Name : Tang Sophearin

Affiliation : Department of Intelligence Science and Technology, Nishida Laboratory

Student ID : 6930-30-3963

Data Mining (3)

1. Analyze the example data, in which train.dat is used for training and test.dat is used for evaluation. Try a variation of kernels and parameters and discuss the results.

Default prediction (default regularization parameter c = 1.0):

```
C:\Users\Phearin\Documents\Assignment\Practice of Information System\Information System Analysis\Data Mining 3>svm_learn train.dat def_model
Scanning examples...done
Reading examples into memory...100..200..300..400..500..600..700..800..900..1000..1100..1200..1300..1400..1500..1600..1700..1800..1900..2000..OK. (2000 examples read)
Setting default regularization parameter C=1.0000
Optimizing..........................................................................................................................................
.......................................................................................................................................................
...........done. (425 iterations)
Optimization finished (5 misclassified, maxdiff=0.00085).
Runtime in cpu-seconds: 0.05
Number of SV: 878 (including 117 at upper bound)
L1 loss: loss=35.67674
Norm of weight vector: |w|=19.55576
Norm of longest example vector: |x|=1.00000
Estimated VCdim of classifier: VCdim<=383.42791
Computing XiAlpha-estimates...done
Runtime for XiAlpha-estimates in cpu-seconds: 0.00
XiAlpha-estimate of the error: error<=5.85% (rho=1.00,depth=0)
XiAlpha-estimate of the recall: recall=>95.40% (rho=1.00,depth=0)
XiAlpha-estimate of the precision: precision=>93.07% (rho=1.00,depth=0)
Number of kernel evaluations: 45954
Writing model file...done

C:\Users\Phearin\Documents\Assignment\Practice of Information System\Information System Analysis\Data Mining 3>svm_classify test.dat def_model def_predict
Reading model...OK. (878 support vectors read)
Classifying test examples..100..200..300..400..500..600..done
Runtime (without IO) in cpu-seconds: 0.00
Accuracy on test set: 97.67% (586 correct, 14 incorrect, 600 total)
Precision/recall on test set: 96.43%/99.00%
```

Prediction using RBF kernel:

```
C:\Users\Phearin\Documents\Assignment\Practice of Information System\Information System Analysis\Data Mining 3>svm_learn -t 2 train.dat rbf_model
Scanning examples...done
Reading examples into memory...100..200..300..400..500..600..700..800..900..1000..1100..1200..1300..1400..1500..1600..1700..1800..1900..2000..OK. (2000 examples read)
Setting default regularization parameter C=0.7910
Optimizing..........................................................................................................................................
.......................................................................................................................................................
...........................................................................................
 Checking optimality of inactive variables...done.
 Number of inactive variables = 773
done. (504 iterations)
Optimization finished (1 misclassified, maxdiff=0.00094).
Runtime in cpu-seconds: 0.85
Number of SV: 1434 (including 257 at upper bound)
L1 loss: loss=58.57970
Norm of weight vector: |w|=24.67613
Norm of longest example vector: |x|=1.00000
Estimated VCdim of classifier: VCdim<=770.81075
Computing XiAlpha-estimates...done
Runtime for XiAlpha-estimates in cpu-seconds: 0.00
XiAlpha-estimate of the error: error<=12.85% (rho=1.00,depth=0)
XiAlpha-estimate of the recall: recall=>92.00% (rho=1.00,depth=0)
XiAlpha-estimate of the precision: precision=>83.87% (rho=1.00,depth=0)
Number of kernel evaluations: 1902541
Writing model file...done

C:\Users\Phearin\Documents\Assignment\Practice of Information System\Information System Analysis\Data Mining 3>svm_classify test.dat rbf_model rbf_predict
Reading model...OK. (1434 support vectors read)
Classifying test examples..100..200..300..400..500..600..done
Runtime (without IO) in cpu-seconds: 0.27
Accuracy on test set: 97.33% (584 correct, 16 incorrect, 600 total)
Precision/recall on test set: 95.81%/99.00%
```

