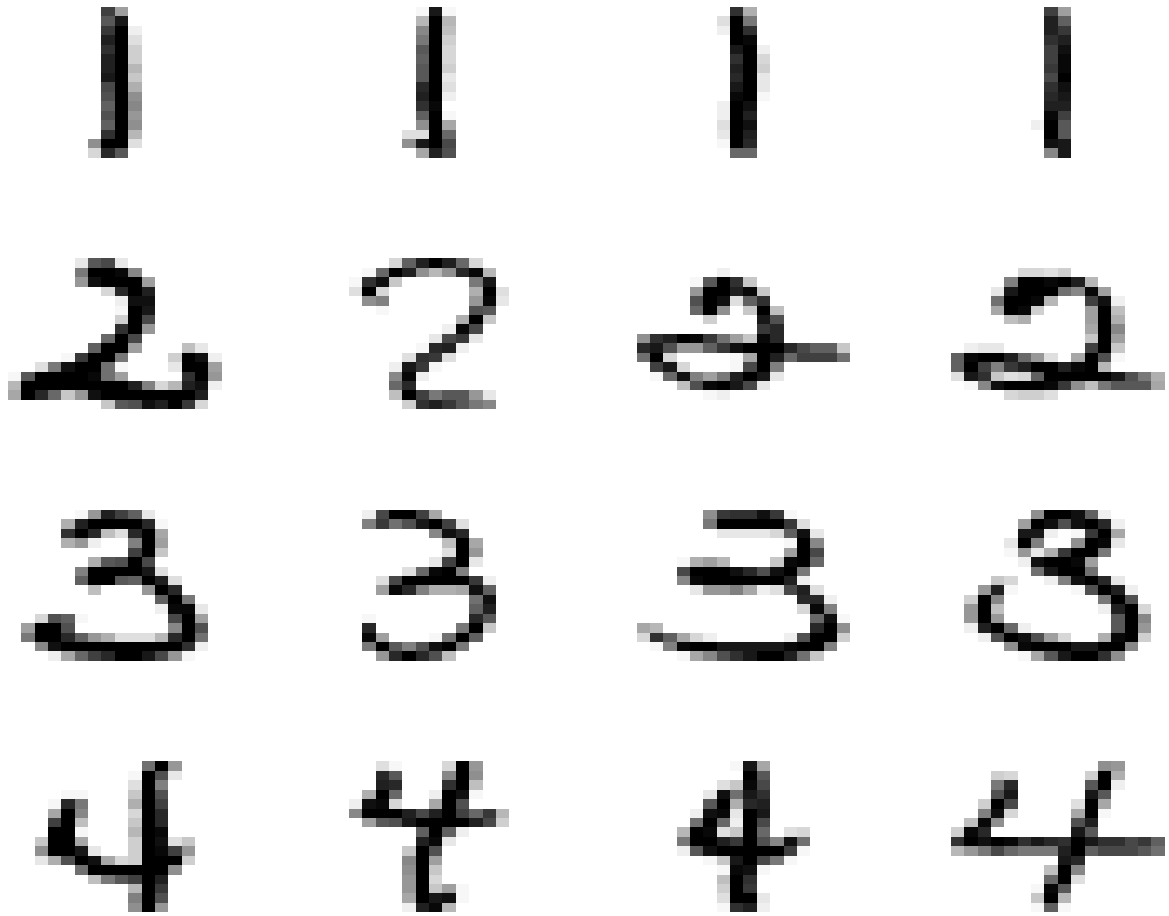
Statistical learning in regularization framework

Yoonkyung Lee
Department of Statistics
The Ohio State University
http://www.stat.ohio-state.edu/∼yklee

Handwritten Digit Recognition

16×16 grayscale images scanned from envelopes



Filtering spam e-mail

Want to predict whether a given email is spam or not.

- percentage of words in the e-mail that match a word: `remove`, `free`, `money`
- percentage of characters in the e-mail that match a character: `$`, `!`
- total number of capital letters in the e-mail
- average length of uninterrupted sequences of capital letters

