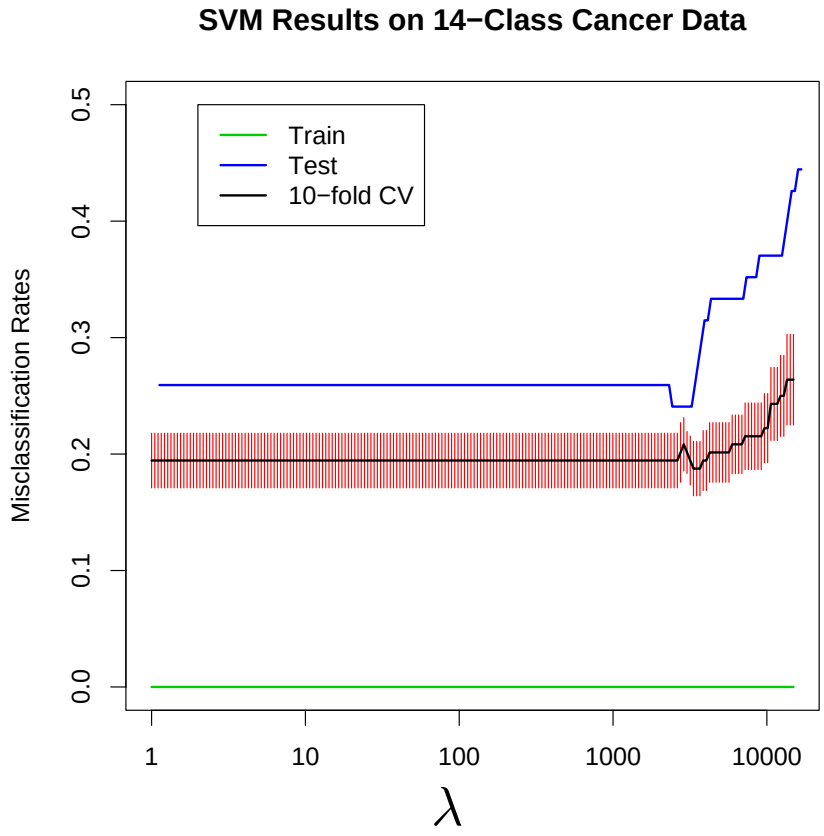
The Entire Regularization Path for the Support Vector Machine

Trevor Hastie

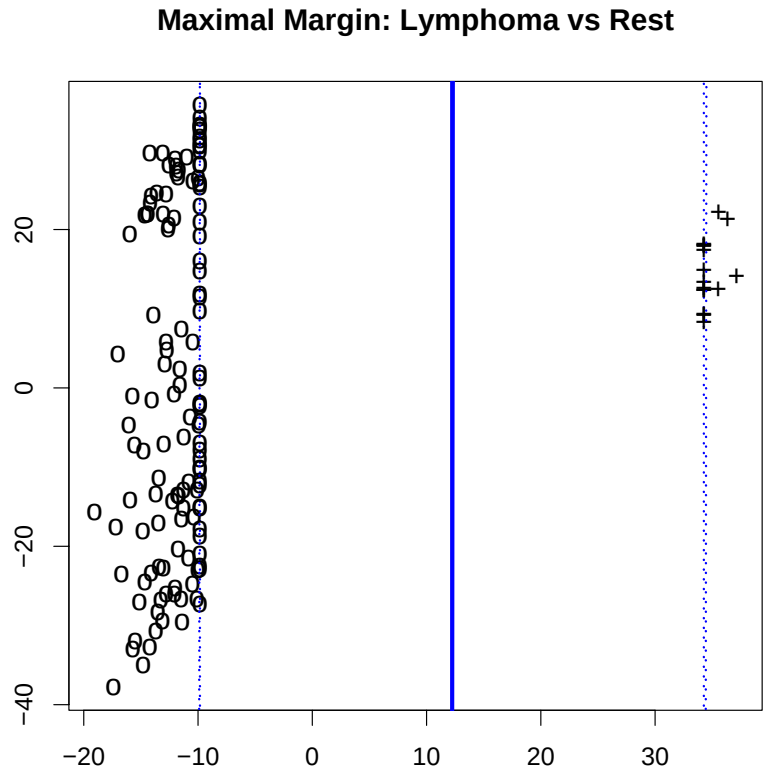
Stanford University

http://www-stat.stanford.edu/∼hastie/Papers

joint work with Saharon Rosset, Ji Zhu, and Rob Tibshirani



- Expression data for 14 cancer classes (Ramaswamy et al, 2001). 144 training observations, 54 test observations, 16000 genes.
- The least regularized SVM solution is not the best.



- When $p \gg N$, maximal margin decision boundaries look "overfitted".
- But this is where SVM's started (and for many have stayed).

Outline

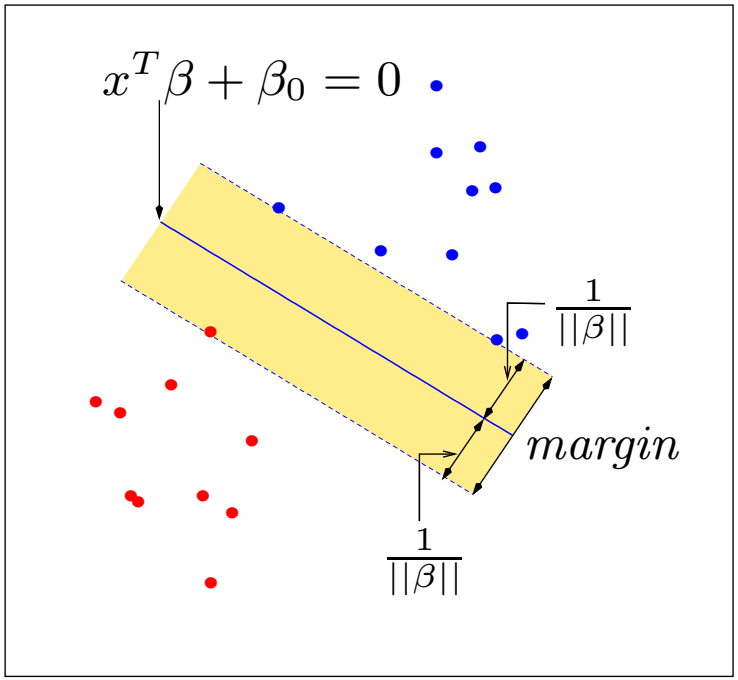
- Optimal margin classifiers from Machine-Learning viewpoint
- An alternative statistical interpretation
- The entire regularization path
- Nonlinear versions via RKHS
- The role and nature of regularization

Maximum Margin Classifier

Boser, Guyon & Vapnik (1992)

Vapnik(1995)

$x_i \in \mathbb{R}^p$, $y_i \in \{-1, 1\}$



$$\min_{\beta,\beta_0} ||\beta||^2$$

$$\text{subject to} \quad y_i(x_i^T\beta + \beta_0) \geq 1,\ i = 1, \ldots, N.$$

Note: $\frac{1}{||\beta||} y_i(x_i^T\beta + \beta_0)$ is distance from x_i to decision boundary.

