Introduction to Support Vector Machines

Eric Nowak

Lear

June 14, 2005

Contents

1 Introduction

SVMs are

- binary
- discriminative
- linear and non-linear
- maximum margin

classifiers

1 Introduction

SVMs are

- binary
- discriminative
- linear and non-linear
- maximum margin

classifiers

- extended to multiclass
- requiring parameters

1.1 Discriminative classifier , SEPARATE 2 classes

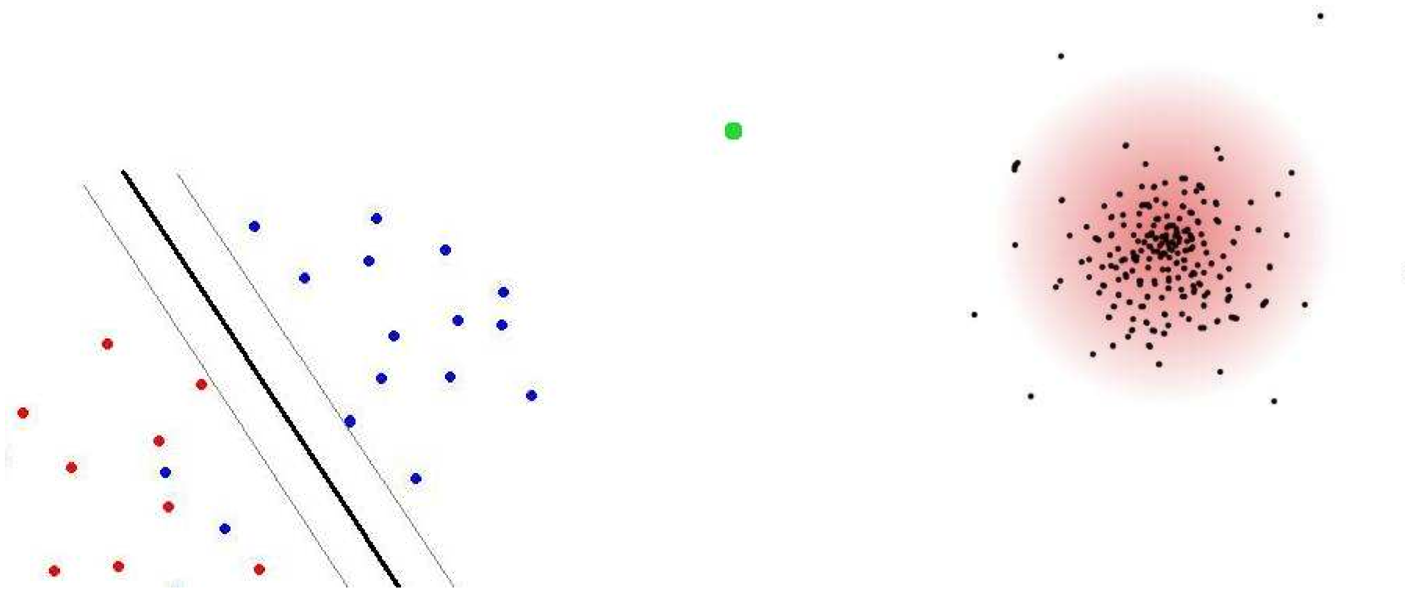
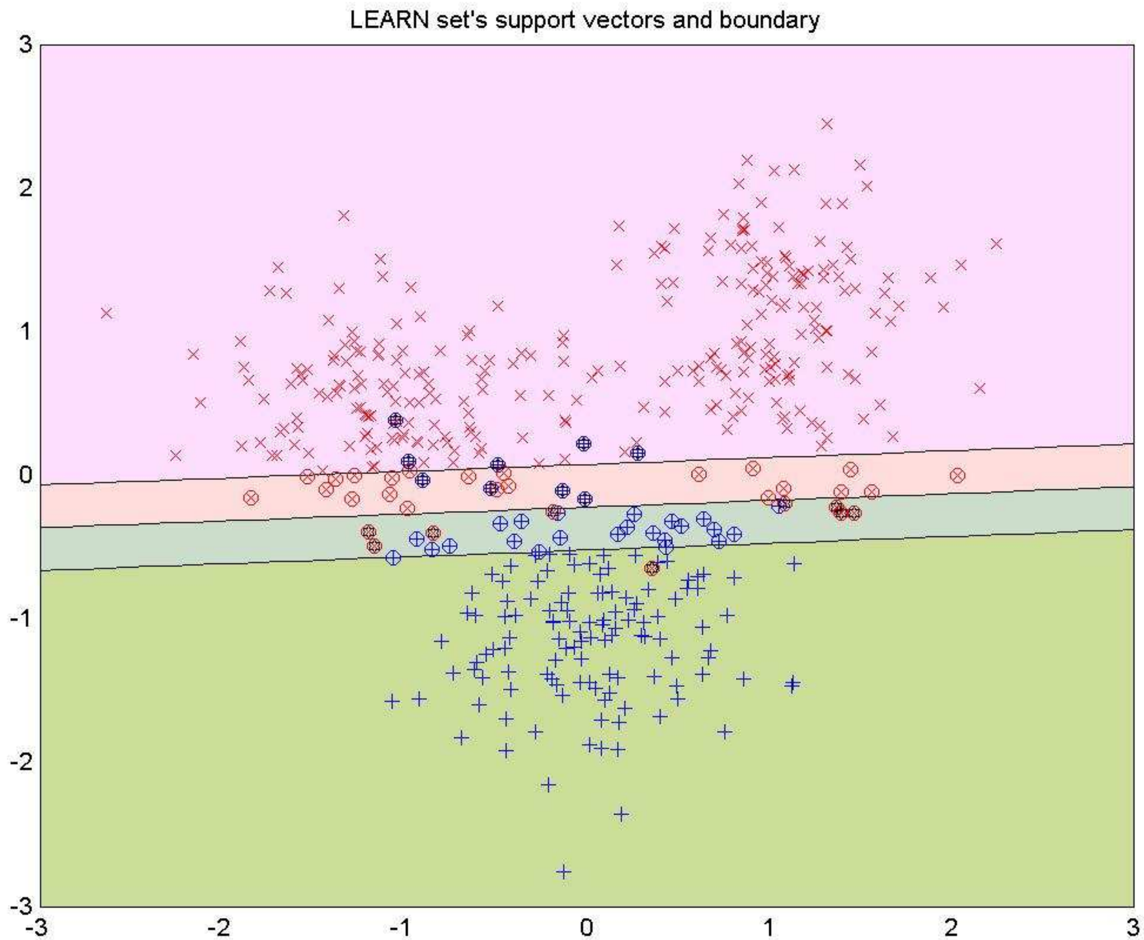