Cancer Diagnosis with Microarray Data

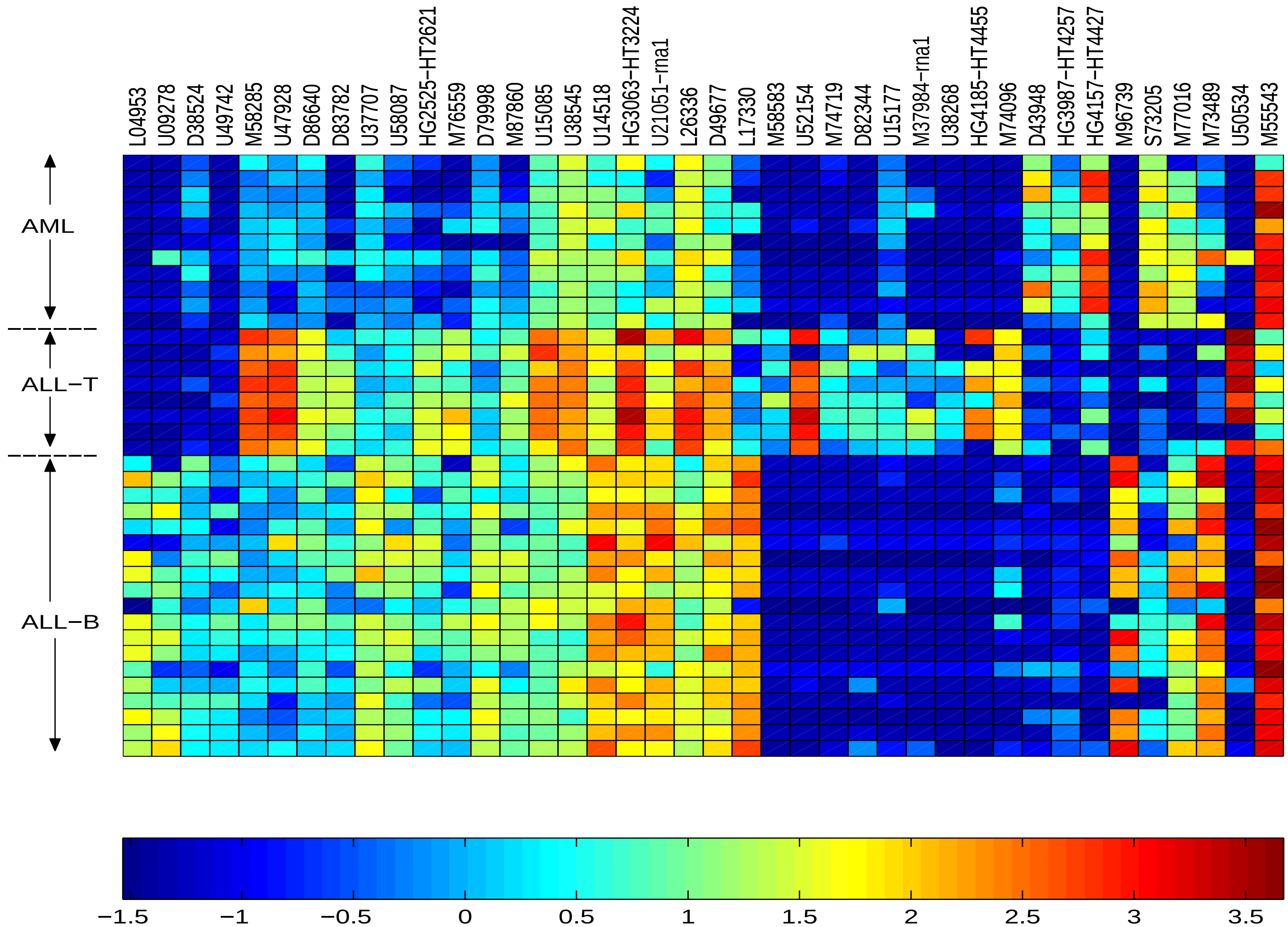
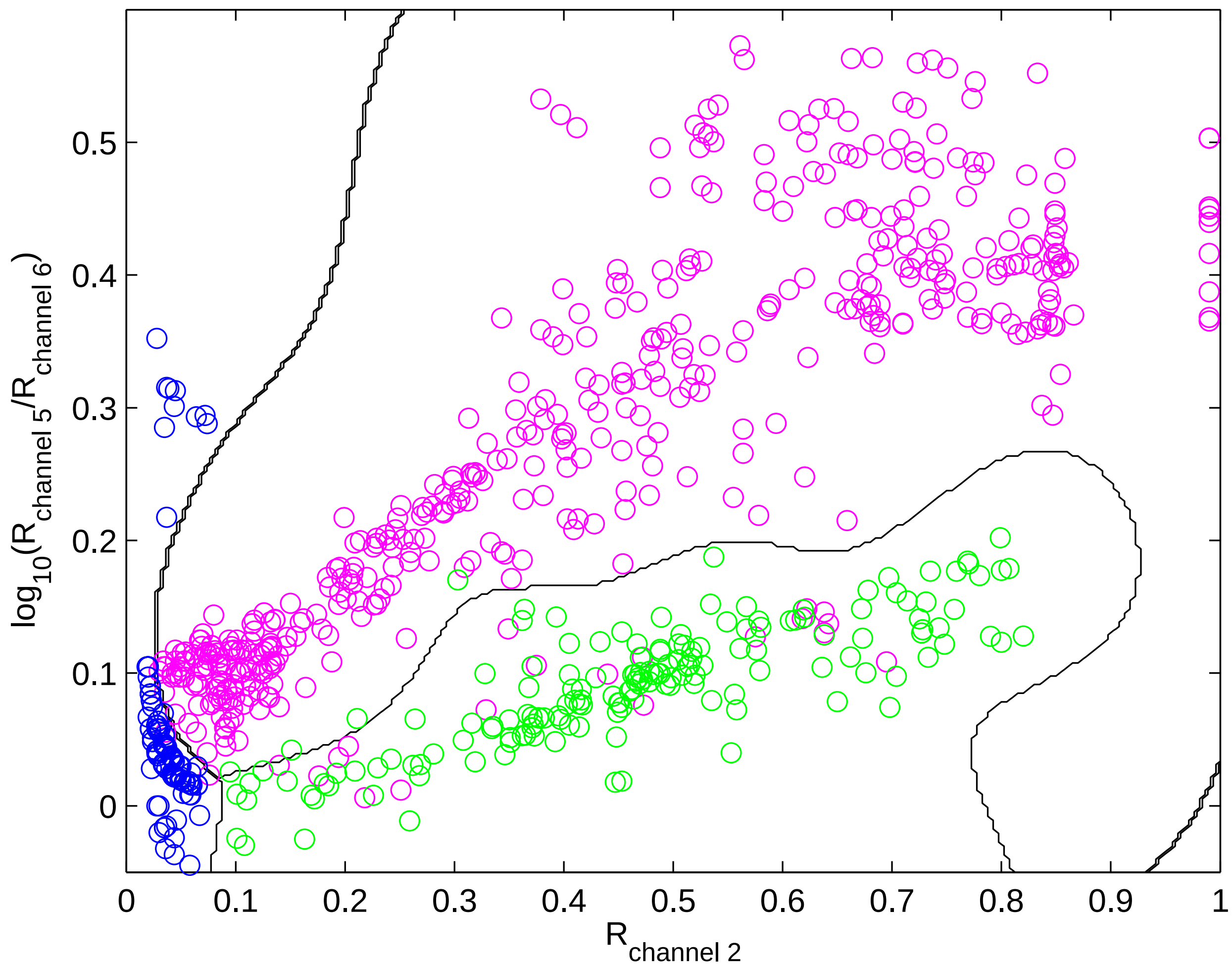


Figure: The heat map shows the expression levels of 40 most important genes for the training samples when they are appropriately standardized. Each row corresponds to a sample, which is grouped into the three classes, and the columns represent genes. The 40 genes are clustered in a way the similarity within each class and the dissimilarity between classes are easily recognized.

Cloud Detection and Classification

MODIS radiance profiles (12 channels) over the Gulf of Mexico in July 2002 (128 clear scenes, 164 water clouds and 476 ice clouds) [*Lee, Wahba, and Ackerman (2003)*]



Predictive learning

- Multivariate function estimation
- A training data set $\{(\boldsymbol{x}_i, y_i), i = 1, \ldots, n\}$
- Learn functional relationship f between $\boldsymbol{x} = (x_1, \ldots, x_p)$ and y from the training data, which can be generalized to novel cases.
  e.g. $f(\boldsymbol{x}) = E(Y|\boldsymbol{X} = \boldsymbol{x})$
- Examples include
  Regression: continuous $y \in R$, and
  Classification: categorical $y \in \{1, \ldots, k\}$.

