1 Logistic Regression Regression

1.1 Introduction

In this section we present a remedial solution to combine our regression models with the pointwise-estimation baseline model, and a more principled solution. Since we are combining the models, our gold standards will now be the original annotated sets composed of with size between two to ten.

1.2 Remedial Solution

First we give brief reminder for our baseline model. We defined two possible events: $\boldsymbol{\Omega} = \{s < t, s > t\}$, and after observing a sequence of comparisons between s and t: $\boldsymbol{S} = \{s < t, s < t, \ldots, s > t \ldots\}$, we can ask what is the probability that the next element we will observe is $s < t$. This is a Bernoulli distribution with parameter p and it is well known that the most likely p is simply:

$$\boldsymbol{Pr}[s < t] = \frac{|\{s < t \in \boldsymbol{S}\}|}{|\boldsymbol{S}|}.$$

In the baseline, if $\boldsymbol{S}$ is empty then we defaulted to $\boldsymbol{Pr}[s < t] = \frac{1}{2}$.

Now we present the remedial solution. Recall in the previous chapter we defined this probability value for the $\hat{y}$ output by elastic net regression:

$$\boldsymbol{Pr}[s < x] = \begin{cases} \frac{1}{2} + \epsilon & \hat{y} < \delta \\ \frac{1}{2} - \epsilon & otherwise, \end{cases}$$

while we used the actual probabilty value p output by the logistic regression model. In the remedial solution, we use the elastic definition defined above, and in

	Elastic Net Regression		l_1-Logistic Regression	
Gold Set	**Pairwise**	**Avg.** τ	**Pairwise**	**Avg.** τ
BCS	90.0%	0.81	93.0%	0.85
Turk	75.0%	0.62	74.0%	0.61
Turk no-tie	81.0%	0.63	81.0%	0.62
Mohit	74.0%	0.61	74.0%	0.61
Mohit no-tie	76.0%	0.52	76.0%	0.53

Table 1: . Results for the two best models combined with pointwise estimation baseline in the remedial fashion. Note how two models performs comparable across all gold sets. In addition, not the gold clusters with no ties enjoyed a higher pairwise accuracy but suffer a lower τ value.

the case of logistic regression, we actually discard the value of p and define:

$$\boldsymbol{Pr}[s < x] = \begin{cases} \frac{1}{2} + \epsilon & p > \frac{1}{2} \\ \frac{1}{2} - \epsilon & otherwise. \end{cases}$$

This captures our intuition that the prediction output by the model is less accruate than that of the actual data. Additionally, we also constructed a version of the Turk and Mohit's clusters where ties are removed. We reasoned that since our models are designed to predict ordering, while ties can be interepreted as synonyms, clusters generated without ties may give a more “fair” representation of how well the models perform. Results are displayed below.

1.3 Solution with Beta Prior

In this section we provide a more formal variant of the remedial solution. The heart of the of the problem is that we have some prior belief about the likelihood that one adjective is weaker than another, and an updated belief after observing some data, be it direct comparison or estimation from a model. Since we are modeling each edge a Bernoulli variable with parameter θ ranging over $[0, 1]$, the prior is then a distribution over the Bernoulli θ, this is the Beta distribution. In this next few paragraphs, we give a brief overview of Beta-Binomial model, in particular how it applies to our problem.

It is well known that the prior the binomial and Bernoulli likelihood function is the Beta distribution with paramters $\beta_1, \beta_2 \in \{1, \ldots\}$, where we have:

$$\begin{aligned}\boldsymbol{Pr}[\theta|\beta_1, \beta_2] &= \frac{\theta^{\beta_1 - 1}(1-\theta)^{\beta_2 - 1}}{\int_0^1 \mu^{\beta_1 - 1}(1-\mu)^{\beta_2 - 1} d\mu} \\ &= \frac{\Gamma(\beta_1 + \beta_2)}{\Gamma(\beta_1)\Gamma(\beta_2)}\theta^{\beta_1 - 1}(1-\theta)^{\beta_2 - 1}.\end{aligned}$$

The exact form of the Γ function is beyond the scope of this introduction but the reader may select any introductory book on statistics for a refresher.