Signed Distance to Hyperplanes

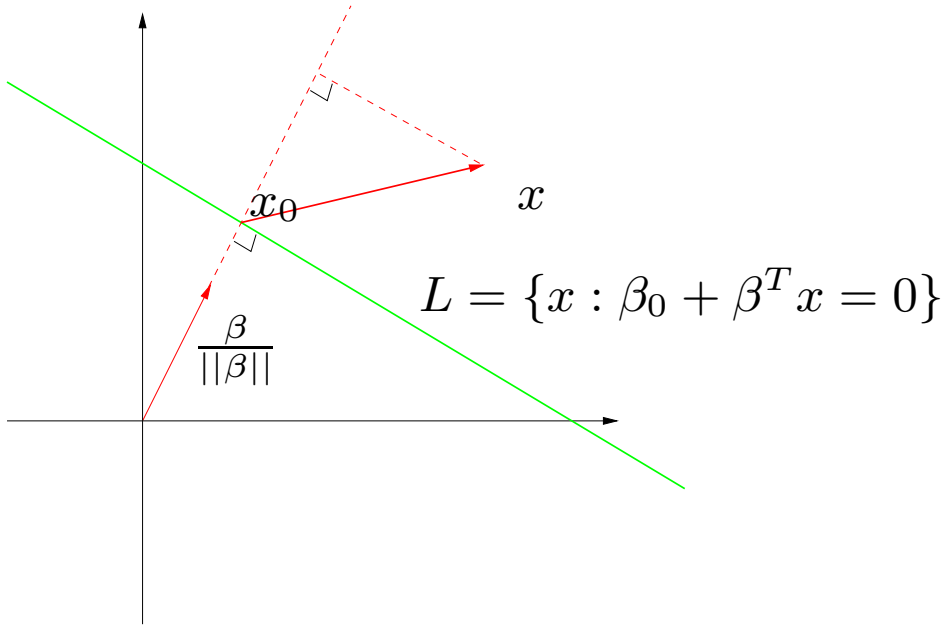
- Hyperplane (affine set) is defined by $L = \{x : f(x) = 0\}$, where $f(x) = \beta_0 + \beta^T x$.
- Signed distance of point x to L is

$$\left\langle \frac{\beta}{||\beta||}, x - x_0 \right\rangle,$$

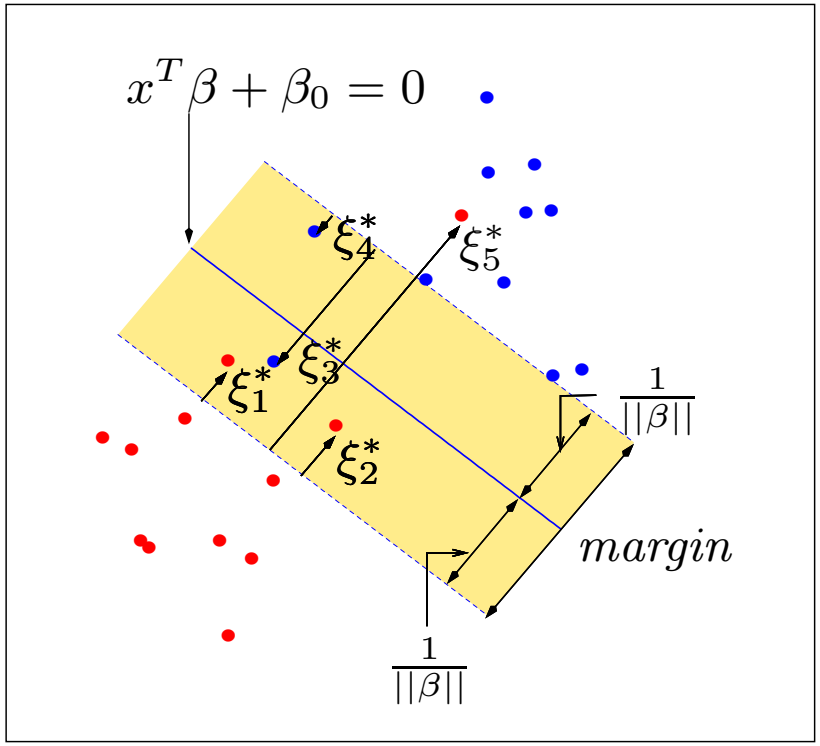
where x_0 is any point in L.

- But this is

$$\frac{1}{||\beta||}(f(x) - f(x_0)) = \frac{1}{||\beta||}(\beta^T x + \beta_0)$$



Overlapping Classes and Soft Margins



$\xi_i^* = \xi_i / ||\beta||$

$$\min_{\beta,\beta_0} ||\beta||^2$$

subject to $y_i(x_i^T\beta + \beta_0) \geq 1 - \xi_i, \;\; \xi_i \geq 0, \;\; \sum_i \xi_i \leq B$

Convex Optimization

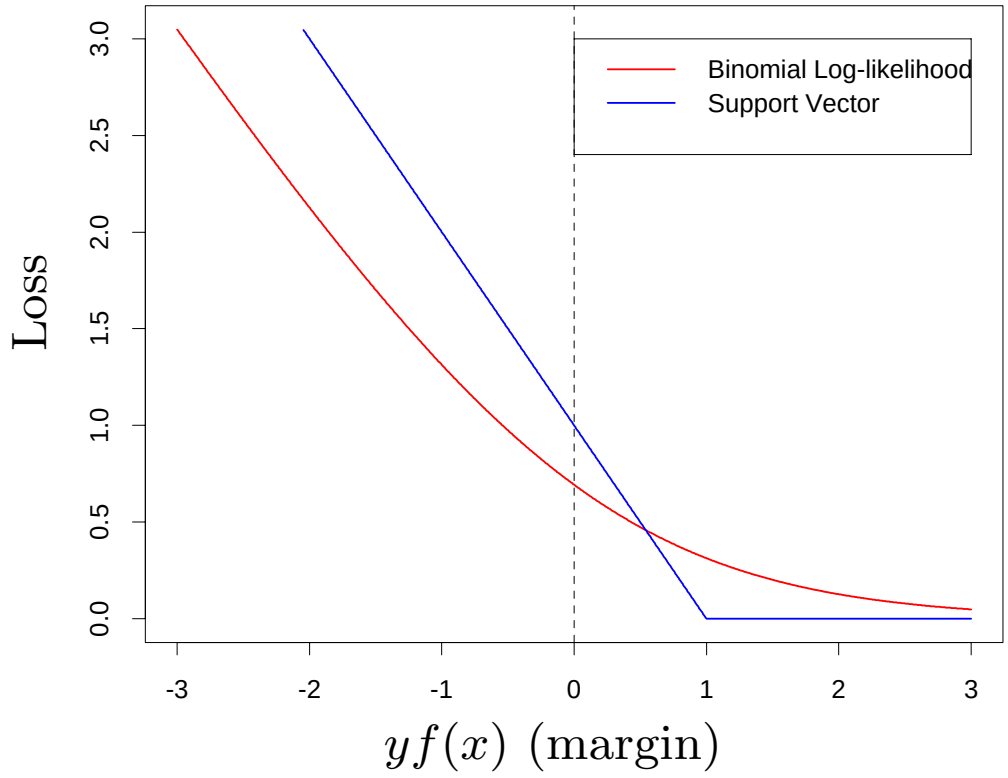
Typically in Machine Learning the linear SVM is formulated as

$$\min_{\beta,\beta_0} \frac{1}{2}\|\beta\|^2 + C\sum_{i=1}^{N}\xi_i$$
$$\text{subject to} \quad \xi_i \geq 0,\ y_i(x_i^T\beta + \beta_0) \geq 1 - \xi_i\ \forall i,$$

Notes:

- $C = C(B)$
- If the data are separable, then for sufficiently large C, we get the maximal margin separator.
- The nature of the regularization via C is not obvious.