Table 1: Disciminative (left) vs Generative (right) classifiers

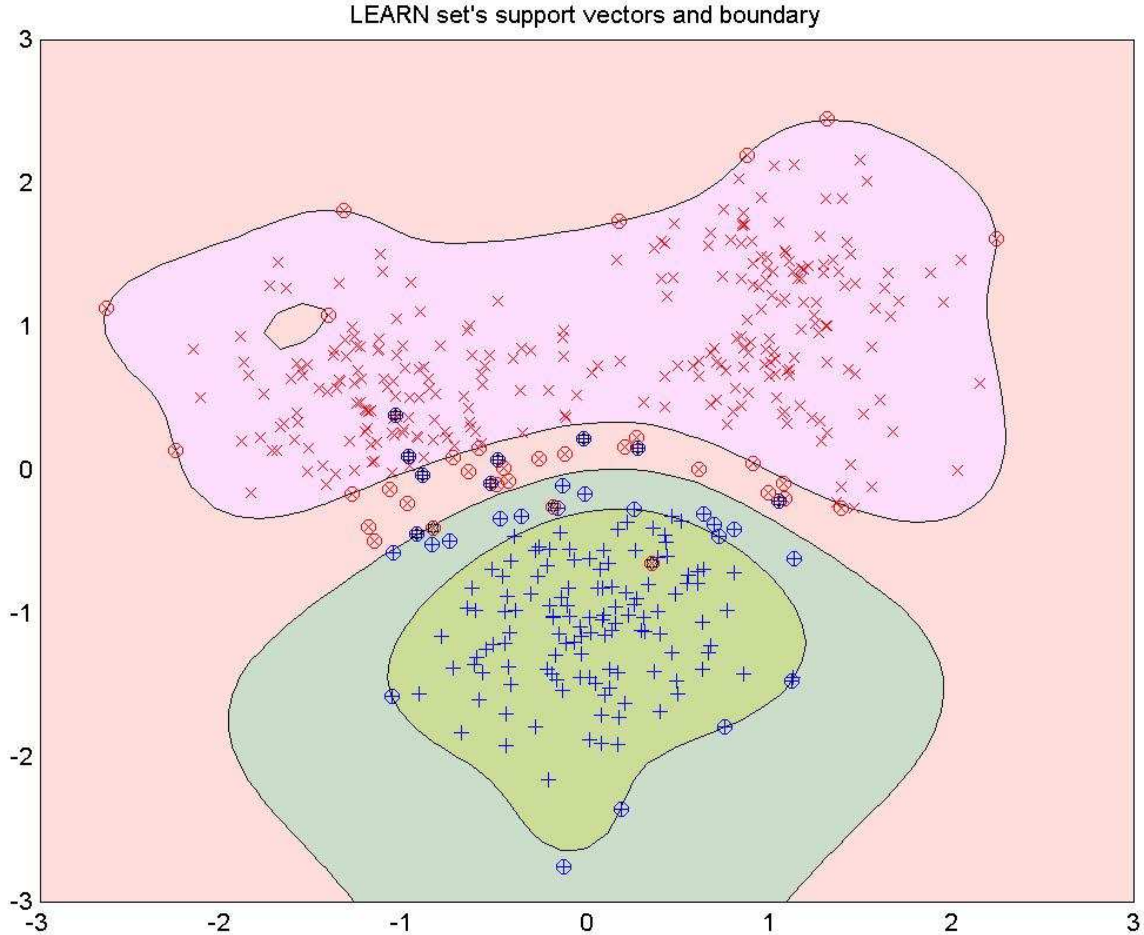
$\Rightarrow$ SVMs answer to *is it A or B ?* but not to *is it A ?*

