Regularization Techniques In Classifier Learning

Songcan Chen Hui Xue

Dept. of Computer Science & Engineering

Nanjing University of Aeronautics & Astronautics

{s.chen;xuehui}@nuaa.edu.cn

http://parnec.nuaa.edu.cn

Outline

- **Regularization Techniques:**

  A Brief Survey

- **Our work:**

  1) Discriminative Regularization

  2) Locality Regularization

Regularization Techniques

- Formulation of problem
- Problem Examples
- Why to regularize learning models
- Regularization techniques

Formulation of problem

1) Given a set of training data

$$\boldsymbol{S} = \{(\boldsymbol{x}_i, \boldsymbol{y}_i) \in \boldsymbol{R}^{\boldsymbol{d}} \times \boldsymbol{Y}, \boldsymbol{i} = 1, 2, \cdots, \boldsymbol{N}\}$$

2) *Given* $\boldsymbol{F} = \{f_\theta \mid \theta \in \Xi\}$ is a hypothesis set from which a desired *f* is derived based on ***S*** and yields good generalization performance for future unseen data, i.e. minimizing

$$R_{gen} = \int_{\Pi \setminus S} (t(\boldsymbol{x}) - f(\boldsymbol{x}))^2 \, dP(\boldsymbol{x})$$

t(*x*) is a true but unknown mapping function from *x* to label *y*. However, the generalization error is *incomputable* due to unknown *P*(x) and *t*(x)!

Problem Examples

- **Regression**
- **Classification**

Regression

Example:

$$F(\boldsymbol{x}) = 0.5 + 0.4\sin(2\pi \boldsymbol{x})$$

adding Gaussian noise to each data, $\sigma = 0.05$

Objective:

$$f_\theta(\boldsymbol{x}, \boldsymbol{w}) = \boldsymbol{w}_0 + \boldsymbol{w}_1 \boldsymbol{x} + \cdots + \boldsymbol{w}_M \boldsymbol{x}^M = \sum_{\boldsymbol{j}=0}^{\boldsymbol{M}} \boldsymbol{w}_j \boldsymbol{x}^j$$

fits $t(x)$, minimizing $R_{emp} = \frac{1}{2}\sum_{\boldsymbol{n}=1}^{\boldsymbol{N}} \left(f_\theta(\boldsymbol{x}_n, \boldsymbol{w}) - \boldsymbol{y}_n \right)^2$

Regression (Cont'd)

Under-fitting

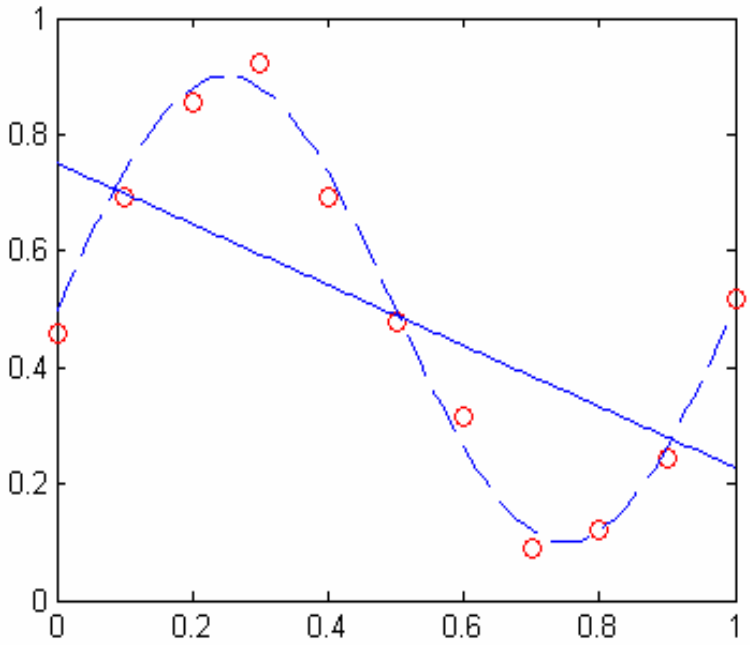


Fig1(a)

One-order fitting polynomial

Good fitting

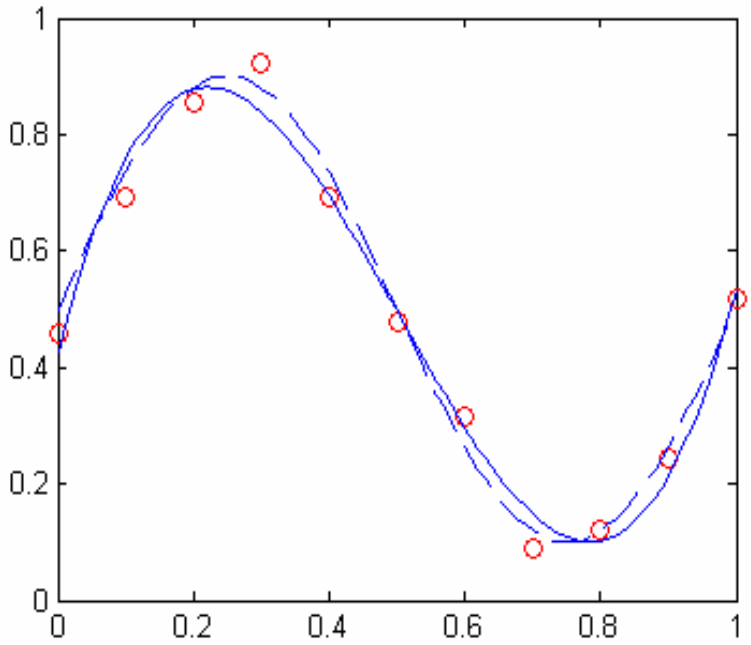


Fig1(b)

Three-order fitting polynomial

Regression (Cont'd)

Overfitting

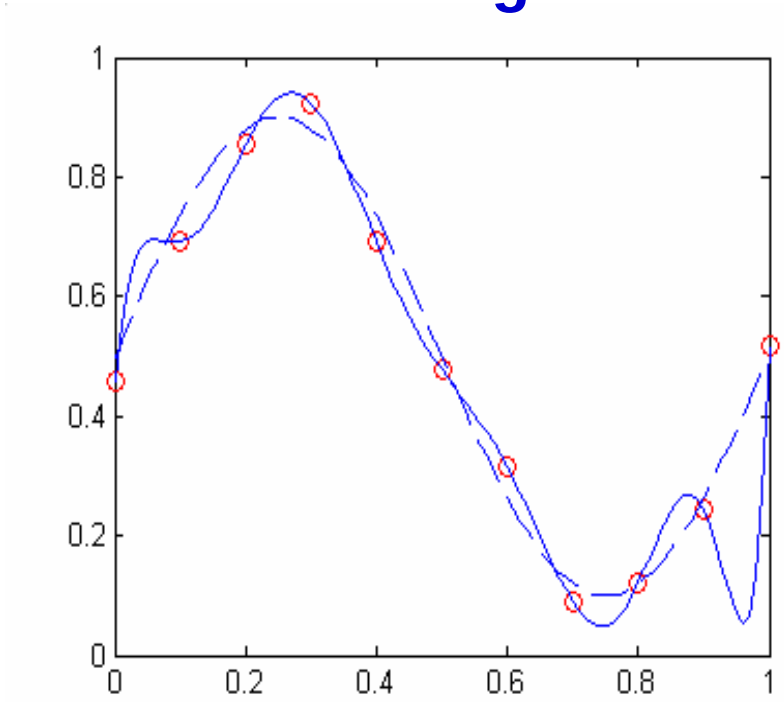


Fig1(c)
Ten-order fitting polynomial

Accuracy

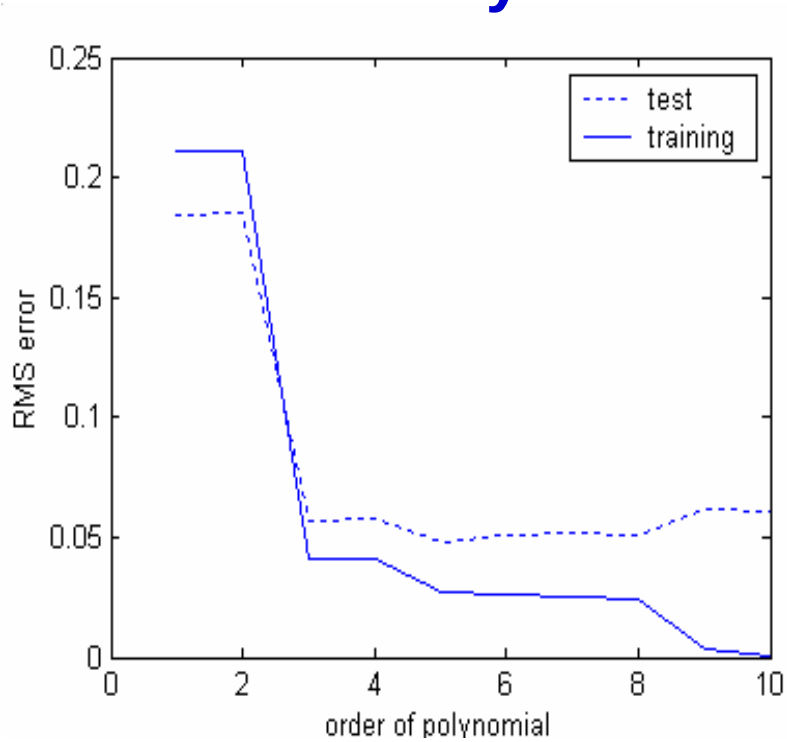
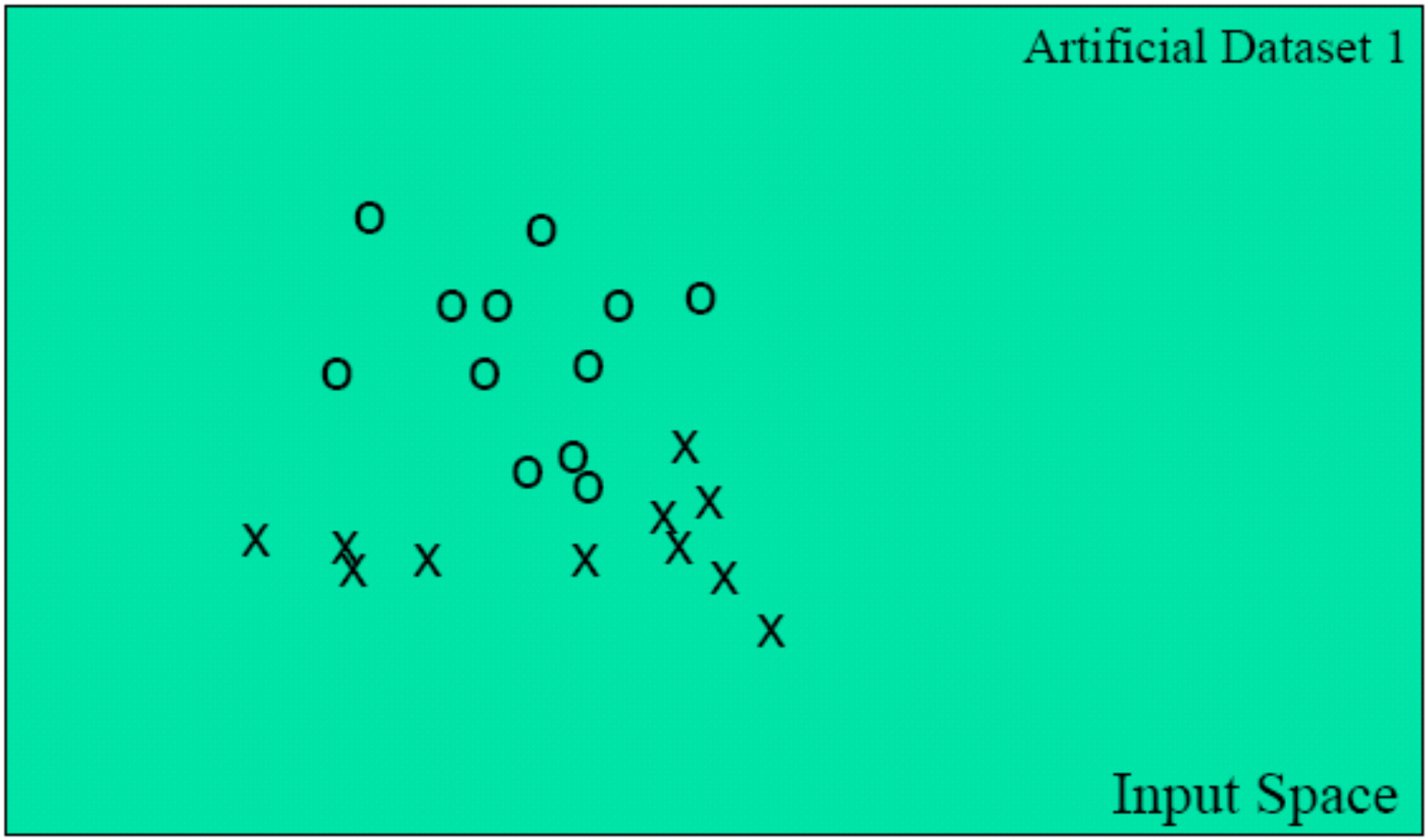