SVM via Loss + Penalty



With $f(x) = x^T\beta + \beta_0$ and $y_i \in \{-1, 1\}$, consider

$$\min_{\beta_0,\,\beta} \sum_{i=1}^{N} [1 - y_i f(x_i)]_+ + \frac{\lambda}{2}\|\beta\|^2$$

This hinge loss criterion is equivalent to the SVM, with $\lambda = 1/C$.

Compare with

$$\min_{\beta_0,\,\beta} \sum_{i=1}^{N} \log\left[1 + e^{-y_i f(x_i)}\right] + \frac{\lambda}{2}\|\beta\|^2$$

This is binomial deviance loss, and the solution is "ridged" linear logistic regression.

Quadratic Programming

$$L_P : \sum_{i=1}^{N} \xi_i + \frac{\lambda}{2}\beta^T\beta + \sum_{i=1}^{N} \alpha_i(1 - y_i f(x_i) - \xi_i) - \sum_{i=1}^{N} \gamma_i \xi_i$$

$$\frac{\partial}{\partial \beta} : \qquad \beta = \frac{1}{\lambda}\sum_{i=1}^{N} \alpha_i y_i x_i$$

$$\frac{\partial}{\partial \beta_0} : \qquad \sum_{i=1}^{N} y_i \alpha_i = 0,$$

along with the KKT conditions

$$\begin{aligned} \alpha_i(1 - y_i f(x_i) - \xi_i) &= 0 \\ \gamma_i \xi_i &= 0 \\ 1 - \alpha_i - \gamma_i &= 0 \end{aligned}$$

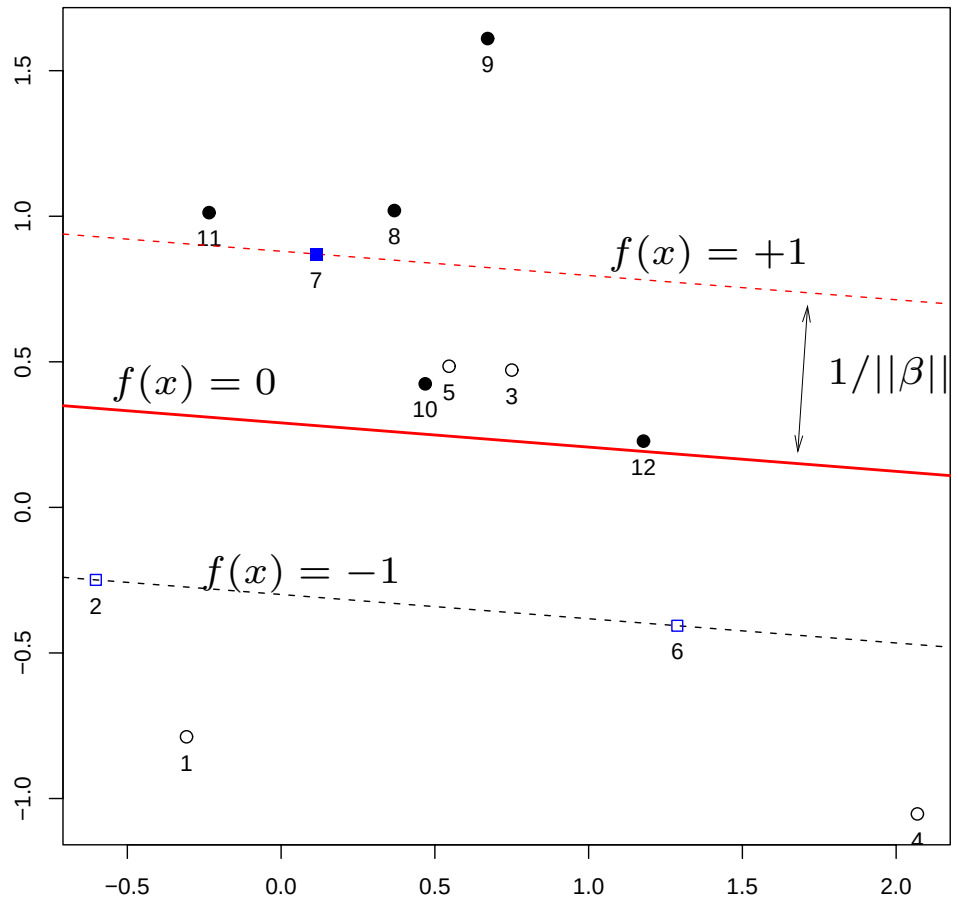
Implications of the KKT conditions

Observations are in one of three states:

- $\mathcal{L} = \{i : y_i f(x_i) < 1,\ \alpha_i = 1\}$, $\mathcal{L}$ for Left of the elbow
- $\mathcal{E} = \{i : y_i f(x_i) = 1,\ 0 \le \alpha_i \le 1\}$, $\mathcal{E}$ for Elbow
- $\mathcal{R} = \{i : y_i f(x_i) > 1,\ \alpha_i = 0\}$, $\mathcal{R}$ for Right of the elbow

- Start with λ large, and the margin very wide. All $\alpha_i = 1$ (if $N_+ = N_-$). As $\lambda \downarrow 0$, the margin gets narrower.
- For the narrowing margin to pass through a point, it's α has to change from 1 to 0 (or from 0 to 1). While this is happening, the point has to linger on the margin. Hence the point moves from $\mathcal{L}$ to $\mathcal{R}$ via $\mathcal{E}$.
- The condition $\sum_i y_i \alpha_i = 0$ demands a certain balance on opposite margins.