Prediction using polynomial kernel:

```
C:\Users\Phearin\Documents\Assignment\Practice of Information System\Information System Analysis\Data Mining 3>svm_learn -t 1 train.dat poly_model
Scanning examples...done
Reading examples into memory...100..200..300..400..500..600..700..800..900..1000..1100..1200..1300..1400..1500..1600..1700..1800..1900..2000..OK. (2000 examples read)
Setting default regularization parameter C=0.1429
Optimizing.........................................................................................................................................................
...................................................................................................................................................................
..............................................................................
 Checking optimality of inactive variables...done.
 Number of inactive variables = 784
done. (493 iterations)
Optimization finished (1 misclassified, maxdiff=0.00099).
Runtime in cpu-seconds: 0.76
Number of SV: 1251 (including 88 at upper bound)
L1 loss: loss=19.30567
Norm of weight vector: |w|=8.66781
Norm of longest example vector: |x|=2.82843
Estimated VCdim of classifier: VCdim<=526.91660
Computing XiAlpha-estimates...done
Runtime for XiAlpha-estimates in cpu-seconds: 0.00
XiAlpha-estimate of the error: error<=4.40% (rho=1.00,depth=0)
XiAlpha-estimate of the recall: recall=>97.10% (rho=1.00,depth=0)
XiAlpha-estimate of the precision: precision=>94.27% (rho=1.00,depth=0)
Number of kernel evaluations: 1787185
Writing model file...done

C:\Users\Phearin\Documents\Assignment\Practice of Information System\Information System Analysis\Data Mining 3>svm_classify test.dat poly_model poly_predict
Reading model...OK. (1251 support vectors read)
Classifying test examples..100..200..300..400..500..600..done
Runtime (without IO) in cpu-seconds: 0.26
Accuracy on test set: 97.33% (584 correct, 16 incorrect, 600 total)
Precision/recall on test set: 95.81%/99.00%
```

Prediction using soft-margin parameter = 10000:

```
C:\Users\Phearin\Documents\Assignment\Practice of Information System\Information System Analysis\Data Mining 3>svm_learn -c 10000 train.dat 10000c_model
Scanning examples...done
Reading examples into memory...100..200..300..400..500..600..700..800..900..1000..1100..1200..1300..1400..1500..1600..1700..1800..1900..2000..OK. (2000 examples read)
Optimizing.........................................................................................................................................................
...................................................................................................................................................................
..............................................................................................done. (509 iterations)
Optimization finished (0 misclassified, maxdiff=0.00096).
Runtime in cpu-seconds: 0.08
Number of SV: 848 (including 0 at upper bound)
L1 loss: loss=0.00000
Norm of weight vector: |w|=22.17302
Norm of longest example vector: |x|=1.00000
Estimated VCdim of classifier: VCdim<=492.64285
Computing XiAlpha-estimates...done
Runtime for XiAlpha-estimates in cpu-seconds: 0.00
XiAlpha-estimate of the error: error<=6.85% (rho=1.00,depth=0)
XiAlpha-estimate of the recall: recall=>93.60% (rho=1.00,depth=0)
XiAlpha-estimate of the precision: precision=>92.77% (rho=1.00,depth=0)
Number of kernel evaluations: 50484
Writing model file...done

C:\Users\Phearin\Documents\Assignment\Practice of Information System\Information System Analysis\Data Mining 3>svm_classify test.dat 10000c_model 10000c_predict
Reading model...OK. (848 support vectors read)
Classifying test examples..100..200..300..400..500..600..done
Runtime (without IO) in cpu-seconds: 0.02
Accuracy on test set: 97.33% (584 correct, 16 incorrect, 600 total)
Precision/recall on test set: 95.81%/99.00%
```

