Theoretical Foundations

Dr. Fayyaz ul Amir Afsar Minhas
PIEAS Biomedical Informatics Research Lab
Department of Computer and Information Sciences
Pakistan Institute of Engineering & Applied Sciences
PO Nilore, Islamabad, Pakistan
http://faculty.pieas.edu.pk/fayyaz/

Topics Covered

- Revision of SVM
- Principles of SVM
  - Structural Risk Minimization
- Introducing the family of large margin classifiers
  - Classification
  - Regression
  - Ranking
  - Multi-class prediction
  - Multi-label prediction
  - Structured output prediction

Primal

$$\min_{\mathbf{w},b} \quad \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{N}\zeta_i$$

$$s.t. \quad y_i\left(\mathbf{w}^T\mathbf{x}^{(i)} + b\right) \geq 1 - \zeta_i$$

$$\zeta_i \geq 0$$

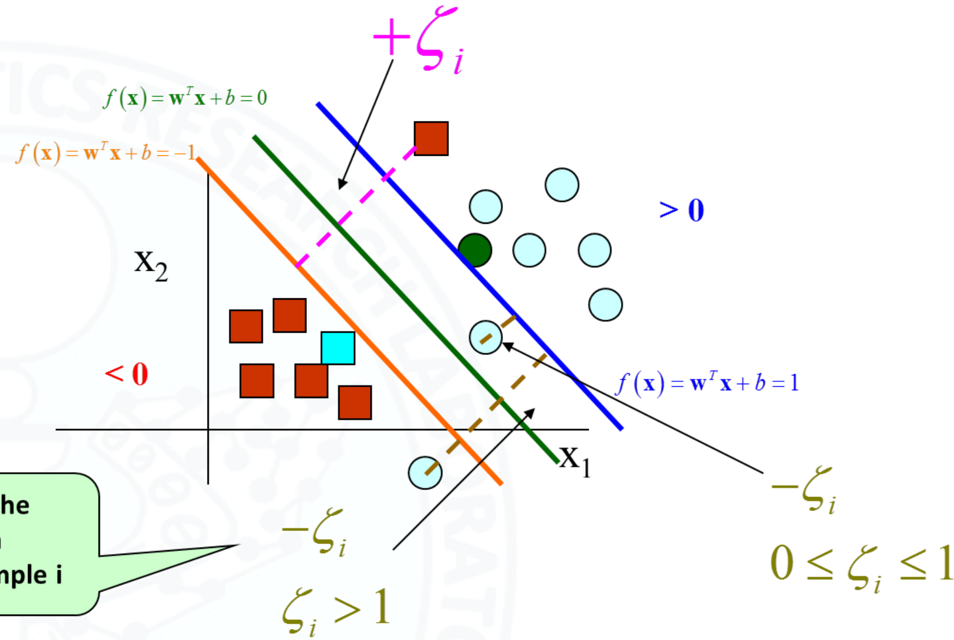
Or

$$\min_{\mathbf{w},b} \quad \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{N}\zeta_i$$

$$s.t. \quad 1 - \zeta_i - y_i\left(\mathbf{w}^T\mathbf{x}^{(i)} + b\right) \leq 0$$

$$-\zeta_i \leq 0$$

ζ_i tells us about the extent of margin violation of example i



- Question: When does an example violate the margin?
  - When: $\zeta_i > 0$
  - Equivalently: $y_i f(\boldsymbol{x_i}) < 1$ or $1 - y_i f(\boldsymbol{x_i}) > 0$
  - Thus, $\zeta_i = \max(0, 1 - y_i f(\boldsymbol{x_i}))$

- What if I remove "C" and multiply λ with $\boldsymbol{w^T w}$?

Dual

- The dual for of the SVM is obtained by substituting the KKT conditions into the primal and inverting the order of the maximization and minimization
  - If the solution exists then, at the optimal point, the value of the primal and the dual are the same

$$\max_{\alpha_i, \beta_i \geq 0} \left\{ \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j \mathbf{x}^{(i)^T} \mathbf{x}^{(j)} \right\}$$