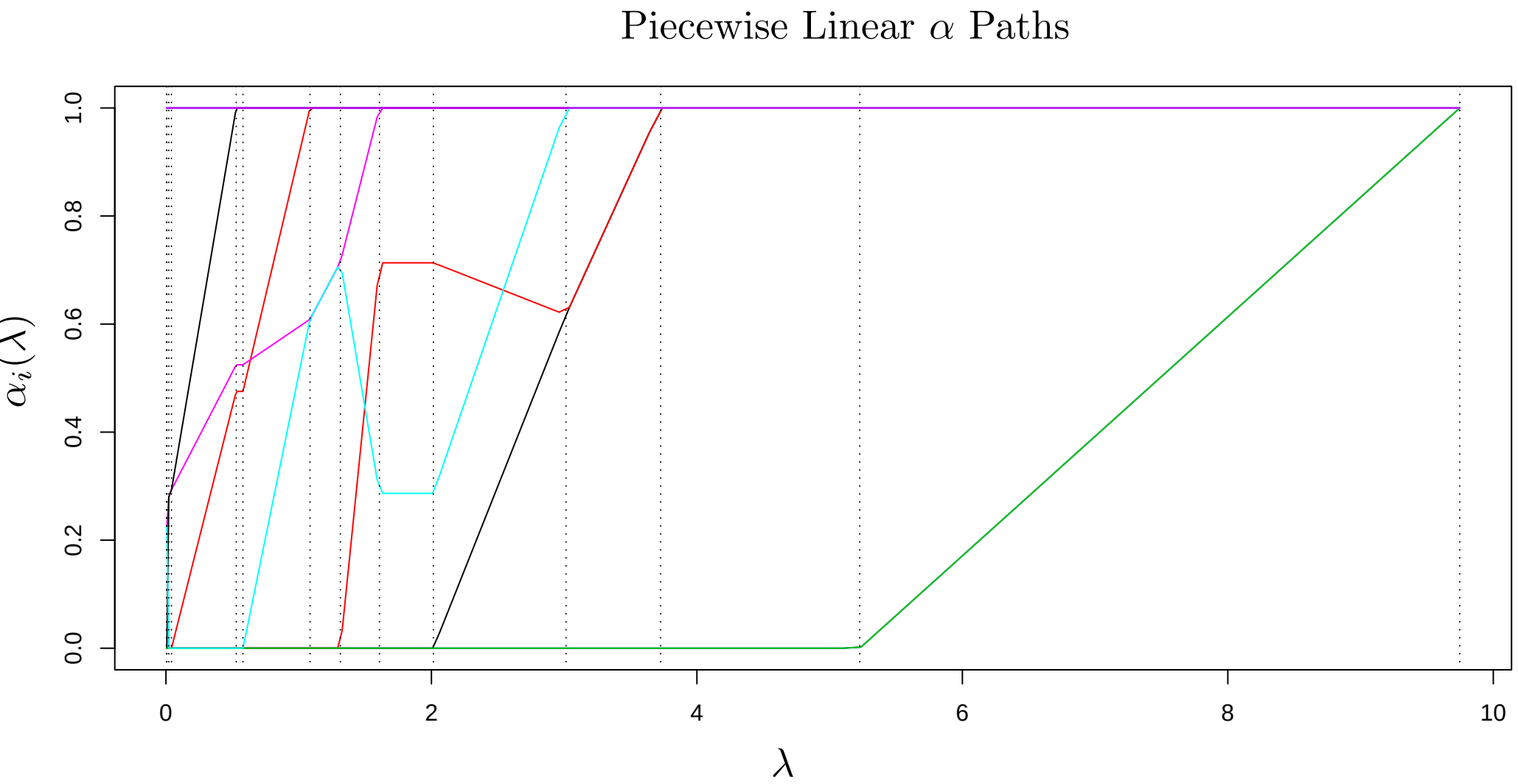
Example

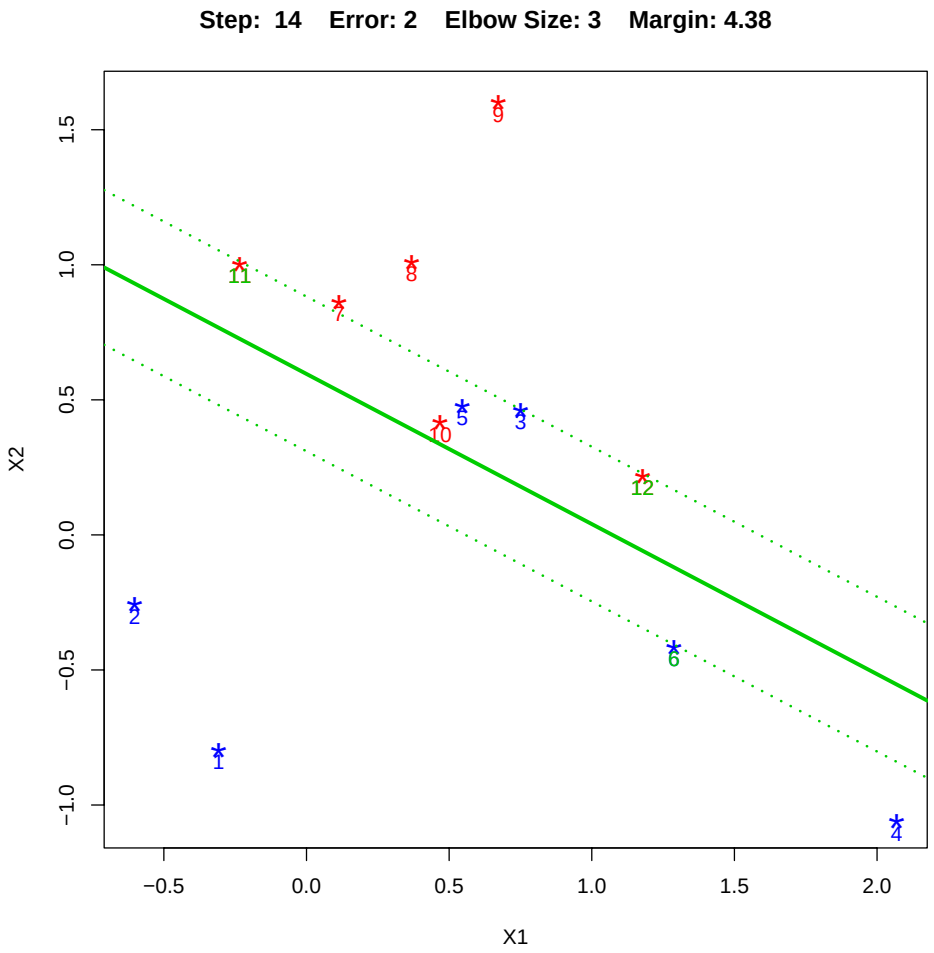


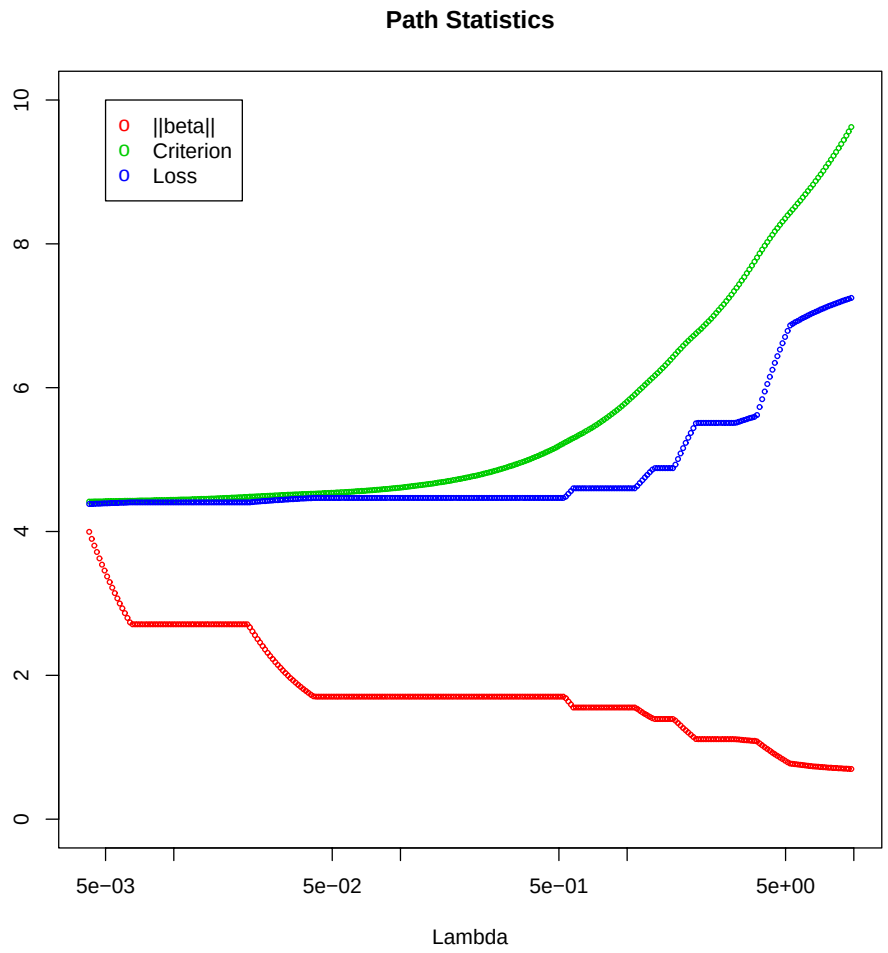
- $\lambda = 0.5$, and the width of the soft margin is $2/||\beta|| = 2 \times 0.587$.
- Two hollow points $\{3, 5\}$ are misclassified, while the two solid points $\{10, 12\}$ are correctly classified, but on the wrong side of their margin $f(x) = +1$; each of these has $\xi_i > 0$.
- The three square shaped points $\{2, 6, 7\}$ are exactly on the margin.