Goodness of a learning method

- Accurate prediction with respect to a given loss $\mathcal{L}(y, f(\boldsymbol{x}))$ e.g. $\mathcal{L}(y, f(\boldsymbol{x})) = (y - f(\boldsymbol{x}))^2$ for regression
- Flexible (nonparametric) and data-adaptive
- Interpretability (e.g. subset selection)
- Computational ease for large p (high dimensional input) and n (large sample)

Method of Regularization (Penalization)

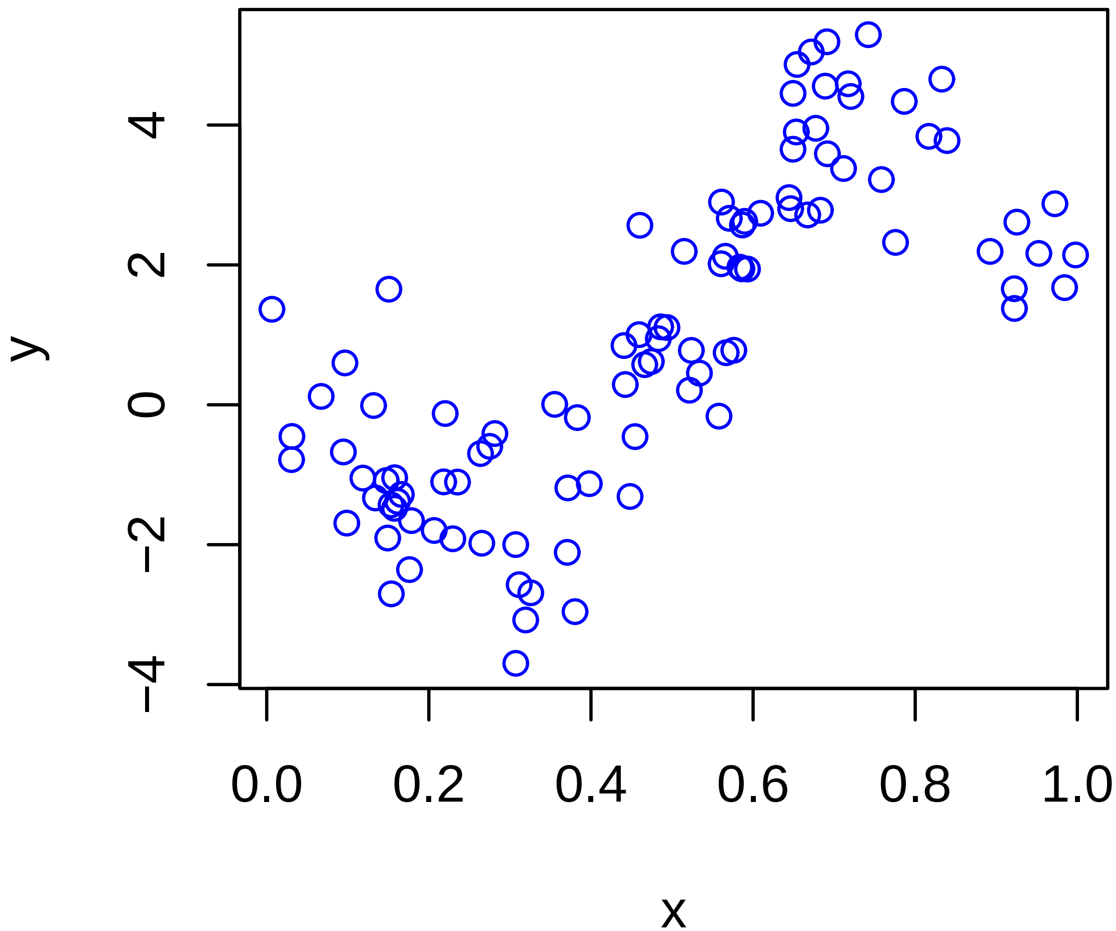
Find $f(\boldsymbol{x}) \in \mathcal{F}$ minimizing

$$\frac{1}{n}\sum_{i=1}^{n}\mathcal{L}(y_i, f(\boldsymbol{x}_i)) + \lambda J(f).$$

- $\mathcal{F}$: a class of candidate functions
- $J(f)$: complexity of the model f
- Without the penalty $J(f)$, ill-posed problem
- $\lambda > 0$: a regularization parameter

Regression

$$y_i = f(x_i) + \epsilon_i \text{ for } i = 1, \ldots, n \text{ where } \epsilon_i \sim N(0, \sigma^2)$$



Smoothing Splines

Wahba (1990), Spline Models for Observational Data.

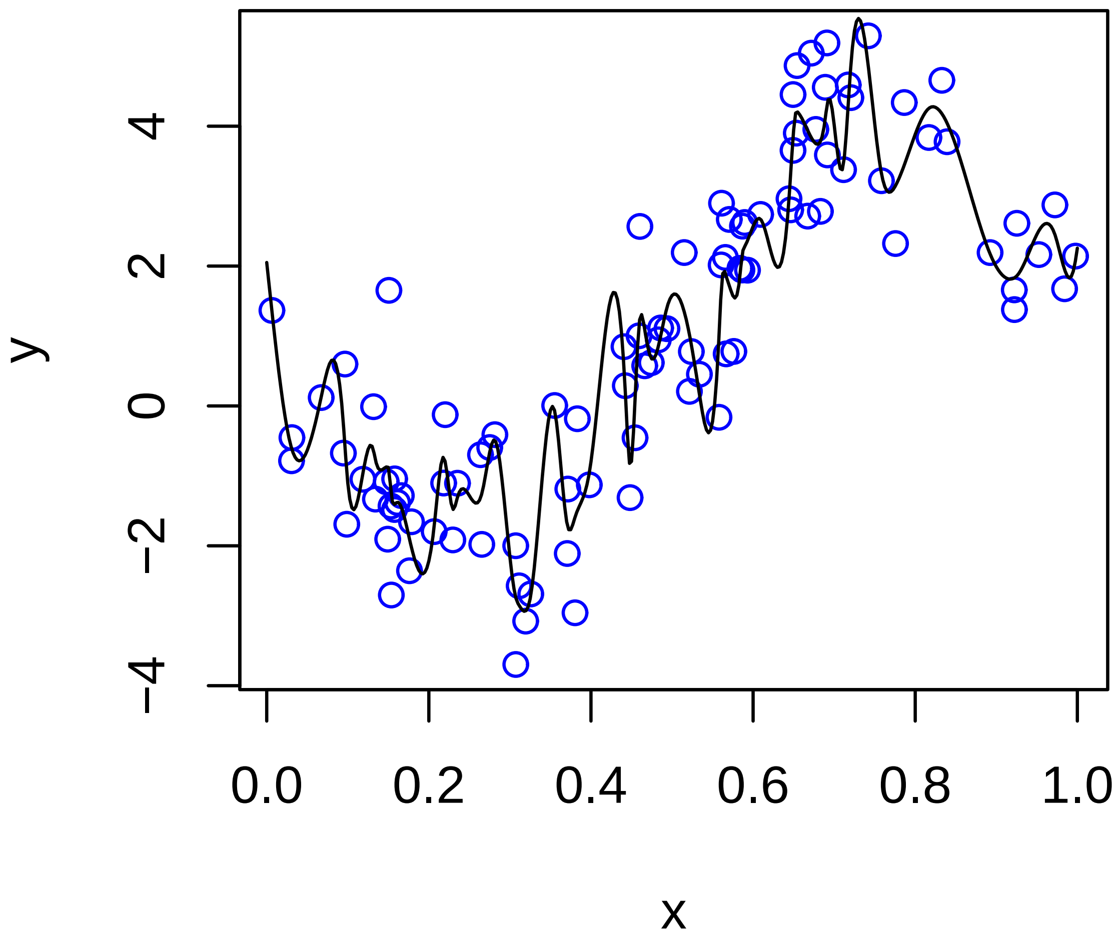
Find $f(x) \in W_2[0,1]$
$= \{f : f, f' \text{ absolutely continuous, and } f'' \in L_2\}$ minimizing