$$\boldsymbol{w} = \sum_{i=1}^{N} \alpha_i y_i \boldsymbol{x}^{(i)}$$

$$s.t \qquad 0 \leq \alpha_i \leq C, \qquad \sum_{i=1}^{N} \alpha_i y_i = 0$$

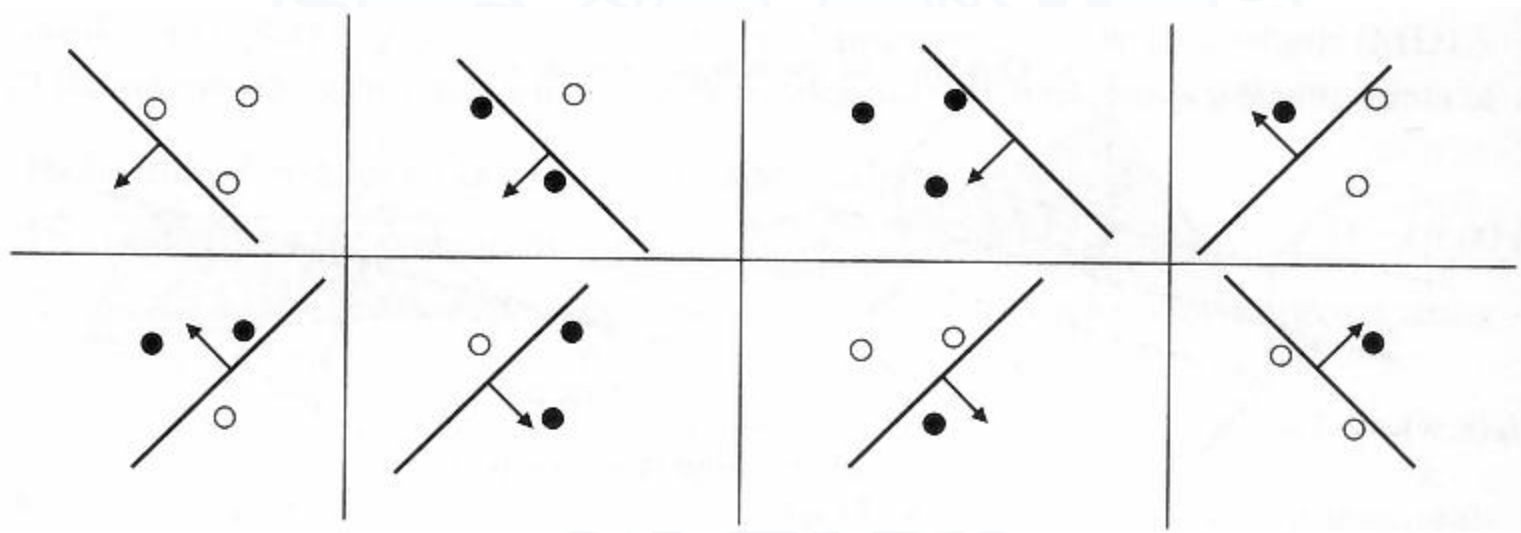
$$f(x) = b + \sum_{i=1}^{N} \alpha_i y_i K\left(\boldsymbol{x}, x^{(i)}\right)$$

C

- Some Observations
  - α_i will be non-zero (positive) only for the points that are support vectors
  - $0 \leq \alpha_i \leq C$
  - C is the weight of the penalty of the term representing margin violation
    - If C is small, then more margin violations will occur
    - If C is large, lesser margin violations will result

VC Dimension

- What’s the maximum number of arbitrarily labeled non-colinear distinct points in space that can always be separated by a linear classifier?

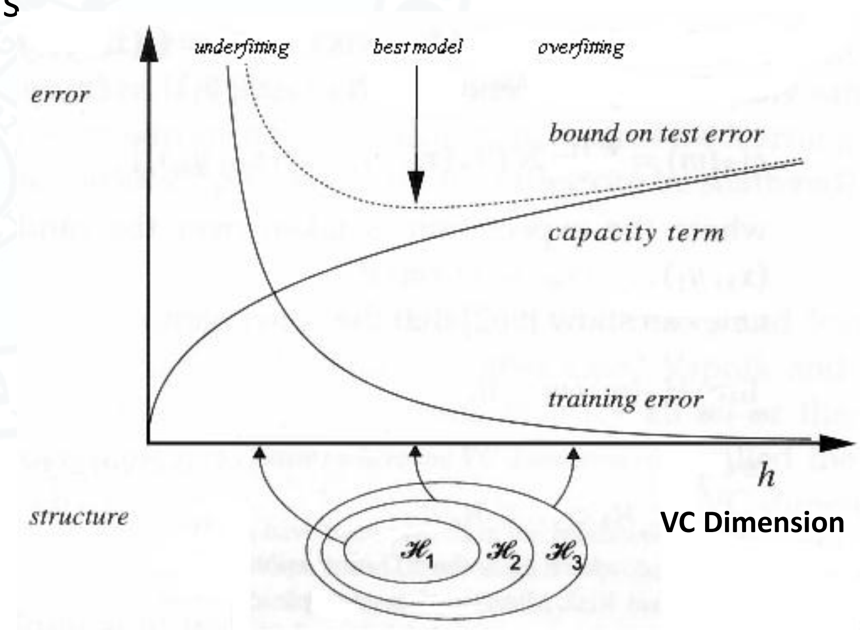


Structural Risk Minimization

- Vapnik and Chervonekis, 1974 showed that

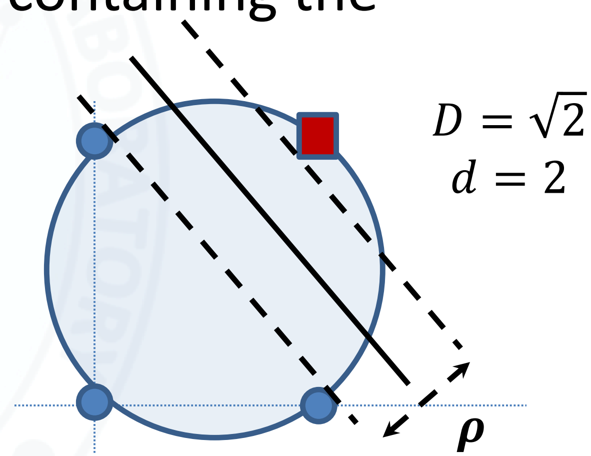
Test Error $\leq$ Training Error + Complexity of set of Models

- To reduce the error on test data we must reduce
  - Training Error
  - Complexity (capacity or freedom) of the model
- However, these are, often, conflicting objectives
- Example
  - Nearest Neighbor
    - Zero Training Error
    - Infinite Complexity/Capacity
  - Linear Classifier
    - Low capacity
    - High training error on data that is not linearly separable



