Machine Learning Assignment 2

June 9, 2016

1 Building a logistic regression classifier by sentence length

1.1 Write down the fitted model equation

The probability of predicting German

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

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 predicting two groups. The odds ratio is small, so this model has seriously bias and tend to predict non-German over German.

$$\text{Odds ration} = \frac{\text{Predicted German}}{\text{Predicted Non-German}} \tag{2}$$

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	0.4586609	0.8430927	-3.0923207	0.3818952	1.7046983	-2.5587346
ADP	-7.41400	0.00000	13.04822	-4.5134628	-0.5841696	5.6995738
ADV	-0.01235211	3.11868558	0.00000000	-0.4124346	1.5624213	-1.8663966
AUX	-2.922434	-4.020015	11.117181	-1.831182	-2.709360	4.911108
CONJ	0.00000000	0.00000000	0.00000000	0.4338546	0.6366664	-1.1394019
DET	-1.118768	12.725544	-22.731433	-1.132230	6.174392	-5.668273
NOUN	0.1414386	-2.6625323	2.9111699	-1.158916	-1.829505	3.446598
NUM	0.00000000	0.00000000	0.00000000	-0.1336489	-0.1516168	0.4538806
PART	4.047479	-1.887776	0.000000	1.3667964	-0.9547400	-0.1924081
PRON	4.675701	0.00000000	-12.512257	2.673749	0.134431	-3.331745
PROPN	1.503901	0.000000	-3.567297	0.6218526	0.2279806	-2.5194832
PUNCT	0.02733586	0.000000	0.000000	0.04666163	-0.03770481	-0.51531662
SCONJ	0.000000	-4.198111	0.000000	0.08411322	-1.31381975	1.48847503
VERB	4.778885	-4.410038	0.000000	2.269673	-1.713162	-0.702764
word_count	-0.01575852	0.00134223	0.04185638	-0.028101196	0.001018213	0.037595042
X	4.50124	-2.96886	0.00000	1.7668118	-1.5761648	-0.3553914
Bias	-0.2307681	-0.4895501	-3.4206723	0.4635337	-0.4296531	-2.8502797

4.2 Briefly explain the differences between the coefficient values

The model with L1 regularization made many weight values become 0, whereas the L2 regularization will not create lots of 0 weight values.

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

Accuracy of L1 and L2

	Total
L1	0.7717532
L2	0.7555691

The model with L2 regularization is lower than L1 regularization.

4.4 Tabulate the confusion matrix

		Predict		
		English	German	Japanese
Language	English	11266	4838	518
	German	4302	11369	223
	Japanese	446	64	9485

5 The 10-fold cross validation

The 10-fold Accuracy is 0.7471471 and the standard error is 0.001944774