MSRI SDP workshop—October 7-11, 2002

Robust Convex Optimization in Classification Problems

Laurent El Ghaoui

Department of EECS, UC Berkeley

elghaoui@eecs.berkeley.edu

goal

- connection between classification and LP, convex QP has a long history (Vapnik, Mangasarian, Bennett, etc)
- recent progresses in **convex** optimization: conic and semidefinite programming; geometric programming; robust optimization
- we'll outline some connections between (robust) convex optimization and classification problems

joint work with: M. Jordan, P. Bartlett, N. Cristianini, G. Lanckriet, C. Bhattacharrya

outline

▷ convex optimization

- SVMs and robust linear programming
- minimax probability machine
- learning the kernel matrix

convex optimization

standard form:

$$\min_x \ f_0(x) \ : \ f_i(x) \leq 0, \ \ i = 1, \ldots, m$$

- arises in many applications
- convexity not always recognized in practice
- can solve large classes of convex problems in polynomial-time (Nesterov, Nemirovski, 1990)

conic optimization

special class of convex problems:

$$\min_x \; c^T x \;\; : \;\; Ax = b, \;\; x \in K$$

where K is a cone, direct product of the following "building blocks":

$K = \mathbf{R}^n_+$	linear programming
$K = \{(y,t) \in \mathbf{R}^{n+1} \;\; : \;\; t \geq \|y\|_2\}$	second-order cone programming, quadratic programming
$K = \{x \in \mathbf{R}^{n \times n} \;\; : \;\; x = x^T \succeq 0\}$	semidefinite programming

fact: can solve conic problems in polynomial-time
(Nesterov, Nemirovski, 1990)

conic duality

dual of conic problem

$$\min_x \; c^T x \;\; : \;\; Ax = b, \;\; x \in K$$

is

$$\max_y \; b^T y \;\; : \;\; c - A^T y \in K^*$$

where

$$K^* = \{z \;\; : \;\; \langle z, x \rangle \geq 0 \; \forall \, x \in K\}$$

is the cone *dual* to K

for the cones mentioned before, and direct products of them, $K = K^*$

robust optimization

conic problem in dual form: $\max_y \; b^T y \;:\; c - A^T y \in K$

$\rightarrow$ what if A is unknown-but-bounded, say $A \in \mathcal{A}$, where $\mathcal{A}$ is given?

robust counterpart: $\max_y \; b^T y \;:\; \forall A \in \mathcal{A},\;\; c - A^T y \in K$

- still convex, but tractability depends on $\mathcal{A}$
- systematic ways to approximate (get lower bounds)
- for special classes of $\mathcal{A}$, approximation is exact

example: robust LP

linear program: $\min_x \; c^T x \;:\; a_i^T x \le b, \;\; i = 1, \ldots, m$

assume a_i's are unknown-but-bounded in ellipsoids

$$\mathcal{E}_i := \left\{ a \;:\; (a - \hat{a}_i)^T \Gamma_i^{-1} (a - \hat{a}_i) \le 1 \right\}$$

where $\hat{a}_i$: center, $\Gamma_i \succ 0$: "shape matrix"

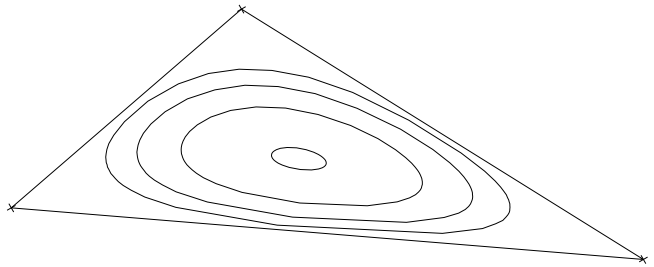
robust LP: $\min_x \; c^T x \;:\; \forall\, a_i \in \mathcal{E}_i, \;\; a_i^T x \le b, \;\; i = 1, \ldots, m$

robust LP: SOCP representation

robust LP equivalent to

$$\min_x \; c^T x \;\; : \;\; \hat{a}_i^T x + \|\Gamma_i^{1/2} x\|_2 \leq b, \;\; i = 1, \ldots, m$$