Now after observing n coin tosses with h heads and t tails for $h + t = n$, the posterior probability over θ given some prior setting of β_1 and β_2 is:

$$\begin{aligned}\boldsymbol{Pr}[\theta|h, t\beta_1, \beta_2] &= \frac{\boldsymbol{Pr}[h|n, \theta]\boldsymbol{Pr}[\theta|, n, \beta_1, \beta_2]}{\boldsymbol{Pr}[h|n, \beta_1 + \beta_2]} \\ &\propto \theta^{h+\beta_1 - 1}(1-\theta)^{t+\beta_2 - 1},\end{aligned}$$

note the posterior distribution is also a beta distribution. Now we a have the distribution $\boldsymbol{Pr}[\theta|h, t\beta_1, \beta_2]$, we can return the the pointwise estimation setting and ask what is the likelihood the next toss lands heads, this is exactly the posteriror mean:

$$\begin{aligned}\boldsymbol{E}[\theta|h, t\beta_1, \beta_2] &= \int_0^1 \theta\boldsymbol{Pr}[\theta|h, t\beta_1, \beta_2] d\theta \\ &= \frac{\beta_1 + h}{\beta_1 + \beta_2 + n}.\end{aligned}$$

Not how the last line appeals strongly to intuition and therefore can be easily used: the expected outcome of the next toss given the prior is simply the prior tosses plus the tosses observed from data.

In our setting, we fix the ratio of β_1 and β_2 so that the prior probability is exactly $1/2$, thus reflecting our ignorance. Note this is consistent with our ad-hoc setting in the remedial solution. The exactly values of β_1 and β_2 is a hyperparameter to be tuned, in practice we set $\beta_1 = \beta_2 = 1$. The coin tosses observed from data and the model are also hyperparameters. Note the more confident we are with the data, the larger the values of h and t should be with respect to β_1 and β_2. We experimented with a variety of values, and settled on the following settings for h and t:

1. If there is an observation, then we use the raw comparison counts between the adjectives as h and t

2. If there is no obervation so we are using the model, we set h to be the probability that the model predicts less than, and $t = 1 - h$.

In informal terms, we are confident in the quality of direct comparisons, if they can be observed, and not very confident in the prediction of the model. Results for Beta-Binomal model is presented below. All in all, the best model uses the beta-binomial model to combine direct observations with $l - 1$-penalized logistic regression model, the regression model uses top the coin toss probability of 10 most connect neighbors as features. This model achieved 75% pairwise accuracy on Mohit's data set and the Turk set, and a Kendall's τ value of 0.61 and 0.62. After adjusting for ties, the pairwise accuracy on Mohit's data set was 76%, which approaches the inter-annotator accuracy of 78%, while the pairwise accuracy on the Turks set was 82% after adjusting for ties.

On the other hand, we observe that although Bansal's method performs well on the gold set they procured, it performs substantially worse across all other data sets: the τ's are close to zero and the pairwise accuracy is comparably low. Such surprisingly low results warrants further inspection. The low score actually hides two sources of error: error from absolute no data for any pair of adjectives, and error from wrong rankings when there is some data. In order to disentangle the two sources of error, we constructed a subset of the Turk set and the base-comparative-superlative sets where at least one pair of words from each cluster must have some data. Recall the original Turk set has 79 clusters, the new set only has 23 clusters, that is to say less than 30% of the Turk clusters appear in the N-gram data set. The story is similarly bleak for the base-comparative-superlative gold set: only 64 out of 238 clusters has any data at all, this is just 27%. After adjusting for data sparsity, we see that that Bansal's method achieves 65% accuracy on the Turk set, which is more in line with how it performed on the Mohit gold set. On the base-comparative-superlative set, Mohit's method only achieved 56% accuracy, this is not as high as the other gold set, but certainly better than random. We also report results for our regression method using N-gram and PPDB data for the two reduced gold sets, and observe the results are comparable two their performance on the full gold set as expected. All in all, it is clear that the increase in performance is primarily due to better coverage due to PPDB data.

