Visualizations for Justifying Machine Learning Predictions

Proposal

David Johnson - davewj@cs.ubc.ca

Domain

Since the user could be approaching this system from many different domains, to discuss the domain of the system requires some generalization. For instance, the user of this system could be using this system to justify machine learning predictions for an NLP task such as sentiment analysis. Alternatively, the system might be used to justify machine learning predictions for a user grappling with a bioinformatics problem. The most accurate (but generalized) domain description of the system is to say that it's within the domain of machine learning.

Dataset

The dataset for this project actually involves two different data sets. First, this project will use a raw dataset. The goal behind the system is that this raw dataset can be arbitrarily chosen by the user (obviously, since the goal of the system is to allow a user to run logistic regression on their own dataset and then explain the predictions to the user via visualizations).

After choosing a raw dataset, It's expected that the user should have fit and serialized a logistic regression model. The user will then come to the system with their serialized fit logistic regression model and use the system to generate visualizations for the fit model. Since this is the intended pipeline, the dataset for this problem is transformed data in a table format where the key value is feature, an item is a particular feature and an attribute value is the effect the feature has on the model.

Task

It’s expected that users will consume the data with intentions of discovery. Users will have transformed raw data into a fit logistic regression model containing some arbitrary number of features. They will be using my system to generate visualizations which justify how their logistic regression model predicts output given the effect and importance of features on the fit model. Since it’s assumed the users are non-domain experts, they will certainly be using the visualizations as a means of discovery -- discovery about the real-world implications of the model itself.

Proposed Solution

I am choosing to abstract the data here in such a way that features are items in a table and the effect on the fit model are attribute values. Given that fit models may contain arbitrary numbers of features, I’m allowing for the possibility that the data may need to be reduced -- particularly by item filtering. Once the item is filtered (if necessary) the data will need to be arranged. Since there will be one key: feature, I believe that a bar chart visualization will be a strong choice here. I also believe there may be value in showing a separate view containing the statistical data in a table form; the statistical table could contain coefficient values, error values, p values etc.. This statistical table should have a button on the GUI allowing it to be hidden/viewable so as to not overwhelm users.

Although this iteration of the system is focusing primarily on selecting features for visualization and generating visualizations the end goal of the system as part of the thesis work (see Personal Expertise section) would be to include natural language generation. The natural language generation would generate a textual justification for the importance of features. Additionally, the system could train the models for the users, rather than requiring users to come to the system having already fit and serialized a model.

Personal Expertise

This project is intended to be a part of my thesis work. The start of this project coincides with the start of my thesis work, so although I don’t have previous expertise from work on the thesis research to draw from, I do have motivation to use this system outside the scope of this particular project. The overall thesis research will involve a larger and more substantial system involving a combination of natural language generation and visualization. I do have an education background in the relevant domain: I have taken Machine Learning, Natural Language Processing, Artificial Intelligence, and Human Computer Interaction -- all of which could be relevant in their own ways to this project. Lastly -- and probably most importantly -- I find the subject very fascinating, and I'm looking forward to getting a chance to spend some time working on the actual project.

Scenario of use

- User opens the system, navigates to the middle partition of the GUI and presses the Load Fit Model button (see Fig 1 in Appendix). The system opens a popup window with an address box allowing the user to navigate to the location of the saved serialized model
- User loads the fit model and the address box closes
- The middle partition of the GUI generates a visualization showing the effect of features and their interpretability in the model (see Fig 2 in Appendix)
- User presses the Unhide button to show the table of statistical data

Implementation Approach

Implementation will be done using Python. Additionally, pandas will be used for data manipulation, matplotlib for generating the visualizations and seaborn for improving the aesthetics of the matplotlib generated visualizations. The system will use scikit-learn for the interacting with the fit logistic regression model. Tkinter will be used for development of the GUI itself within Python.

Milestones

- March 20: Interface v1.0 developed (Interface can load fit serialized models, unpack them, and access model). Narrowed down possible vis idioms to 2-3 choices
- March 27: Experimentation on vis idioms complete, a decision on vis idiom made
- April 10: Interface feature selection implementation complete
- April 17: Interface generates visualizations. Interface 2.0 is complete, has all the functionality to allow for the scenario of use given in proposal to actually be completed. Interface should be presentable
- April 24: Feature freeze. All testing/bug fixing complete. Project is in presentable form

Previous Work

Biran et al (2015) have written about difficulties involved in generating justifications -- difficulties particularly relevant to my proposed project. They suggest generating justifications is a task composed of two problems: selecting relevant features, and discussing both the state of the model and the real world implications of the features. These issues will be the main two issues I expect to encounter during development. Selecting relevant features is an implementation issue, one that will require experimentation during development to determine the best approach for selecting features. The efficacy of showing the user the state of the model and implications of features will depend on the visualizations I decide to use.