Fig1(d)
Test and Training Accuracies

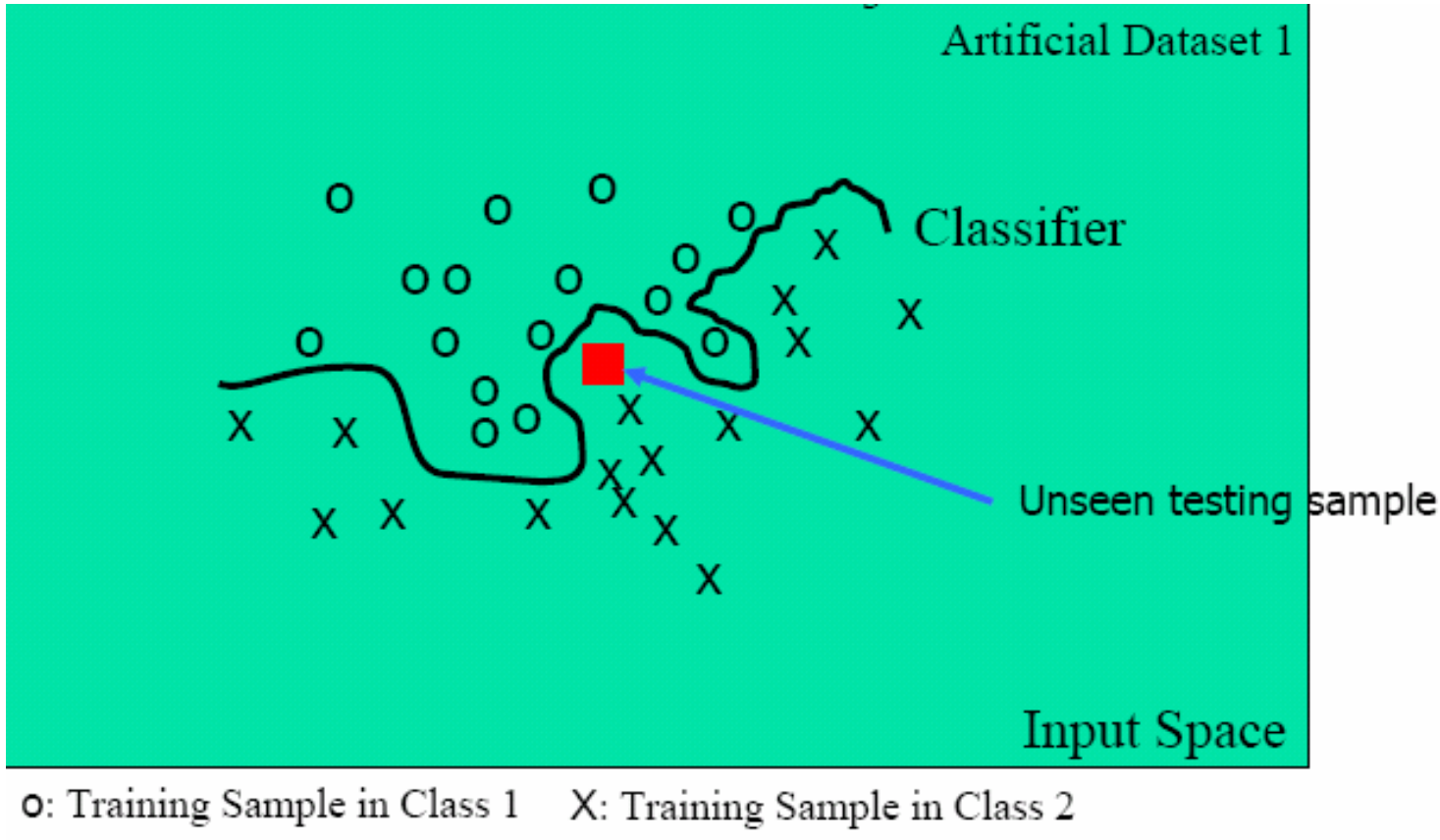
Classification



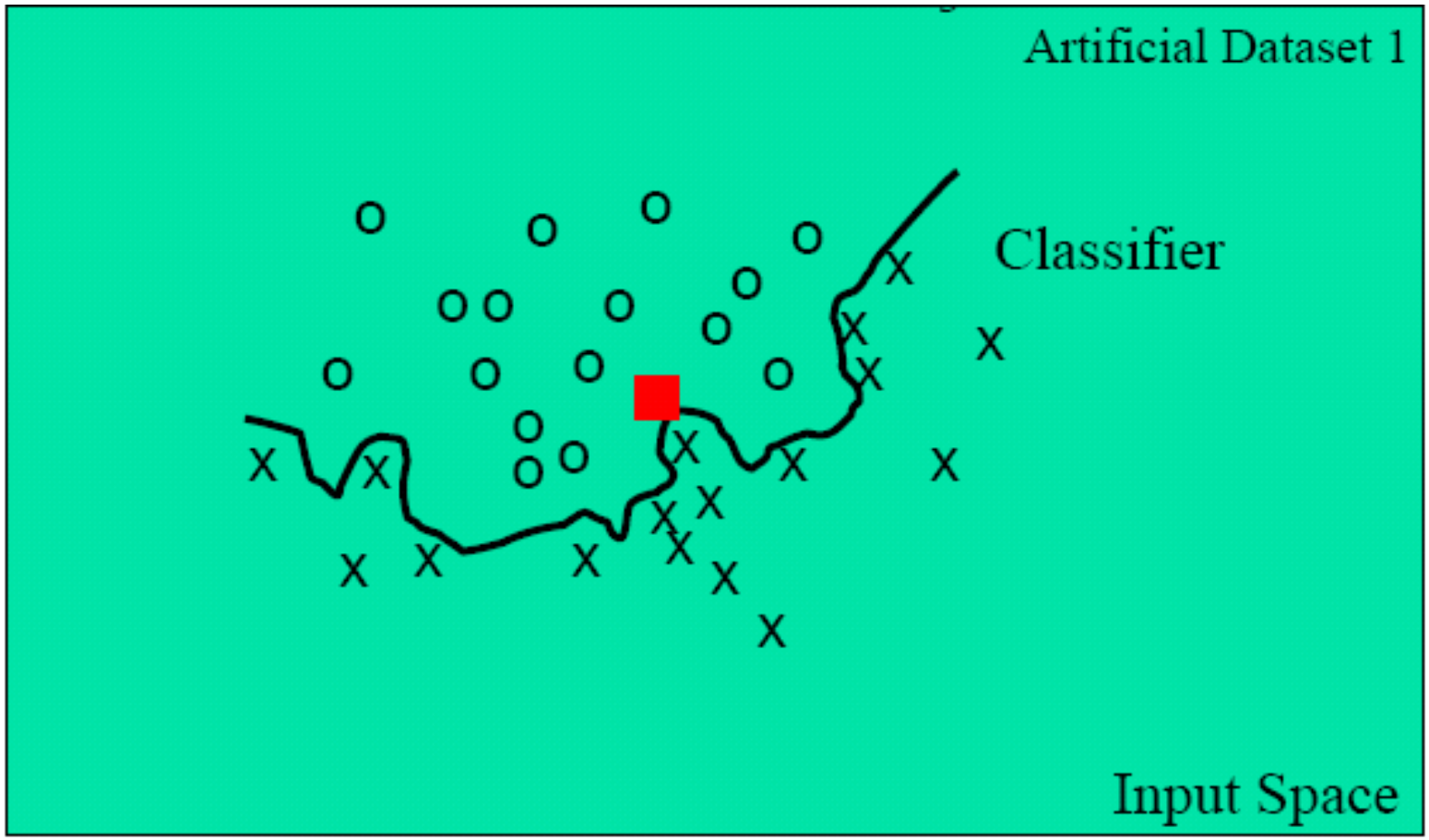
o: Training Sample in Class 1 X: Training Sample in Class 2

Classification (Cont'd)

- Over-fitting, high generalization error: A small emprical risk R_{emp} does not imply small generalization error R_{gen}



