Assignment 3

Learning Objectives:

1. Gain insight into trade-offs between Naïve Bayes and SVM
2. Analyze the effect of skewness of distribution of class attribute values on classification performance
3. Learn about smoothing

Description:

In this assignment you will work with two data sets. One will be the Play Tennis data set we have worked with several times, and the other will be a data set constructed from chat data originally collected in Chinese, with each contribution assigned one of 17 topic codes. For the purpose of this assignment, it is not important to understand what the features represent or what the specific topics are, so the attribute names have been replaces with word1-word1000. The class attribute is called Topic, and its values are of the form c<number>-<number>.

The Topic dataset is provided in ".arff" format. The data from the Play Tennis data set is included here for your reference:

outlook	temperature	humidity	windy	play
sunny	hot	high	FALSE	no
sunny	hot	high	TRUE	no
overcast	hot	high	FALSE	yes
rainy	mild	high	FALSE	yes
rainy	cool	normal	FALSE	yes
rainy	cool	normal	TRUE	no
overcast	cool	normal	TRUE	yes
sunny	mild	high	FALSE	no
sunny	cool	normal	FALSE	yes
rainy	mild	normal	FALSE	yes
sunny	mild	normal	TRUE	yes
overcast	mild	high	TRUE	yes
overcast	hot	normal	FALSE	yes
rainy	mild	high	TRUE	no

Step-by-Step Guide:

1. List all of the counts you would store for a simple Naïve Bayes model trained from the Play Tennis data set. You may wish to express your results in tabular form using MS Excel. Note that for this assignment we are adding an additional possible value for the Humidity feature, namely low, which never occurs in the training data.

Outlook			Temperature			Humidity			Windy			Play	
	yes	no		yes	no		yes	no		yes	no	yes	no
Sunny	2	3	Hot	2	2	High	3	4	False	6	2	9	5
Overcast	4	0	Mild	4	2	Normal	6	1	True	3	3		
rainy	3	2	cool	3	1	Low	0	0					
Sunny	2/9	3/5	Hot	2/9	2/5	High	3/9	4/5	False	6/9	2/5	9/14	5/14
Overcast	4/9	0/5	Mild	4/9	2/5	Normal	6/9	1/5	True	3/9	3/5		
rainy	3/9	2/5	cool	3/9	1/5	Low	0/9	0/5					

2. Now construct a model from the same data using a form of smoothing where you simply add 1 to all counts.

Outlook			Temperature			Humidity			Windy			Play	
	yes	no		yes	no		yes	no		yes	no	yes	no
Sunny	3	4	Hot	3	3	High	4	5	False	7	3	10	6
Overcast	5	1	Mild	5	3	Normal	7	2	True	4	4		
rainy	4	3	cool	4	2	Low	1	1					
Sunny	3/12	4/8	Hot	3/12	3/8	High	4/12	5/8	False	7/11	3/7	10/16	6/16
Overcast	5/12	1/8	Mild	5/12	3/8	Normal	7/12	2/8	True	4/11	4/7		
rainy	4/12	3/8	cool	4/12	2/8	Low	1/12	1/8					

3. Compute your prediction for the following test instances using both models above, and show all of your work. Comment on the impact of zero counts on predictions. Comment on the impact of smoothing on prediction.

Outlook	Temperature	Humidity	Windy	Play
overcast	hot	normal	TRUE	
rainy	hot	high	FALSE	
overcast	cool	normal	TRUE	
rainy	mild	low	FALSE	

Non-Smoothed Models:

.1. Overcast, hot, normal, true

Likelihood of yes = 4/9 * 2/9 * 6/9 * 3/9 * 9/14 = 0.0141
Likelihood of no = 0/5 * 2/5 * 1/5 * 3/5 * 5/14 = 0

Probability of yes = 0.0141/(0.0141 + 0) * 100% = 100%
Probability of no = 0/ (0.0141 + 0) * 100% = 0

2. Rainy, hot, high, false

Likelihood of yes = 3/9 * 2/9 * 3/9 * 6/9 * 9/14 = 0.0106
Likelihood of no = 2/5 * 2/5 * 4/5 * 2/5 * 5/14 = 0.0183

Probability of yes = 0.0106/(0.0106 + 0.0183) * 100% = 37%
Probability of no = 0.0183/(0.0106 + 0.0183) * 100% = 63%

3. Overcast, cool, normal, true

Likelihood of yes = 4/9 * 3/9 * 6/9 * 3/9 * 9/14 = 0.0211
Likelihood of no = 0/5 * 1/5 * 1/5 * 3/5 * 5/14 = 0

Probability of yes = 0.0211/(0.0211 + 0) * 100% = 100%
Probability of no = 0/(0.0211 + 0) * 100% = 0