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - f(x_i))^2 + \lambda \int_0^1 (f''(x))^2 dx.$$

- $J(f) = \int_0^1 (f''(x))^2 dx = \|P^1 f\|^2$: curvature of f
- $\lambda \to 0$: interpolation
- $\lambda \to \infty$: linear fit
- $0 < \lambda < \infty$: piecewise cubic polynomials with two continuous derivatives

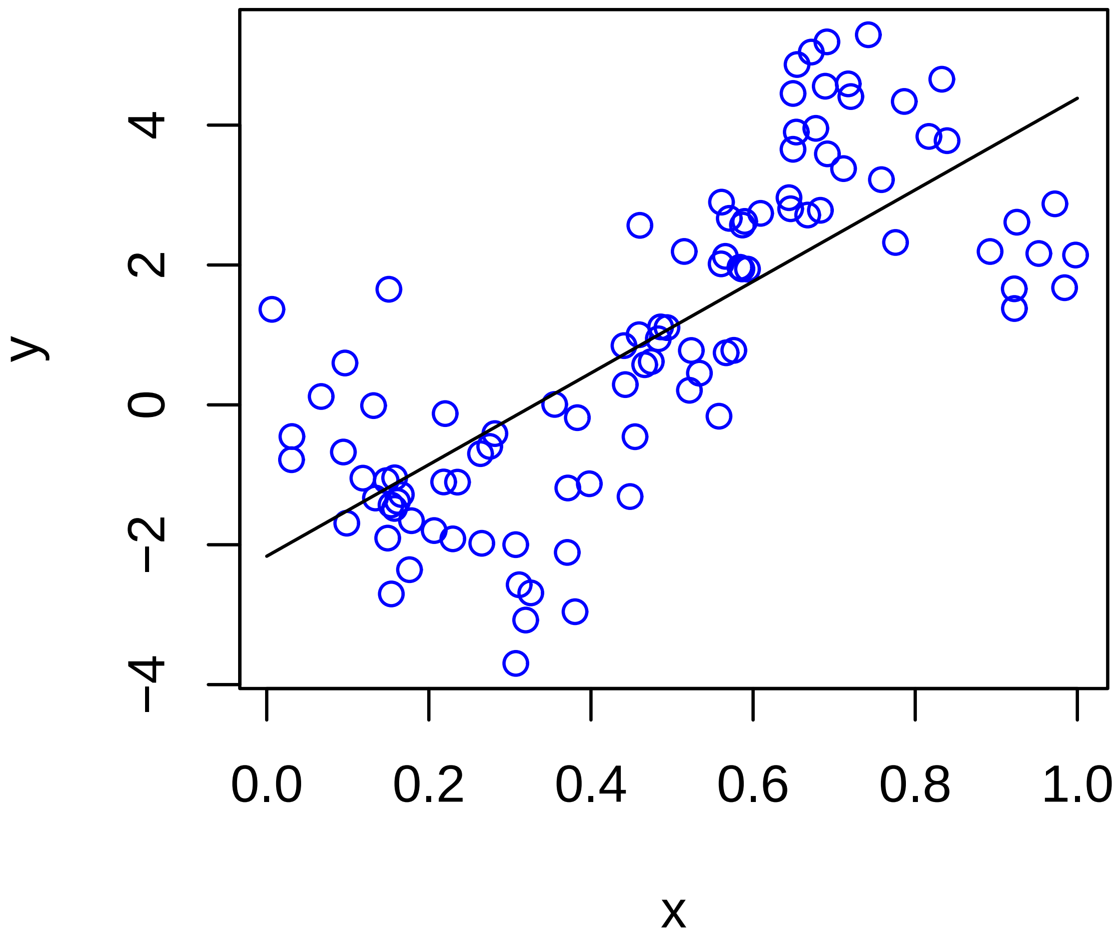
Small λ: overfit

$\lambda \to 0$: interpolation



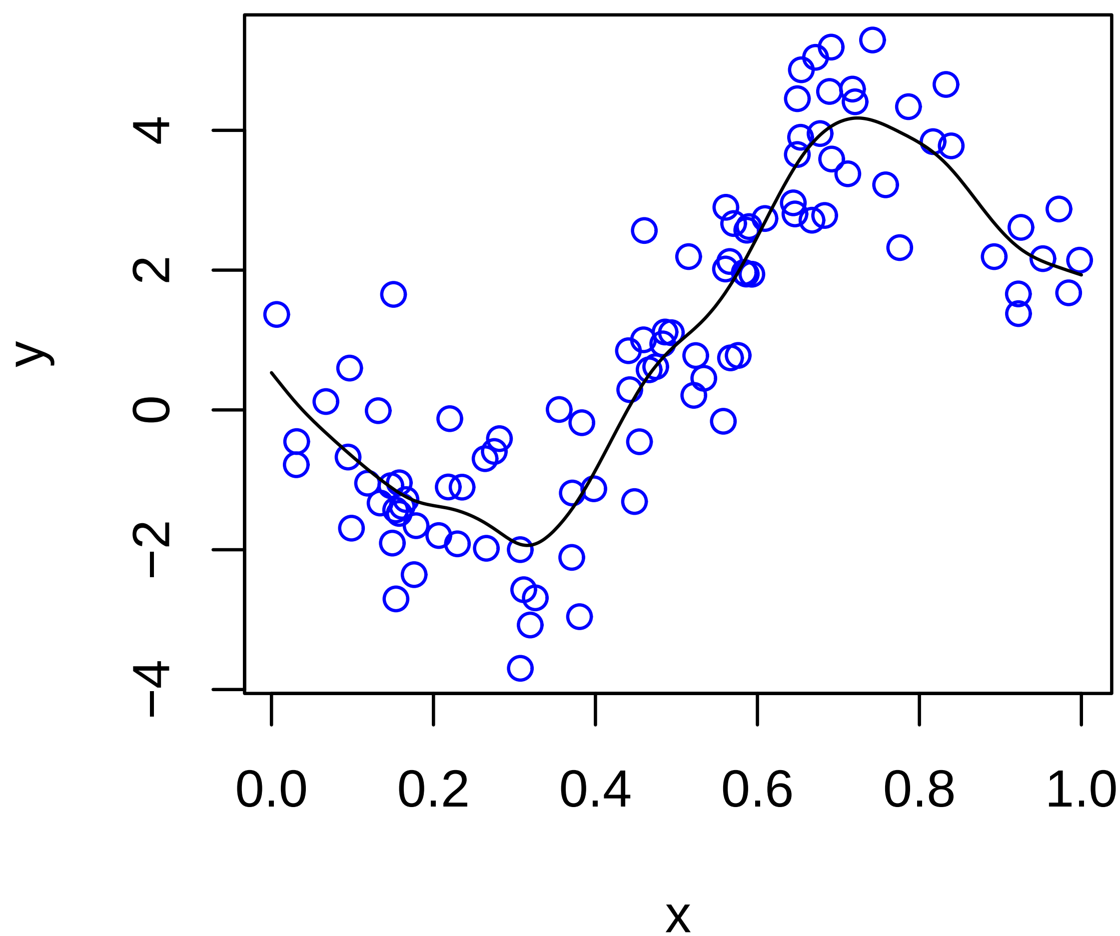
Large λ: underfit

$\lambda \to \infty$: linear fit

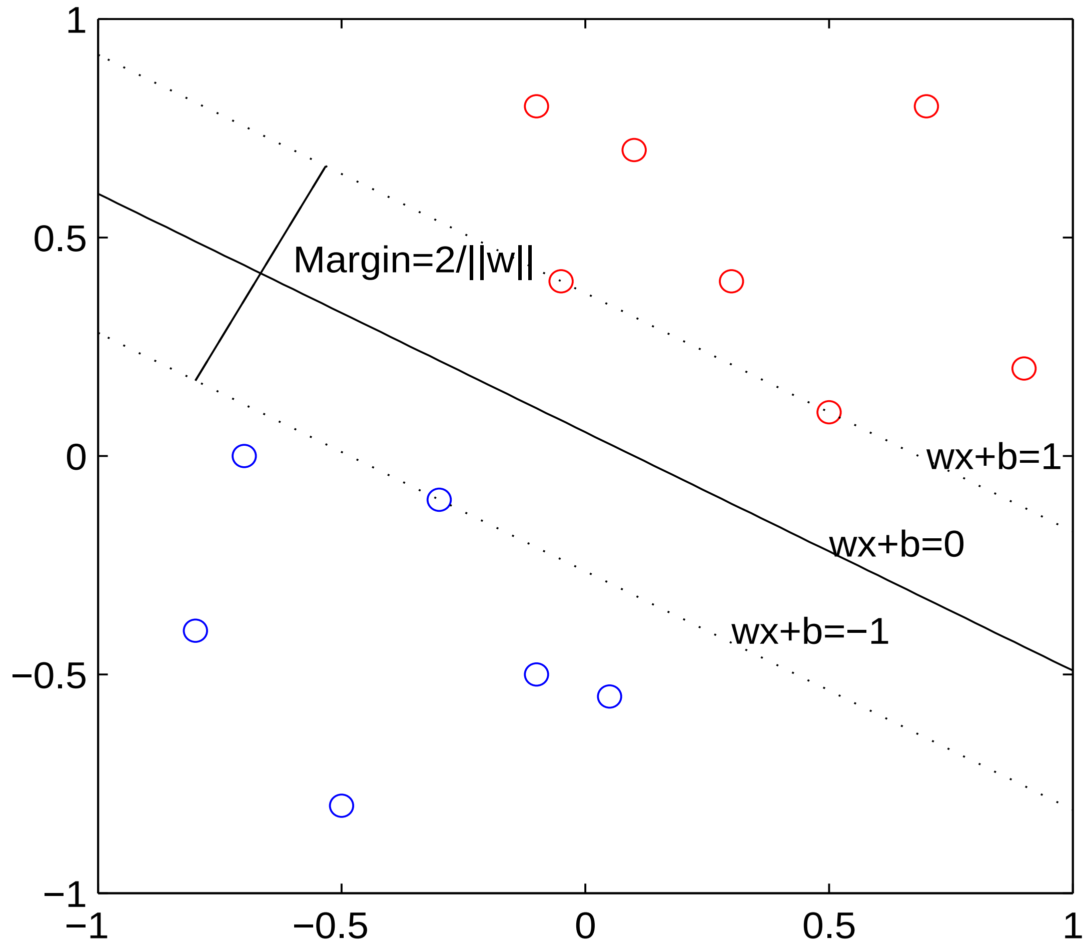


Moderate λ

$0 < \lambda < \infty$: piecewise cubic polynomials with two continuous derivatives