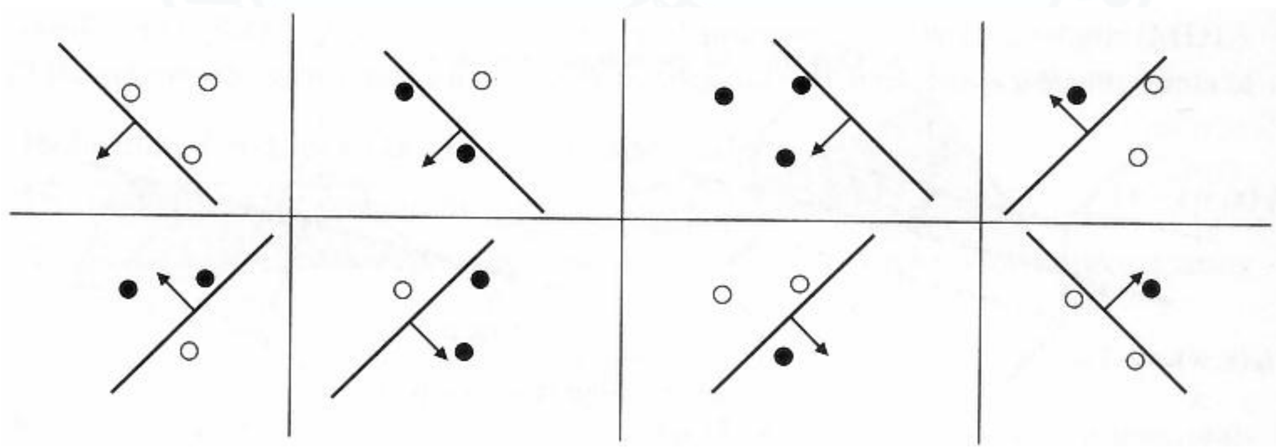
Margin and VC Dimensions

- The VC Dimension of the hyper-plane in d-dimensional space with margin ρ is bounded by:
  - $\mathrm{h}(\rho) \leq min\left(1 + \frac{D^2}{4\boldsymbol{\rho}^2}, d\right)$
  - D is the diameter of the smallest sphere containing the training data points
  - Vapnik (1998, theorem 8.4)
- High VC Dimensions
  - Large d
  - Large D (can be normalized)
  - Small margin
- **This is the basis for the theory for why margin needs to be controlled, especially in high dimensions**



Margin and VC Dimensions

- SVM separates points using a “slab” of a certain width called the margin
- Large margin means
  - Large “slab” Which means
    - Lesser freedom to move the line
      - Lesser points can be shattered
        - » Smaller VC Dimension
  - It’s more difficult to cross a bigger moat by accident



Understanding SRM

Vapnik Proved that

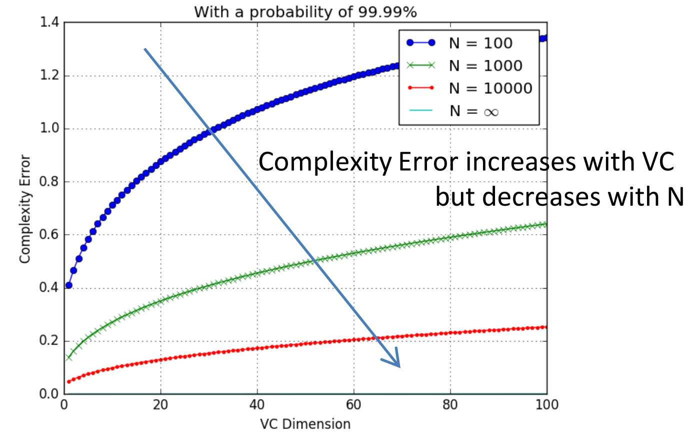
Test Error $\leq$ Training Error + Complexity of set of Models

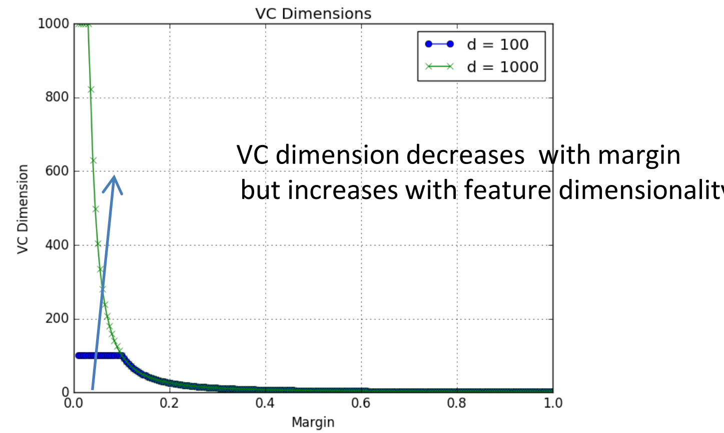
Specifically,

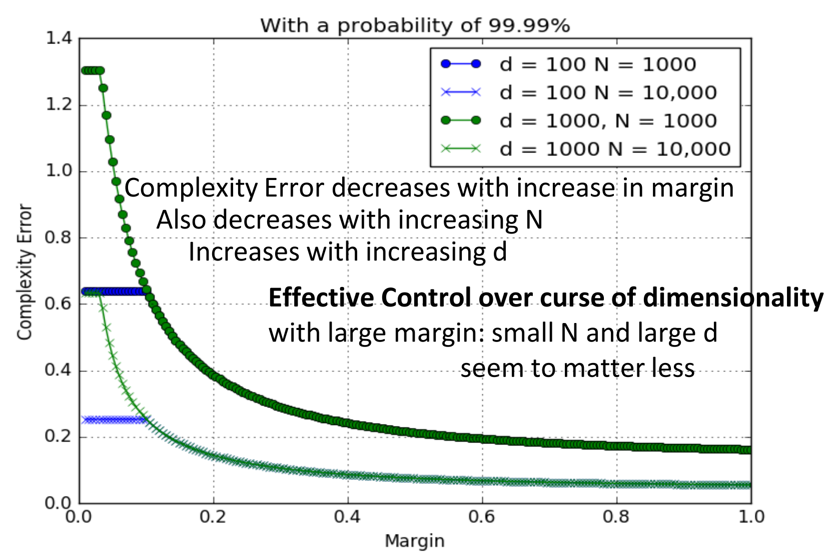
Given N training points, then with probability $1 - p$

$$Test\ Error \leq Training\ Error + \sqrt{\frac{log(2m)+1}{m} - \frac{log(p/4)}{N}}$$

