SVM: Algorithms of Choice for Challenging Data

Boriana Milenova, Joseph Yarmus, Marcos Campos
Data Mining Technologies
ORACLE Corp.

Overview

☒SVM theoretical framework

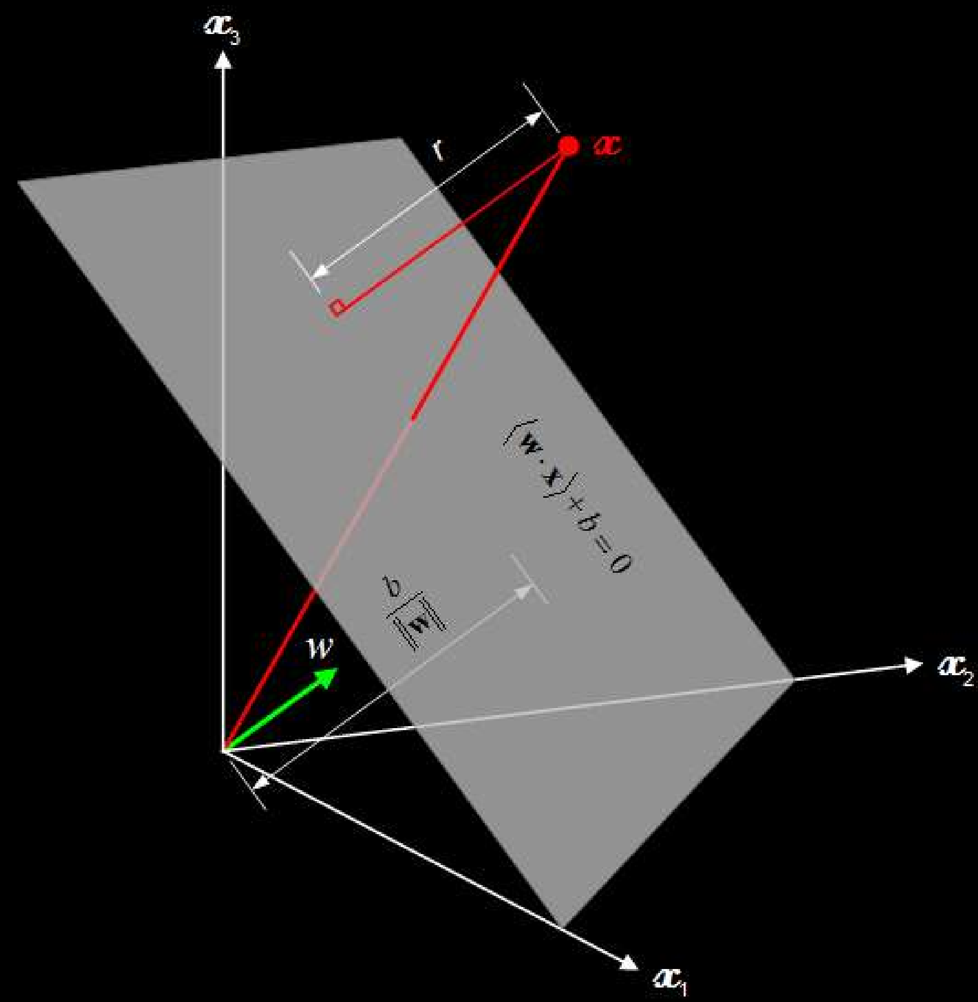
☒ ORACLE data mining technology

- SVM parameter estimation
- SVM optimization strategy

☒SVM on challenging data

SVM Model Defines a Hyperplane

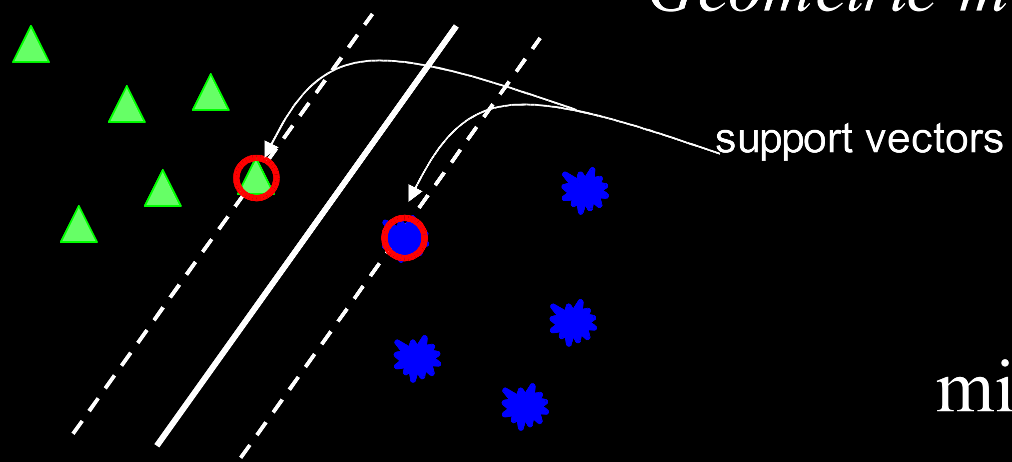
- ☒Linear models in feature space
- ☒Hyperplane defined by a set of coefficients and a bias term