i.e. set label A or B to all points of the space

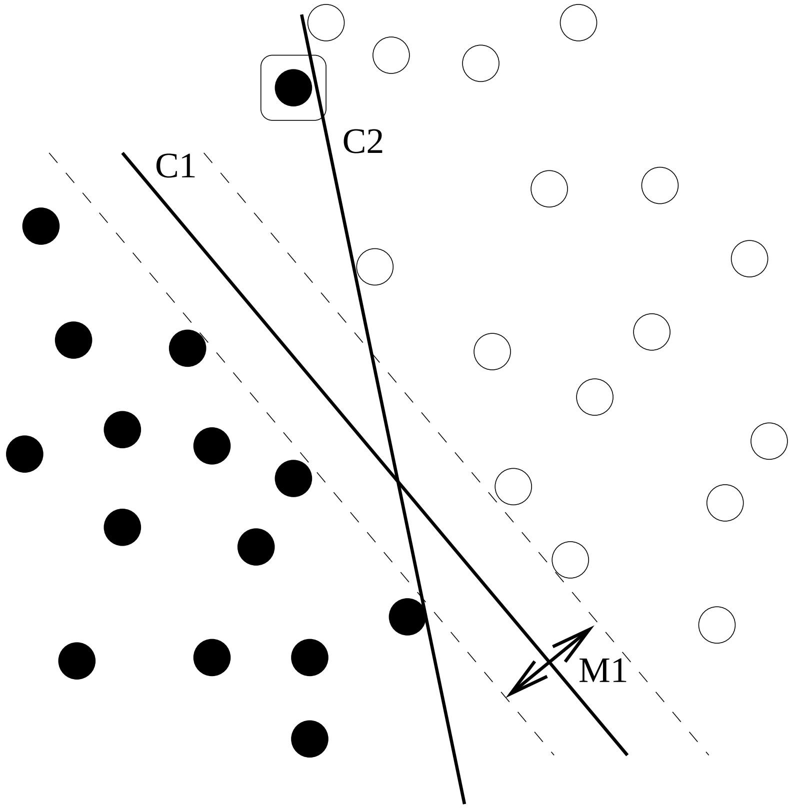
1.2 Linear and Non Linear boundary



1.3 Linear and Non Linear boundary



1.4 Maximum margin classifier

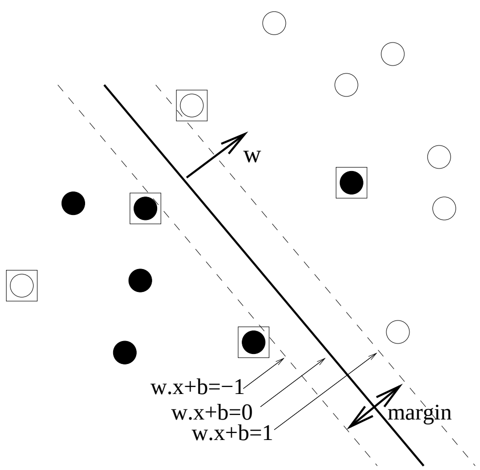


Trade-off Training set performance / Margin

Margin $\Rightarrow$ Generalization

2 Theory

2.1 constraints



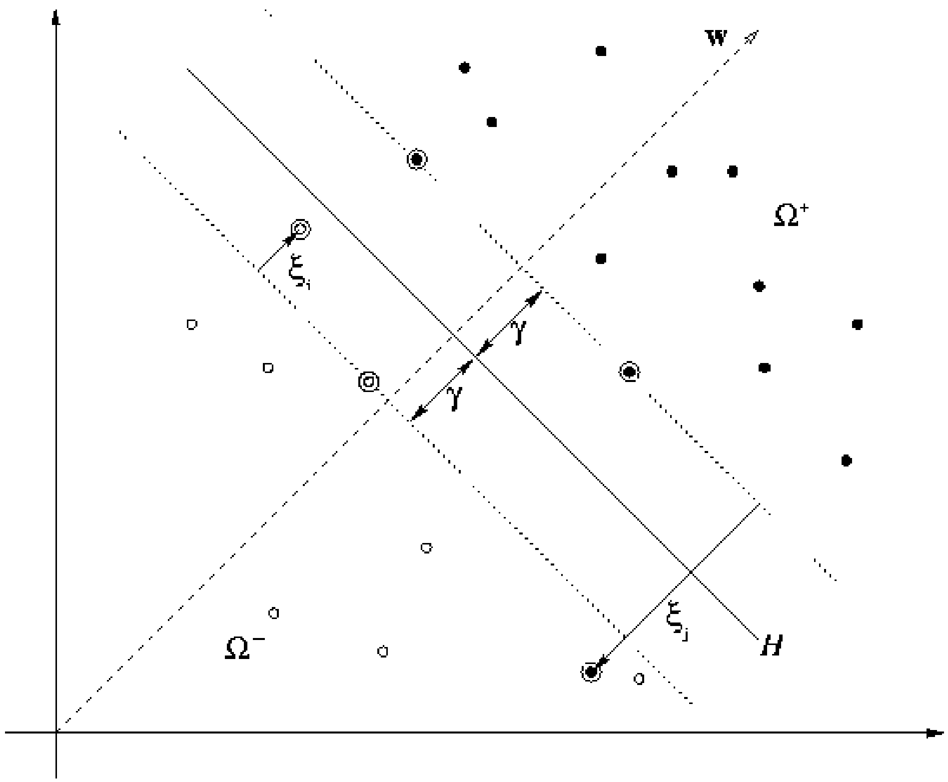
- $\mathbf{x}_i$: ith sample , $y_i = \{0, 1\}$: ith label
- $g(x) = \mathbf{w}\mathbf{x} + b = \sum w_j x_j + b$
- $\frac{1}{\|\mathbf{w}\|}$ = margin : maximized
- $\forall i \; y_i(\mathbf{w}\mathbf{x}_i + b) \geq 1$