Prediction using soft-margin parameter = 100:

```
C:\Users\Phearin\Documents\Assignment\Practice of Information System\Information System Analysis\Data Mining 3>svm_learn -c 100 train.dat 100c_model
Scanning examples...done
Reading examples into memory...100..200..300..400..500..600..700..800..900..1000..1100..1200..1300..1400..1500..1600..1700..1800..1900..2000..OK. (2000 examples read)
Optimizing.........................................................................................................................................................
...................................................................................................................................................................
..............................................................................................done. (509 iterations)
Optimization finished (0 misclassified, maxdiff=0.00096).
Runtime in cpu-seconds: 0.07
Number of SV: 848 (including 0 at upper bound)
L1 loss: loss=0.00000
Norm of weight vector: |w|=22.17302
Norm of longest example vector: |x|=1.00000
Estimated VCdim of classifier: VCdim<=492.64285
Computing XiAlpha-estimates...done
Runtime for XiAlpha-estimates in cpu-seconds: 0.00
XiAlpha-estimate of the error: error<=6.85% (rho=1.00,depth=0)
XiAlpha-estimate of the recall: recall=>93.60% (rho=1.00,depth=0)
XiAlpha-estimate of the precision: precision=>92.77% (rho=1.00,depth=0)
Number of kernel evaluations: 50484
Writing model file...done

C:\Users\Phearin\Documents\Assignment\Practice of Information System\Information System Analysis\Data Mining 3>svm_classify test.dat 100c_model 100c_predict
Reading model...OK. (848 support vectors read)
Classifying test examples..100..200..300..400..500..600..done
Runtime (without IO) in cpu-seconds: 0.00
Accuracy on test set: 97.33% (584 correct, 16 incorrect, 600 total)
Precision/recall on test set: 95.81%/99.00%
```

Result comparison

	Runtime in cpu-seconds	Number of SV	L1 loss	Training error	Training recall	Training precision	Testing accuracy	Testing precision/recall
Default	0.05	878	35.67674	5.85%	95.40%	93.07%	97.67%	96.43%/99.00%
RBF	0.85	1434	58.57970	12.85%	92.00%	83.87%	97.33%	95.81%/99.00%
Polynomial	0.76	1251	19.30567	4.40%	97.10%	94.27%	97.33%	95.81%/99.00%
c = 10000	0.08	848	0.00000	6.85%	93.60%	92.77%	97.33%	95.81%/99.00%
c = 100	0.07	848	0.00000	6.85%	93.60%	92.77%	97.33%	95.81%/99.00%

According to the results above, we can conclude that by using default prediction, we can generalize the best on the testing data since we receive highest testing accuracy, precision and recall. Moreover, it also has fastest learning time among other models with only 0.05 cpu-seconds. We can also see that c=10000 and c=100 models overfit training data (L1 loss = 0) but yield no better result when the models are used in training set.

2. Explain your original data in detail.

For this assignment, I used partial FIFA 18 soccer players' dataset as features for each player. Then I labelled each player with +1 or -1 as whether I like the player or not respectively. I used 14 attributes as player features including:

- Age
- Overall performance
- Potential
- Acceleration
- Aggression
- Agility
- Balance
- Ball control
- Composure
- Crossing
- Curve
- Dribbling
- Finishing
- Free kick accuracy

Each attribute has range from 0 – 100. 300 instances (players) were extracted from the dataset. After labelling preference class to each player (+1 or -1), I split the data into training sets (80% → 240 instances) and testing sets (20% → 60 instances). Finally, the data were formatted appropriately and trained using SVM-light.