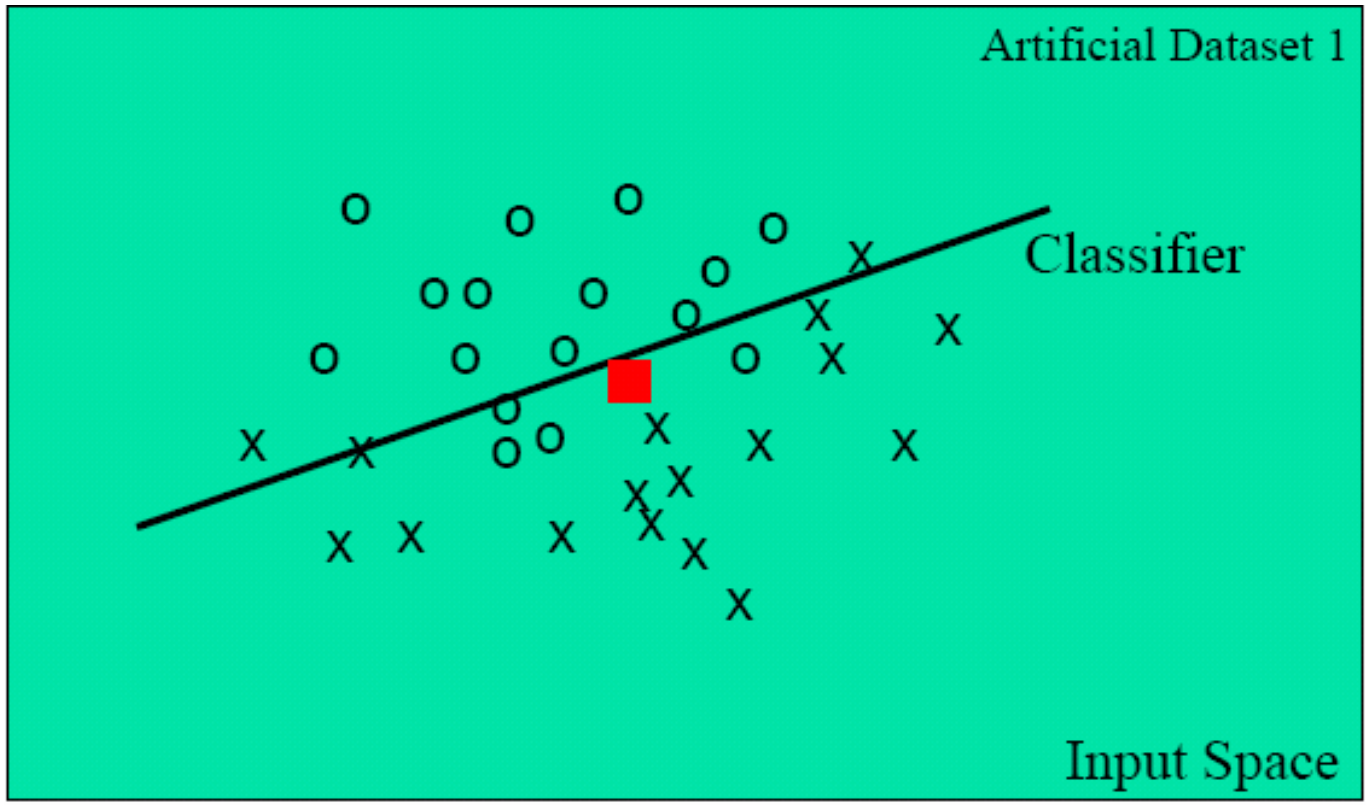
Classification (Cont'd)



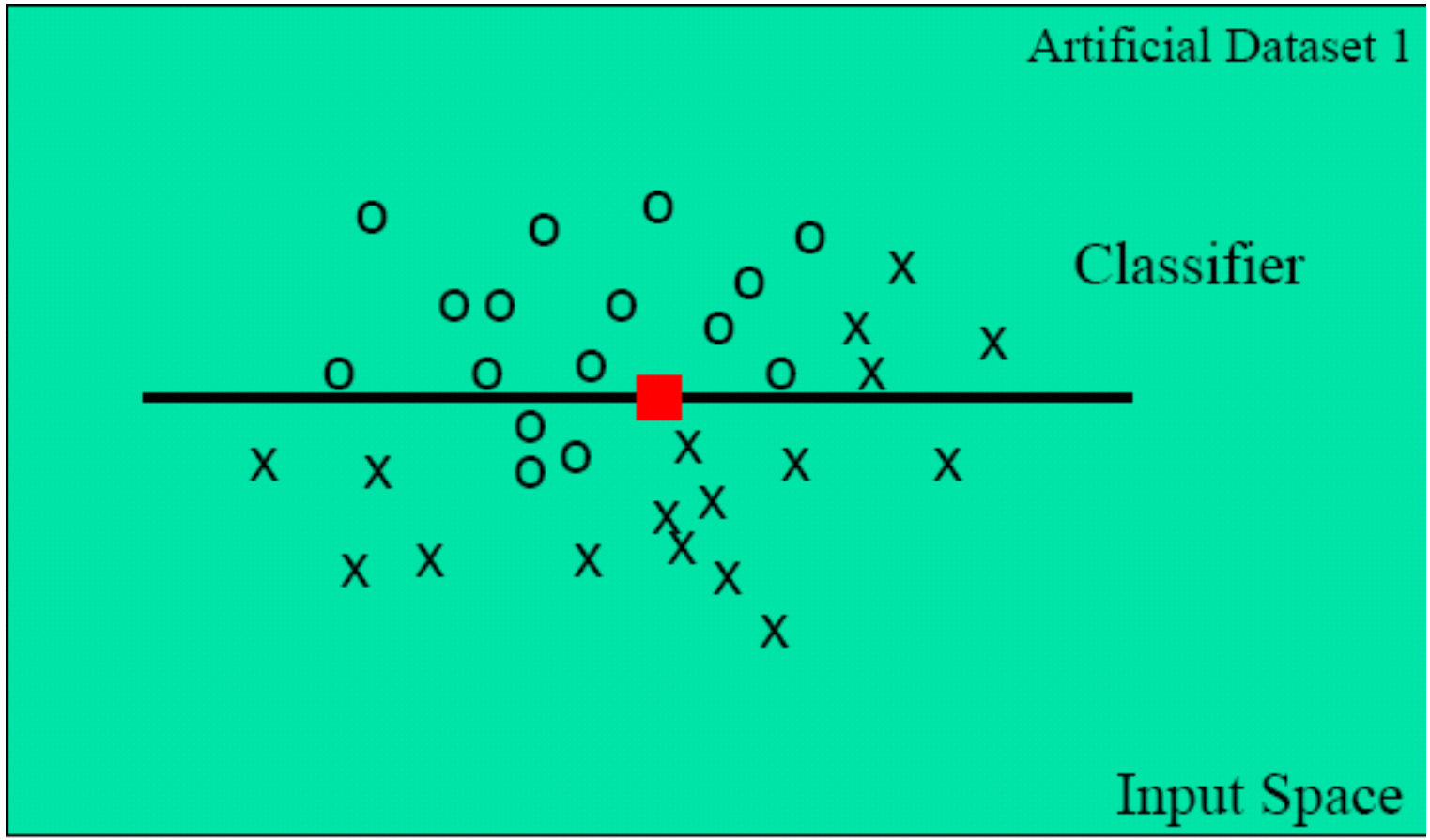
o: Training Sample in Class 1 X: Training Sample in Class 2

Classification (Cont'd)

- Under-fitting

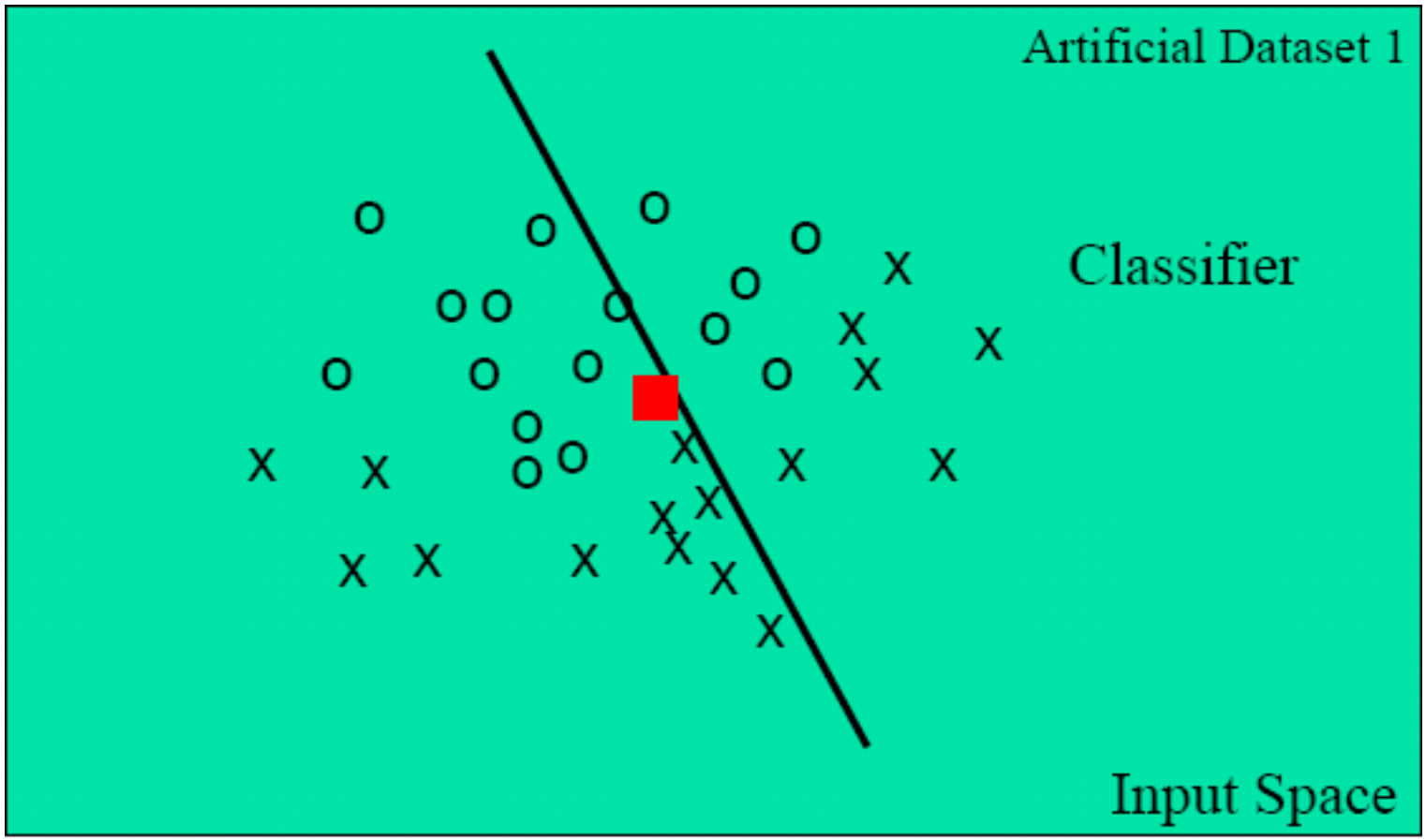


Classification (Cont'd)