4. Rainy, mild, Low, False

Likelihood of yes = 3/9 * 4/9 * 0/9 * 6/9 * 9/14 = 0
Likelihood of no = 2/5 * 2/5 * * 0/5 * 2/5 * 5/14 = 0

Probability of yes = 0%
Probability of no = 0%

Smoothed Models:

1. Overcast, hot, normal, true

Likelihood of yes = 5/12 * 3/12 * 7/12 * 4/11 * 10/16 = 0.0138
Likelihood of no = 1/8 * 3/8 * 2/8 * 4/7 * 6/16 = 0.0025

Probability of yes = 0.0138/(0.0138 + 0.0025) * 100% = 85%
Probability of no = 0.0025/(0.0138 + 0.0025) * 100% = 15%

2. Rainy, hot, high, false

Likelihood of yes = 4/12 * 3/12 * 4/12 * 7/11 * 10/16 = 0.011
Likelihood of no = 3/8 * 3/8 * 5/8 * 3/7 * 6/16 = 0.0141

Probability of yes = 0.011/(0.011 + 0.0141) * 100% = 44%
Probability of no = 0.0141/(0.011 + 0.0141) * 100% = 56%

3. Overcast, cool, normal, true

Likelihood of yes = 5/12 * 4/12 * 7/12 * 4/11 * 10/16 = 0.0184
Likelihood of no = 1/8 * 2/8 * 2/8 * 4/7 * 6/16 = 0.0017

Probability of yes = 0.0184/(0.0184 + 0.0017) * 100% = 92%
Probability of no = 0.0017/(0.0184 + 0.0017) * 100% = 8%

4. Rainy, mild, Low, False

Likelihood of yes = 4/12 * 5/12 * 1/12 * 7/11 * 10/16 = 0.0046
Likelihood of no = 3/8 * 3/8 * 1/8 * 3/7 * 6/16 = 0.0028

Probability of yes = 0.0046/(0.0046 + 0.0028) * 100% = 62%
Probability of no = 0.0028/(0.0046 + 0.0028) * 100% = 38%

Optional: Load the Topic dataset into weka and run a cross-validation experiment using Naïve Bayes (from the bayes folder) and then one using SVM (called SMO under the functions folder). Which one performed better? Why do you think this was the case?

Results using Naïve Bayes

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         833               68.1669 %
Incorrectly Classified Instances       389               31.8331 %
Kappa statistic                          0.6151
Mean absolute error                      0.0454
Root mean squared error                  0.1657
Relative absolute error                 45.2608 %
Root relative squared error             74.0523 %
Total Number of Instances             1222

=== Detailed Accuracy By Class ===

TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
  0.889     0.007      0.833     0.889     0.86       0.936    c4-3
  0.698     0.031      0.726     0.698     0.711      0.943    c4-2
  0.793     0.022      0.822     0.793     0.807      0.964    c3-0
  0.8       0.009      0.828     0.8       0.814      0.953    c2-1
  0.667     0.014      0.652     0.667     0.659      0.972    c1-1
  0.395     0.008      0.625     0.395     0.484      0.842    c2-0
  0.895     0.001      0.971     0.895     0.932      0.97     c3-1
  0.815     0.005      0.786     0.815     0.8        0.946    c4-7
  0.815     0.232      0.617     0.815     0.703      0.874    na
  0.759     0.005      0.786     0.759     0.772      0.996    c4-4
  0.357     0.034      0.575     0.357     0.441      0.832    c6-1
  0.476     0          1         0.476     0.645      0.98     c5-3
  0         0.001      0         0         0          0.874    c4-9
  0.6       0.017      0.574     0.6       0.587      0.896    c1-2
  0         0          0         0         0          0.86     c4-10
  0.292     0.019      0.463     0.292     0.358      0.794    c6-2
```

0.5	0.001	0.667	0.5	0.571	0.781	c4-6

=== Confusion Matrix ===

a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	<-- classified as
40	0	0	1	0	0	0	1	2	0	1	0	0	0	0	0	0	a = c4-3
1	90	0	0	0	1	0	0	23	3	3	0	1	0	0	7	0	b = c4-2
0	0	111	0	4	0	0	0	23	0	1	0	0	0	0	1	0	c = c3-0
2	2	0	48	0	2	0	0	6	0	0	0	0	0	0	0	0	d = c2-1
0	0	2	0	30	0	0	0	8	1	4	0	0	0	0	0	0	e = c1-1
1	4	2	0	0	15	0	0	13	0	0	0	0	0	0	3	0	f = c2-0
0	1	1	0	0	0	34	0	1	0	1	0	0	0	0	0	0	g = c3-1
0	1	0	0	1	0	0	22	3	0	0	0	0	0	0	0	0	h = c4-7
1	8	9	1	7	2	0	3	313	0	19	0	0	12	0	8	1	i = na
0	0	0	0	0	0	0	0	5	22	1	0	0	1	0	0	0	j = c4-4
0	5	10	1	2	1	1	1	58	1	50	0	0	7	0	3	0	k = c6-1
0	1	0	0	0	0	0	0	10	0	0	10	0	0	0	0	0	l = c5-3
1	0	0	0	0	0	0	0	7	0	0	0	0	0	0	0	0	m = c4-9
0	3	0	0	2	0	0	0	8	0	5	0	0	27	0	0	0	n = c1-2
0	2	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	o = c4-10
2	6	0	7	0	2	0	1	26	1	1	0	0	0	0	19	0	p = c6-2
0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	2	q = c4-6

Results using SVM

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances	898	73.4861 %
Incorrectly Classified Instances	324	26.5139 %
Kappa statistic	0.6795	
Mean absolute error	0.1045	
Root mean squared error	0.2246	
Relative absolute error	104.2081 %	
Root relative squared error	100.3751 %	
Total Number of Instances	1222	

=== Detailed Accuracy By Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
0.844	0.006	0.844	0.844	0.844	0.914	c4-3
0.806	0.027	0.776	0.806	0.791	0.93	c4-2
0.871	0.009	0.924	0.871	0.897	0.976	c3-0
0.717	0.009	0.811	0.717	0.761	0.961	c2-1
0.711	0.008	0.78	0.711	0.744	0.955	c1-1
0.395	0.009	0.577	0.395	0.469	0.788	c2-0
0.947	0.001	0.973	0.947	0.96	0.985	c3-1
0.889	0.004	0.828	0.889	0.857	0.976	c4-7
0.872	0.196	0.671	0.872	0.759	0.843	na
0.759	0.001	0.957	0.759	0.846	0.989	c4-4
0.479	0.042	0.598	0.479	0.532	0.781	c6-1
0.857	0.002	0.857	0.857	0.857	0.992	c5-3
0.25	0	1	0.25	0.4	0.677	c4-9
0.444	0.015	0.526	0.444	0.482	0.886	c1-2
0	0	0	0	0	0.816	c4-10
0.277	0.009	0.643	0.277	0.387	0.787	c6-2
0.5	0	1	0.5	0.667	0.754	c4-6

=== Confusion Matrix ===