Classification: separable case

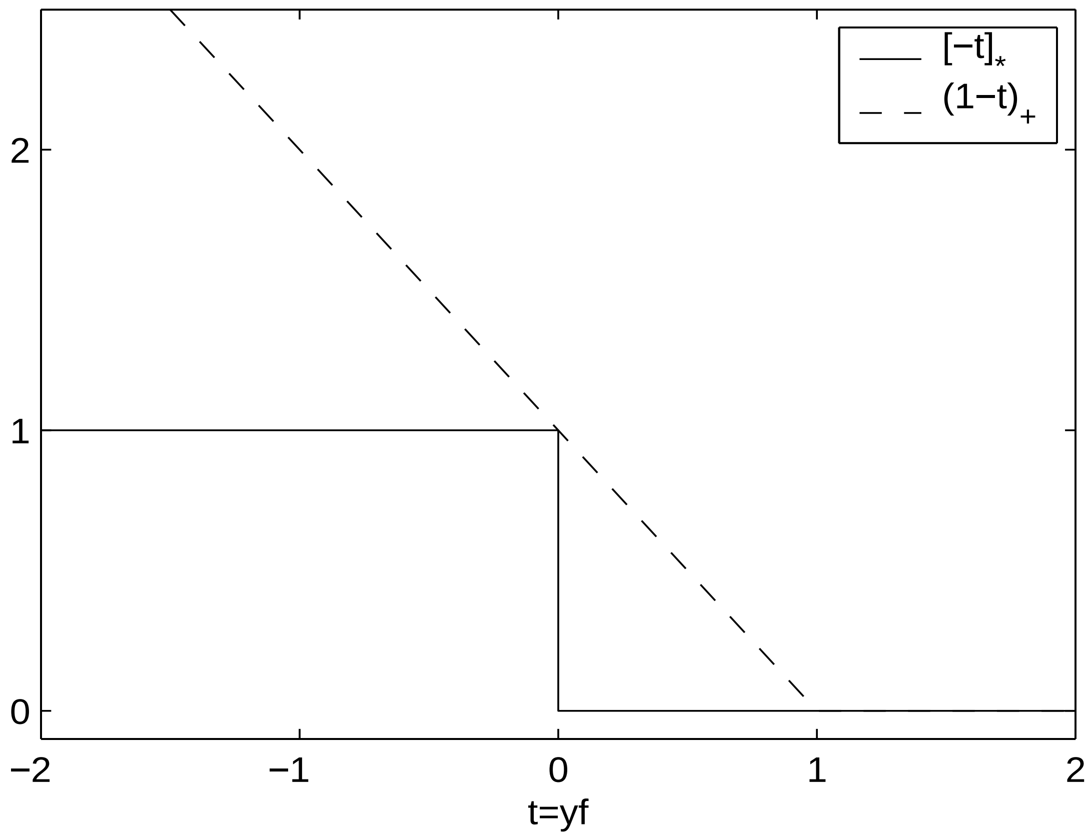


Support Vector Machines

Vapnik (1995), The Nature of Statistical Learning Theory.
http://www.kernel-machines.org

- $y_i \in \{-1, 1\}$, class labels in binary case
- $f(\boldsymbol{x})$: a bit dubious
- Classification rule : $\phi(\boldsymbol{x}) = sign(f(\boldsymbol{x}))$
- $\mathcal{L}(y, f(\boldsymbol{x})) = (1 - yf(\boldsymbol{x}))_+$

Hinge loss



$(1 - yf(\boldsymbol{x}))_+$ is an upper bound of the misclassifi cation loss function $I(y \neq \phi(\boldsymbol{x})) = [-yf(\boldsymbol{x})]_* \leq (1 - yf(\boldsymbol{x}))_+$ where $[t]_* = I(t \geq 0)$ and $(t)_+ = \max\{t, 0\}$.

Linear SVM

Find $f \in \mathcal{F} = \{f(\boldsymbol{x}) = \boldsymbol{w}^\top \boldsymbol{x} + b \mid \boldsymbol{w} \in R^p \text{ and } b \in R\}$ minimizing

$$\frac{1}{n}\sum_{i=1}^{n}(1 - y_i f(\boldsymbol{x}_i))_+ + \lambda\|\boldsymbol{w}\|^2,$$

where $J(f) = J(\boldsymbol{w}^\top \boldsymbol{x} + b) = \|\boldsymbol{w}\|^2$.

Regularization in RKHS

Find $f = \sum_{\nu=1}^{M} d_\nu \phi_\nu + h$ with $h \in \mathcal{H}_K$ minimizing

$$\frac{1}{n}\sum_{i=1}^{n} \mathcal{L}(y_i, f(\boldsymbol{x}_i)) + \lambda \|h\|^2_{\mathcal{H}_K}.$$

- $\mathcal{H}_K$: a reproducing Kernel Hilbert space of functions defined on an arbitrary domain
- $K(\boldsymbol{x}, \boldsymbol{x}')$: reproducing kernel (positive definite) s.t.
  i) $K(\boldsymbol{x}, \cdot) \in \mathcal{H}_K$ for each $\boldsymbol{x}$
  ii) $f(\boldsymbol{x}) =< K(\boldsymbol{x}, \cdot), f(\cdot) >_{\mathcal{H}_K}$ for all $f \in \mathcal{H}_K$, so
  $K(\boldsymbol{x}, \boldsymbol{x}') =< K(\boldsymbol{x}, \cdot), K(\boldsymbol{x}', \cdot) >_{\mathcal{H}_K}$
- The null space spanned by $\{\phi_\nu\}_{\nu=1}^{M}$
- $J(f) = \|h\|^2_{\mathcal{H}_K}$: penalty

Cubic smoothing splines

Find $f(x) \in W_2[0,1]$
$= \{f : f, f' \text{ absolutely continuous, and } f'' \in L_2\}$ minimizing

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - f(x_i))^2 + \lambda \int_0^1 (f''(x))^2 dx.$$

- The null space: $M = 2$, $\phi_1(x) = 1$, and $\phi_2(x) = x$.
- The penalized space: $\mathcal{H}_K = W_2^0[0,1] =$ $\{f \in W_2[0,1] : f(0) = 0, f'(0) = 0\}$ is an RKHS with
  i) $\|f\|^2 = \int_0^1 (f''(x))^2 dx$
  ii) $K(x,x') = \int_0^1 (x-u)_+(x'-u)_+ du$.

SVM in general

Find $f(\boldsymbol{x}) = b + h(\boldsymbol{x})$ with $h \in \mathcal{H}_K$ minimizing

$$\frac{1}{n}\sum_{i=1}^{n}(1 - y_i f(\boldsymbol{x}_i))_+ + \lambda\|h\|^2_{\mathcal{H}_K}.$$

- The null space: $M = 1$ and $\phi_1(\boldsymbol{x}) = 1$
- Linear SVM: $\mathcal{H}_K = \{h(\boldsymbol{x}) = \boldsymbol{w}^\top\boldsymbol{x} \mid \boldsymbol{w} \in R^p\}$ with $K(\boldsymbol{x}, \boldsymbol{x}') = \boldsymbol{x}^\top\boldsymbol{x}'$ and $\|h\|^2_{\mathcal{H}_K} = \|\boldsymbol{w}^\top\boldsymbol{x}\|^2_{\mathcal{H}_K} = \|\boldsymbol{w}\|^2$
- Nonlinear SVM: $K(\boldsymbol{x}, \boldsymbol{x}') = (1 + \boldsymbol{x}^\top\boldsymbol{x}')^d$, $\exp(-\|\boldsymbol{x} - \boldsymbol{x}'\|^2/2\sigma^2)$

Representer Theorem

Kimeldorf and Wahba (1971)

- The minimizer $f = \sum_{\nu=1}^{M} d_\nu \phi_\nu + h$ with $h \in \mathcal{H}_K$ of

$$\frac{1}{n}\sum_{i=1}^{n} \mathcal{L}(y_i, f(\boldsymbol{x}_i)) + \lambda \|h\|^2_{\mathcal{H}_K}$$