o: Training Sample in Class 1 X: Training Sample in Class 2

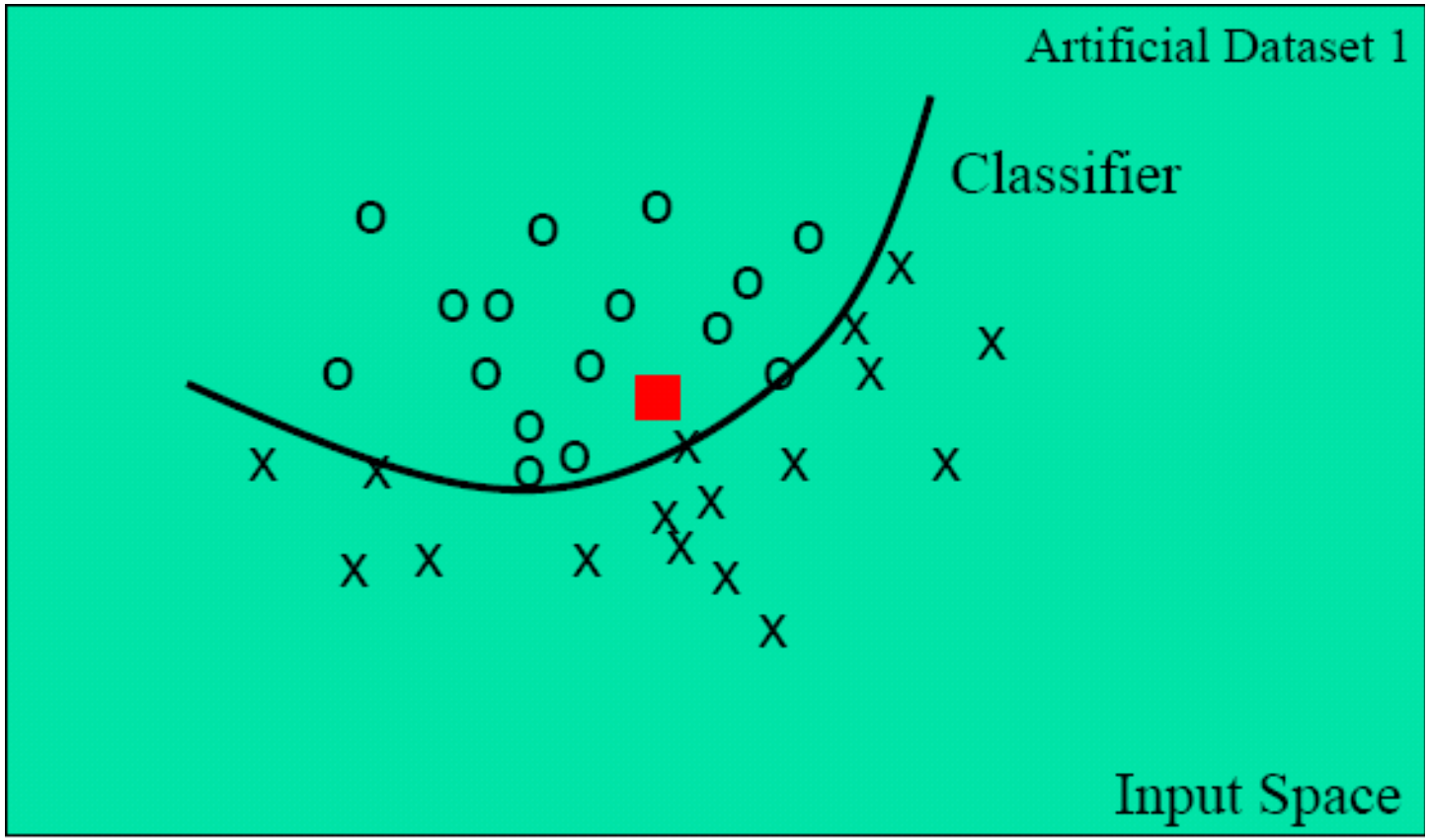
Classification (Cont'd)



o: Training Sample in Class 1 X: Training Sample in Class 2

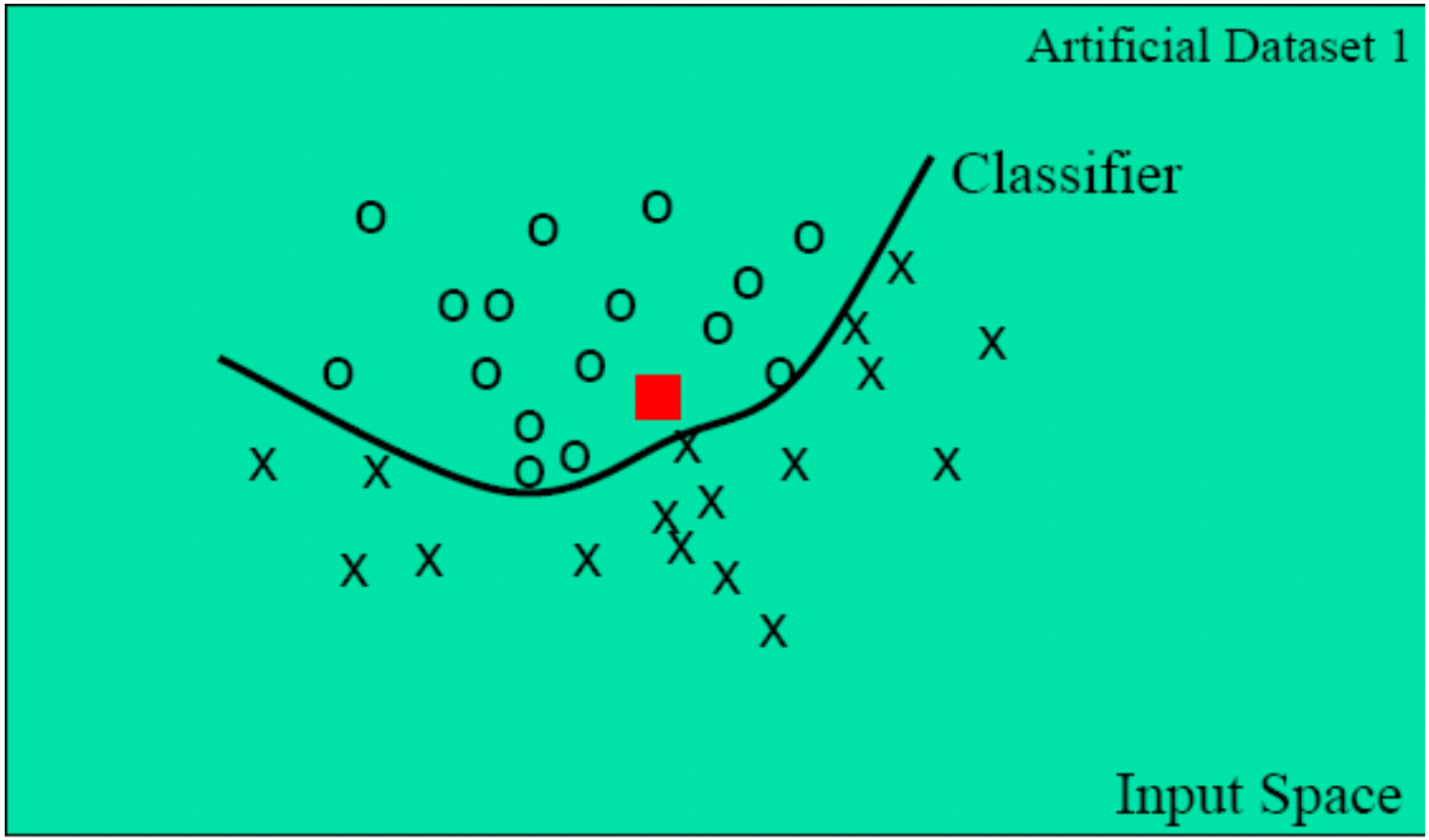
Classification (Cont'd)

- Good fitting



o: Training Sample in Class 1 X: Training Sample in Class 2

Classification (Cont'd)



o: Training Sample in Class 1 X: Training Sample in Class 2

Common Characteristics

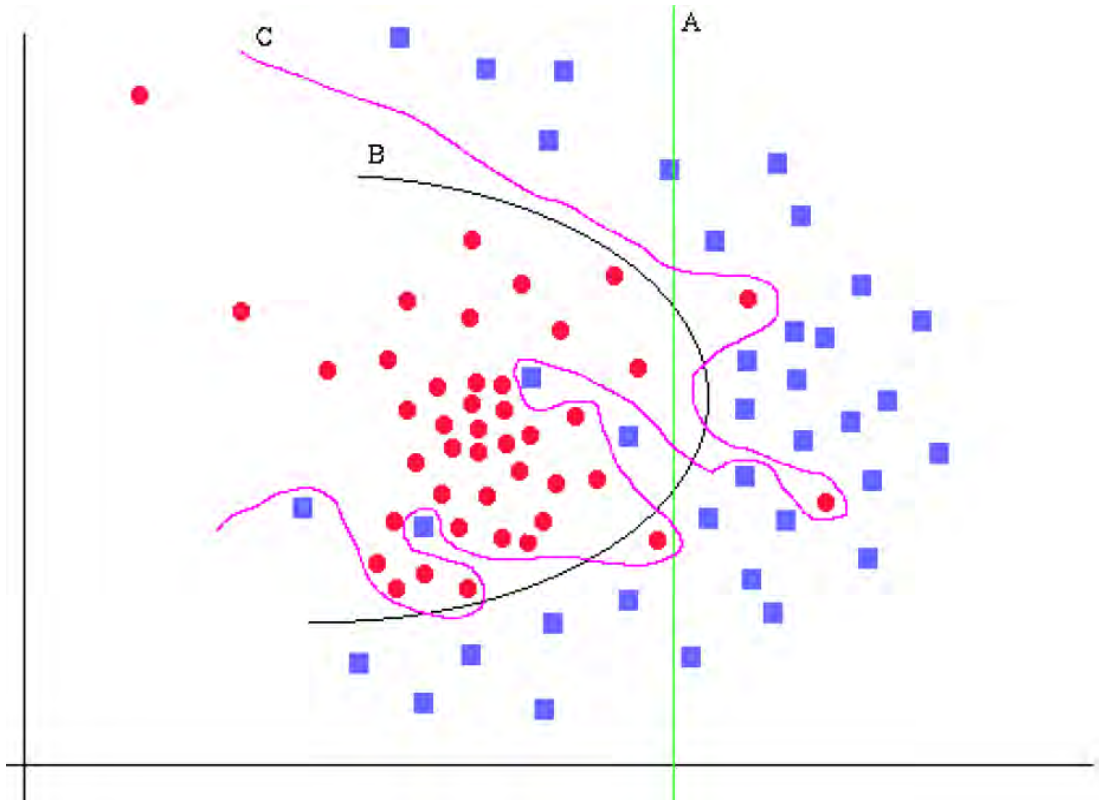
In the above Problems

- **Limited training samples**
- **Large types of fitting functions (Complexity)**

Our focus is on how to make the fitting function yield good prediction on unseen data, i.e., good generalization (or fitting);

Difficulty: a ill-posed problem, i.e., there are lots of fitting functions which can yield very small training error but just a few can have good generalization!