Maximum Margin Models

$$Functional\ margin = \min(y_i f(x_i))$$

$$Geometric\ margin = \min\left(\frac{y_i f(x_i)}{\|\mathbf{w}\|}\right) = \frac{1}{\|\mathbf{w}\|}$$



$$\min\|\mathbf{w}\| \Rightarrow \max(margin)$$

SVM Optimization Problem

Minimize ||**w**|| subject to $y_i f(x_i) \geq 1$

Lagrangian in primal space:

$$L_p(\mathbf{w}) = \frac{1}{2}\langle \mathbf{w} \cdot \mathbf{w} \rangle - \sum \alpha_i \left[y_i \left(\langle \mathbf{w} \cdot \mathbf{x}_i \rangle + b \right) - 1 \right]$$

subject to $\alpha_i \geq 0$

$$\frac{\partial L_p}{\partial \mathbf{w}} = 0 \qquad \mathbf{w} = \sum \alpha_i y_i \mathbf{x}_i$$

$$\frac{\partial L_p}{\partial b} = 0 \qquad \sum \alpha_i y_i = 0$$

Duality

Lagrangian in dual space:

$$L_D = \sum \alpha_i - \frac{1}{2} \sum \alpha_i \alpha_j y_i y_j \left\langle \mathbf{x}_i \cdot \mathbf{x}_j \right\rangle$$

subject to $\alpha_i \geq 0 \quad \sum \alpha_i y_i = 0$

Dot products!

- dimension-insensitive optimization
- generalized dot products via non-linear map ϕ

$$K(\mathbf{x}_i, \mathbf{x}_j) = \left\langle \phi(\mathbf{x}_i) \cdot \phi(\mathbf{x}_j) \right\rangle$$

Towards Higher Dimensionality via Kernels

1. Transform data via non-linear mapping ϕ to an inner product feature space
2. Train a linear machine in the new feature space

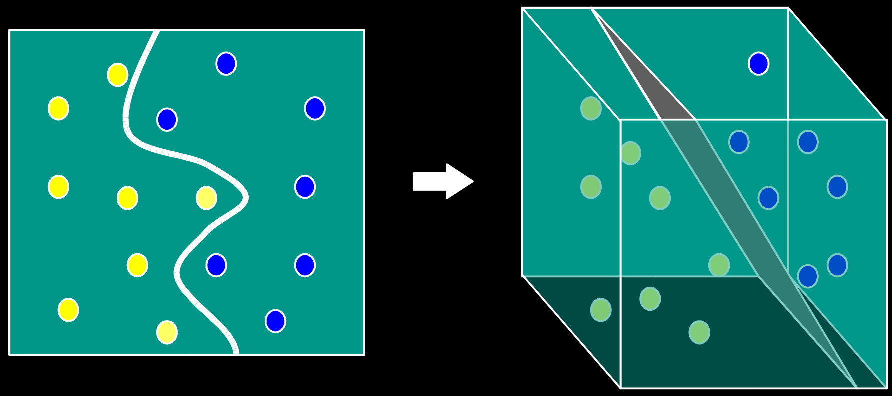
Mercer's kernels:

- symmetry

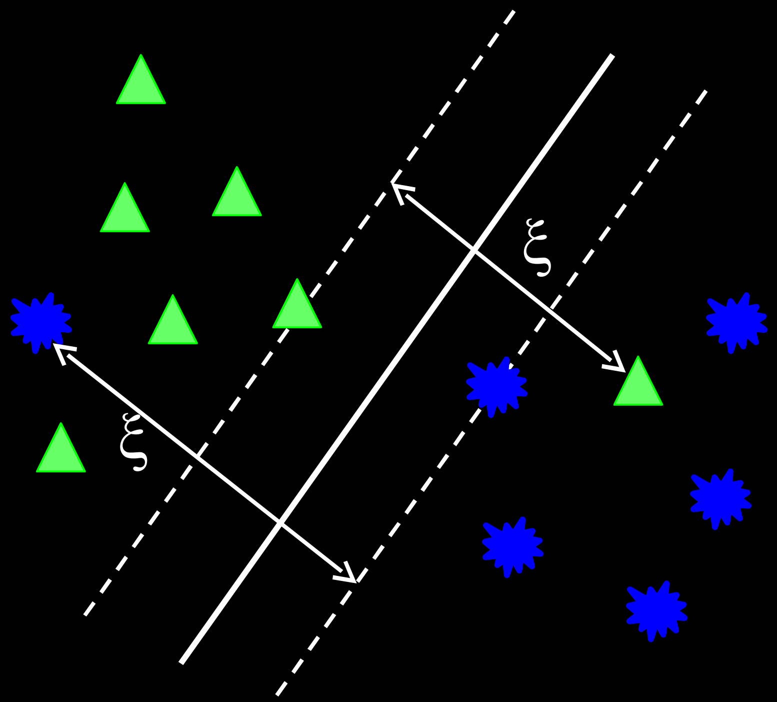
  $$K(\mathbf{x}_i, \mathbf{x}_j) = K(\mathbf{x}_j, \mathbf{x}_i)$$

- positive semi-definite kernel matrix
- reproducing property

  $$\langle K(\mathbf{x}_i,.) \cdot K(\mathbf{x}_j,.) \rangle = K(\mathbf{x}_i, \mathbf{x}_j)$$



Soft Margin: Non-Separable Data



$$L_p(\mathbf{w}) = \frac{1}{2}\langle \mathbf{w} \cdot \mathbf{w} \rangle + C\sum \xi^k$$

subject to

$$y_i\left(\langle \mathbf{w} \cdot \mathbf{x}_i \rangle + b\right) \geq 1 - \xi_i$$

Capacity parameter *C*

trades off complexity and empirical risk

1-Norm Dual Problem

Lagrangian in dual space:

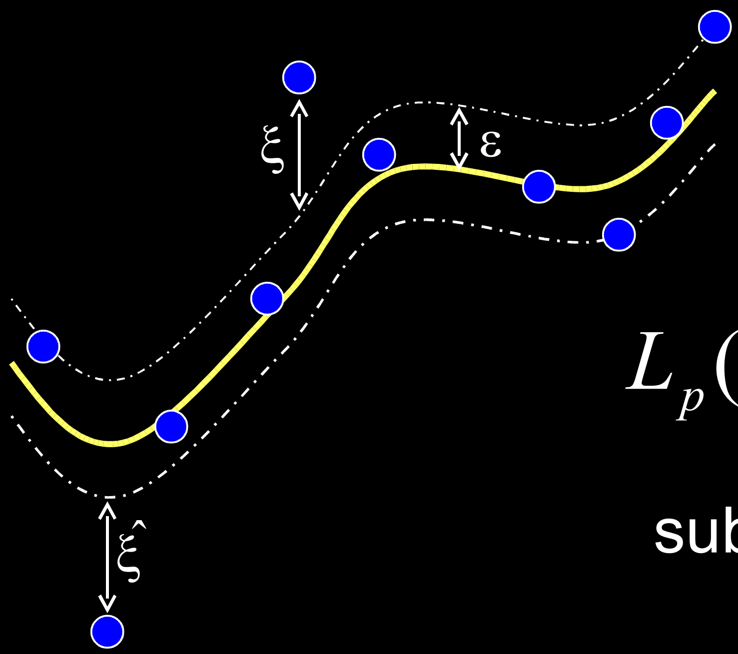
$$L_D = \sum \alpha_i - \frac{1}{2} \sum \alpha_i \alpha_j y_i y_j K(\mathbf{x}_i, \mathbf{x}_j)$$

subject to $0 \le \alpha_i \le C \quad \sum \alpha_i y_i = 0$

Quadratic problem

- linear and inequality constraints

SVM Regression



$$L_p(\mathbf{w}) = \frac{1}{2}\langle \mathbf{w}\cdot\mathbf{w}\rangle + C\sum(\xi^k + \hat{\xi}^k)$$

subject to

$$\left(\langle \mathbf{w}\cdot\mathbf{x}_i\rangle + b\right) - y_i \le \varepsilon + \xi_i$$

$$y_i - \left(\langle \mathbf{w}\cdot\mathbf{x}_i\rangle + b\right) \le \varepsilon + \hat{\xi}_i$$

SVM Fundamental Properties

- ☒Convexity
  - single global minimum
- ☒Regularization
  - trades off structural and empirical risk to avoid overfitting
- ☒Sparse solution
  - usually only a fraction of training data become support vectors
- ☒Not probabilistic

