1 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.

	Elastic Net Regression		l_1-Logistic Regression	
Gold Set	**Pairwise**	**Avg.** τ	**Pairwise**	**Avg.** τ
BCS	00.0%	0.00	00.0%	0.00
Turk	00.0%	0.00	00.0%	0.00
Turk no-tie	00.0%	0.00	00.0%	0.00
Mohit	00.0%	0.00	00.0%	0.00
Mohit no-tie	00.0%	0.00	00.0%	0.00

Table 2: . Results for the two best models combined with pointwise estimation baseline using Beta-Binomial model.

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.