- Where
  - $m = \frac{N}{h}$, $h = VC\ Dimension$
  - For SVM: $h(\rho) \leq min\left(1 + \frac{D^2}{4\rho^2}, d\right)$, d is dimensionality
  - $Training\ Error = \frac{1}{2N}\sum_{i=1}^{N}|y_i - f(\boldsymbol{x_i};\theta)|$
  - $Test\ Error = \frac{1}{2}\int|y - f(\boldsymbol{x};\theta)|\ dP(\boldsymbol{x}, y)$
    - Generalization Error







Structural Risk Minimization

- Thus, to optimize the generalization of a classifier
  - Reduce Training Error
  - Reduce Classifier Complexity
    - By margin maximization
- Mathematically,

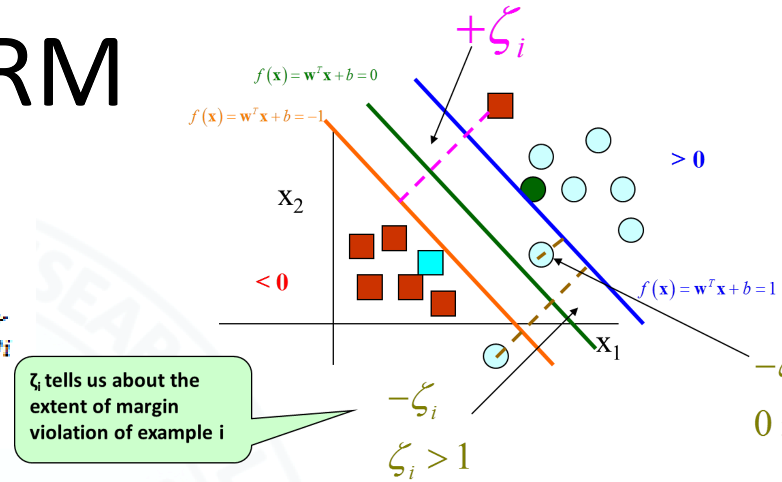
$$f^* = argmin_f \{\underbrace{L(X,Y;f) + \lambda g(\|f\|)}_{\text{Structural Risk}}\}$$

Regularization
Classifier Complexity
(smoothing) term

Empirical Loss (or risk)
term

X, Y is the training data
f is the learning function

SVM and SRM



- For SVM

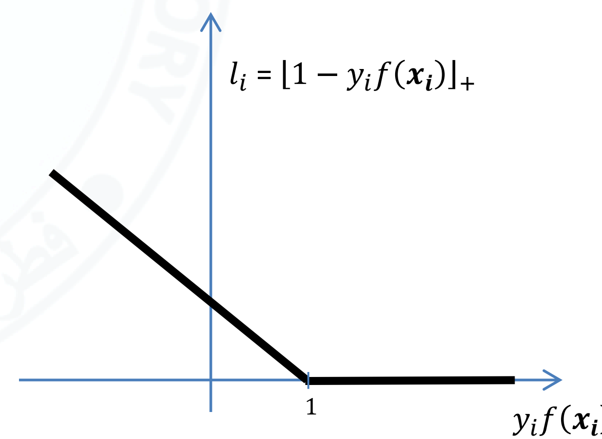
$$\min_{\mathbf{w},b} \quad \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{N}\zeta_i$$

$$s.t. \quad y_i\left(\mathbf{w}^T\mathbf{x}^{(i)} + b\right) \geq 1 - \zeta_i$$

$$\zeta_i \geq 0$$

- We can represent It as:
  - $\boldsymbol{min_{w,b}} \ \frac{\lambda}{2}\|\boldsymbol{w}\|^2 + \sum_{i=1}^{N} \boldsymbol{max}\left(0, 1 - y_i\left(\boldsymbol{w}^T\boldsymbol{x}_i + b\right)\right)$

- SVM can be mapped to the form : $f^* = argmin_f\,\{L(X,Y;f) + \lambda g(\|f\|)\}$
  - $f(\boldsymbol{x}) = \langle \boldsymbol{w}, \boldsymbol{x}\rangle$ is the discriminant function
    - kernel trick allows non-linear classification
  - $L(X,Y;f) = \sum_{i=1}^{N}[1 - y_i\langle \boldsymbol{w}, \boldsymbol{x}_i\rangle]_+$
    - $[\theta]_+ = \theta \ for \ \theta \geq 0 \ and \ 0 \ otherwise$
    - Hinge loss function
  - $g(\|f\|) = \|\boldsymbol{w}\|_2^2$
  - This setting allows for convex optimization
    - Single local and global minima