Ill Posed Problem



- **Note:** Small loss as well as Smooth curves

How to Boost Generalization

Four Basic Categories of Methods

1. **Model Selection**
2. **Regularization (✓)**
3. **Model Combination or Ensemble**
4. **Multiview approach on single dataset**

D. Schuurmans, F. Southey. Metric-based methods for adaptive model selection and regularization. ML, 48, 51-84, 2002

Why to Use Regularization?

The previously stated problem is ill-posed but a *well-posed* (适定或良态) one refers to such a problem if it satisfies three conditions below

- **Existence**
- **Uniqueness**
- **Continuity (Stability)**

S. Haykin. Neural Networks: A Comprehensive Foundation. Tsinghua University Press, 2002

Regularization Techniques

- Tikhonov Regularization
- Typical Regularization Methods

Regularization is a means of controlling the complexity of the fitting function being learnt.

Tikhonov Regularization

In 1963, Tikhonove proposed a new method called *Regularization* for solving ill-posed problems.

The basic idea of Regularization or Motivation:

Stabilize the solution by means of some auxiliary nonnegative functional that embeds prior information about the solution.

Essence

Incorporating Prior Information to develop model

The most common of prior information involves the assumption that the input-output mapping function is *smooth*, in the sense that *similar inputs correspond to similar outputs*

Related Concepts:

Tikhonov Functional

Green's function

Tikhonov Functional

Given the training input-output pairs $\{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$

$$R(f) = R_{emp}(f) + \lambda R_{reg}(f)$$

$$= \frac{1}{2}\sum_{i=1}^{N}[y_i - f(\mathbf{x}_i)]^2 + \frac{1}{2}\lambda \|Df\|^2$$

where $R_{emp}(f)$ and $R_{reg}(f)$ are the empirical risk term and the regularization term respectively

Tikhonov Functional (Cont'd)

- λ is a regularization parameter that controls the trade-off between the fitting goodness of data and the roughness (complexity) of the solution, its selection is, up to now, an open problem!

- D denotes a linear differential operator, which is defined as the Fréchet differential of Tikhonov functional.

Geometrically, D is interpreted as a local linear approximation of the manifold in high-dimensional space. The smoothness prior *implicitly* involved in D makes the solution stable.

Z. Chen, S. Haykin. On different facets of regularization theory. Neural Computation, 14(2): 2791-2846, 2002

Green's function

The solution of the classical Tikhonov regularization problem can be represented by the expansion:

$$f_\lambda(\mathbf{x}) = \sum_{i=1}^{N} w_i G(\mathbf{x}, \mathbf{x}_i)$$

where $w_i = [y_i - f(\mathbf{x}_i)]/\lambda$, and $G(\mathbf{x}, \mathbf{x}_i)$ is the Green's function, RBF is its special case.

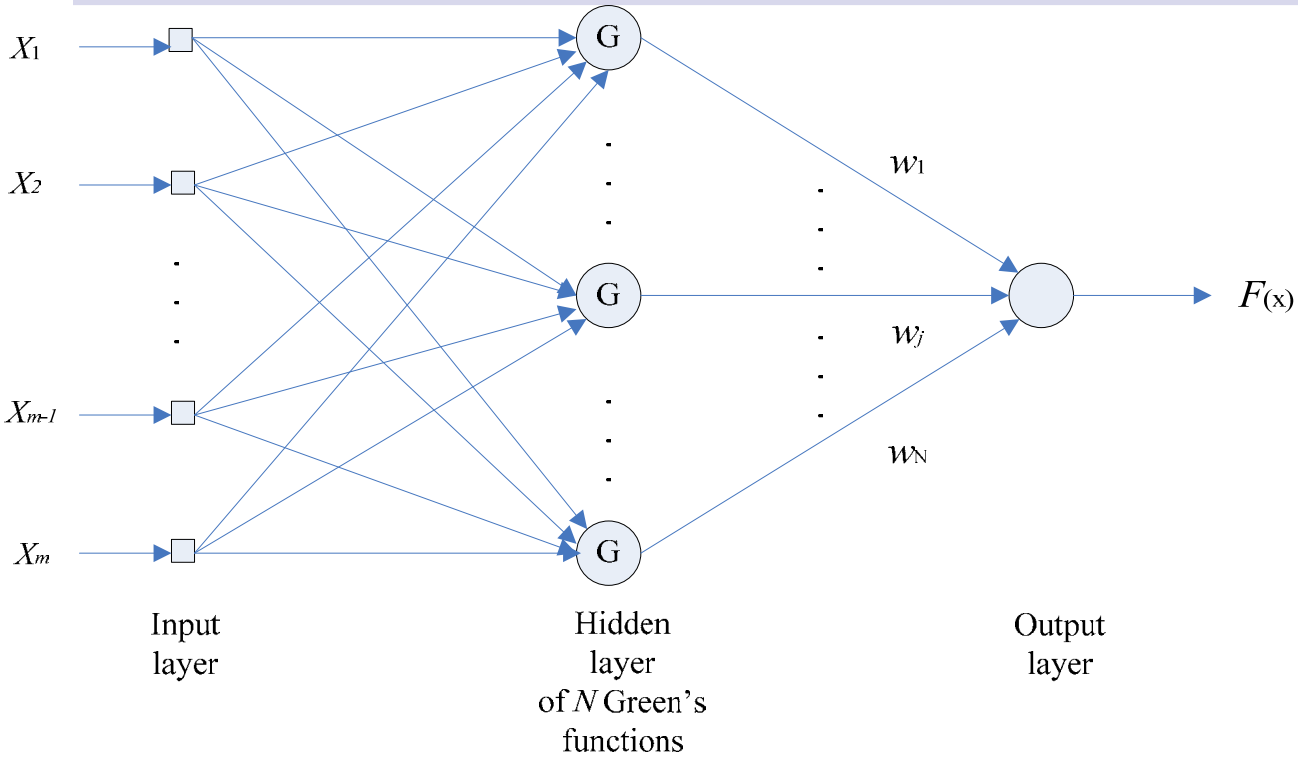
→ form a (neural) network called regularization network (RN)

Typical Regularization Methods

- **Regularization Networks (RN)**
- **Generalized Radial Basis Function Networks (GRBFN)**
- **Support Vector Machines (SVM)**
- **Manifold Regularization (MR), etc.**

View the classifier learning as a multivariate functional fitting problem

Regularization Networks (RN)

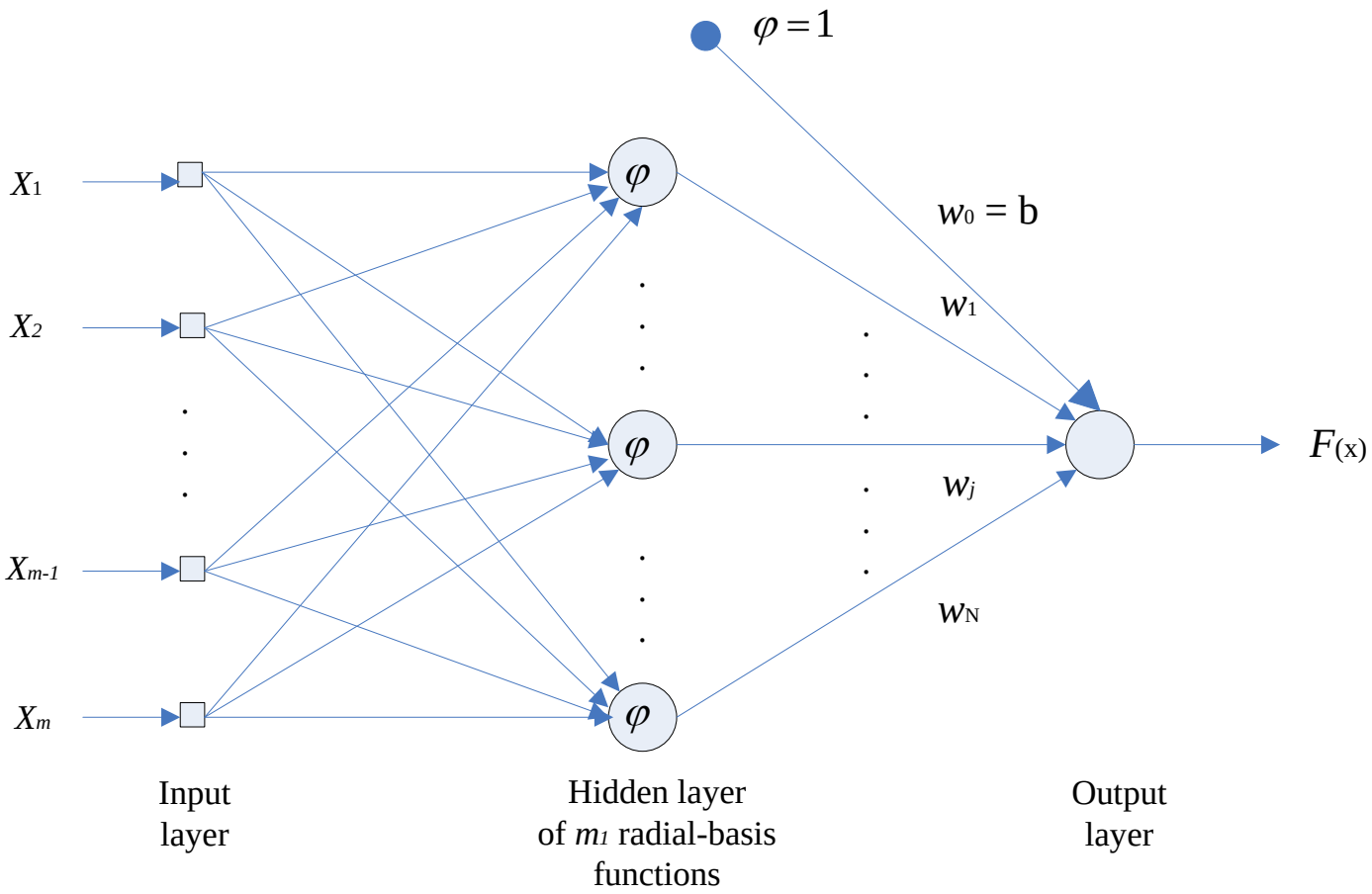


If the multivariate Gaussian function is selected as the Green's function, the solution will be *an optimal interpolant* in the sense that it minimizes the Tikhonov functional.