One notable decision for my proposed project is that it focuses exclusively on explaining logistic regression. Although this is primarily due to the heavy time constraint, there is also reason to believe logistic regression is a good choice due to its interpretability. Lipton (2016) discussed model interpretability, suggesting that interpretable models are transparent and have strong post-hoc interpretability. Logistic regression fits these requirements well, suggesting it is a good choice for this type of project.

In contrast to the idea of focusing on one interpretable model, Ribeiro et al (2016) argue that explanations of predictions should be model agnostic, i.e. that explanations should be separate from the models themselves. Rather than focusing on beginning with interpretable models, they treat the original models as black boxes and instead learn interpretable models on the predictions of the black boxes. Using this approach, they developed LIME (Local Interpretable Model-agnostic Explanations) (2016, August) which generates both text and visualization explanations of their learned models. Although this is quite an interesting approach, it’s substantially more challenging and unrealistic for my project given the time constraints.

Among previous works in developing actual systems, the most relevant to my proposed project is PreJu (Prediction Justifier) by Biran et al. PreJu (2014) determines the effect of each feature in a logistic regression classifier and the importance (expected effect) of a feature to determine key features. Key features are then used to generate a justification narrative with both text and visualization in an attempt to justify to the user the importance of features to the model.

Bibliography

[1] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016, August). Why Should I Trust You?: Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1135-1144). ACM.

[2] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). Model-agnostic interpretability of machine learning. In *Human Interpretability in Machine Learning workshop*.

[3] Biran, O., & McKeown, K. (2014). Justification narratives for individual classifications. In *Proceedings of the AutoML workshop at ICML* (Vol. 2014).

[4] Biran, O., & McKeown, K. (2015). Generating justifications of machine learning predictions. http://www.cs.columbia.edu/~orb/papers/d2t_2015.pdf.

[5] Lipton, Z. C. (2016). The mythos of model interpretability. In *ICML Workshop on Human Interpretability in Machine Learning*.