The Path

- The α_i are piecewise-linear in λ (or $1/C$) [MOVIES].
- The points in $\mathcal{E}$ characterize these paths, since points must stay on the margin ($y_i f(x_i) = 1$) while their α_i lie in $(0, 1)$.
- Points can revisit the margin more than once.
- The coefficients β_0 and β are piecewise-linear in $C = 1/\lambda$. Recall LARS (Efron et. al.2002): quadratic criterion, L_1 constraint.
- The margins can stay wedged while their α_i change, if they are "loaded to capacity".
- For non-separable data, the loss $\sum_i \xi_i$ achieves a minimum value, with a positive margin.


Piecewise Linear α Paths
$\alpha_i(\lambda)$
0.0
0.2
0.4
0.6
0.8
1.0
0
2
4
6
8
10
λ


Step: 14 Error: 2 Elbow Size: 3 Margin: 4.38
X2
X1
−1.0
−0.5
0.0
0.5
1.0
1.5
−0.5
0.0
0.5
1.0
1.5
2.0


Path Statistics
||beta||
Criterion
Loss
Lambda
0
2
4
6
8
10
5e−03
5e−02
5e−01
5e+00

Kernels

SVMs also fit nonlinear models in function spaces H_K generated by p.s.d. kernels $K(x, z)$ (RKHS). The radial kernel is very popular:

$$K(x, z) = \exp(-\gamma||x - z||^2)$$

- $H_K = \text{span}_{z \in \mathbb{R}^p}\{K(x, z)\}$
- H_K is endowed with a norm $||f||_{H_K}$; for most kernels this behaves like a roughness functional.
- This leads to the more general SVM problem

$$\min_{\beta_0, g \in H_K} \sum_{i=1}^{N} [1 - y_i f(x_i)]_+ + \frac{\lambda}{2}\|g\|^2_{H_K},$$

with $f(x) = \beta_0 + g(x)$.

Kernels and Solution Paths

- Results from RKHS show that the solution is finite dimensional, with $g(x) = \sum_{j=1}^{N} \theta_j K(x, x_j)$.
- With $\mathbf{K}$ the $N \times N$ gram matrix with elements $K(x_i, x_j)$, the criterion becomes

$$\min_{\beta_0, \boldsymbol{\theta} \in \mathbb{R}^p} \sum_{i=1}^{N} [1 - y_i f(x_i)]_+ + \frac{\lambda}{2} \boldsymbol{\theta}^T \mathbf{K} \boldsymbol{\theta},$$

- Can show that $\hat{\theta}_j = \alpha_j y_j / \lambda$, and hence

$$g(x) = \frac{1}{\lambda} \sum_{j=1}^{N} \alpha_j y_j K(x, x_j),$$

  with the KKT conditions the same as before.
- The α path algorithms work just like before [MOVIE].

Kernels and Feature Spaces

The popular ML view is that the kernel provides an implicit map $h(x) : \mathbb{R}^p \mapsto \mathbb{R}^M$, into a high-dimensional feature space.

- The kernel computes inner-products in this space:

$$K(x, z) = \langle h(x), h(z) \rangle.$$

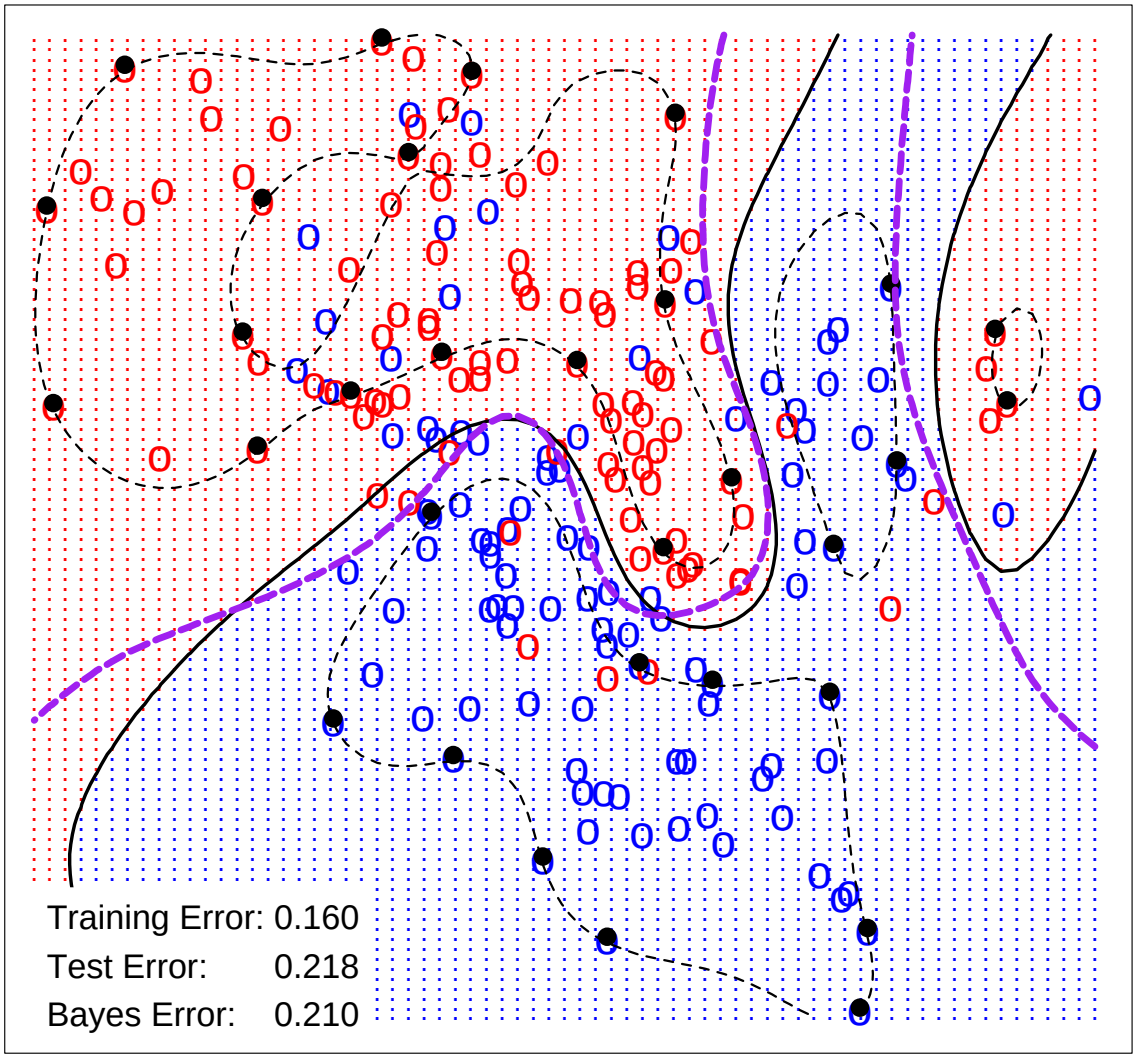
- A linear SVM in this feature space would have