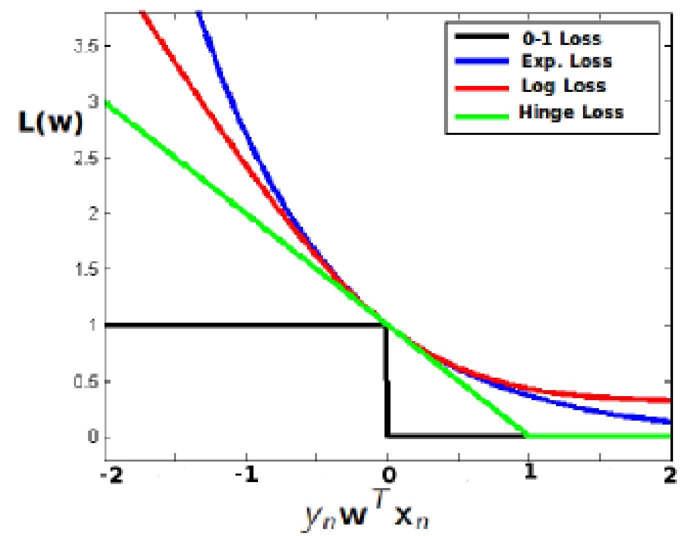


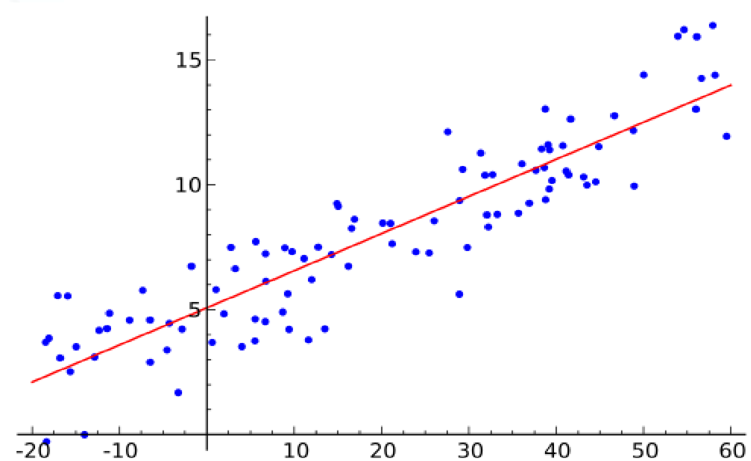
SVM, SRM, Margin and Complexity

- SVMs are learning machines that reduce structural risk (Vapnik, 1961) by
  - Reducing the empirical error $\sum_{i=1}^{N} max\left(0, 1 - y_i(w^T x_i + b)\right)$
  - Controlling complexity of the classifier
    - Through the control on the upper bound on the VC dimension (remember Vapnik, 1995: $\mathbf{VC}(\rho) \leq \frac{D^2}{4\rho^2}$)
      - By maximizing the margin: $\frac{1}{\|w\|^2}$
- The reduction in structural risk minimizes the upper bound on the generalization error (Vapnik and Chervonekis, 1974)
  - This leads to <u>good generalization performance</u>

What can we do with SRM?

- The principal of SRM allows us to develop a family of large margin learning machines by changing its components
- Example
  - SVM: $\boldsymbol{min}_{\boldsymbol{w},\boldsymbol{b}} \ \frac{\lambda}{2}\|\boldsymbol{w}\|^2 + \sum_{\boldsymbol{i}=\boldsymbol{1}}^{\boldsymbol{N}}[1 - y_i f(\boldsymbol{x_i})]_+$
  - Regularized least square regression
    - $\boldsymbol{min}_{\boldsymbol{w},\boldsymbol{b}} \ \frac{\boldsymbol{\lambda}}{\boldsymbol{2}}\|\boldsymbol{w}\|^{\boldsymbol{2}} + \sum_{\boldsymbol{i}=\boldsymbol{1}}^{\boldsymbol{N}}\left(y_i - f(\boldsymbol{x_i})\right)^{\boldsymbol{2}}$
  - Support Vector Regression
    - $\boldsymbol{min}_{\boldsymbol{w},\boldsymbol{b}} \ \frac{\boldsymbol{\lambda}}{\boldsymbol{2}}\|\boldsymbol{w}\|^{\boldsymbol{2}} + \sum_{\boldsymbol{i}=\boldsymbol{1}}^{\boldsymbol{N}}[|y_i - f(\boldsymbol{x_i})| - \epsilon]_+$
  - Feature selection
    - $\boldsymbol{min}_{\boldsymbol{w},\boldsymbol{b}} \ \frac{\lambda}{2}\|\boldsymbol{w}\|_1^2 + \sum_{\boldsymbol{i}=\boldsymbol{1}}^{\boldsymbol{N}}[1 - y_i f(\boldsymbol{x_i})]_+$





Regularizers

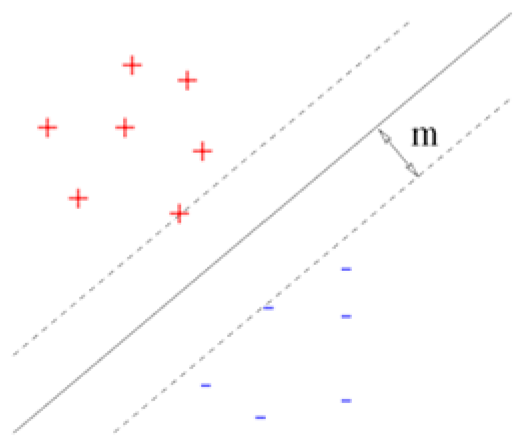
- There are also other regularizers
  - $\|\boldsymbol{w}\|_{\boldsymbol{2}}^{\boldsymbol{2}} = w_1^{\boldsymbol{2}} + w_2^{\boldsymbol{2}} + \cdots + w_d^{\boldsymbol{2}}$
    - Convex, Smooth
  - $\|\boldsymbol{w}\|_{\boldsymbol{1}}^{\boldsymbol{1}} = |\boldsymbol{w_1}| + |\boldsymbol{w_2}| + \cdots + |\boldsymbol{w_d}|$
    - Used for feature reduction
    - "1-norm Support Vector Machine", Zhu et al. (2004)
  - $\|\boldsymbol{w}\|_{\boldsymbol{0}} = \boldsymbol{number\ of\ non - zero\ elements\ in\ w}$
    - Minimization of this norm will lead to feature selection
    - "Use of the Zero-Norm with Linear Models and Kernel Methods", JMLR, Weston et al., (2003)

Generalization Performance bounds

- The leave one out cross validation error of an SVM is bounded as:

$$L.O.O.E. \leq \frac{\min(\#\ support\ vect., D^2/m^2)}{n+1}$$

where D is the diameter of a ball containing all x_i, $i \leq n+1$ and m is the margin of an optimal hyperplane.