$\longrightarrow$ a second-order cone program!



interpretation: smoothes boundary of feasible set

LP with Gaussian coefficients

assume $a \sim \mathcal{N}(\hat{a}, \Gamma)$, then for given x,

$$\mathbf{Prob}\{a^T x \leq b\} \geq 1 - \epsilon$$

is **equivalent** to:

$$\hat{a}^T x + \kappa \|\Gamma^{1/2} x\|_2 \leq b$$

where $\kappa = \Phi^{-1}(1-\epsilon)$ and Φ is the c.d.f. of $\mathcal{N}(0,1)$

hence,

- can solve LP with Gaussian coefficients using **second-order cone** programming
- resulting SOCP is similar to one obtained with ellipsoidal uncertainty

LP with random coefficients

assume $a \sim (\hat{a}, \Gamma)$, *i.e.* distribution of a has mean $\hat{a}$ and covariance matrix Γ, but is otherwise **unknown**

Chebychev inequality:

$$\mathbf{Prob}\{a^T x \leq b\} \geq 1 - \epsilon$$

is **equivalent** to:

$$\hat{a}^T x + \kappa \|\Gamma^{1/2} x\|_2 \leq b$$

where

$$\kappa = \sqrt{\frac{1-\epsilon}{\epsilon}}$$

leads to SOCP similar to ones obtained previously

outline

- convex optimization
- ▷ SVMs and robust linear programming
- minimax probability machine
- kernel optimization

SVMs: setup

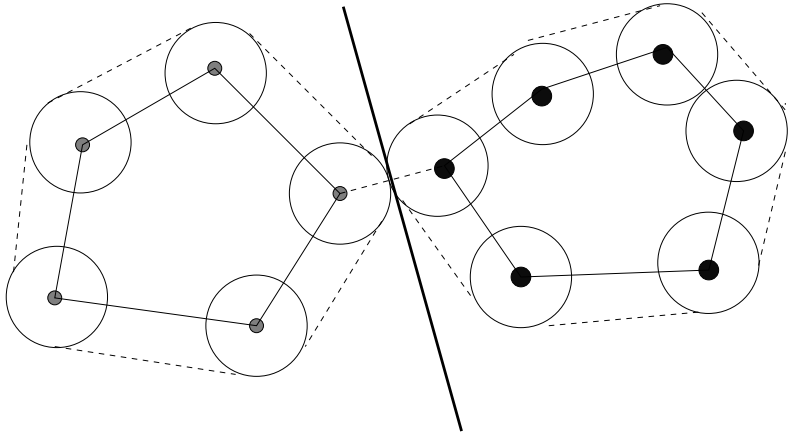
given data points x_i with labels $y_i = \pm 1$, $i = 1, \ldots, N$

two-class linear classification with support vector:

$$\min \ \|a\|_2 \ : \ y_i(a^T x_i - b) \geq 1, \ \ i = 1, \ldots, N$$

- problem is feasible iff there exists a separating hyperplane between the two classes
- if so, amounts to select *one* separating hyperplane among the many possible

SVMs: robust optimization interpretation



interpretation: SVMs are a way to handle noise in data points

- assume each data point is unknown-but-bounded in a sphere of radius ρ and center x_i
- find the largest ρ such that separation is still possible between the two classes of perturbed points

variations

can use other data noise models:

- hypercube uncertainty (gives rise to LP)
- ellipsoidal uncertainty ($\rightarrow$ QP)
- probabilistic uncertainty, Gaussian or Chebychev ($\rightarrow$ QP)