$$\beta = \frac{1}{\lambda} \sum_{j=1}^{N} \alpha_j y_j h(x_j), \qquad \text{and hence}$$

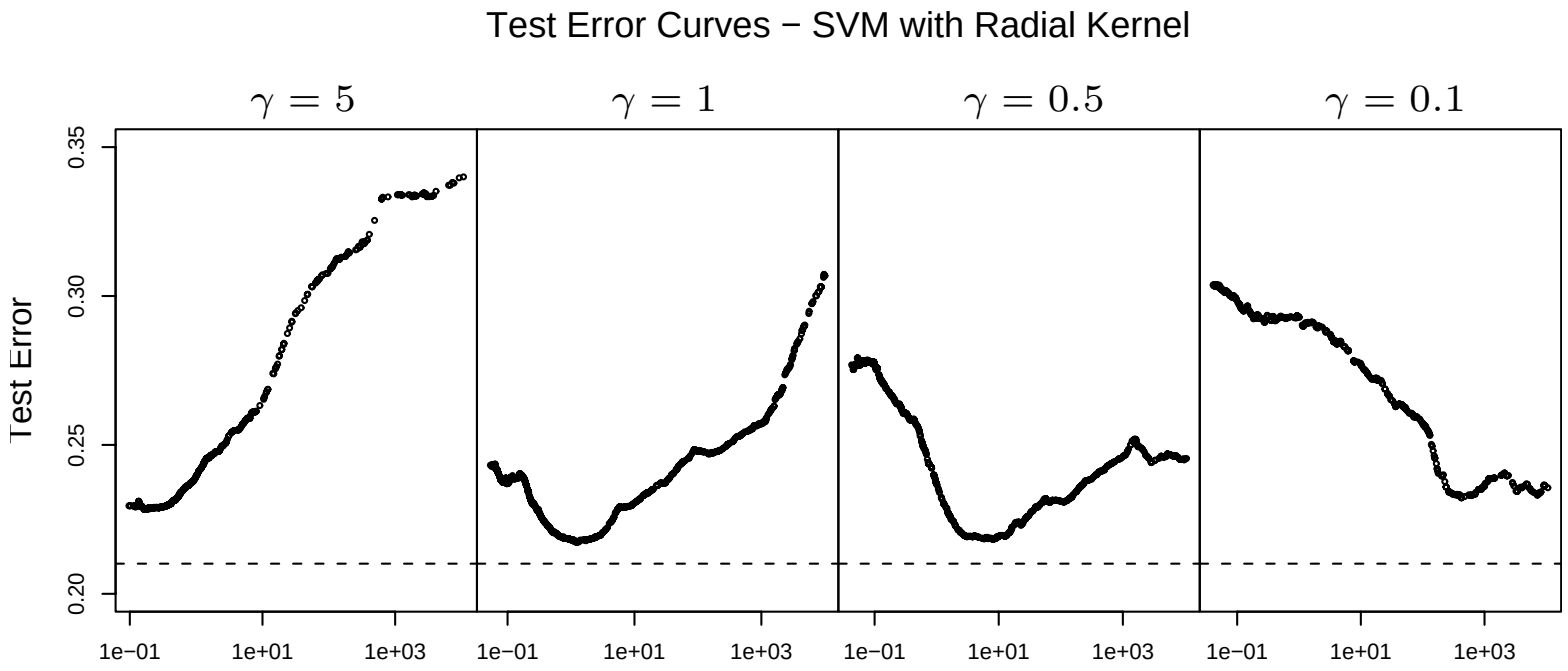
$$\beta^T x = g(x) = \frac{1}{\lambda} \sum_{j=1}^{N} \alpha_j y_j \langle h(x_j), h(x) \rangle = \frac{1}{\lambda} \sum_{j=1}^{N} \alpha_j y_j K(x_j, x)$$

- This hides the nature of the regularization in this feature space.

