Interpretability

March 8, 2017; due March 19, 2017 (11:59pm)

1 (everyone)

The models that we have considered in class differ strongly in the degree to which their results are "interpretable." For each of the models below, characterize their level of transparency in comparison to alternate models. (For the three classification algorithms, assume each data point is featurized by 10,000 features.) What strategies does each model give us to control (and possibly improve upon) its interpretability?

- Decision trees (binary classification)
- Logistic regression (binary classification)
- Naive Bayes (binary classification)
- Topic models

Deliverable: one-page paper (single-spaced).

2 (choose either 2.1 or 2.2)

2.1 Implementation

One strategy for interpreting an *instance-level* classification proposed by Martens and Provost (2014) and Chen et al. (2015) is as follows. Assume that we have a model $\mathcal{M}$ that has been trained on some training data. Let's assume a new data point x is represented by F features $\{x_1, \ldots, x_F\}$, and we make a prediction ($\hat{y}$) for this data point using $\mathcal{M}$ (in this case, perhaps $\hat{y} = 1$). One way of assessing which qualities of x are most responsible for that prediction is to identify a *minimal set* of features of x that, if removed from x (e.g., by setting them equal to 0), would lead to the opposite prediction ($\hat{y} = 0$). This minimal set contains the *smallest* number of features that would create this change, and is known as the "evidence counterfactual."

In our case, let $\mathcal{M}$ be binary logistic regression ($\mathcal{Y} = \{0, 1\}$), where given a vector of learned parameters $\beta \in \mathbb{R}^F$ and input data point x, the probability of the positive class is given as follows:

$$P(y = 1 \mid x, \beta) = \frac{\exp\left(\sum_{x=1}^{F} x_i \beta_i\right)}{1 + \exp\left(\sum_{x=1}^{F} x_i \beta_i\right)}$$

If $P(y = 1 \mid x, \beta) \geq 0.5$, we predict $\hat{y} = 1$ and 0 otherwise. The folder (`http://courses.ischool.berkeley.edu/i290-dds/s17/hw/dds_s17_hw3/`) includes a trained model for the task of predicting the sentiment of movies (the trained model in this case is a set of feature coefficients). It also includes a set of 10 new data points (not used to train the model). Given this information, identify the following for each of those 10 points:

- The prediction $\hat{y}$ and $P(y = 1 \mid x, \beta)$ under the logistic regression model
- The evidence counterfactual that, if removed from x, would lead to the opposite prediction $\neg\, \hat{y}$

Deliverable: Code to calculate the evidence counterfactual from the data in the folder above, along with a README file providing the prediction and evidence counterfactual for each of the 10 input data points.

2.2 Critique

The term "interpretability" is often used to denote the degree to which the results of a *model* are understandable by a user, but many other aspects of data analysis can be subject to evaluations of interpretability as well, such as the selection criteria for data collection, the choice of representation for the input data, and so on — all of these involve separate questions of transparency from understanding the model on its own, but significantly contribute to what we can learn from data analysis. In this section, consider the following questions:

- What other elements of data analysis (beyond the model results) need to be interrogated for interpretability?
- What criteria would you propose to assess the degree to which those components are interpretable?
- How does the interpretability of those components contribute to the overall interpretability of the analysis?

Deliverable: one-page paper (single-spaced)

References

Daizhuo Chen, Samuel P. Fraiberger, Robert Moakler, and Foster Provost. Enhancing transparency and control when drawing data-driven inferences about individuals. In *SSRN*, 2015.

David Martens and Foster Provost. Explaining data-driven document classifications. *MIS Q.*, 38(1):73–100, March 2014. ISSN 0276-7783. URL `http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2282998`.