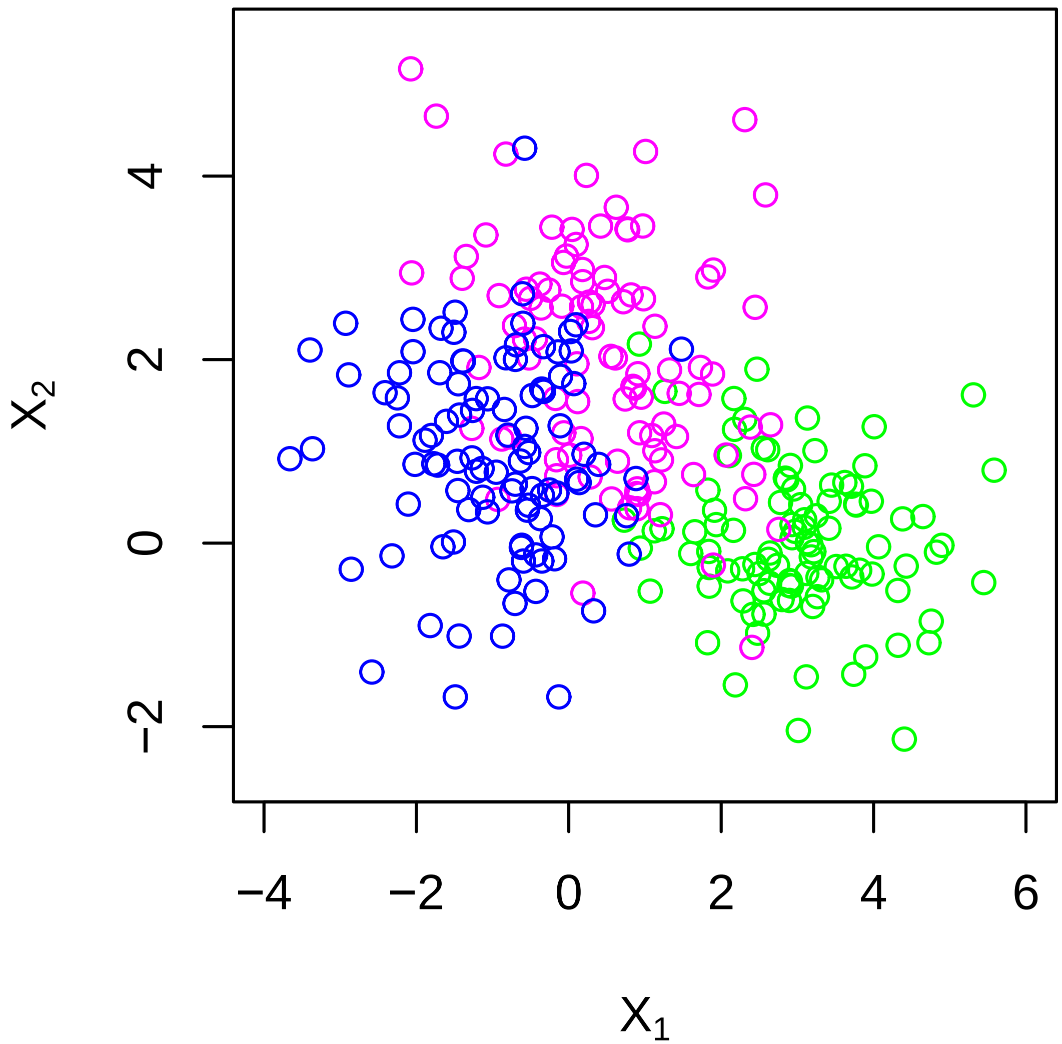
  has a representation of the form

$$\hat{f}(\boldsymbol{x}) = \sum_{\nu=1}^{M} \hat{d}_\nu \phi_\nu(\boldsymbol{x}) + \sum_{i=1}^{n} \hat{c}_i K(\boldsymbol{x}_i, \boldsymbol{x}).$$

- $\|h\|^2_{\mathcal{H}_K} = \sum_{i,j} \hat{c}_i \hat{c}_j K(\boldsymbol{x}_i, \boldsymbol{x}_j)$.