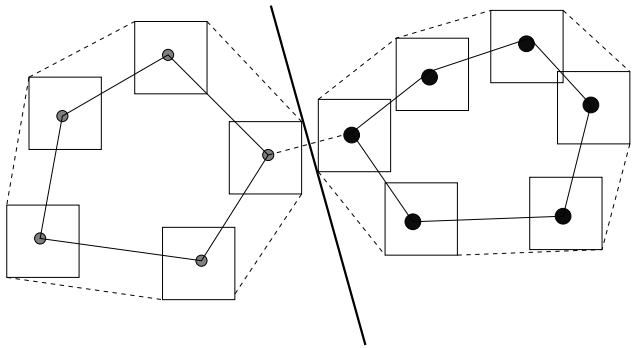
separation with hypercube uncertainty

assume each data point is unknown-but-bounded in an hypercube $\mathcal{C}_i$:

$$x_i \in \mathcal{C}_i := \{\hat{x}_i + \rho P u \ : \ \|u\|_\infty \leq 1\}$$

where centers $\hat{x}_i$ and "shape matrix" P are given

robust separation:



leads to **linear** program

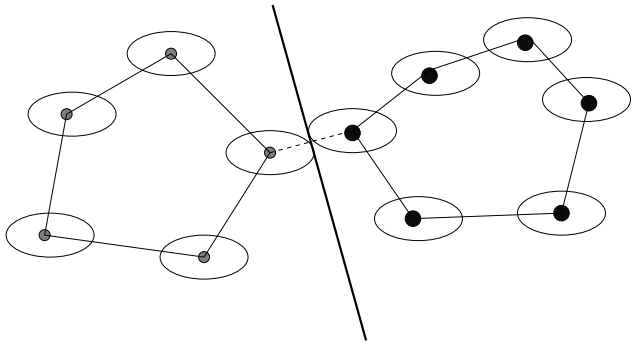
$$\min \|Pa\|_1 \ : \ y_i(a^T \hat{x}_i - b) \geq 1, \ \ i = 1, \ldots, N$$

separation with ellipsoidal uncertainty

assume each data point is unknown-but-bounded in an ellipsoid $\mathcal{E}_i$:

$$x_i \in \mathcal{E}_i := \{\hat{x}_i + \rho P u \ : \ \|u\|_2 \leq 1\}$$

where center $\hat{x}_i$ and "shape matrix" P are given



robust separation leads to QP

$$\min \|Pa\|_2 \ : \ y_i(a^T \hat{x}_i - b) \geq 1, \ \ i = 1, \ldots, N$$

outline

- convex optimization
- SVMs and robust linear programming
- ▷ minimax probability machine
- kernel optimization

minimax probability machine

goal:

- make assumptions about the data generating process
- do not assume Gaussian distributions
- use second-moment analysis of the two classes

let $\hat{x}_\pm, \Gamma_\pm$ be the mean and covariance matrix of class $y = \pm 1$

MPM: maximize ϵ such that there exists (a, b) such that

$$\begin{aligned} \inf_{x \sim (\hat{x}_+, \Gamma_+)} \mathbf{Prob}\{a^T x \leq b\} &\geq 1 - \epsilon \\ \inf_{x \sim (\hat{x}_-, \Gamma_-)} \mathbf{Prob}\{a^T x \geq b\} &\geq 1 - \epsilon \end{aligned}$$

MPMs: optimization problem

$\longrightarrow$ two-sided, multivariable Chebychev inequality:

$$\inf_{x \sim (\hat{x}, \Gamma)} \mathbf{Prob}\{a^T x \leq b\} = \frac{(b - a^T \hat{x})_+^2}{(b - a^T \hat{x})_+^2 + a^T \Gamma a}$$

MPM problem leads to **second-order cone program**:

$$\min_a \ \|\Gamma_+^{1/2} a\|_2 + \|\Gamma_-^{1/2} a\|_2 \ \ : \ \ a^T(\hat{x}_+ - \hat{x}_-) = 1$$

complexity is the same as standard SVMs

link with Fisher discriminant analysis

Fisher's discriminant analysis (FDA) solves

$$\min_a \ \|\Gamma_+^{1/2} a\|_2^2 + \|\Gamma_-^{1/2} a\|_2^2 \ : \ a^T(\hat{x}_+ - \hat{x}_-) = 1$$