A tighter bound on LOO Error: Vapnik 2000

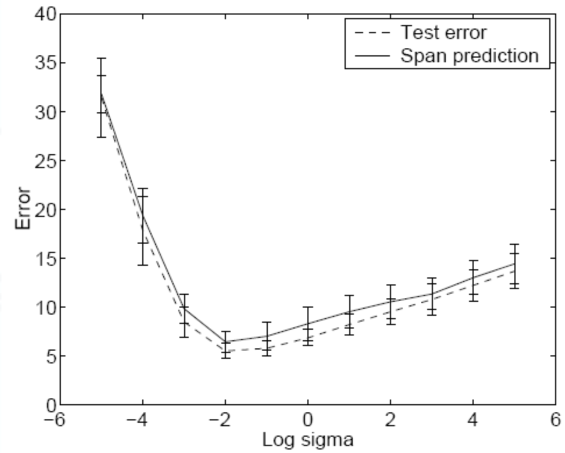
The expectation of the probability of error $p_{error}^{\ell-1}$ for a SVM trained on the training data of size $\ell - 1$ has the bound

$$Ep_{error}^{\ell-1} \leq E\left(\frac{S\max(D, 1/\sqrt{C})\sum_{i=1}^{n^*}\alpha_i^0 + m}{\ell}\right),$$

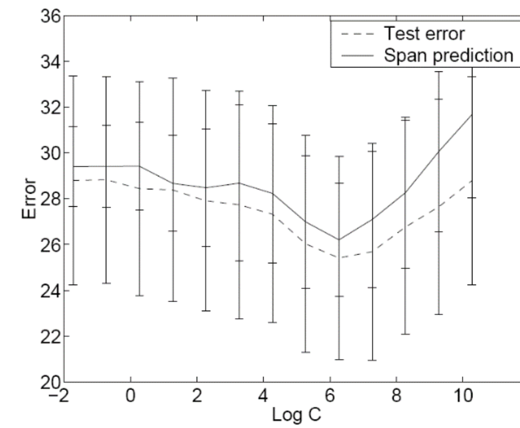
- Error is the leave one out error
- E indicates the expected value
- n = m + n* is the number of support vectors
  - n* = number of support vectors for which $\boldsymbol{0 < \alpha < C}$
  - m = number of support vectors for which $\boldsymbol{\alpha = C}$
- l = number of examples
- S: The span is the largest distance of any support vector from its approximation based on a constrained linear combination of all other support vectors
- D is the diameter of the largest sphere containing the data points
- C is the margin violation parameter

Impact of bounds

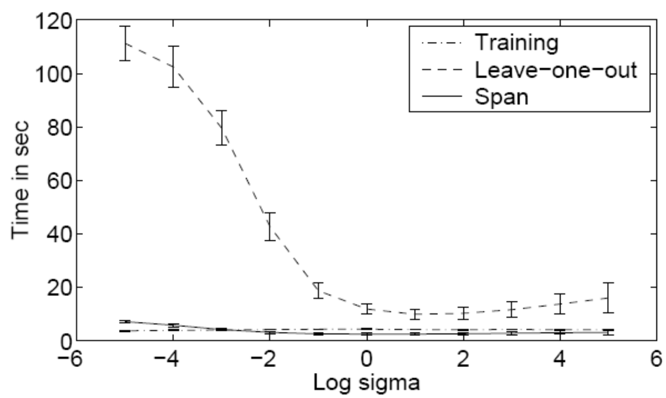
- These bounds can be used for model selection
  - Choosing the value of sigma in the RBF kernel or the ‘C’ parameter
- The time required to compute the span is not prohibitive and is a good alternative to computing the true leave one out error for model selection purposes



(a) choice of σ in the postal database



(b) choice of C in the breast-cancer database



V. Vapnik and O. Chapelle, “Bounds on Error Expectation for Support Vector Machines,” *Neural Comput*, vol. 12, no. 9, pp. 2013–2036, Sep. 2000.

What are kernels?

- Kernels are
  - A way of Data representation
  - Inner Products
  - Measures of similarity
  - Measures of function regularity

Data Representation: Kernels

- Advantages of kernel representation
  - Nonlinear Feature Mapping
  - Always of size n x n
    - Computationally very attractive
  - No need of explicit feature representation
    - Example
      - No obvious way to represent protein sequences as vectors in a biologically relevant way
      - Meaningful pairwise sequence comparison methods exist
      - Is the Local alignment score a kernel?

Representer theorem

- (Schölkopf, LNCS, 2001)
- Any problem represented as follows with a strictly monotonically increasing function 'g' and an arbitrary risk function 'L'

$$f^* = argmin_f \{L(X,Y;f) + \lambda g(\|f\|)\}$$

- Has a solution of the form
  - $f^*(x) = \sum_{i=1}^{N} \alpha_i k(x, x_i)$
  - $k(x, x_i)$ is the kernel function used
- How do SVMs satisfy the Representer theorem?