Classification

$$y_i \in \{1 : \textit{green},\ 2 : \textit{magenta},\ 3 : \textit{blue}\}$$



Test error rates

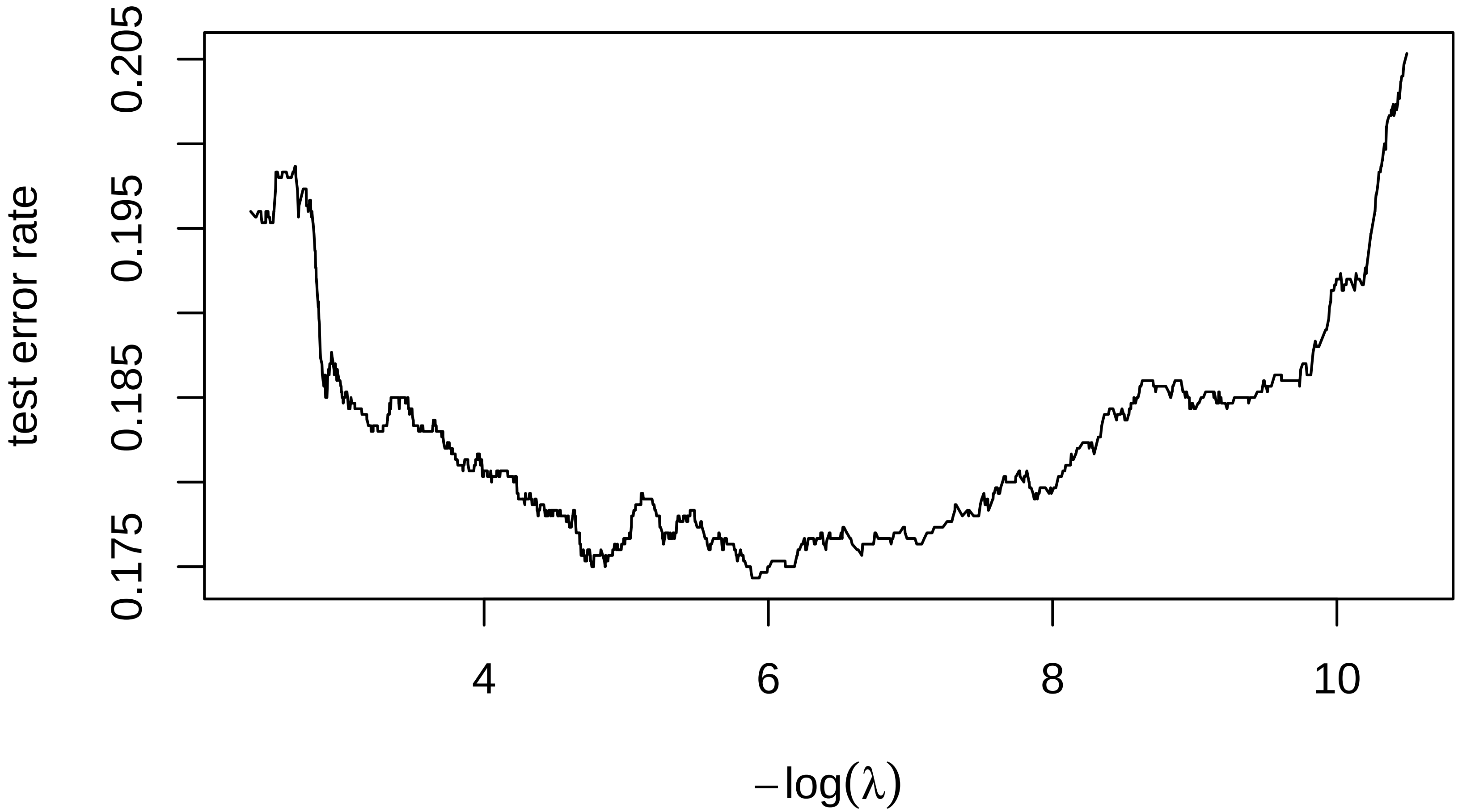
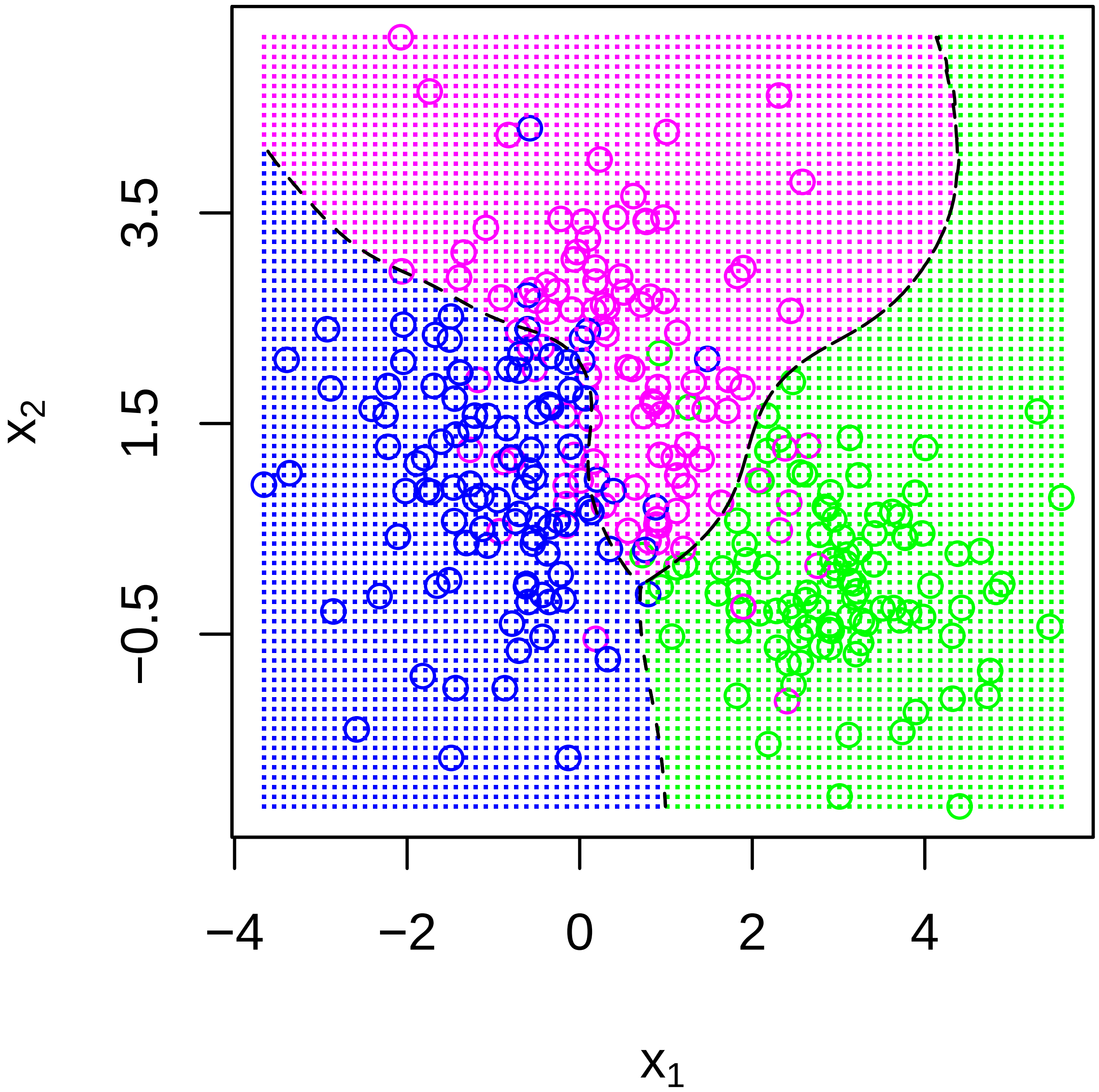


Figure: Test error rate as a function of λ.

Estimated classification boundaries



Statistical Issues

- Risk or generalization error estimation
- Model selection - choice of tuning parameter(s)
- Variable (feature) selection
- Computation:
  - Efficient algorithm
  - Understanding the characteristics of solutions
- Large sample theory:
  - Consistency
  - Rate of convergence
- Understanding and incorporating data geometry
- Dealing with nonstandard data structure and domains (text, sequence, graph,...)

My research

- Statistical understanding of the SVM
- Optimal extension to the multiclass case
- Feature selection
- Characterization of the entire solution path

Concluding remarks

- RKHS method provides a unified framework for statistical model building.
- It can solve a wide range of statistical learning problems in a principled way.
- There are many interesting research problems in the interface of statistics and machine learning!

If you want to learn more ...

- The slides of this talk can be downloaded from my webpage.
- STAT 760 (Winter 2007):
  Introduction to Statistical Learning
- STAT 763 (Spring 2007):
  Nonparametric Function Estimation
  (with emphasis on smoothing splines)
- STAT 881 (Spring 2007):
  Advanced Statistical Learning