CS 594 – Homework 2 **Due: 12/03**

Huayi Li, Xiaokai Wei, Sihong Xie

(a) Describe your choice of feature functions f for your implementation. How did you choose these features f?

Answer:

- We model the pairwise dependency between Y_i ($i = 4, 5, 6$) and Y_i ($i = 1, 7, 8, 9$) ($Y_{1:10}$ correspond to $Q_{11:20}$), respectively. The dependency structure in shown in Figure 1. They are connected in the graph because either they have strong correlation or they belong to similar topic. We also identify the most relevant X_i for Y_i ($i = 4, 5, 6$) based on pearson correlation (The top correlated pairs are listed in Figure 2). According to both correlation calculation and human inspection, Y_i ($i = 1, 7, 8, 9$) have little correlation with $X_{1:10}$. Therefore, we only consider the pairwise dependency among Y_i for them.
- Y_i ($i = 2, 3, 10$) does not have obvious correlation with other Y_i. So they solely depend on $X_{1:10}$.

The longest chain in this model is of length 4 (i.e., Y_i ($i = 1, 7, 8, 9$)), which makes the inference readily tractable. Besides, to prevent over-fitting, we add $l2$ regularization in the objective to be optimized.

```
isolated y's
12
13
20

chain without x's

11----17----18-----19


chain with x's

5    9    2    3    2    4
 \  /      \  /      \  /
  16  ----- 15 -------14
```

Figure 1: Dependency structure among X_i and and Y_i

(b) Provide your source code for your implemented predictive model.

Attached in email.

(c) What is the 5-fold cross-validated log-loss of your predictive model?

Answer:

We calculate the log-loss based on the following formula:

```
['Q18', 'Q19', 0.85212390016094319]
['Q2', 'Q4', 0.78942283080558362]
['Q6', 'Q13', 0.70444562350345785]
['Q2', 'Q10', 0.68791916049010982]
['Q1', 'Q10', 0.6392092266028]
['Q2', 'Q3', 0.62868167722568358]
['Q11', 'Q17', 0.60880636534164101]
['Q4', 'Q10', 0.58820877128290372]
['Q3', 'Q13', 0.56358738395530616]
['Q9', 'Q16', 0.55377492419453833]
['Q2', 'Q15', 0.54103244033623132]
['Q14', 'Q15', 0.5284397406288498]
['Q11', 'Q19', 0.5274273340898028]
['Q5', 'Q16', 0.52468477931712387]
['Q11', 'Q18', 0.52291812058888365]
['Q3', 'Q15', 0.50567122132317732]
```

Figure 2: Pearson correlation of question pairs

$$
\begin{aligned}
logloss &= -\frac{1}{N}\sum_{j=1}^{N}\mathcal{I}(y^j_{1:10})\log P(y^j_{1:10}|x^j_{1:10}) \\
&= -\frac{1}{N}\sum_{j=1}^{N}\mathcal{I}(y^j_{1:10})\log[P(y^j_{12}|x^j_{1:10}) * P(y^j_{3}|x^j_{1:10}) * P(y^j_{10}|x_{1:10^j}) * \\
&\quad P(y^j_{4,5,6}|x^j_{1:10}) * P(y^j_{1,7,8,9})] \\
&= -\frac{1}{N}\sum_{j=1}^{N}\mathcal{I}(y^j_{1:10})[\log P(y^j_{2}|x^j_{1:10}) + \log P(y^j_{3}|x^j_{1:10}) + \log P(y^j_{10}|x^j_{1:10}) + \\
&\quad \log P(y^j_{4,5,6}|x^j_{1:10}) + \log P(y^j_{1,7,8,9})]
\end{aligned}
$$

where j is the index of the data points and $N = 22/5 \approx 4$. The 5-fold cross-validation average log-loss is 10.255007.

(d) How does this 5-fold cross-validation log-loss change as a function of the complexity of conditional random field structure? (Either using a smaller set of features, e.g., $\theta_i = 0$ or parameter regularization)?

Answer:

We do experiments on models with different levels of complexity and report the log-loss below.

- Over-simplified case: do not model the dependency structure among Y_i. In this case, the model degenerates into logistic regression. The average log loss is 10.572548
- Overcomplicated case: add more dependency among Y_i. All Y's line up to form a chain which yields an average log loss 10.583492

As we can observe from the results above, when the model gets too complicated (e.g., with more pairwise features between Y_i), the cross-validtion log-loss suffers; when the model gets over-simplified (e.g. without modeling the dependency among Y_i), the log loss also degrades.