Representer theorem

- The Representer theorem allows us to represent the decision function of any learning machine with a regularizer and an empirical loss function in terms of the data points alone

$$f^*(x) = \sum_{i=1}^{N} \alpha_i k(x, x_i)$$

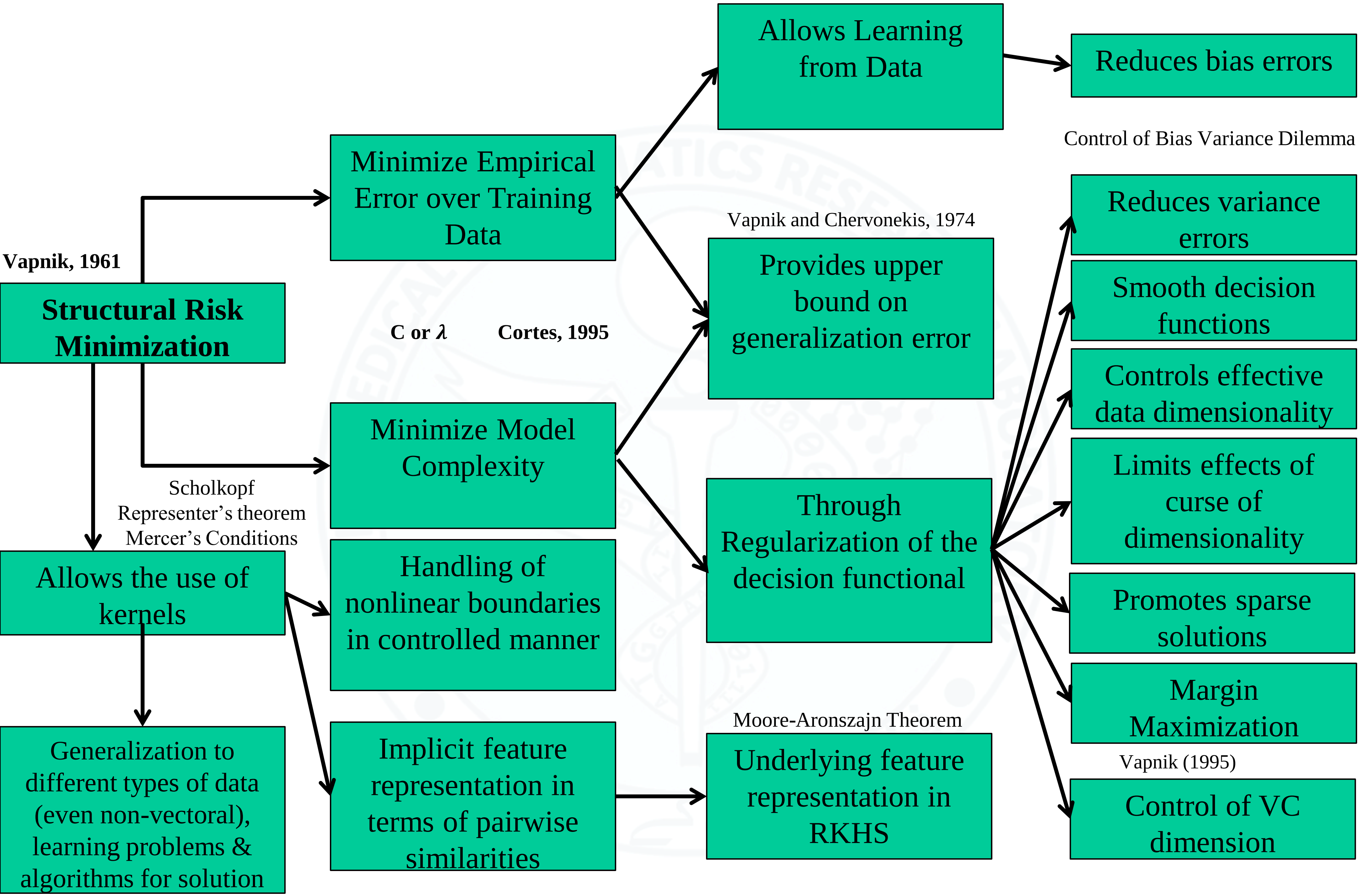
- This, has great implications in solving the optimization problems posed by different learning algorithms

Kernels, Regularization and the Curse of dimensionality

- How does the SVM control the curse of dimensionality?
- We know that we can also express the function as

$$\boldsymbol{f}(\boldsymbol{x}) = \sum_{\boldsymbol{i}=\mathbf{1}}^{\boldsymbol{N}} \boldsymbol{\alpha}_{\boldsymbol{i}} \boldsymbol{K}(\boldsymbol{x}_{\boldsymbol{i}}, \boldsymbol{x})$$

- Thus, $\|\boldsymbol{f}\| = \boldsymbol{\alpha}^{\boldsymbol{T}} \boldsymbol{K} \boldsymbol{\alpha}$
- Now, regularization requires that we minimize $\|\boldsymbol{f}\|$, which will, in essence, produce a small set of non-zero $\boldsymbol{\alpha}$
- The number of positive $\boldsymbol{\alpha}$ controls the number of effective dimensions in the kernel space
- Thus, regularization allows effective control over the curse of dimensionality when using kernels
  - This is one reason why the generalization performance of SVMs is dependent on the number of support vectors


Vapnik, 1961
Structural Risk Minimization
Minimize Empirical Error over Training Data
Allows Learning from Data
Reduces bias errors
Control of Bias Variance Dilemma
Vapnik and Chervonekis, 1974
Provides upper bound on generalization error
C or λ
Cortes, 1995
Minimize Model Complexity
Through Regularization of the decision functional
Reduces variance errors
Smooth decision functions
Controls effective data dimensionality
Limits effects of curse of dimensionality
Promotes sparse solutions
Margin Maximization
Vapnik (1995)
Control of VC dimension
Scholkopf
Representer's theorem
Mercer's Conditions
Allows the use of kernels
Handling of nonlinear boundaries in controlled manner
Generalization to different types of data (even non-vectoral), learning problems & algorithms for solution
Implicit feature representation in terms of pairwise similarities
Moore-Aronszajn Theorem
Underlying feature representation in RKHS