Solvable in polynomial time…

SVM in the Database

ORACLE Data Mining (ODM)

- commercial SVM implementation in the database
- product targets application developers and data mining practitioners
- focuses on ease of use and efficiency

Challenges:

- effective and inexpensive parameter tuning
- computationally efficient SVM model optimization

SVM Out-Of-The-Box

Inexperienced users can get dramatically poor results

LIBSVM examples:

	Out-of-the-box correct rate	After tuning correct rate
Astroparticle Physics	0.67	0.97
Bioinformatics	0.57	0.79
Vehicle	0.02	0.88

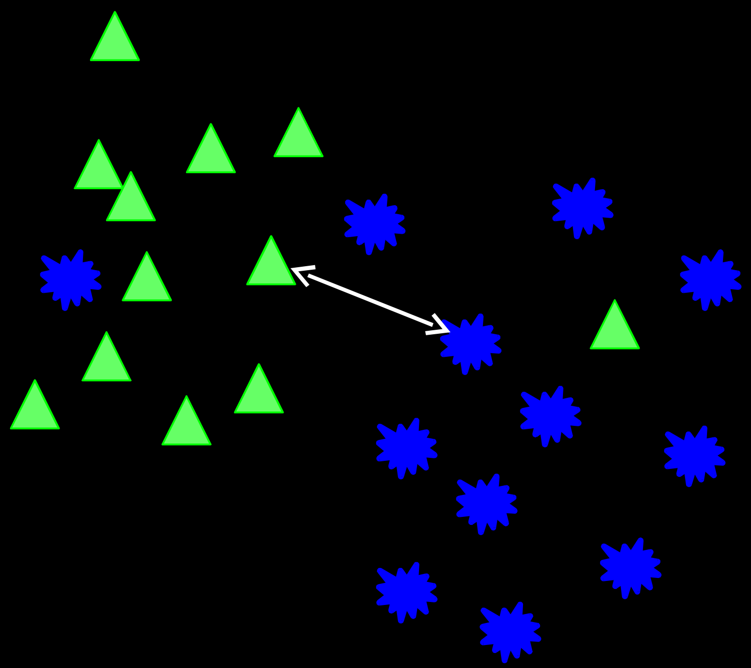
SVM Parameter Tuning

- Grid search (+ cross-validation or generalization error estimates)
  - naive
  - guided (Keerthi & Lin, 2002)
- Parameter optimization
  - gradient descent (Chapelle et al., 2000)
- Heuristics

ODM On-the-Fly Estimates

- Standard deviation for Gaussian kernel
  - single kernel parameter
  - kernel has good numeric properties
    - bounded, no overflow
- Capacity
  - key to good classification generalization
- Epsilon estimate for regression
  - key to good regression generalization

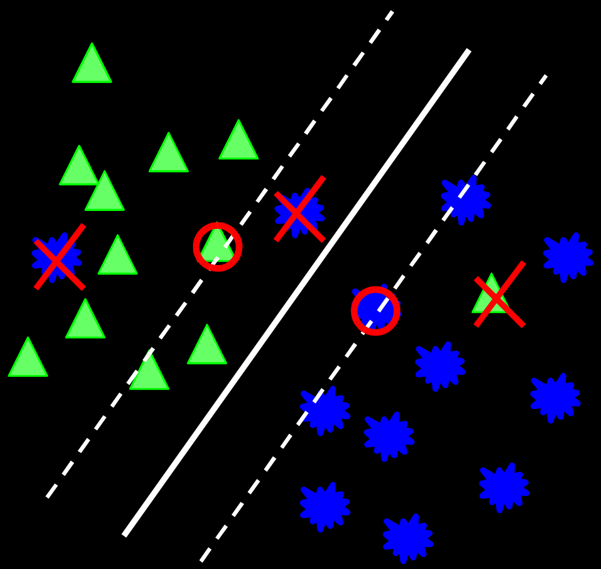
ODM Standard Deviation Estimate



Goal: Estimate distance between classes

3. Pick random pairs from opposite classes
4. Measure distances
5. Order descending
6. Exclude tail (90^{th} percentile)
7. Select minimum distance

