Machine Learning Assignment 2

June 8, 2016

1 Building a logistic regression classifier by sentence length

1.1 Write down the fitted model equation

The probability of predicting German

$$F(x) = \frac{1}{1 + exp(-(-0.4125701 + (-0.0053695) * words\_count} \quad (1)$$

The R results

```
glm(formula = German ~ word_count, family = "binomial", data =
    Language.words)

Coefficients:
                    Estimate          Std. Error           z
          value Pr(>|z|)
(Intercept)   -0.4125701      0.0182046     -22.663       < 2e-16 **
    *
word_count  -0.0053695     0.0007967     -6.739        1.59e-11 ***
---
    Null deviance: 56199  on 42510  degrees of freedom
Residual deviance: 56153  on 42509  degrees of freedom
AIC: 56157
```

1.2 Interpreting

This coefficient represents the odds ratio of two groups. The odds ratio is small, so this model has seriously bias and tend to predict non-German over German.

$$\text{Odds ration} = \frac{German}{Non - German} \quad (2)$$

1.3 accuracy, precision, recall and F1-score

Threshold was set at 0.5

Accuracy	Precision	Recall	F-score
0.6261203	0.0000000	0.0000000	0.0000000

2 Try probability threshold other than 0.5

2.1 Find the threshold to maximize the F-score

When the threshold reached 0.350, the maximize F-score was 0.551706609

2.2 Discrimination function

$$\begin{cases} f(X) > 0.350 & 1 \quad \text{Label as true value} \\ f(X) < 0.350 & 0 \quad \text{Label as false value} \end{cases} \tag{3}$$

2.3 Accuracy, precision, recall and F1-score

Accuracy	Precision	Recall	F-score
0.4191621	0.38774052	0.9559582232	0.551706609

3 Building a logistic regression classifier by sentence length and 15 POS tags

3.1 The model

```
glm(formula = as.factor(language) ~ ., family = "binomial", data
    = All.language.set)
```

Coefficients: (1 not defined because of singularities)

	Estimate	*Std.Error*	*z*	*valuePr*(> |*z*|)
(*Intercept*)	−17.061351	0.837509	−20.372	$< 2e-16 **$
ADJ	17.716542	0.847298	20.909	$< 2e-16 **$
ADP	17.852731	0.855065	20.879	$< 2e-16 **$
ADV	21.209763	0.858209	24.714	$< 2e-16 **$
AUX	11.635816	0.872059	13.343	$< 2e-16 **$
CONJ	20.280064	0.899407	22.548	$< 2e-16 **$
DET	31.363951	0.864858	36.265	$< 2e-16 **$
NOUN	12.418114	0.845375	14.689	$< 2e-16 **$
NUM	17.360289	0.862973	20.117	$< 2e-16 **$
PART	6.950780	0.964719	7.205	$5.81e-13 **$
PRON	19.295260	0.863053	22.357	$< 2e-16 **$
PROPN	16.059863	0.838902	19.144	$< 2e-16 **$
PUNCT	17.367860	0.853688	20.345	$< 2e-16 **$
SCONJ	−0.769437	1.081114	−0.712	0.477
VERB	9.260456	0.874992	10.583	$< 2e-16 **$
word_count	0.008460	0.001185	7.140	$9.32e-13 **$
X	*NA*	*NA*	*NA*	*NA*

3.2 Accuracy,precision, recall and F1-score

Accuracy	Precision	Recall	F-score
0.7921950	0.7288344	0.7073739	0.7179438

3.3 Comments

The model is not seriously suffer from the imbalance class scenario. And the scores that evaluate the model are all better than the previous model.

4 Two three-way classifiers predicting the language

4.1 Fit two model L1 and L2 regularization

L1-regularized logistic regression In the model English were label as 1, German were label as 2, Japanese were label as 3.

Table 1: The weights of L1 and L2 regularized logistic regression

	L1			L2		
	English	German	Japanese	English	German	Japanese
ADJ	1.0237669	0.7890187	-14.9408574	0.6857224	1.9886923	-5.7071440
ADP	-9.1574879	0.8148334	17.3049650	-7.5130140	-0.7236038	14.4031352
ADV	-2.362436	4.409166	-12.633060	-0.7832083	1.8059409	-4.0984844
AUX	-5.950119	-5.580809	17.651757	-3.028476	-3.144421	13.035045
CONJ	0.6777368	3.3818634	-11.6251249	0.7378444	0.7341410	-2.9635377
DET	-3.255645	14.201862	-46.809332	-1.924565	7.122533	-15.065589
NOUN	0.2328024	-4.1722805	-1.2321694	-1.156510	-2.184627	4.220319
NUM	0.02271107	0.46700548	-2.79667839	-0.1383791	-0.1861638	2.0201765
PART	10.9002724	-10.5199639	-0.2396408	2.254829954	-1.089416445	0.003708113
PRON	5.558718	1.838501	-36.379638	4.4464282	0.1642955	-9.2919527
PROPN	1.5139127	-0.7192767	-13.3201763	-0.0087899	0.3484264	-4.9457591
PUNCT	-0.3312016	0.3454577	-3.2736490	-0.01843138	-0.03923988	1.07009851
SCONJ	1.526599	-18.451875	22.526309	0.1497023	-1.5100842	4.9089908
VERB	5.502277	-6.439820	-12.663024	3.801662	-1.958400	-1.149571
word_count	-0.017022811	0.008632529	0.066097922	-0.019900851	0.004098783	0.042046788
X	13.466554	-16.610988	-9.177153	2.9900801	-1.8183289	-0.8029581
Bias	0.08844401	-0.24518582	0.29294469	0.4948953	-0.4902554	-4.3635235

4.2 Briefly explain the differences between the coefficient values

The differences of coefficient values among three languages in L2 regularization are smaller than L1 regularization. For example. coefficient value of ADJ in L1 regularization model, Japanese has the value which much smaller than other languages. However, the ADJ coefficient values have smaller differences among three languages in L2 regularization model. In other words, models with L1 regularization has more serious class imbalance issues than with L2 regularization .

4.3 Calculate and compare accuracy of both L1 and L2 regularized models

Accuracy of L1 and L2

	Total
L1	0.7853262
L2	0.7662017

4.4 Tabulate the confusion matrix

		Predict		
		English	German	Japanese
Language	English	11577	4591	454
	German	4366	11290	238
	Japanese	233	57	9705

5 The 10-fold cross validation

The 10-fold Accuracy is 0.7662949 and the standard error is 0.00200215