submission.
6. Sentiment of the submission’s title.
7. Sentiment of the comment.
8. Cosine similarity between the Sentiment of the submission and the Sentiment of the comment.
9. Number of submissions that the submitter wrote.
10. Number of submissions that the comment author wrote.
11. Number of comments that the submitter wrote.
12. Number of comments that the comment author wrote.
13. Cosine similarity between the submitter’s texts in l2 normalized tfidf representation and the commenter’s texts in l2 normalized tfidf representation in all of CMV, while idf was learned on all data.
14. % of the Adjective in the comment.
15. Topic modeling feature set, vector of topics for each comment/ submission.
16. If this is the first comment of this user in this submission.
17. Number of comments of this user in this submission until the unit UTC.
18. Time between the submission and the comment.
19. Time between the submission and the first comment.
20. Time between the first comment and the comment**.**
21. Ratio of features 18 & 19.
22. Number of times the submitter responded to the comment user in this submission.
23. Number of comments of the submitter in this submission’s discourse.
24. Number of times the submitter responded in total.
25. Number of times the submitter responded to the comment user in this submission/total number of comments to the submission.
26. Number of times the submitter responded to the comment user in this submission/ Number of times the submitter responded in total.
27. Number of times the submitter responded in total/ number of comments to the submission.
28. The seniority of the commenter in change my view (how many days passed since his first comment in the data).
29. The seniority of the submitter in change my view (how many days passed since his first comment in the data).

**Work phases**

* Data import & preparation – using Reddit’s API and building a scraper and a parser.

The relevant data is drawn from CMV subreddit and from logs and DBs of CMV delta mechanism. Then, a data understanding phase is performed in order to build the optimal parser for preparing the data.

The outcome is built based on a union of two sources:

1. The deltaBot log data.
2. If the comment’s child is written by the submitter, with length greater than 50 characters (CMV rule) and contains a symbol from the set: {'&amp;#8710;', '&#8710;', '&#916;', '&amp;916;', '∆', '!delta', 'Δ', '&delta;'}.

**Modules**: reddit\_api.py, comments\_statistics.py

* Build a feature representation for all units.

Each unit is represented by a feature vector of length 50, while sentiment features are 3 columns each <positive, negative, objective)>, topics feature composed of 15 topics, similarity feature is based on cosine similarity between l2 normalized tfidf vector of all the text written by the commenter and the submitter until the (while the idf is learned on all the data), and finally the treatment is recognized in a comment if the quoting symbol appears: “>” together with the entire quote is taken from the submission or from the parent comment as long as the parent is the submitter.

**Modules:** createFeatures\_delta.py, sim\_feature.py

* Analyze statistics.

For each feature calculates seven statistic measures for treatment group and control group separately as an indication before and after the matching to understand the matching’s effect.

The ones constitute the strongest indications for overlap improvement for each of the features are:

1. Standardized difference:
   1. and are the mean of the treated and control units accordingly in a specific feature.
   2. and are the variance of the treated and control units accordingly in a specific feature.
2. % reduction in imbalance:

**Modules:** Covariate\_imbalance.py

* Propensity method.

Estimates the propensity score based on three methods- logistic regression, liner regression and probit. Creates a histogram for each group. Chosen propensity is based on logistic regression.

**Modules:** propensity\_score\_hist.py

* Common support and KNN matching method based on Propensity score.

In order for overlap between treatment and control to exist, combined with the second assumption of common support - , min/max thresholds are calculated according to the units’ propensity score and all units crossing the thresholds are removed.

Then, KNN is performed with the desired settings i.e. number of chosen neighbors, sampling with replacement etc., between treatment and control in order to match units.

**Modules**: matching.py

* In order to be able to measure the ATE of any feature on the outcome, there is a need to prepare the settings for it, therefore a machine learning model is trained, for choosing the best model, we trained 16 different ones, including random forest, multinomial Naïve Bayes, perceptron etc. with Kfold = 14 and chose the model that yielded the best AUC, which was logistic regression.

**Modules**: Classifier\_delta\_by\_features\_treatment.py

* Finally, to Measure the ATE using the covariate adjustment formula, the chosen model predicts on two artificial datasets, one is the control set when given to the model as treatment, while the other is the opposite meaning the treatment set given as control. Then, the gold label of the unit is compared to the prediction of that unit and the normalized sum represent the ATE estimator.

**Modules**: ATE.py

**Results and conclusions**

* Analyze statistics results:

The following table exhibits the statistics results before and after the matching:

As shown in the table in the % reduction imbalance column show that the standardized difference