Mixture Example — $C = 1$, $\gamma = 1$

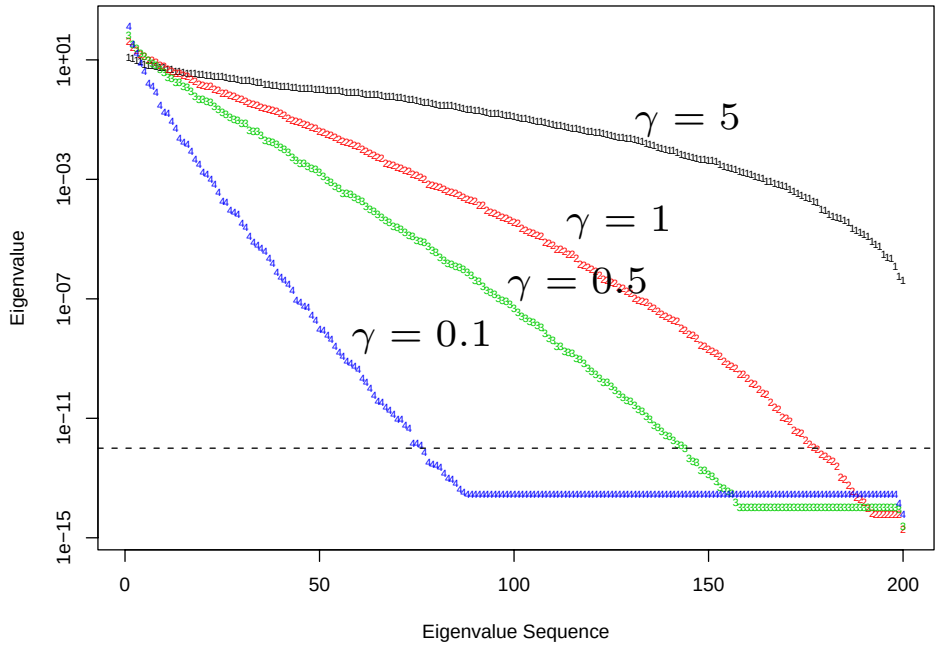


The Need for Regularization



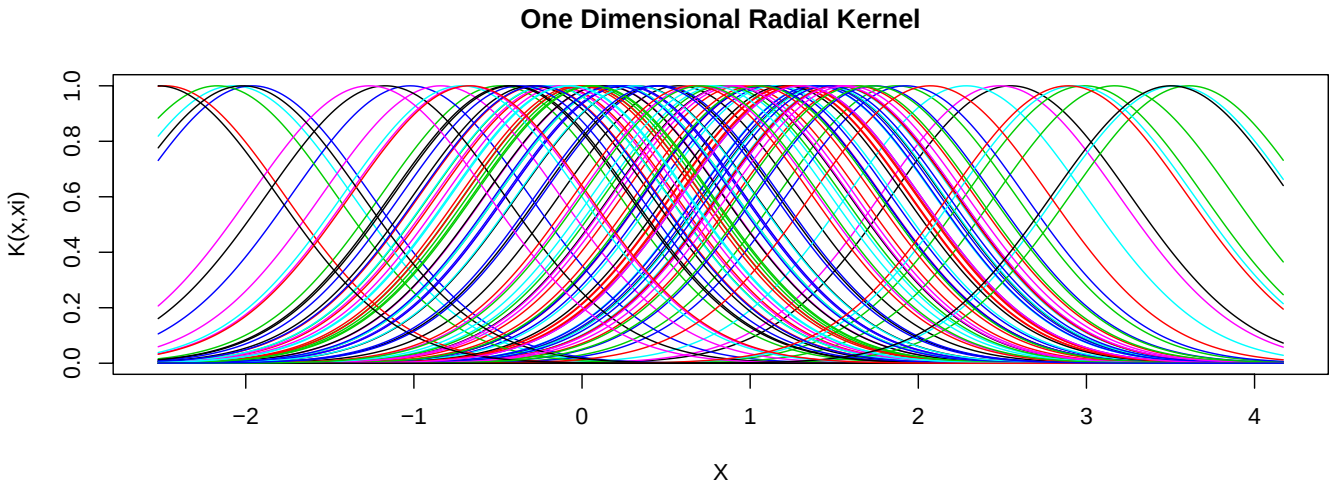
- γ is a kernel parameter: $K(x, z) = \exp(-\gamma||x - z||^2)$.
- λ (or C) are regularization parameters, which have to be determined using some means like cross-validation.

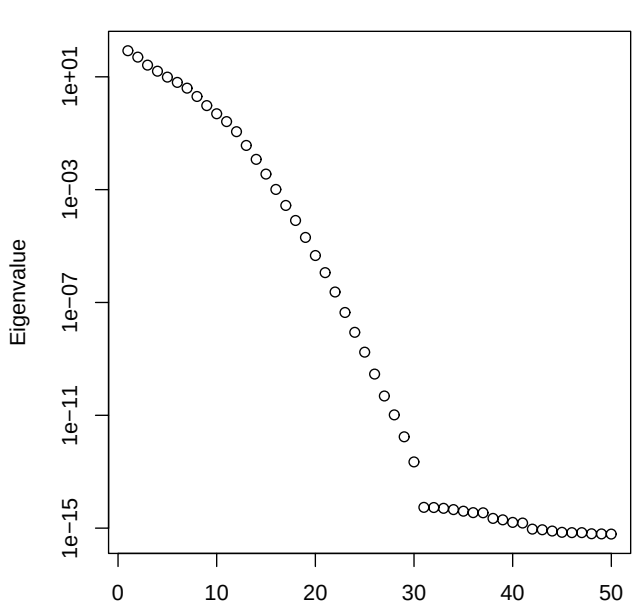
γ	5	1	0.5	0.1
Training Errors	0	12	21	33
Effective Rank	200	177	143	76

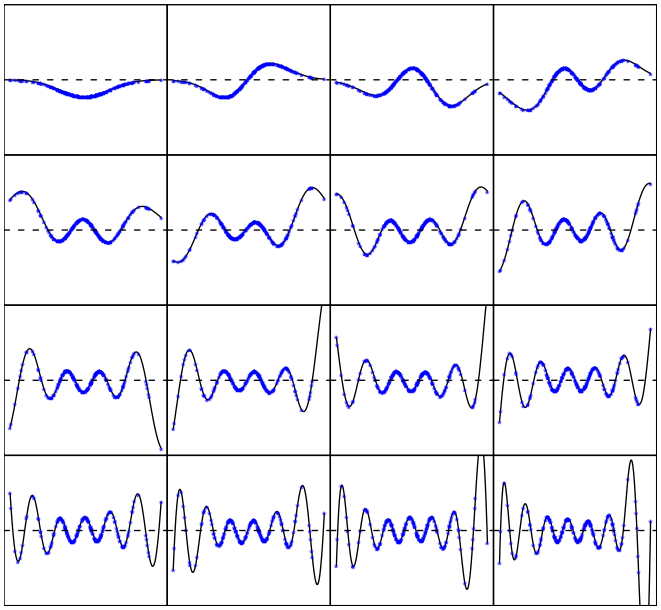


The role of γ

- Commonly thought that radial kernel induces an infinite dimensional feature space.
- As we see here, the feature space has effective dimension typically less than N.
- Many of the "features" (eigenvectors of $\mathbf{K}$), are squashed down dramatically by their eigenvalues.