Deficiencies of RN

- The one-to-one correspondence between the training input data and the Green's function causes expensive computational cost, especially for large N ⟹ GRBFN
- Only emphasize the smoothness of the classifier and do not incorporate any prior *intra-class* and *inter-class* information into the formulation which is vital for classification

Generalized Radial Basis Function Networks (GRBFN)



GRBFN (Cont'd)

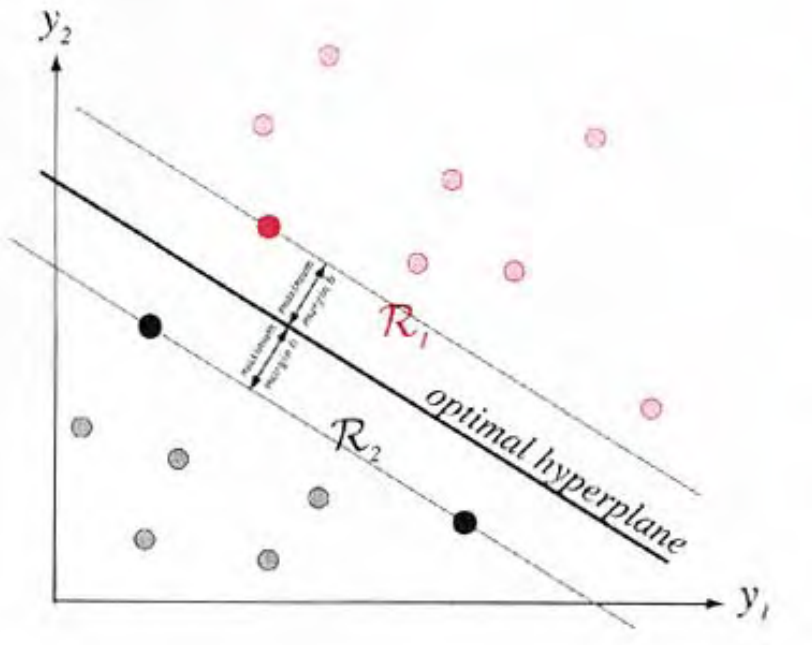
- Apply clustering strategies to determine the parameters of hidden neurons and then adopt the regularization technique to optimize a least squares error criterion to derive a classifier
- Incorporate the *intra-class* information generated from clustering into the traditional regularization
- However, still neglect the *inter-class* information

Support Vector Machines (SVM)

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

$$s.t. \quad y_i(w^T x_i + b) \geq 1 - \xi_i \qquad i = 1, \cdots, N$$

$$\xi_i \geq 0 \qquad i = 1, \cdots, N$$



Richard O. Duda et al. Pattern Classification. Wiley, 2001

SVM (Cont'd)

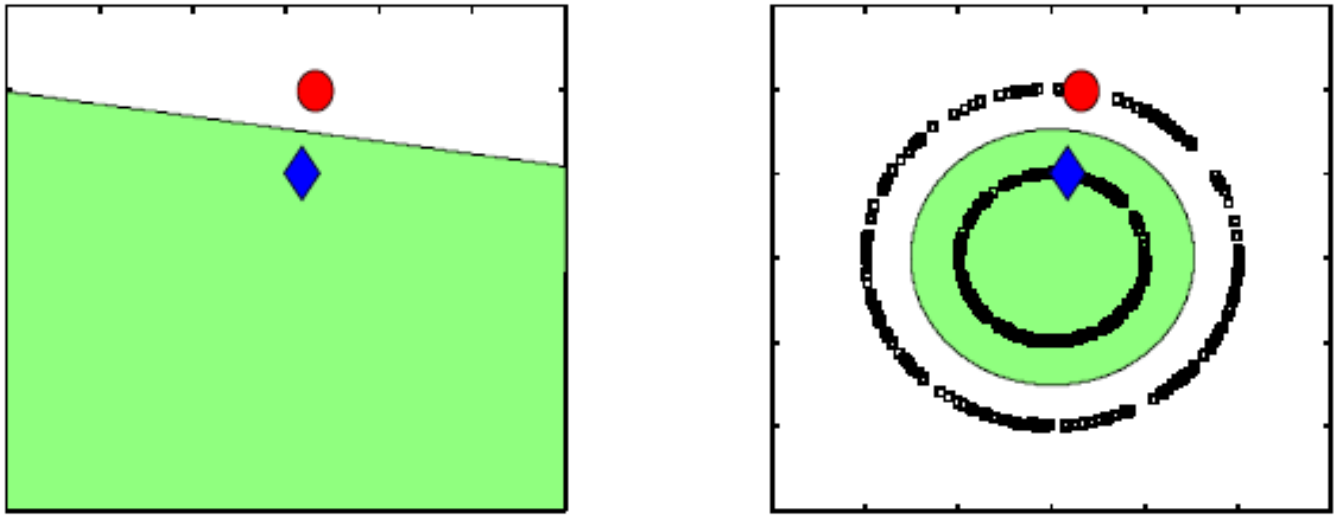
- Use the hinge-loss function as the $R_{emp}(f)$ instead of the common square-loss function, which more emphasizes the prior *inter-class* discriminative knowledge than traditional regularization
- However, does not take the *intra-class* information into account and thus does not make sufficient use of the prior class structural knowledge

Manifold Regularization (MR)

$$\min_{f \in H} \left\{ \frac{1}{N} \sum_{i=1}^{N} V\left(y_i, f(\mathbf{x}_i)\right) + \gamma_A \left\| f \right\|_K^2 + \gamma_I \left\| f \right\|_I^2 \right\}$$

where the regularization term $\left\| f \right\|_K^2$ controls the complexity of the classifier and the other regularization term $\left\| f \right\|_I^2$ controls the complexity measured by the manifold geometry of the sample distribution

MR (Cont'd)



M. Belkin, P. Niyogi, and V. Sindhwani. Manifold regularization: A geometric framework for learning from examples. Department of Computer Science,University of Chicago,Tech.Rep, TR-2004-06, 2004.

LapRLS and LapSVM

- **LapRLS**

$$\min_{f \in H_K} \frac{1}{l} \sum_{i=1}^{l} (y_i - f(x_i))^2 + \gamma_A \| f \|_K^2 + \frac{\gamma_I}{(u+l)^2} f^T L f$$

- **LapSVM**

$$\min_{f \in H_K} \frac{1}{l} \sum_{i=1}^{l} (1 - y_i f(x_i))_+ + \gamma_A \| f \|_K^2 + \frac{\gamma_I}{(u+l)^2} f^T L f$$

Supervised MR

- **MR is a semi-supervised framework**
- **Its supervised Version**

$$\min_{f \in H} \left\{ \frac{1}{N} \sum_{i=1}^{N} V\left(y_i, f(\mathbf{x}_i)\right) + \gamma_A \|f\|_K^2 + \sum_{j=1}^{c} \gamma_{I_j} \|f\|_{I_j}^2 \right\}$$

Where c is the number of classes.

The number of regularizers equals to c

Deficiencies of Supervised MR

Construct a graph for each class, i.e. different $\| f \|_I^2$ correspond to different classes, which undoubtedly leads to the appearance of many free regularization parameters in the formulation, especially for the multi-class problems.

As a result, the computational complexity in training of MR will increase sharply**, coined as "curse of the number of regularizers"**

Comparison on Use of Prior Information

Method	Prior Information	
	Inter-class (Discriminant information)	Intra-class (Structural information)
RN	× (Least Squares loss)	×
GRBFN	× (Least Squares loss)	√
SVM	√ (Hinge loss)	×
MR	√ (Hinge loss)	√ (Manifold)

Our Work

- **Discriminative Regularization for classification**

  -- RN, GRBFN, SVM, MR

Motivation

Deficiencies of Traditional methods

- Essentially derived from multivariate functional fitting or regression problems, however, classification is just its special case;
- Data independent;
- (Consequently,) in classifier design, these methods give more concerns to the smoothness of the classifier

S. Haykin, Neural Networks: A Comprehensive Foundation. Tsinghua University Press, 2001.
Z. Chen, S. Haykin. On different facets of regularization theory. Neural Computation, 14(2): 2791-2846, 2002
T. Poggio, F. Girosi. Networks for approximation and learning. Proc. Of the IEEE. 78: 1481-1497, 1990a
T. Poggio, F. Girosi. Regularization algorithms for learning that are equivalent to multilayer networks. Science, 247, 978-982, 1990b

Deficiencies (Cont'd)

- In classification, similar inputs near the discriminant boundaries are more likely to belong to different classes, implying that the same global smoothness constraint imposed on the whole domain may not sufficiently formulate the discrimination among classes.

- The main goal of classification is to separate the samples of different classes in the output space as far as possible. Hence, *the prior discriminative information* is crucial for classification.

Characteristics of DR

- The "No Free Lunch" Theorem

  A model is good when incorporating prior knowledge of the problem at hand

- Directly introduce the *intra-class* and *inter-class* information as the new *discriminative regularization term* ***to seek an more effective classifier***
- Integrate the prior *geometrical* information into the single regularization term
- Analytic solutions

Discriminative Regularization Term

$$\min_{f \in F} \left\{ \frac{1}{2} \sum_{i=1}^{N} [y_i - f(\mathbf{x}_i)]^2 + \frac{1}{2} R_{disreg}(f, \eta) \right\}$$

A general definition for $R_{disreg}(f, \eta)$

$$R_{disreg}(f, \eta) = \eta A(f) - (1 - \eta) B(f)$$

DR Term (Cont'd)

Use the generalized variance (GV) in statistics, similar to Maximum Margin Criterion (MMC)

$$A(f) = S_b = \sum_{k=1}^{c} \frac{1}{N_k} \sum_{i=1}^{N_k} \left\| f(\mathbf{x}_i^{(k)}) - \frac{1}{N_k} \sum_{j=1}^{N_k} f(\mathbf{x}_j^{(k)}) \right\|^2$$

$$B(f) = S_w = \sum_{k=1}^{c} \sum_{l \neq k} \left\| \frac{1}{N_k} \sum_{i=1}^{N_k} f(\mathbf{x}_i^{(k)}) - \frac{1}{N_l} \sum_{j=1}^{N_l} f(\mathbf{x}_j^{(l)}) \right\|^2$$

--- yielding a DR-GV learning model.