```
   a   b   c   d   e   f   g   h   i   j   k   l   m   n   o   p   q   <-- classified as
  38   0   0   2   0   0   0   1   2   0   1   0   0   0   0   1   0 |   a = c4-3
   1 104   0   0   0   2   0   0  17   0   3   1   0   0   0   1   0 |   b = c4-2
   0   0 122   0   1   1   0   0  12   0   4   0   0   0   0   0   0 |   c = c3-0
   2   2   0  43   0   1   0   0   7   0   3   0   0   0   0   2   0 |   d = c2-1
   0   0   0   0  32   0   0   0   8   0   5   0   0   0   0   0   0 |   e = c1-1
   0   4   2   1   0  15   0   0  13   0   1   0   0   1   0   1   0 |   f = c2-0
   0   0   0   0   0   0  36   0   0   0   1   0   0   1   0   0   0 |   g = c3-1
   0   0   0   0   0   0   0  24   3   0   0   0   0   0   0   0   0 |   h = c4-7
   2  10   1   2   2   2   0   2 335   0  17   1   0   8   0   2   0 |   i = na
   0   0   0   0   0   0   0   0   4  22   2   0   0   0   0   1   0 |   j = c4-4
   0   3   6   1   4   2   1   1  46   0  67   0   0   8   0   1   0 |   k = c6-1
   0   1   0   0   0   0   0   0   2   0   0  18   0   0   0   0   0 |   l = c5-3
   0   0   0   0   0   0   0   0   5   0   0   0   2   0   0   1   0 |   m = c4-9
   0   1   0   0   1   0   0   0  17   0   6   0   0  20   0   0   0 |   n = c1-2
   0   2   0   0   0   1   0   0   1   0   0   0   0   0   0   0   0 |   o = c4-10
   2   6   1   4   0   2   0   1  27   1   2   1   0   0   0  18   0 |   p = c6-2
   0   1   0   0   1   0   0   0   0   0   0   0   0   0   0   0   2 |   q = c4-6
```

In addition to the fact that the performance for Naïve Bayes was lower overall than for SVM, you can see it is more true of Naïve Bayes than SVM that the classes that had the lowest performance were the most infrequent ones. So skewness was more of an issue for Naïve Bayes than for SVM. We know that in cases where the class value distribution is skewed naïve bayes falls prey to overpredicting the majority class unless the features are really strong. But SVM does not have this problem.

Deliverables:

1. Submit your answers for Steps 1-3
2. Optional: Submit your answers for Step 4

Miscellaneous Notes:

1. If you have not increased your heap size yet in your computer, please increase it now!
2. The experiments involving support vector machines take more than 10 minutes depending on your computer.