Reasons for using a SVM

- Use of structural risk minimization
  - Reduction of empirical error
  - Reduction of complexity
- Tunable Complexity
- Effects of curse of dimensionality is reduced due to the control over complexity
- Can have linear or non-linear boundaries through kernels
- Kernels allow use of implicit feature representation
- Absence of local minima
- Sparseness of the solution: Not every data point is required – only support vectors determine the solution
- Guaranteed error bounds
- Very flexible to different types of problems in machine learning
  - Multiple instance, ranking, multi-view, regression, clustering, structured output learning …
  - Can be molded to explicitly optimize performance metrics such as AUC, $\boldsymbol{AUC_{\alpha \to \beta}}$
- Allow for large scale learning:
  - O(n) algorithms exist for linear boundaries – n is the number of examples
  - $\boldsymbol{O\left(\frac{d}{\lambda \epsilon}\right)}$ algorithms exist for linear boundaries to achieve an $\boldsymbol{\epsilon}$ accurate solution
    - d is the number of non-zero feature vectors, $\boldsymbol{\lambda}$ is the regularization parameter

Issues with SVMs

- Perhaps the biggest limitation of the support vector approach lies in choice of the kernel."
  Burgess (1998)
- "A second limitation is speed and size, both in training and testing."
  Burgess (1998)
- "Discete data presents another problem..."
  Burgess (1998)
- "...the optimal design for multiclass SVM classifiers is a further area for research."
  Burgess (1998)
- "Although SVMs have good generalization performance, they can be abysmally slow in test phase, a problem addressed in (Burges, 1996; Osuna and Girosi, 1998)."
  Burgess (1998)
- "Besides the advantages of SVMs - from a practical point of view - they have some drawbacks. An important practical question that is not entirely solved, is the selection of the kernel function parameters - for Gaussian kernels the width parameter [sigma] - and the value of [epsilon] in the [epsilon]-insensitive loss function...[more]"
  Horváth (2003) in Suykens et al.
- "However, from a practical point of view perhaps the most serious problem with SVMs is the high algorithmic complexity and extensive memory requirements of the required quadratic programming in large-scale tasks."
  Horváth (2003) in Suykens et al. p 392
- Kernels determine regularization so the choice of a kernel makes the SVM prone to over-fitting

Required Reading

- C. J. C. Burges, "A Tutorial on Support Vector Machines for Pattern Recognition," *Data Min Knowl Discov*, vol. 2, no. 2, pp. 121–167, Jun. 1998.
  - Cited by: 14248
- Optional

- Zhang, ChunHua, YingJie Tian, and NaiYang Deng. "The New Interpretation of Support Vector Machines on Statistical Learning Theory." *Science in China Series A: Mathematics* 53, no. 1 (January 28, 2010): 151–64. doi:10.1007/s11425-010-0018-6.
- Abe, Prof Dr Shigeo. "Variants of Support Vector Machines." In *Support Vector Machines for Pattern Classification*, 163–226. Advances in Pattern Recognition. Springer London, 2010. http://link.springer.com/chapter/10.1007/978-1-84996-098-4_4.
- Section-III: Constructing Kernels in J. Shawe-Taylor and N. Cristianini, Kernel Methods for Pattern Analysis. New York, NY, USA: Cambridge University Press, 2004.
- J. Vert, K. Tsuda, and B. Scholkopf, "A primer on kernel methods," in *Kernel Methods in Computational Biology*, MIT Press, 2004, pp. 35–70.

Assignment (20 Marks)

1. If an example does not violate the margin and does not lie on the margin what will be its α_i? Why? [1]
2. If an example violates the margin, what will be its α_i? Why? [1]
3. If an example lies exactly on the margin, what will be its α_i? Why? [1]
4. What is the relationship between α_i of the dual and $\boldsymbol{\xi_i}$ of the primal for an example $\boldsymbol{i}$? [1]
5. How has the Representer's theorem been used to solve the SVM problem in the primal in the paper given below? [2]

   O. Chapelle, "Training a Support Vector Machine in the Primal," *Neural Comput*, vol. 19, no. 5, pp. 1155–1178, May 2007.
6. What is the Gram matrix? [1]
7. Why should a kernel satisfy the Mercer's conditions? [1]
8. If a kernel doesn't satisfy Mercer's conditions, will the corresponding SVM be convex? [1]
9. What happens we choose $||w||_1$ or $||w||_0$ instead of $||w||_2$ in the optimization? How? [1]
10. What happens if we choose $\sum_i \xi_i^2$ instead of $\sum_i \xi_i$ in the optimization and remove the constraints $\boldsymbol{\xi} \geqslant \mathbf{0}$? [2]
11. How does margin maximization lead to better generalization? Explain in terms of Structural Risk Minimization. [1]
12. How is the Representer Theorem Useful? [1]
13. What are the error bounds of an SVM useful for? [1]
14. How can you create a bias-less SVM Without any loss in accuracy? [1]
15. What is meant by a regularization path? [1]
16. How can you make a kernelized Nearest Neighbor Classifier? [1.5]
17. How can you make a regularized nearest neighbor classifier? [1.5]
18. Why does the generalization error bound increase with increase in the number of support vectors? [1]

End of Lecture

I believe that learning has just started, because whatever we did before, it was some sort of a classical setting known to classical statistics as well. Now we come to the moment where we are trying to develop a new philosophy which goes beyond classical models.

- Vapnik

http://www.learningtheory.org/learning-has-just-started-an-interview-with-prof-vladimir-vapnik/