Focus on the *global* class relationship between the samples and thus fail to sufficiently characterize the *local* manifold structure of the data

H. Li, T. Jiang, and K. Zhang. Efficient and robust feature extraction by maximun margin criterion. IEEE Trans. on Neural Networks, vol.17(1), 157-165, 2006.

Manifold Structure

For each sample $\boldsymbol{x}_i$, in terms of Locality Sensitive Discriminant Analysis (LSDA), we first divide the nearest neighborhood *ne(i)* into two nonoverlapping subsets

$$ne_b(i) = \left\{ \mathbf{x}_i^j \,\middle|\, if \quad \mathbf{x}_i^j \quad and \quad \mathbf{x}_i \quad belong \;\; to \;\; same \;\; class, \;\; 1 \le j \le k \right\}$$

$$ne_w(i) = \left\{ \mathbf{x}_i^j \,\middle|\, if \quad \mathbf{x}_i^j \quad and \quad \mathbf{x}_i \quad belong \;\; to \;\; different \;\; classes, \;\; 1 \le j \le k \right\}$$

D. Cai, X. He, K. Zhou, J. Han, and H. Bao. Locality sensitive discriminant analysis. IJCAI, 708-713, 2007.

Weight Matrices

Define the two weight matrices of G_b and G_w respectively

$$W_{b,ij} = \begin{cases} 1 & if\ \mathbf{x}_j \in ne_b(i)\ or\ \mathbf{x}_i \in ne_b(j) \\ 0 & otherwise \end{cases}$$

$$W_{w,ij} = \begin{cases} 1 & if\ \mathbf{x}_j \in ne_w(i)\ or\ \mathbf{x}_i \in ne_w(j) \\ 0 & otherwise \end{cases}$$

DR Term (DRT)

Characterize the *intra-class* compactness from the *intra-class* graph

$$\tilde{S}_b = \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\left\| f(\mathbf{x}_i) - f(\mathbf{x}_j)\right\|^2 W_{b,ij}$$

Characterize the *inter-class* separability from the *inter-class* graph

$$\tilde{S}_w = \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\left\| f(\mathbf{x}_i) - f(\mathbf{x}_j)\right\|^2 W_{w,ij}$$

The Optimization Problem

$$\min_{f \in F} \left\{ \frac{1}{2} \sum_{i=1}^{N} \left[y_i - f(\mathbf{x}_i) \right]^2 + \frac{1}{2} \left[\eta \tilde{S}_b - (1-\eta) \tilde{S}_w \right] \right\}$$

where η is the parameter that regulates the relative significance of the intra-class compactness versus the inter-class separability, $0 \leq \eta \leq 1$

Intuitive Interpretation

An Intuitive interpretation for the regularizing term

$$\min \frac{1}{2}[\eta \tilde{S}_b - (1-\eta)\tilde{S}_w] \Leftrightarrow \max \frac{1}{2}[(1-\eta)\tilde{S}_w - \eta \tilde{S}_b]$$

Hence, it maximizes the average margin between classes and plays a similar role in SVM.

A large average margin indicates that patterns in the output space are close to each other if they are from the same class but are far from each other if they are from different classes.

Haifeng Li, Tao Jiang, and Keshu Zhang, Efficient and Robust Feature Extraction by Maximum Margin Criterion, IEEE TNN, VOL. 17, NO. 1, JAN. 2006, 157-165

Linear Classifiers

$$f(\mathbf{x}) = \mathbf{w}^T \mathbf{x} + b$$

$$\tilde{S}_b = \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\left[f(\mathbf{x}_i) - f(\mathbf{x}_j)\right]^2 W_{b,ij}$$

$$= \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\left(\mathbf{w}^T\mathbf{x}_i - \mathbf{w}^T\mathbf{x}_j\right)^2 W_{b,ij}$$

$$= \mathbf{w}^T\mathbf{X}(\mathbf{D}_b - \mathbf{W}_b)\mathbf{X}^T\mathbf{w}$$

$$= \mathbf{w}^T\mathbf{X}\mathbf{L}_b\mathbf{X}^T\mathbf{w}$$

$$\tilde{S}_w = \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\left[f(\mathbf{x}_i) - f(\mathbf{x}_j)\right]^2 W_{w,ij}$$

$$= \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\left(\mathbf{w}^T\mathbf{x}_i - \mathbf{w}^T\mathbf{x}_j\right)^2 W_{w,ij}$$

$$= \mathbf{w}^T\mathbf{X}(\mathbf{D}_w - \mathbf{W}_w)\mathbf{X}^T\mathbf{w}$$

$$= \mathbf{w}^T\mathbf{X}\mathbf{L}_w\mathbf{X}^T\mathbf{w}$$

Relationship between DR and Dimensionality Reduction Methods

$$\min_{\mathbf{w},b}\left\{\frac{1}{2}\sum_{i=1}^{N}\left[y_i-\left(\mathbf{w}^T\mathbf{x}_i+b\right)\right]^2+\frac{1}{2}\mathbf{w}^T\mathbf{X}\left[\eta\mathbf{L}_b-(1-\eta)\mathbf{L}_w\right]\mathbf{X}^T\mathbf{w}\right\}$$

- Find an orientation for which the projected samples are well separated, similar to the intuitive motivation in DR
- Any similar supervised dimensionality reduction methods can be also embedded in DR as the regularization term
- Provide a brand-new viewpoint to *combine dimensionality reduction with classification*

Differences of DR and MR

DR: $$\min_{f \in F} \left\{ \frac{1}{2} \sum_{i=1}^{N} \left[y_i - f(\mathbf{x}_i) \right]^2 + \frac{1}{2} \left[\eta \tilde{S}_b - (1-\eta) \tilde{S}_w \right] \right\}$$

⟹ Only have one adjustable regularization parameter

MR: $$\min_{f \in H} \left\{ \frac{1}{N} \sum_{i=1}^{N} V\left(y_i, f(\mathbf{x}_i) \right) + \gamma_A \| f \|_K^2 + \sum_{j=1}^{c} \gamma_{I_j} \| f \|_{I_j}^2 \right\}$$

⟹ The number of regularization parameters depends on the class number

Effectively avoid the potential ***"Curse of the number of the regularizers"*** of MR in the optimization

Analytic Solutions to DR

- Many traditional regularization methods use conjugate gradient algorithms
  - Converge slowly
  - Can not guarantee to converge to the global optimum
- DR obtains the solutions directly from solving a set of linear equations
  - Simple
  - Stable

J.A.K. Suykens and J. Vandewalle. Least squares support vector machine classifiers. Neural Processing Letters, 9: 293-300, 1999.

Analytic Solutions (Cont'd)

- Many existing learning machines decompose the multiclass classification problem into multiple two-class classification problems
- DR solves directly both two-class and multi-class problems in a unified framework in terms of the simple analytic solution

B.K. Natarajan, Machine Learning: A Theoretical Approach. Morgan Kaufmann, Los Alamitos, CA, 1991.
E. Gelenbe, K.F. Hussain. Learning in the multiple class random neural network. IEEE Trans. on Neural Networks, 13(6): 1257-1267, 2002.
G. Ou, Y.L. Murphey. Multi-class pattern classification using neural networks. PR, 40(1): 4-18, 2007.

Vector Labeled Outputs

Code the class labels following the *one-of-c* rule, i.e. if x_i belongs to the kth class, then

$$\mathbf{y}_i = [0,\cdots,1,\cdots,0]^T \in R^c$$

where the kth element is 1 and the other elements are 0, $\forall i = 1,\cdots,N$

$$f(\mathbf{x}) = \mathbf{W}^T\mathbf{x} + \mathbf{b}$$

where $\mathbf{W} \in R^{n\times c}$, $\mathbf{b} \in R^c$

The Equality Constraints

$$\min_{\mathbf{W},\mathbf{b}} \left\{ \frac{1}{2} \sum_{i=1}^{N} \left\| \mathbf{e}_i \right\|^2 + \frac{1}{2} \left[\eta \tilde{S}_b - (1-\eta) \tilde{S}_w \right] \right\}$$

subject to

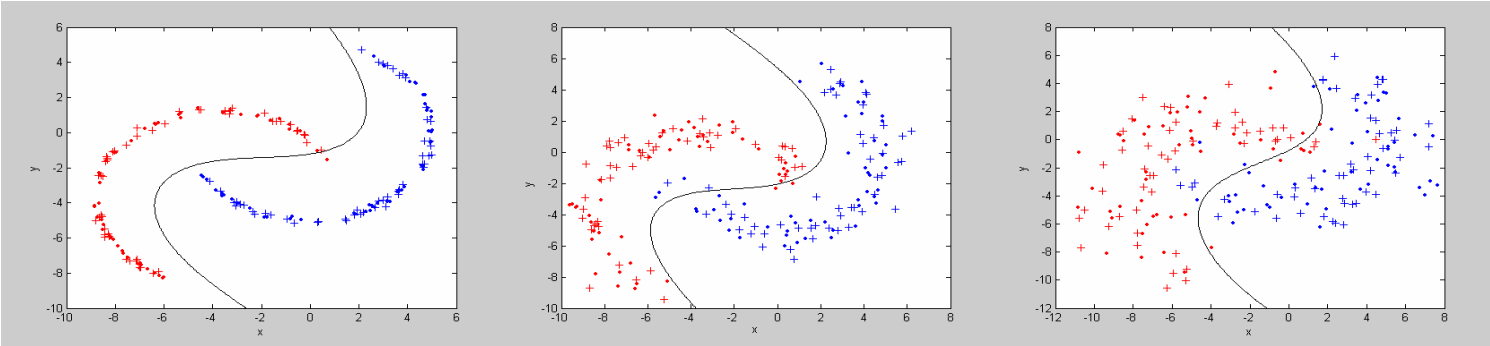
$$\mathbf{W}^T \mathbf{x}_i + \mathbf{b} + \mathbf{e}_i = \mathbf{y}_i \ , \quad \forall i = 1, \cdots, N$$

Given the parameter $\eta \in [0,1]$, the global solution is characterized by the dual linear system with dual variables $\boldsymbol{\alpha} \in R^{c \times N}$