	Elastic Net Regression		l_1-Logistic Regression		MILP	
Gold Set	**Pairwise**	**Avg.** τ	**Pairwise**	**Avg.** τ	**Pairwise**	**Avg.** τ
BCS	90.0%	0.81	**93.0%**	**0.85**	18.0%	0.02
Turk	75.0%	0.62	**75.0%**	**0.62**	25.0%	0.13
Mohit	74.0%	0.61	**75.0%**	**0.61**	69.6%	0.57
Turk no-tie	81.0%	0.63	**82.0%**	**0.63**	19.0%	0.12
Mohit no-tie	76.0%	0.52	**76.0%**	**0.53**	68.0%	0.46
BCS has-data	00.0%	0.00	00.0%	0.00	56.0%	0.13
Turk has-data	00.0%	0.00	00.0%	0.00	65.0%	0.46

Table 2: . The first two columns show results for the two best models combined with pointwise estimation baseline using Beta-Binomial model. The third column displays Mohit's MILP method using N-gram data only. The results shows that l_1-logistic regression outperformed elastic net regression on most data sets by a small (possibly insignificant) margin, otherwise they are equivalent. In particular, oberve how logistic regression performs just as well on Mohit's set as it does no the Turk set. Furthermore, both model outperform MILP by a non-trivial amount on all gold sets. Finally, note how well the MILP method performs on Mohit's gold cluster, versus how poorly it performs on other gold standards. The bottom two rows report results across all methods for the subsets of the gold clusters where there is some N-gram data.

1.4 Conclusion and Future Work

We will close this chapter by reiterating our assumption, and offer some possible avenues of exploration. First, we review the fundamental hypothesis underlying all methods we have seen so far: there is a positive relationship between how words are used on average in the context of intensifiers such as adverbs or N-gram patterns, and how an annotator ranks pairs of such words in isolation. The fact that all successful methods we have seen so far achieved close to or above 70% pairwise accuracy confirms this hypothesis. However, we must caution those who wish to improve upon this results, inter-annotator accuracy on Mohit's set is 78%. In fact, many of the rankings are subjective and we do not believe it is meaningful to achieve, for example 90%, on the gold standards procured by Mohit and us.

In fact, consider this cluster:

- interesting
- entertaining
- fascinating
- intriguing
- amusing
- exciting,

the gold set suggests

$$\text{interesting} < \text{intriguing} < \text{amusing} = \text{entertaining} < \text{fascinating} = \text{exciting},$$

but an equally plausible ranking is:

interesting < amusing < entertaining < fascinating < intriguing < exciting.

Penalized elastic net regression output:

interesting < entertaining < fascinating < intriguing < amusing < exciting.

We leave it to the reader to judge which cluster makes the most sense, if any. This example suggests an immediate improvement upon our work: procure better clusters so the results are more meaningful. There are two ways to improve the quality of each cluster: prune the clusters for synonyms, and resist the temptation for large clusters. The first suggestion is self explanatory, since none of the models account for synonyms, clusters that contain words annotators would consider to be synonyms will unfairly depress the score. Eschewing large clusters is important because the larger the cluster, the more likely polysemy will

References

Gerard de Melo and Mohit Bansal. 2013. Good, great, excellent: Global inference of semantic intensities. *Transactions of the Association for Computational Linguistics*, 1:279–290.

Joo-Kyung Kim and Marie-Catherine de Marneffe. 2013. Deriving adjectival scales from continuous space word representations. In *EMNLP*, pages 1625–1630.

Peter F Schulam and Christiane Fellbaum. 2010. Automatically determining the semantic gradation of german adjectives. In *KONVENS*, pages 163–167. Citeseer.

Vera Sheinman and Takenobu Tokunaga. 2009. Adjscales: Visualizing differences between adjectives for language learners. *IEICE TRANSACTIONS on Information and Systems*, 92(8):1542–1550.

Vera Sheinman, Takenobu Tokunaga, Isaac Julien, Peter Schulam, and Christiane Fellbaum. 2012. Refining wordnet adjective dumbbells using intensity relations. In *Sixth International Global Wordnet Conference*, pages 330–337.