3. Analyze the data and discuss the result.

Default prediction (default regularization parameter c = 0.0):

```
C:\Users\Phearin\Documents\Assignment\Practice of Information System\Information System Analysis\Data Mining 3>svm_learn train.dat def_model
Scanning examples...done
Reading examples into memory...100..200..OK. (240 examples read)
Setting default regularization parameter C=0.0000
Optimizing.....................................done. (40 iterations)
Optimization finished (33 misclassified, maxdiff=0.00074).
Runtime in cpu-seconds: 0.00
Number of SV: 71 (including 62 at upper bound)
L1 loss: loss=65.41548
Norm of weight vector: |w|=0.00293
Norm of longest example vector: |x|=323.59852
Estimated VCdim of classifier: VCdim<=1.89989
Computing XiAlpha-estimates...done
Runtime for XiAlpha-estimates in cpu-seconds: 0.01
XiAlpha-estimate of the error: error<=26.67% (rho=1.00,depth=0)
XiAlpha-estimate of the recall: recall=>0.00% (rho=1.00,depth=0)
XiAlpha-estimate of the precision: precision=>0.00% (rho=1.00,depth=0)
Number of kernel evaluations: 4758
Writing model file...done

C:\Users\Phearin\Documents\Assignment\Practice of Information System\Information System Analysis\Data Mining 3>svm_classify test.dat def_model def_predict
Reading model...OK. (71 support vectors read)
Classifying test examples..done
Runtime (without IO) in cpu-seconds: 0.00
Accuracy on test set: 83.33% (50 correct, 10 incorrect, 60 total)
Precision/recall on test set: -1.#J%/0.00%
```

Prediction using RBF kernel:

```
C:\Users\Phearin\Documents\Assignment\Practice of Information System\Information System Analysis\Data Mining 3>svm_learn -t 2 train.dat rbf_model
Scanning examples...done
Reading examples into memory...100..200..OK. (240 examples read)
Setting default regularization parameter C=0.5000
Optimizing...........................................................................................................done. (106 iterations)
Optimization finished (33 misclassified, maxdiff=0.00090).
Runtime in cpu-seconds: 0.01
Number of SV: 240 (including 33 at upper bound)
L1 loss: loss=46.86134
Norm of weight vector: |w|=3.09278
Norm of longest example vector: |x|=1.00000
Estimated VCdim of classifier: VCdim<=20.13053
Computing XiAlpha-estimates...done
Runtime for XiAlpha-estimates in cpu-seconds: 0.00
XiAlpha-estimate of the error: error<=13.75% (rho=1.00,depth=0)
XiAlpha-estimate of the recall: recall=>0.00% (rho=1.00,depth=0)
XiAlpha-estimate of the precision: precision=>-1.#J% (rho=1.00,depth=0)
Number of kernel evaluations: 37815
Writing model file...done

C:\Users\Phearin\Documents\Assignment\Practice of Information System\Information System Analysis\Data Mining 3>svm_classify test.dat rbf_model rbf_predict
Reading model...OK. (240 support vectors read)
Classifying test examples..done
Runtime (without IO) in cpu-seconds: 0.00
Accuracy on test set: 83.33% (50 correct, 10 incorrect, 60 total)
Precision/recall on test set: -1.#J%/0.00%
```

Prediction using polynomial kernel:

```
C:\Users\Phearin\Documents\Assignment\Practice of Information System\Information System Analysis\Data Mining 3>svm_learn -t 1 train.dat poly_model
Scanning examples...done
Reading examples into memory...100..200..OK. (240 examples read)
Setting default regularization parameter C=0.0000
Optimizing.........................done. (25 iterations)
Optimization finished (33 misclassified, maxdiff=0.00002).
Runtime in cpu-seconds: 0.01
Number of SV: 70 (including 65 at upper bound)
L1 loss: loss=62.17509
Norm of weight vector: |w|=0.00000
Norm of longest example vector: |x|=33886427.67200
Estimated VCdim of classifier: VCdim<=14.29271
Computing XiAlpha-estimates...done
Runtime for XiAlpha-estimates in cpu-seconds: 0.00
XiAlpha-estimate of the error: error<=27.92% (rho=1.00,depth=0)
XiAlpha-estimate of the recall: recall=>0.00% (rho=1.00,depth=0)
XiAlpha-estimate of the precision: precision=>0.00% (rho=1.00,depth=0)
Number of kernel evaluations: 20289
Writing model file...done

C:\Users\Phearin\Documents\Assignment\Practice of Information System\Information System Analysis\Data Mining 3>svm_classify test.dat poly_model poly_predict
Reading model...OK. (70 support vectors read)
Classifying test examples..done
Runtime (without IO) in cpu-seconds: 0.00
Accuracy on test set: 83.33% (50 correct, 10 incorrect, 60 total)
Precision/recall on test set: -1.#J%/0.00%
```