ODM Capacity Estimate



Goal: Allocate sufficient capacity to separate typical examples

2. Pick m random examples per class
3. Compute y_i assuming α = C

$$y_i = \sum_{j=1}^{2m} C y_j K(\mathbf{x}_j, \mathbf{x}_i)$$

5. Exclude noise (incorrect sign)
6. Scale C, $y_i = \pm 1$ (non bounded sv)

$$C = y_i / \sum_{j=1}^{2m} y_j K(\mathbf{x}_j, \mathbf{x}_i)$$

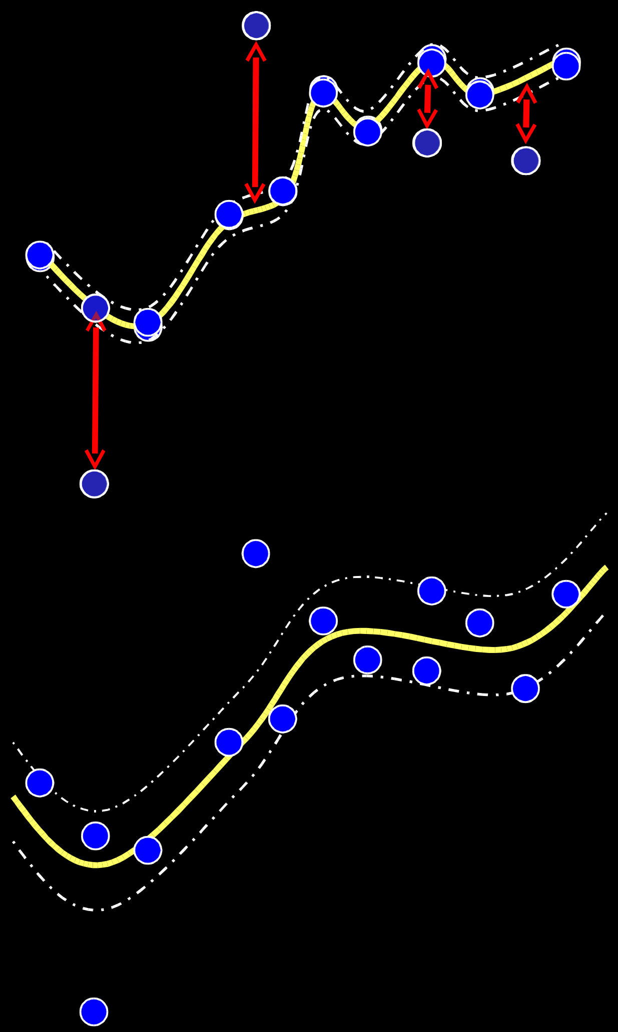
8. Order descending
9. Exclude tail (90th percentile)
10. Select minimum value

Some Comparison Numbers

LIBSVM examples:

	Out-of-the-box	On-the-fly estimates	Grid search + xval
Astroparticle Physics	0.67	0.97	0.97
Bioinformatics	0.57	0.84	0.85
Vehicle	0.02	0.71	0.88

ODM Epsilon Estimate



Goal: estimate target noise by fitting a preliminary model

3. Pick m random examples
4. Train SVM model with $\varepsilon \rightarrow 0$
5. Compute residuals on remaining data
6. Scale $\varepsilon_t = (\varepsilon_{t-1} + \sigma_n)/2$
7. Retrain

Comparison Numbers Regression

	On-the-fly estimates RMSE	Grid search RMSE
Boston housing	6.57	6.26
Computer activity	0.35	0.33
Pumadyn	0.02	0.02

Optimization Approaches

- QP solvers
  - MINOS, LOQO, quadprog (Matlab)
- Gradient descent methods
  - Sequentially update one α coefficient at a time
- Chunking and decomposition
  - optimize small “working sets” towards global solution
  - analytic solution possible (SMO - Platt, 1998)

Chunking strategy

```
/* WS working set */
select initial WS randomly;
while (violations)
{
  Solve QP on WS;
  Select new WS;
}
```

ODM Working Set Selection

- ☒Avoid oscillations
  - overlap across chunks
  - retain non-bounded support vectors
- ☒Choose among violators
  - add large violators
- ☒Computational efficiency
  - avoid sorting

Who to Retain?