One Dimensional Radial Kernel
K(x,xi)
0.0
0.2
0.4
0.6
0.8
1.0
−2
−1
0
1
2
3
4
X


Eigenvalue
1e+01
1e−03
1e−07
1e−11
1e−15
0
10
20
30
40
50



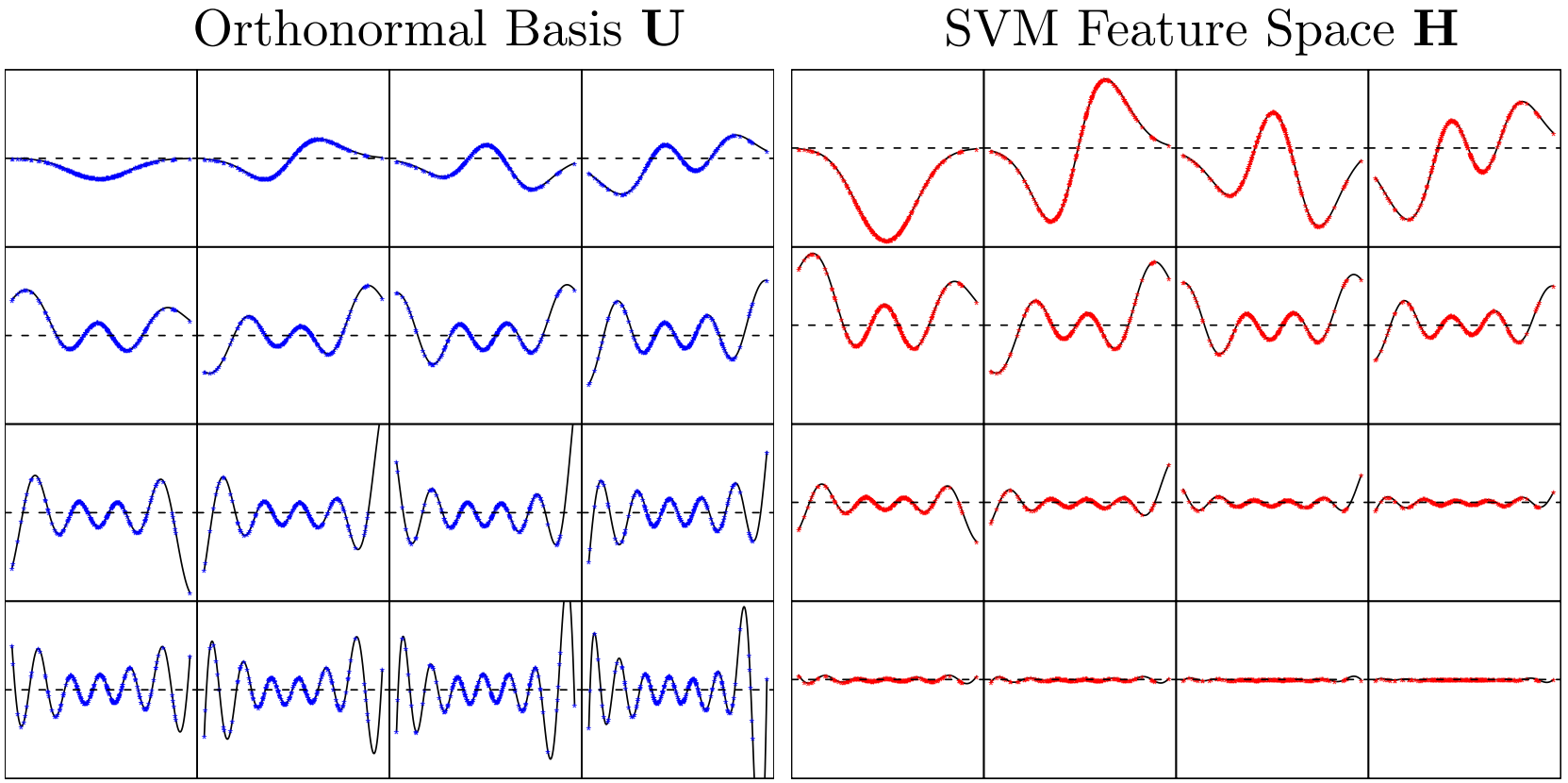
The Nature of the Regularization

$$\min_{\beta_0,\boldsymbol{\theta}} L[\mathbf{y}, \mathbf{K}\boldsymbol{\theta}] + \frac{\lambda}{2}\boldsymbol{\theta}^T\mathbf{K}\boldsymbol{\theta}$$

- reparametrize using the eigen-decomposition of $\mathbf{K} = \mathbf{U}\mathbf{D}\mathbf{U}^T$. Let $\mathbf{K}\boldsymbol{\theta} = \mathbf{U}\boldsymbol{\theta}^*$ where $\boldsymbol{\theta}^* = \mathbf{D}\mathbf{U}^T\boldsymbol{\theta}$.
- Equivalent criterion is

$$\min_{\beta_0,\boldsymbol{\theta}^*} L[\mathbf{y}, \mathbf{U}\boldsymbol{\theta}^*] + \frac{\lambda}{2}\boldsymbol{\theta}^{*T}\mathbf{D}^{-1}\boldsymbol{\theta}^*.$$

- The coefficients θ_j^* of the unit norm features (columns $\mathbf{u}_j$ of $\mathbf{U}$) are differentially penalized (small d_j get penalized more)
- The $h(x)$ parametrization (recall $K(x,z) = \langle h(x), h(z)\rangle$) corresponds to $\mathbf{H} = \mathbf{U}\mathbf{D}^{\frac{1}{2}}$
- Hence $||\mathbf{h}_j|| = d_j^{\frac{1}{2}}$ — many of the h_j are norm challenged!


Orthonormal Basis **U**
SVM Feature Space **H**

A Curiosity at the Most Regularized End

In most regularized setting, kernel SVM behaves like kernel density classification! This is easier to see in the "no intercept" case.

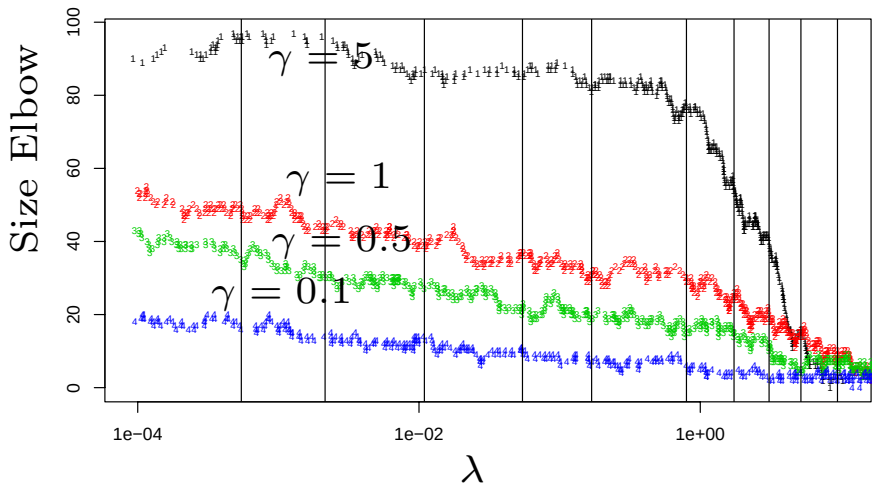
- When λ is large, all $\alpha_i = 1$.

$$\begin{aligned} f(x) &= \frac{1}{\lambda}\sum_{i=1}^{N}\alpha_i y_i K_\gamma(x, x_i) \\ &= \frac{n}{\lambda}\cdot\left(\frac{n_+}{n}\cdot\frac{1}{n_+}\sum_{j\in\mathcal{I}_+}K_\gamma(x, x_j) - \frac{n_-}{n}\cdot\frac{1}{n_-}\sum_{j\in\mathcal{I}_-}K_\gamma(x, x_j)\right) \\ &\propto \hat{\pi}_+\hat{d}_+(x) - \hat{\pi}_-\hat{d}_-(x). \end{aligned}$$

- γ is now a regularization parameter.
- With a "linear" kernel, this becomes nearest-centroid classification.

Computations

- The computations at step ℓ are $O(m_\ell^3 + N m_\ell)$, where $m_\ell = |\mathcal{E}_\ell|$.
- If the average size of the elbow is m, with updating this reduces to $O(m^2 + Nm)$ per average step.



- The algorithm seems to take $O(cN)$ steps, where $c \approx 4 - 6$ in our experience, hence overall the algorithm costs about $O(cN^2 m)$, which is the same order as the SVM (single fit).
- Benchmark experiments suggest that the cost is about a factor of 1.5 times that for a typical single fit using `libsvm`.

Related Work and Caveats

- So far experience with relatively small data-sets; $N \leq 1000$; not sure how well algorithm scales. Do steps get too many and too small?
- Related work in literature: Fine and Scheinberg (2002), Diehl and Cauwenberghs (2003). Use active set methods for online learning, and also for parameter perturbation.
- The Fine et. al. algorithms work on medium to large scale problems, so there is hope for scalability.

Summary

- Regularization (choice of C or λ) can be crucial for SVM, as well as kernel parameters like γ.
- Regularization and its role is under-appreciated in SVM circles.
- Fitting the entire path makes regularization automatic.
- `library(svmpath)` soon to appear in R.