2.2 optimization

- $g(x) = \mathbf{w}\mathbf{x} + b = \sum w_j x_j + b$
- Lagrange :
  - $\mathbf{w} = \sum_i \alpha_i y_i \mathbf{x}_i$
  - $L_D = \sum_i \alpha_i - 0.5 \sum \alpha_i \alpha_j y_i y_j \mathbf{x}_i \mathbf{x}_j$

  Conclusion :
  - quadratic optimization problem : find the α_i
  - problem defined in the $\mathbf{x}_i \mathbf{x}_j$

2.3 non separable data



- $\forall i \; y_i(\mathbf{w}\mathbf{x}_i + b) \geq 1 - \xi_i \;,\; \xi_i \geq 0$
- minimize $\|w\| + C \sum_i \xi_i$
- C is a parameter $\Rightarrow$ 1D grid search

2.4 non separable data

minimize $\|w\| + C\sum_i \xi_i$

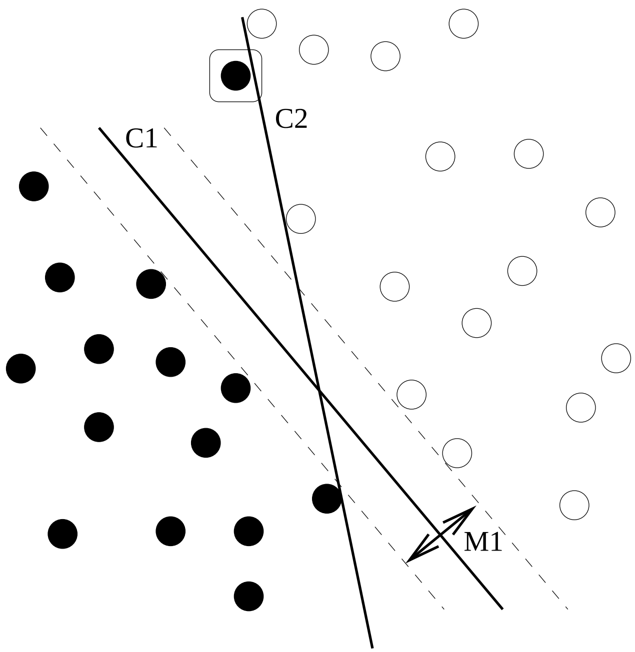


Figure 1: Classifier 1 : C=0 - Classifier 2 : C=inf

2.5 non linear svm

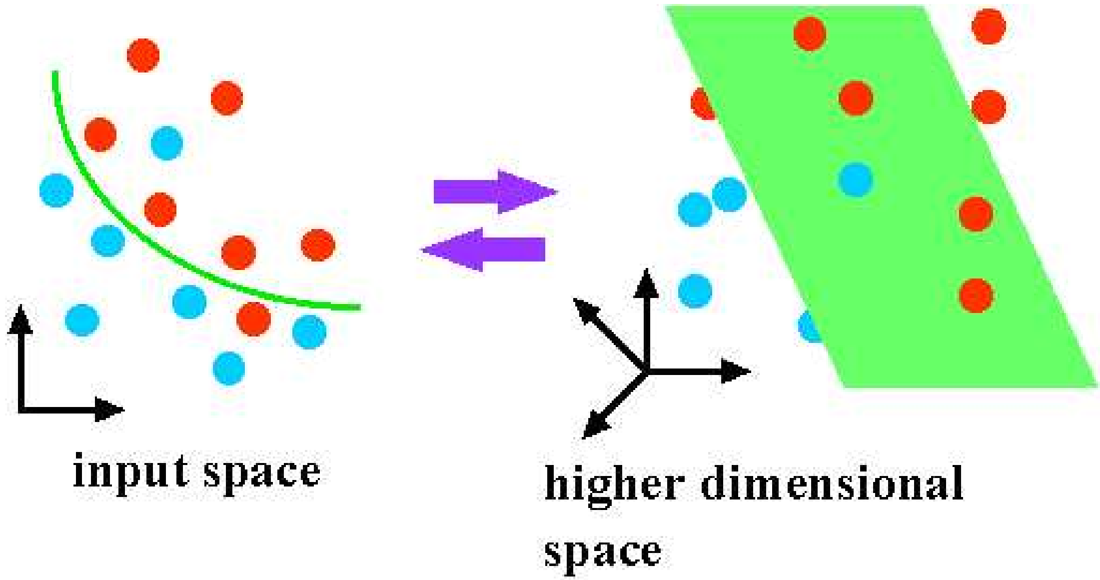


Figure 2: In a higher dimensional space , it is easier to separate the classes

2.6 non linear svm

- $L_D = \sum_i \alpha_i - 0.5 \sum \alpha_i \alpha_j y_i y_j \mathbf{x}_i \mathbf{x}_j$
- Throw the x_i into a higher dimensional space with $\Phi : E^n \to E^N,\ N >> n$
- $L_D \to \Phi(\mathbf{x}_i)\Phi(\mathbf{x}_j) \to K(\mathbf{x}_i, \mathbf{x}_j)$
- Kernels : no need to know Φ, K is enough
  - Linear : $K(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i \mathbf{x}_j$
  - Polynomial: $K(\mathbf{x}_i, \mathbf{x}_j) = (1 + \mathbf{x}_i \mathbf{x}_j)^d$
  - RBF (+) : $K(\mathbf{x}_i, \mathbf{x}_j) = e^{-\frac{\|x-y\|^2}{\sigma^2}}$