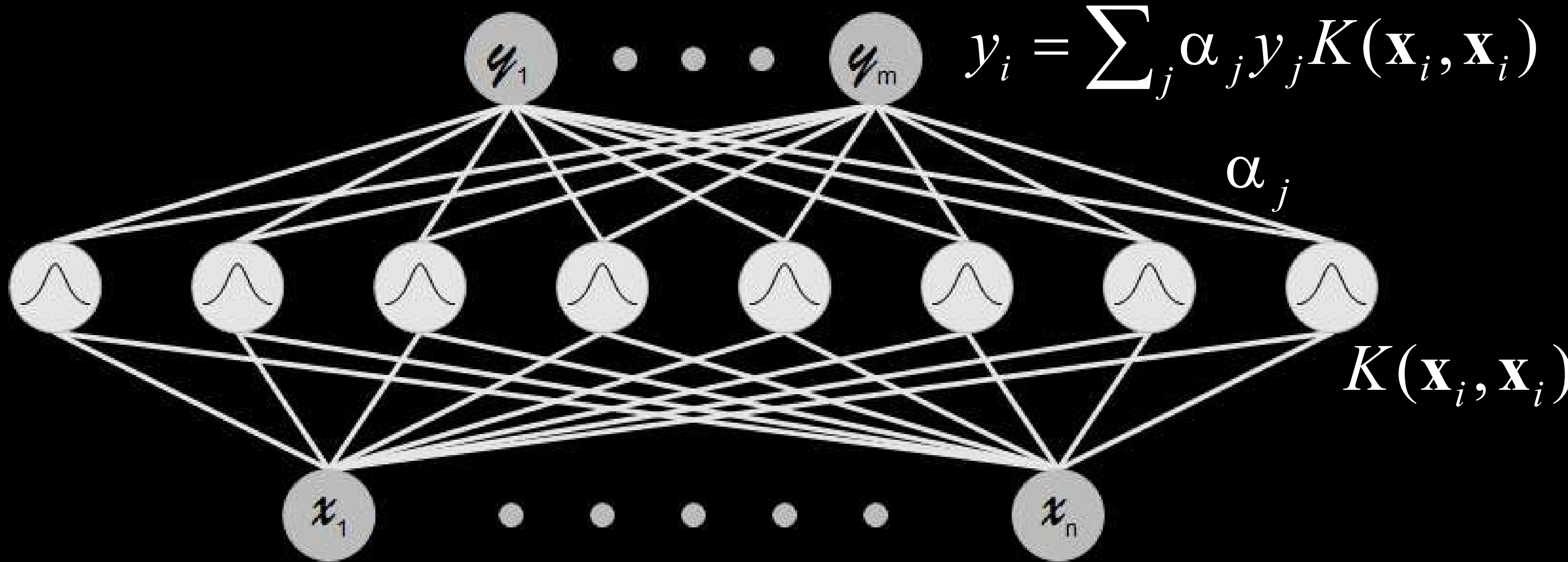
```
/* Examine previous working set */
if (non-bounded sv < 50%)
{
  retain all non-bounded sv;
  add other randomly selected up to 50%;
}
else
{
  randomly select non-bounded sv;
}
```

Who to Add?

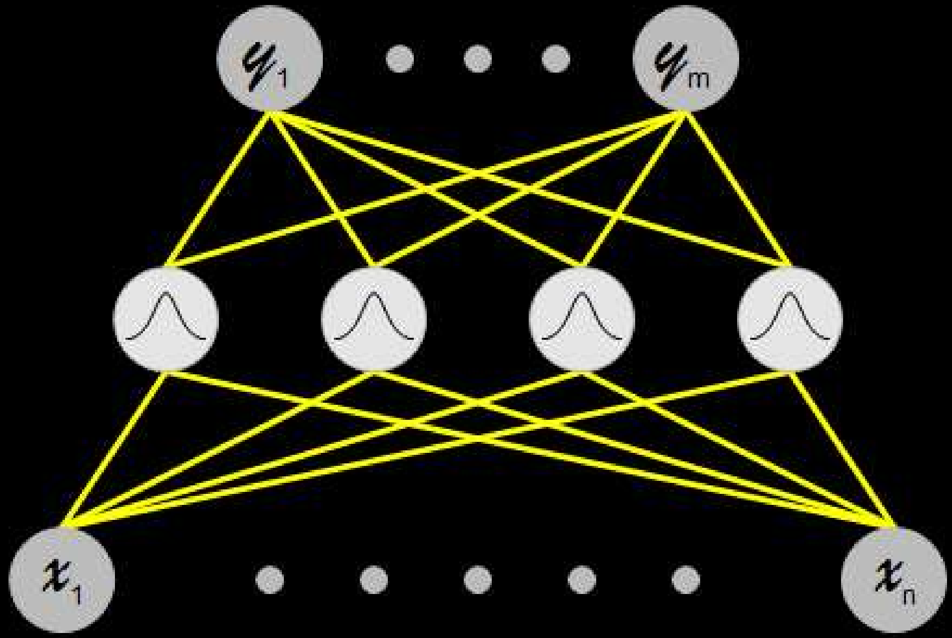
```
create violator list;
/* Scan I - pick largest violators */
while (new examples < 50% AND WS Not Full)
{
  if (violation > avg_violation)
      add to WS;
}

/* Scan II - pick other violators */
while (new examples < 50% AND WS Not Full)
{
      add randomly selected violators to WS;
}
```