Appendix

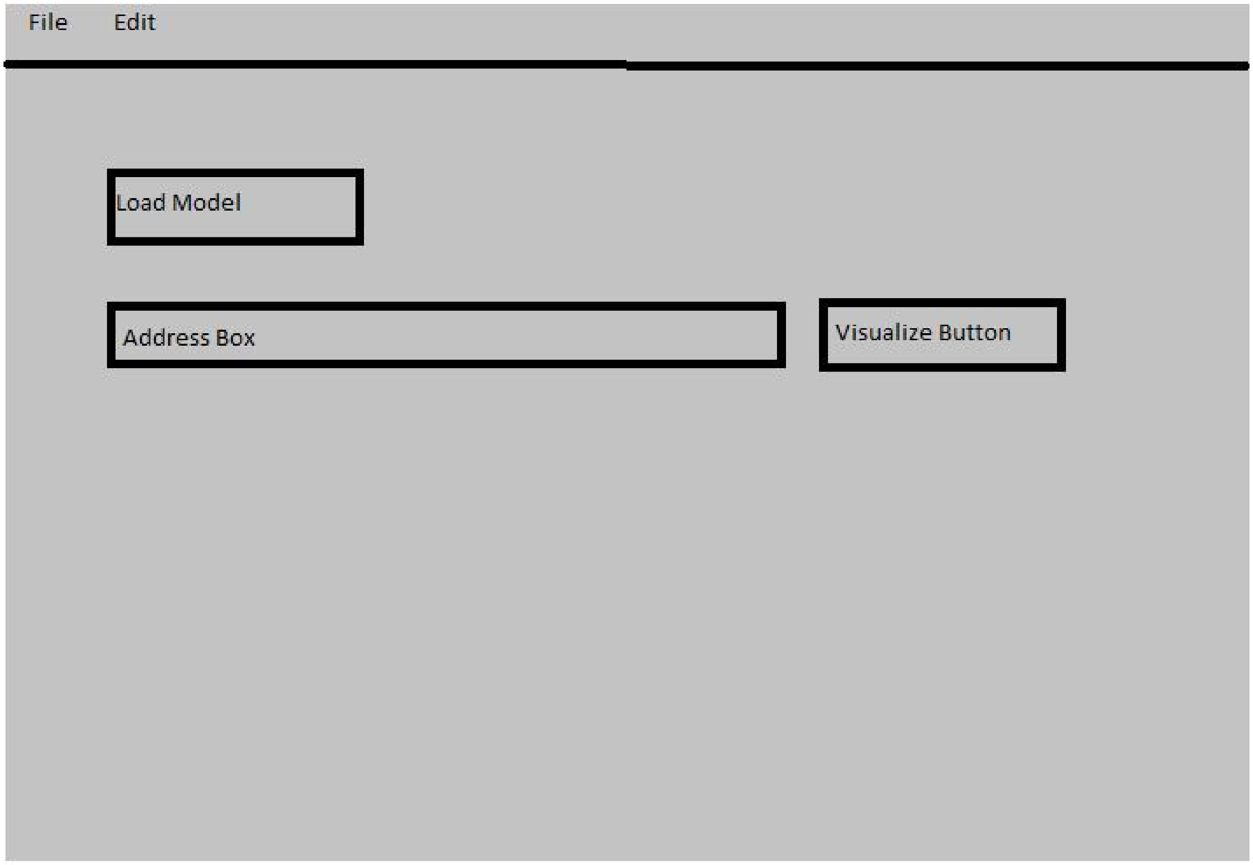

Fig 1. Shows the basic interface. This is the first screen that users will be presented with in which they can load their fit model, an address box will show the disk location the model is being loaded from, and the visualize button will run the system and generate visualizations

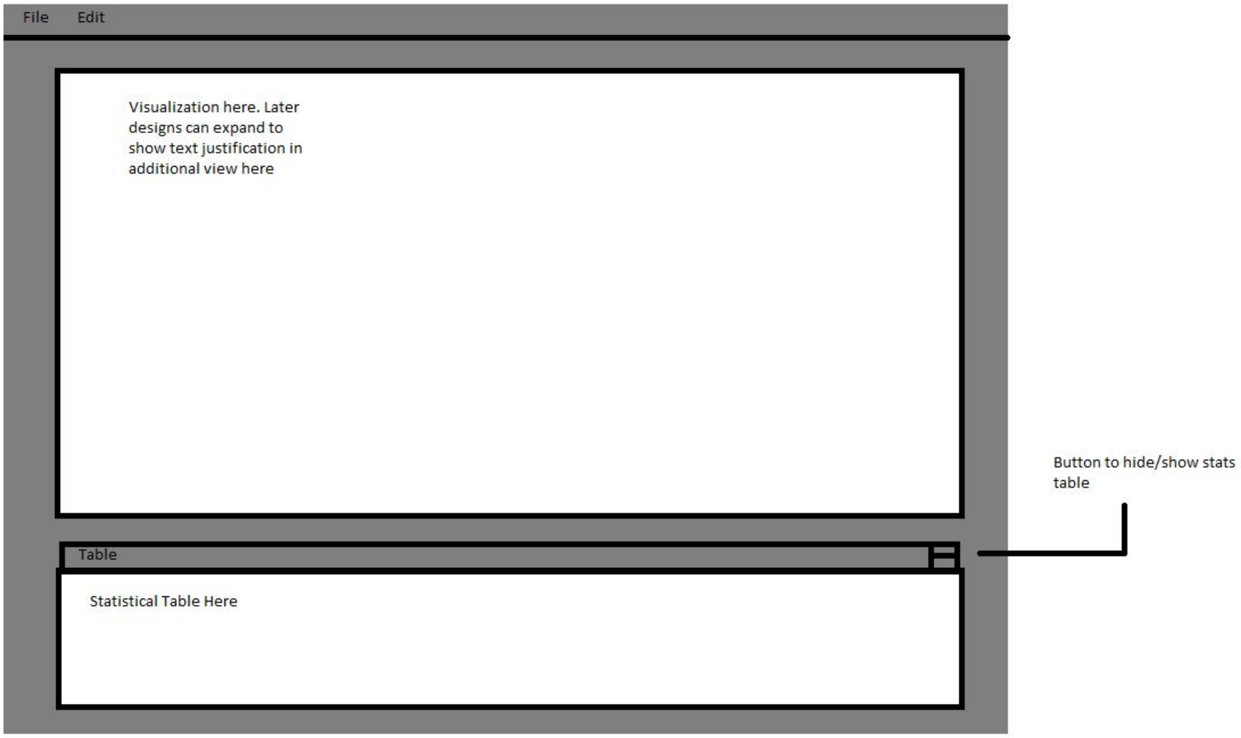
File
Edit
Visualization here. Later designs can expand to show text justification in additional view here
Button to hide/show stats table
Table
Statistical Table Here

Fig 2. Shows what the interface will look like after generating some visualizations for the user. The top box will contain the actual visualizations, the button will contain the additional statistical table information. There is a button to hide or show the button table.