  Parameters : d or σ AND $C \Rightarrow$ 2D grid search

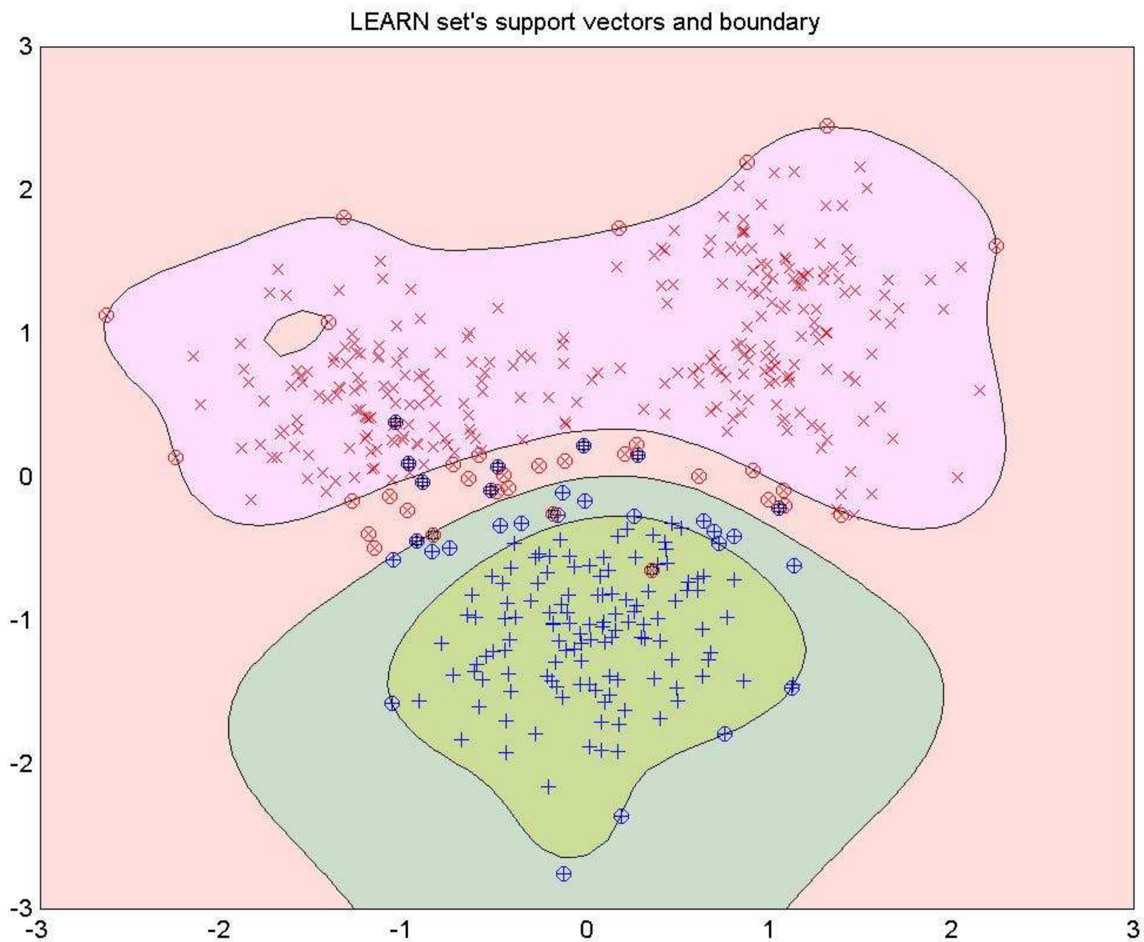


Figure 3: RBF Kernel : medium sigma

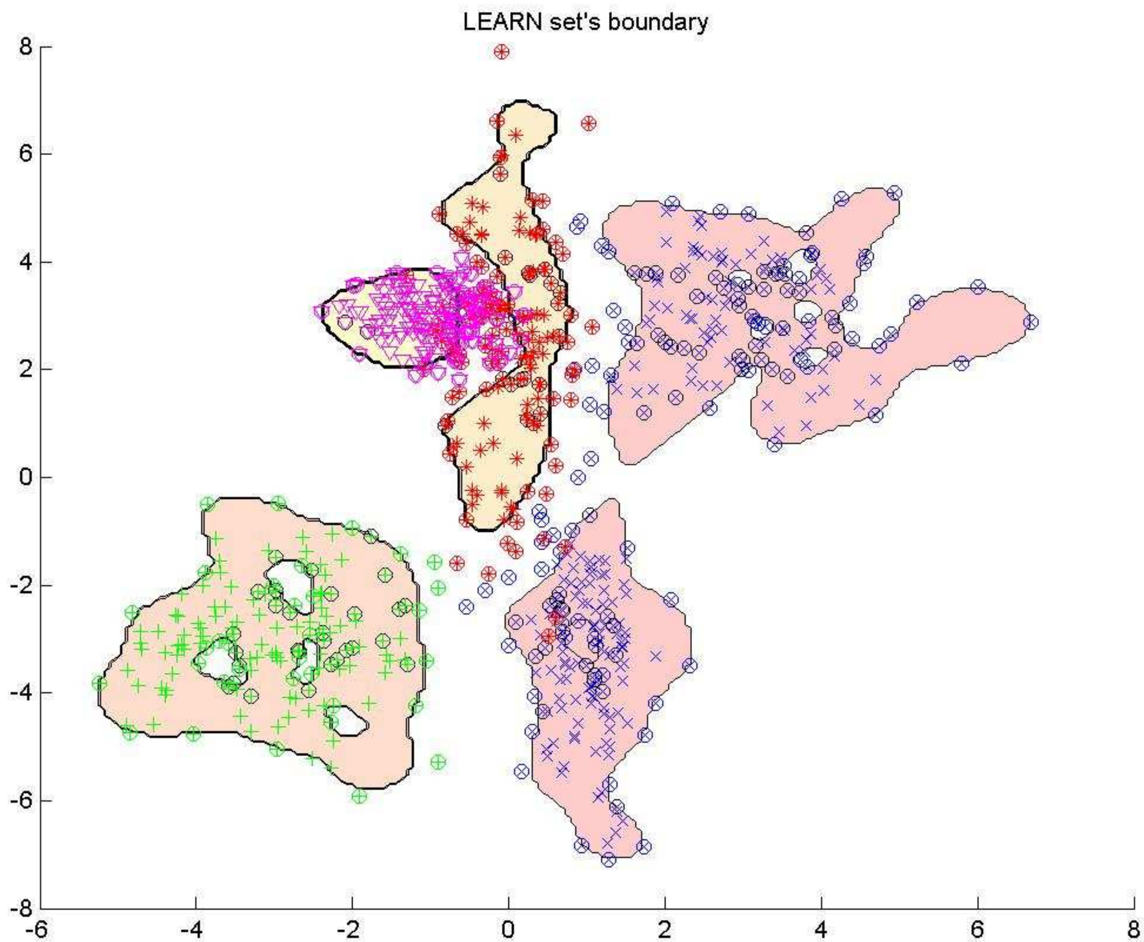


Figure 4: RBF Kernel : small sigma

2.7 multiclass SVMs

round	1-2	1-3	1-4	1-5	2-3	2-4	2-5	3-4	3-5	4-5
predict.	1	1	2	1	2	4	5	3	3	5

Table 2: One Versus One (1vs1) approach : max nb of votes - Cx(C-1)/2 classifiers - Real lab=1 - Predicted lab=1

2.7 multiclass SVMs

round	1-2	1-3	1-4	1-5	2-3	2-4	2-5	3-4	3-5	4-5
predict.	1	1	2	1	2	4	5	3	3	5

Table 2: One Versus One (1vs1) approach : max nb of votes - Cx(C-1)/2 classifiers - Real lab=1 - Predicted lab=1

round	1-r	2-r	3-r	4-r	5-r
predict.	2.3	-0.8	-10	1.2	-1.5

Table 3: One versus Rest (1vsR) approach : max prediction value - C classifiers - Real lab=1 - Predicted lab=1