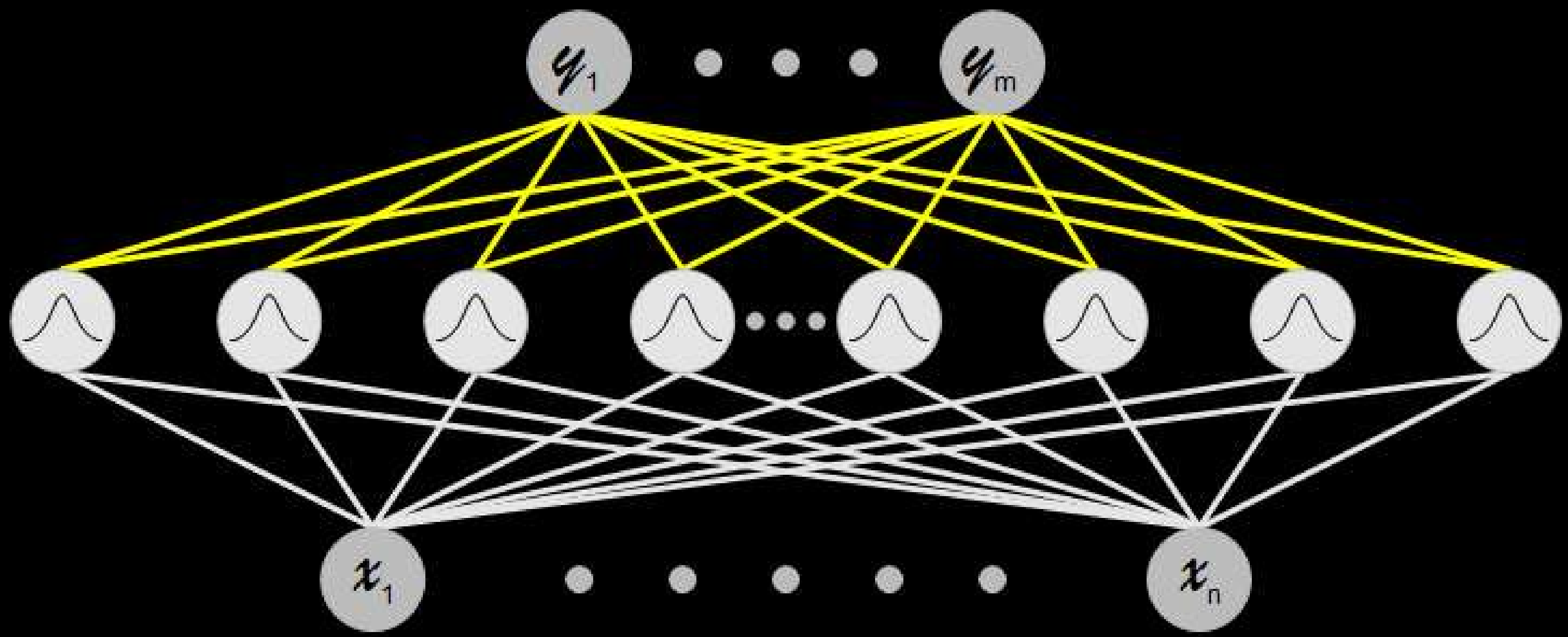
SVM in Feed-Forward Framework



DOF in Neural Nets / RBF



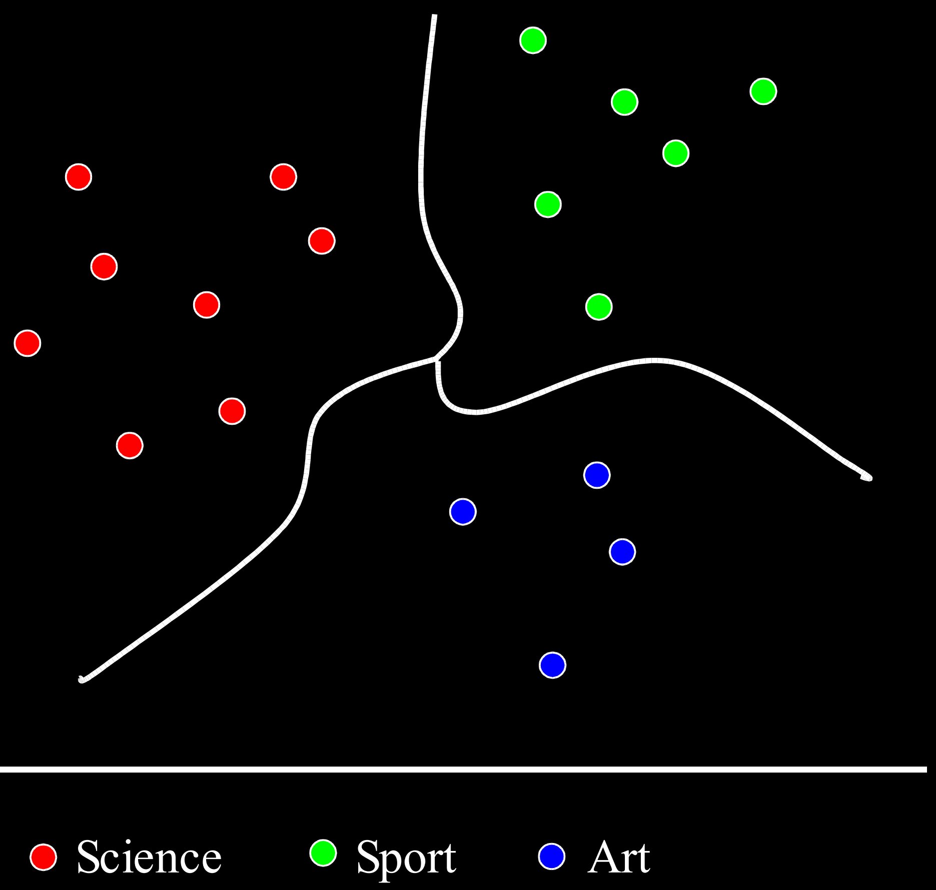
DOF in SVM



SVM vs. Neural Net / RBF

	SVM	NN / RBF
Regularization	✓	–
Global minimum	✓	–
Compact model	–	✓

Text Mining



Domain characteristics:

- thousands of features
- hundreds of topics
- sparse data

SVM in Text Mining

Reuters corpus

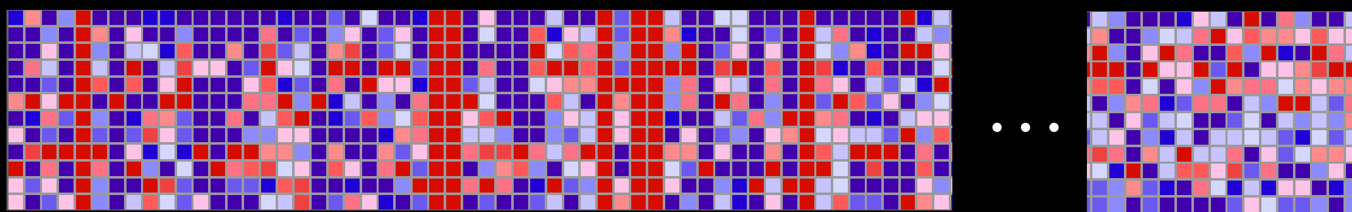
~10K documents, ~10K terms, 115 classes

Accuracy: recall / precision breakeven point

Naive Bayes	Rocchio	C4.5	K-NN	SVM linear	SVM non-linear
0.72	0.80	0.79	0.82	0.84	0.86

Joachims, 1998

Biomining



microarray data

Domain characteristics:

- thousands of features
- very few data points
- dense data

SVM on Microarray Data

Multiple tumor types

- 144 samples, 16063 genes, 14 classes
- Accuracy: correct rate

Naive Bayes	Weighted voting	K-NN	SVM linear
0.43	0.62	0.68	0.78

Ramaswamy et al., 2001

Other domains

High dimensionality problems:

- image (color and texture histograms)
- satellite remote sensing
- speech

Linear kernels sufficient in most cases

- data separability
- single parameter tuning (capacity)
- small model size

Final Note

- ☒SVM classification and regression algorithms available in ORACLE 10G database
- ☒Two APIs
  - JAVA (J2EE)
  - PL/SQL

References

Chapelle, O., Vapnik, V., Bousquet, O., & Mukherjee, S. (2001). Choosing Multiple Parameters for Support Vector Machines.

Hsu C., Chang C., & Lin, C. (2003). A Practical Guide to Support Vector Classification.

Joachims, T. (1998). Text Categorization with Support Vector Machines: Learning with Many Relevant Features.

Keerthi, S. & Lin, C. (2002). Asymptotic Behaviors of Support Vector Machines with Gaussian Kernel.

Platt, J. (1998). Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines.

Ramaswamy, S., Tamayo, P., Rifkin, R., Mukherjee, S., Yeang, C., Angelo, M., Ladd, C., Reich, M., Latulippe, E., Mesirov, J., Poggio, T., Gerald, W., Loda, M., Lander, E., Golub, T. (2001). Multi-Class Cancer Diagnosis Using Tumor Gene Expression Signatures.

ORACLE®