The Subjectivity of Data

January 31, 2016; due February 10, 2016 (11:59pm)

For this homework, complete section 1 and either 2.1 or 2.2.

1 (everyone)

In this homework, we'll attend to the source of data in many supervised classification tasks: human judgments. We'll consider the task of *sentiment analysis*, a term used to cover the assessment of both the attitude of a writer toward a specific target, such as a movie [Pang et al., 2002] and also the general tone of text [Dodds et al., 2011].

1.1 Annotation

Imagine that you have been given the following instructions:

> We'd like you to create an "emotional trajectory" of Hamlet, rating each scene for the extent to which it describes events that you might characterize as being emotionally positive ("happy") or negative ("sad"). How you decide what counts as positive and negative is entirely up to you.
>
> For a description of what happens in each of the 20 scenes, consult a scene-by-scene summary here: `http://utminers.utep.edu/ajkline/hamlet.htm`
>
> Then rate each scene below on a scale from -5 to 5 – a score of 5 corresponds to extremely positive, -5 to extremely negative, and 0 to neutral. For each scene, please also describe why you've assigned the score you've provided – what's your rationale?

a.) First, annotate the 20 scenes of Hamlet according your judgment of the sentiment-as-tone expressed within each scene, using the instructions above. Submit your annotations as a tab-separated file in the following format:

Scene	Judgment	Rationale
1.1		
1.2		
1.3		
…	…	…
5.2		

Table 1: TSV submission

b.) We'll measure the inter-annotator agreement rate for this task over all submissions; what kind of agreement might we expect (high, good, moderate, fair, poor), and why? How would you either make the instructions more precise or change the task to encourage higher agreement? [Deliverable: 100-200 words]

2 (choose either 2.1 or 2.2)

2.1 Implementation

1. Cohen's κ is one method to measure the agreement between two annotators, defined as the difference between the observed agreement P_o and the expected agreement P_e, normalized by the total agreement possible under that correction for chance.

$$\frac{P_o - P_e}{1 - P_e}$$

   Given a confusion matrix, the observed agreement rate P_o is simply the total fraction of observed agreements; the expected agreement rate P_e between annotators A and B over a total of K categories can be calculated as:

$$P_e = \sum_{k=1}^{K} P(A = k)P(B = k)$$

   Where again we can calculate $P(A = k)$ (the probability that annotator A assigns label k) as the observed fraction of A's assignments with with label k. Implement Cohen's κ for the data here: `http://courses.ischool.berkeley.edu/i290-dds/s17/hw/hw1.part1.1.txt`

2. One drawback of Cohen's κ is that it is only valid when measuring the agreement rate between exactly *two* annotators, and both annotators provide judgments for exactly the same items. An alterative when the number of annotators is greater than two is Fleiss' κ, again defined as the difference between the observed agreement P_o and the expected agreement P_e–but here we measure both quantities as agreement rates among *pairs* of annotators. Using the description of Fleiss' κ described in Fleiss [1971] (pp. 378-379), implement for the data here: `http://courses.ischool.berkeley.edu/i290-dds/s17/hw/hw1.part1.2.txt`.

   For both implementations, code your calculation of κ yourself from scratch (don't just use scikit-learn's implementation, for example). [Deliverables: code and κ values]

2.2 Critique

The following two papers attempt to tackle extremely difficult problems by creating data in the form of human judgments that can be used to train predictive classifiers:

- Identifying dogmatism in tweets [Fast and Horvitz, 2016]
- Irony detection [Wallace et al., 2014]

Discuss both papers' use of human judgments as data, and how that relates to their overall methodology. What strategies do both papers use to encourage the creation of high-quality data? In what real-world circumstances would we expect a model trained on this data to perform well, and where might it fail?

Deliverable: one-page paper (single-spaced)

References

Peter Sheridan Dodds, Kameron Decker Harris, Isabel M. Kloumann, Catherine A. Bliss, and Christopher M. Danforth. Temporal patterns of happiness and information in a global social network: Hedonometrics and twitter. *PLOS ONE*, 6(12):1–1, 12 2011. doi: 10.1371/journal.pone.0026752. URL `http://dx.doi.org/10.1371%2Fjournal.pone.0026752`.

Ethan Fast and Eric Horvitz. Identifying dogmatism in social media: Signals and models. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 690–699, Austin, Texas, November 2016. Association for Computational Linguistics. URL `https://aclweb.org/anthology/D16-1066`.

Joseph Fleiss. Measuring nominal scale agreement among many raters. In *Psychological Bulletin*, 1971.

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. Thumbs up?: Sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing - Volume 10*, EMNLP '02, pages 79–86, Stroudsburg, PA, USA, 2002. Association for Computational Linguistics. doi: 10.3115/1118693.1118704. URL `https://doi.org/10.3115/1118693.1118704`.

Byron C. Wallace, Do Kook Choe, Laura Kertz, and Eugene Charniak. Humans require context to infer ironic intent (so computers probably do, too). In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 512–516, Baltimore, Maryland, June 2014. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/P/P14/P14-2084`.