1vs1 better : redundancy, specialization

3 In practice

3.1 Papers

- Vapnik 95
- C. J. C. Burges. A Tutorial on Support Vector Machines for Pattern Recognition. Knowledge Discovery and Data Mining, 2(2), 1998.

3.2 Softwares

http://kernel-machines.org/software.html

- C++ : SVMLight (interface - ; features +), LibSvm (contrary)
- Matlab : OsuSvm

3.3 Action !

```
// fL, fT     : f.(i,j)= feat val j of object i
// labL, labT : l.(i)  = label of object i

ScaleInfo si        = fL.scaleEachDim(0,1);
fT.scaleEachDim(si);

ParamLin param(c);
//ParamRBF param(c,g);

Model model         = param.train(fL,labL);
pred                = model.predict(fT);
evalAccuracy(labT, pred);
```

4 Conclusion

Discriminative

4 Conclusion

Discriminative

Linear , non Linear : Occam's Razor

4 Conclusion

Discriminative

Linear , non Linear : Occam's Razor

Mutliclass strategy

4 Conclusion

Discriminative

Linear , non Linear : Occam's Razor

Mutliclass strategy

Margin parameter : C

4 Conclusion

Discriminative

Linear , non Linear : Occam's Razor

Mutliclass strategy

Margin parameter : C

RBF Kernel parameter : σ

4 Conclusion

Discriminative

Linear , non Linear : Occam's Razor

Mutliclass strategy

Margin parameter : C

RBF Kernel parameter : σ

Parameter(s) : grid search

4 Conclusion

Discriminative

Linear , non Linear : Occam's Razor

Mutliclass strategy

Margin parameter : C

RBF Kernel parameter : σ

Parameter(s) : grid search

Tricks

- Feature matrix normalization
- Never eval perf on training set
- ...