Prediction using soft-margin parameter = 10000:

```
..........................................................................................done. (540252 iterations)
Optimization finished (0 misclassified, maxdiff=0.00097).
Runtime in cpu-seconds: 30.67
Number of SV: 15 (including 0 at upper bound)
L1 loss: loss=0.00000
Norm of weight vector: |w|=110.14572
Norm of longest example vector: |x|=323.59852
Estimated VCdim of classifier: VCdim<=1131668271.81067
Computing XiAlpha-estimates...done
Runtime for XiAlpha-estimates in cpu-seconds: 0.00
XiAlpha-estimate of the error: error<=6.25% (rho=1.00,depth=0)
XiAlpha-estimate of the recall: recall=>75.76% (rho=1.00,depth=0)
XiAlpha-estimate of the precision: precision=>78.13% (rho=1.00,depth=0)
Number of kernel evaluations: 29716250
Writing model file...done

C:\Users\Phearin\Documents\Assignment\Practice of Information System\Information System Analysis\Data Mining 3>svm_classify test.dat 10000c_model 10000c_predict
Reading model...OK. (15 support vectors read)
Classifying test examples..done
Runtime (without IO) in cpu-seconds: 0.00
Accuracy on test set: 95.00% (57 correct, 3 incorrect, 60 total)
Precision/recall on test set: 88.89%/80.00%
```

Prediction using soft-margin parameter = 100:

```
..........................................................................................done. (23096 iterations)
Optimization finished (1 misclassified, maxdiff=0.00093).
Runtime in cpu-seconds: 1.02
Number of SV: 17 (including 2 at upper bound)
L1 loss: loss=5.19300
Norm of weight vector: |w|=11.70353
Norm of longest example vector: |x|=323.59852
Estimated VCdim of classifier: VCdim<=13262928.47495
Computing XiAlpha-estimates...done
Runtime for XiAlpha-estimates in cpu-seconds: 0.00
XiAlpha-estimate of the error: error<=7.08% (rho=1.00,depth=0)
XiAlpha-estimate of the recall: recall=>69.70% (rho=1.00,depth=0)
XiAlpha-estimate of the precision: precision=>76.67% (rho=1.00,depth=0)
Number of kernel evaluations: 1272676
Writing model file...done

C:\Users\Phearin\Documents\Assignment\Practice of Information System\Information System Analysis\Data Mining 3>svm_classify test.dat 100c_model 100c_predict
Reading model...OK. (17 support vectors read)
Classifying test examples..done
Runtime (without IO) in cpu-seconds: 0.00
Accuracy on test set: 93.33% (56 correct, 4 incorrect, 60 total)
Precision/recall on test set: 87.50%/70.00%
```

Result comparison

	Runtime in cpu-seconds	Number of SV	L1 loss	Training error	Training recall	Training precision	Testing accuracy	Testing precision/recall
Default	0.00	71	65.41548	26.67%	0.00%	0.00%	83.33%	-1.#J%/0.00%
RBF	0.01	240	46.86134	13.75%	0.00%	-1.#J%	83.33%	-1.#J%/0.00%
Polynomial	0.01	70	62.17509	27.92%	0.00%	0.00%	83.33%	-1.#J%/0.00%
c = 10000	30.67	15	0.00000	6.25%	75.76%	78.13%	95.00%	88.89%/80.00%
c = 100	1.02	17	5.19300	7.08%	69.70%	76.67%	93.33%	87.50%/70.00%

According to the result above, we can conclude that using c = 100 is a reasonable choice for SVM model. Although the performance is slightly lower than that of c = 10000 model, but it is 30 times faster than the c = 10000 model. Other models have quite low performances despite having quick learning speed.