

CS769 Advanced Natural Language Processing  
Spring 2010 Homework 4  
Reference Answer

March 10, 2010

**Answer 1**

The SVM light format has the label of the example first, followed by term-count pairs where each term is given as a number. Depending on how you ordered the terms in your vocabulary the values of the terms may be different, but each word in the joke occurs only once, so there should be 16 pairs.

The first document in the positive.txt is represented in SVM-light format by: +1 1:1 2:1 3:1 4:1 5:1 6:1 7:1 8:1 9:1 10:1 11:1 12:1 13:1 14:1 15:1 16:1  
 Note: the words in vocabulary are ordered by their occurrence order in the positive.txt and negative.txt.

You may also get something like this which is also OK: 13:1 16:1 273:1 338:1 396:1 639:1 734:1 735:1 748:1 973:1 1003:1 1029:1 1338:1 1403:1 1469:1 1535:1

**Answer 2**

Any random way of splitting the positive and negative documents into 10 equal sized folds. Each fold has  $400/10 = 40$  documents. Of these, 20 are positive and 20 are negative documents.

Its not a good idea to simply split the data without randomizing the order at all because you will get 5 folds with all positive test cases and 5 with all negative. Since some algorithms have different Type I (false positive) and Type II (false negative) errors, and others assume priors from the training; data each fold should have a similar ratio of positive and negative examples as the entire dataset.

**Answer 3**

Accuracies are generally between 60-90%

**Answer 4**

Accuracies are generally between 60-90%

**Answer 5**

It is definitely best to check your answers to this with the Matlab ttest or ttest2 function. The result you get depends on how you chose your folds, but the most likely result should be that there is a significant difference at the 5% level, and the polynomial kernel should be performing worse. There is also a chance that you wont be able to reject the null hypothesis.

**Answer 6**

Any new feature , like number of words, number of repeated words, etc.

**Answer 7**

Mostly, the new feature did not improve classification.

**Answer 8** The original primal SVM problem is

$$\min_{w,b,\xi} \frac{1}{2} \| w \|^2 + C \sum_{i=1}^n \xi_i$$

such that

$$\begin{aligned} y_i(w^T x_i + b) &\geq 1 - \xi_i, i = 1, 2, \dots, n \\ \xi_i &\geq 0 \end{aligned}$$

Must link:  $x_{n+1}$  and  $x_{n+2}$  belong to the same class.

i.e. They lie on the same side of the classification line. So, add

$$(w^T x_{n+1} + b)(w^T x_{n+2} + b) > 0$$

Cannot link:  $x_{n+3}$  and  $x_{n+4}$  belong to different classes. So, add

$$(w^T x_{n+3} + b)(w^T x_{n+4} + b) < 0$$

Taking into account the new  $\xi_i$ 's in the object function, the primal SVM problem with the must-link and cannot-link is:

$$\min_{w, b, \xi} \frac{1}{2} \|w\|^2 + C_1 \sum_{i=1}^n \xi_i + C_2 \sum_{i=n+1}^{n+4} \xi_i$$

such that

$$\begin{aligned} y_i(w^T x_i + b) &\geq 1 - \xi_i, i = 1, 2, \dots, n \\ |(w^T x_i + b)| &\geq 1 - \xi_i, i = n + 1, \dots, n + 4 \\ \xi_i &\geq 0, i = 1, 2, \dots, n + 4 \\ (w^T x_{n+1} + b)(w^T x_{n+2} + b) &> 0 \\ (w^T x_{n+3} + b)(w^T x_{n+4} + b) &< 0 \end{aligned}$$

### Answer 9

Assume we can get the initial value for the parameters, then repeat the following two steps:

**E-Step:** For  $i = 1, \dots, n$ ,  $k = 1, 2$ , compute  $\gamma_{ik} = p(y_i = k | x_i, \Theta)$ . where

$$p(y_i = 1 | x_i, \Theta) = \frac{p(x_i | \mu_1, \sigma_1^2) p(\pi_1)}{p(x_i | \mu_1, \sigma_1^2) p(\pi_1) + p(x_i | \mu_2, \sigma_2^2) p(\pi_2)}$$

and

$$p(y_i = 2 | x_i, \Theta) = \frac{p(x_i | \mu_2, \sigma_2^2) p(\pi_2)}{p(x_i | \mu_1, \sigma_1^2) p(\pi_1) + p(x_i | \mu_2, \sigma_2^2) p(\pi_2)}$$

**M-Step:** Compute  $\Theta$ .

$$\begin{aligned}\pi_k &= \frac{\sum_{i=1}^n \gamma_{ik}}{n}, k = 1, 2 \\ \mu_k &= \frac{\sum_{i=1}^n x_i \gamma_{ik}}{\sum_{i=1}^n \gamma_{ik}}, k = 1, 2 \\ \sigma_k^2 &= \frac{\sum_{i=1}^n \gamma_{ik} (x_i - \mu_k)^2}{\sum_{i=1}^n \gamma_{ik}}, k = 1, 2\end{aligned}$$

**Answer 10**

#Iter.	log likelihood	$\pi_1$	$\mu_1$	$\sigma_1$	$\pi_2$	$\mu_2$	$\sigma_2$
0	-23.8728	0.5000	-1.4781	1.0000	0.5000	0.8800	1.0000
1	-17.9800	0.5501	-0.9954	0.4466	0.4499	0.6256	0.9199
2	-16.6554	0.5939	-1.0138	0.3372	0.4061	0.8273	0.7935
3	-15.0710	0.6349	-1.0063	0.3365	0.3651	1.0210	0.5638
4	-11.5108	0.6638	-0.9941	0.3383	0.3362	1.1714	0.2333
5	-11.0412	0.6667	-0.9924	0.3386	0.3333	1.1865	0.1665
6	-11.0412	0.6667	-0.9924	0.3386	0.3333	1.1865	0.1665

**Answer 11**

Since  $\exp(z) > 0$  and  $\log_2(1 + \exp(z)) > \log_2(1) = 0$ , so

$$\begin{aligned}c(\theta) &= \sum_{i=1}^n \log_2(1 + \exp(-y_i \theta^\top x_i)) \\ &\geq \sum_{i=1}^n \log_2(1 + \exp(-y_i \theta^\top x_i)) \cdot I_{\{sgn(\theta^\top x_i) \neq y_i\}}\end{aligned}$$

where  $I$  is the indicator function. If  $sgn(\theta^\top x_i) \neq y_i$ , then  $-y_i \theta^\top x_i > 0$  and  $\log_2(1 + \exp(-y_i \theta^\top x_i)) > \log_2(1 + 1) = 1$

so  $c(\theta) > \sum_{i=1}^n I_{\{sgn(\theta^\top x_i) \neq y_i\}}$  = the number of classification mistakes.