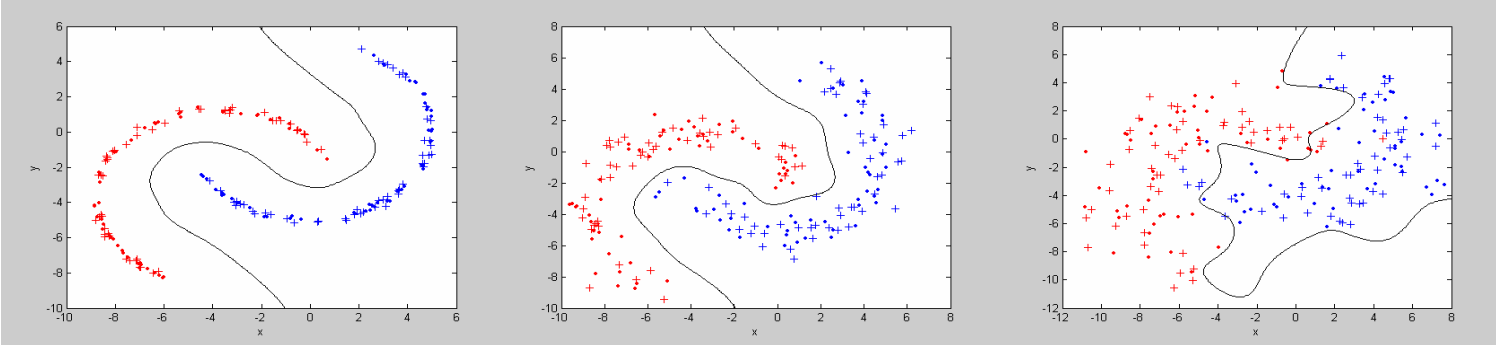
$$\begin{bmatrix} \mathbf{b} & \boldsymbol{\alpha} \end{bmatrix} \begin{bmatrix} 0 & \mathbf{1}_N^T \\ \mathbf{1}_N & \boldsymbol{\Omega}_\eta + \mathbf{I}_N \end{bmatrix} = \begin{bmatrix} \mathbf{0}_c & \mathbf{Y} \end{bmatrix}$$

Toy Problems

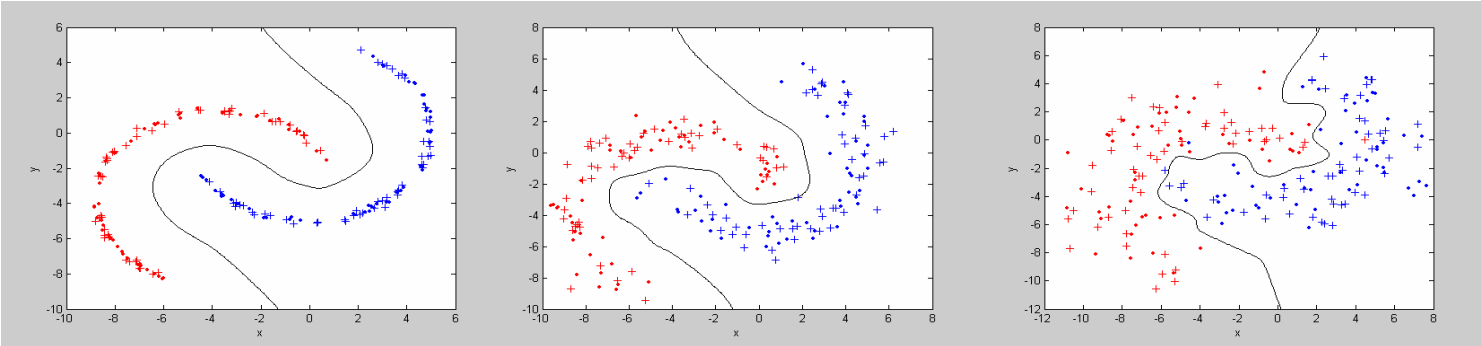
RN:



DR-GV:



DR:



Overfitting?
Training and Testing Accuracies

- **Training accuracies (%) compared between RN, DRGV and DRLSC in the three Two-Moon datasets**

	(A)	(B)	(C)
RN	99.00	95.00	90.00
DRGV	**100.0**	**100.0**	98.00
DRLSC	**100.0**	**100.0**	**99.00**

Testing accuracies (%) compared between RN, DRGV and DRLSC in the three Two-Moon datasets

	(A)	(B)	(C)
RN	**100.0**	98.00	93.00
DRGV	**100.0**	**100.0**	93.00
DRLSC	**100.0**	**100.0**	**98.00**

Linear Kernel:

DR:

10 /20

Dataset	Number of classes	Dimension	Classification accuracy			
			SVM	MR	DR-GV	DR
Ionosphere	2	34	87.78	85.00*	85.17*	**88.01**
Sonar	2	60	77.88	**79.87***	70.77*	76.83
Water	2	38	**98.64***	97.98*	94.58*	96.27
Wdbc	2	30	94.98*	94.42*	95.65*	**96.21**
Bupa	2	6	66.99*	68.90*	66.94*	**70.23**
Pid	2	8	73.96*	69.19*	76.93*	**78.26**
Diabetes	2	8	75.05*	69.40*	77.24*	**78.72**
Wine	3	13	95.67*	97.11*	**99.00**	**99.00**
Lenses	3	4	74.62*	82.85*	82.31*	**84.62**
Tae	3	5	50.39*	51.58*	51.97*	**56.58**
New_thyroid	3	5	**95.65***	90.00	89.81*	91.11
Iris	3	4	94.53*	**96.53***	86.13*	87.07
Cmc	3	9	**55.68***	52.38	50.93*	51.96
Balance_scale	3	4	87.86*	**89.20***	87.83*	88.75
Soybean_small	4	35	**100.0**	**100.0**	99.17	99.58
Vehicle	4	18	**79.81***	79.72*	77.48*	78.16
Dermatology	6	33	96.74*	98.21	96.63*	**98.26**
Ecoli	6	6	**86.96***	83.04*	85.48	85.71
Glass	6	9	62.57	61.65*	61.93*	**63.76**
Yeast	10	8	52.35*	55.54*	**56.66***	56.27

RBF Kernel

DR:

16 /20

Dataset	Classification accuracy					
	RN	GRBFN	SVM	MR	DR-GV	DR
Ionosphere	89.60*	86.88*	95.11*	98.30*	94.26*	**99.43**
Sonar	82.88*	77.02*	85.00*	92.31*	87.50*	**94.23**
Water	95.59*	91.02*	90.51*	98.31*	98.31*	**99.32**
Wdbc	93.12*	94.28*	94.25*	95.09*	95.51*	**96.60**
Bupa	72.43*	73.64*	73.06*	78.03*	73.29*	**81.73**
Pid	76.25*	77.84	76.56*	**80.73***	77.42*	78.36
Diabetes	77.08*	75.16*	77.08*	77.37*	78.91*	**79.07**
Wine	73.67*	76.11*	77.78*	83.56*	96.45*	**97.56**
Lenses	75.38*	70.00*	79.23*	81.54*	86.15	**87.69**
Tae	52.63*	47.37*	54.34*	58.29*	56.58*	**61.32**
New_thyroid	93.33*	90.83*	96.02	**97.31***	94.81*	96.02
Iris	96.80*	96.80*	98.27	98.67	96.67*	**98.80**
Cmc	55.33*	56.29	56.41	56.36	55.43*	**56.82**
Balance_scale	91.28*	91.21*	**92.04**	91.63	91.25*	91.82
Soybean_small	**100.0**	82.92*	62.50*	**100.0**	**100.0**	**100.0**
Vehicle	73.35*	70.66*	74.76*	73.70*	81.82	**82.61**
Dermatology	97.37*	96.36*	97.28*	98.70	98.53	**98.91**
Ecoli	88.81	89.11	89.17	**89.70***	88.39	88.63
Glass	70.37	65.69*	72.75	**76.24***	70.73	71.65
Yeast	60.56	59.22*	60.58*	**61.57***	60.28*	60.56

Dataset	Classification accuracy boost (DR)	
	Linear Kern	RBF Kernel
Ionosphere	**88.01**	**99.43**
Sonar	76.83	**94.23**
Water	96.27	**99.32**
Wdbc	**96.21**	**96.60**
Bupa	**70.23**	**81.73**
Pid	**78.26**	78.36 (80.73MR)
Diabetes	**78.72**	**79.07**
Wine	**99.00**	**97.56**
Lenses	**84.62**	**87.69**
Tae	**56.58**	**61.32**
New_thyroid	91.11	96.02 (97.31MR)
Iris	87.07	**98.80**
Cmc	51.96	**56.82**
Balance_scale	88.75	91.82 (92.04SVM)
Soybean_small	99.58	**100.0**
Vehicle	78.16	**82.61**
Dermatology	**98.26**	**98.91**
Ecoli	85.71	88.63 (89.70MR)
Glass	**63.76**	71.65 (76.24MR)
Yeast	56.27	60.56 (61.57MR)

Image Recognition (AR base)

	G3/P11	G5/P9	G7/P7
RN	71.73	77.56	92.14
GRBFN	10.73	12.00	24.43
SVM	64.18	71.78	91.43
MR	68.45	72.22	91.00
DR-GV	70.00	78.78	92.00
DR	**74.27**	**80.89**	**93.57**

Coil-20 Image Recognition (Cont'd)

	G9/P63	G18/P54	G36/P36
RN	96.03	97.41	98.19
GRBFN	59.92	66.39	58.47
SVM	96.98	98.98	99.44
MR	97.38	98.33	98.75
DR-GV	98.10	99.07	99.44
DR	**98.33**	**99.17**	**99.72**

USPS Character Recognition

Available at: http://www.cs.toronto.edu/~roweis/data.html

Experimental Results

G10/P1090	Classification accuracy					
	RN	GRBFN	SVM	MR	DR-GV	DR
1 vs. 7	94.77	87.39	95.69	95.73	95.78	**96.97**
2 vs. 3	94.54	94.04	95.69	95.00	94.45	**96.79**
2 vs. 7	96.61	95.83	96.65	96.83	96.38	**97.71**
3 vs. 8	92.57	92.43	92.75	92.98	91.47	**93.58**
4 vs. 7	98.35	94.95	98.62	98.53	98.39	**99.08**

G100/P1000	Classification accuracy					
	RN	GRBFN	SVM	MR	DR-GV	DR
1 vs. 7	99.75	96.90	99.85	**99.95**	99.85	**99.95**
2 vs. 3	98.00	96.45	98.10	98.35	98.15	**98.40**
2 vs. 7	99.55	98.95	**99.70**	**99.70**	99.60	**99.70**
3 vs. 8	97.70	95.85	98.40	97.90	98.20	**98.50**
4 vs. 7	99.30	98.30	**99.70**	99.50	99.40	**99.70**

In Summary

Regularization	Loss Function		Regularization Term			Dependent on the number of classes
	Square-Loss Function	Hinge-Loss Function	$\|f\|_K^2$	$\|f\|_I^2$	$R_{disreg}(f,\eta)$	
RLSC	✓		✓			
LapRLSC	✓		✓	✓		✓
SVM		✓	✓			
LS-SVM	✓	✓	✓			
LapSVM		✓	✓	✓		✓
DR	✓				✓	

Future work

- Generalization error bound;
- Semi-supervised (discriminant) framework;
- Sparse solutions;
- Parameter selection;
- Structured DR framework;
- Applications;

etc.

Thanks a lot!

Q&A