- reduces to a (linearly constrained) least-squares problem (hence, widely used)
- no clear way to compute the "bias" term b
- no clear probabilistic interpretation of FDA
- MPM approach has same complexity

dual problem

express problem as unconstrained min-max problem:

$$\min_{a} \max_{\|u\|_2 \leq 1,\, \|v\|_2 \leq 1} u^T \Gamma_+^{1/2} a - v^T \Gamma_-^{1/2} a + \lambda(1 - a^T(x_+ - x_-))$$

exchange min and max, and set $\kappa := 1/\lambda$:

$$\min_{\kappa,u,v} \ \rho \ : \ x_+ + \Gamma_+^{1/2} u = x_- + \Gamma_-^{1/2} v, \ \ \|u\|_2 \leq \kappa, \ \ \|v\|_2 \leq \kappa$$

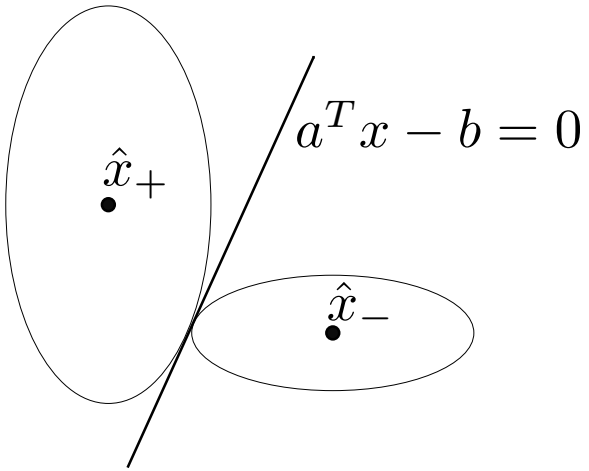
robust optimization interpretation

assume data with label $+$ generated arbitrarily in ellipsoid

$$x_+ \in \mathcal{E}_+(\rho) := \left\{ \hat{x}_+ + \Gamma_+^{1/2} u \ : \ \|u\|_2 \leq \rho \right\}$$

and similarly for data with label $-$

MPM finds largest ρ for which robust separation is possible

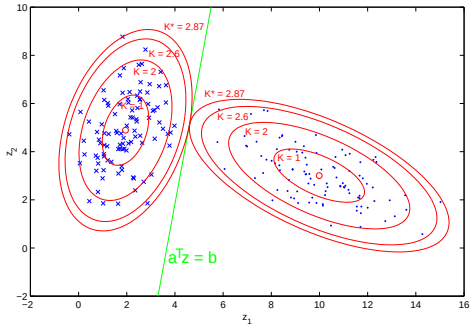


geometric interpretation

define the two ellipsoids

$$\mathcal{E}_{\pm}(\rho) := \left\{ \hat{x}_{\pm} + \Gamma_{\pm}^{1/2} u \ : \ \|u\|_2 \leq \kappa \right\}$$

and find largest κ for which ellipsoids intersect



problem amounts to minimize the maximum of the Mahalanobis distances to the two classes

$$\kappa_* = \min_{\mathbf{z}} \max \left(\|\Gamma_+^{-1/2}(\mathbf{z} - \hat{x}_+)\|_2, \|\Gamma_-^{-1/2}(\mathbf{z} - \hat{x}_-)\|_2 \right).$$

optimal upper bound on misclassification error: $1 - \alpha_* = 1/(1 + \kappa_*^2)$

robustness to estimation errors

in practice, the first and second moment of the classes have to be estimated ...

how does this affect the MPM classifier?

we will seek an MPM classifier that is robust to estimation errors

robust MPM

assume that the first- and second- moment of each class are only known with bounds

robust MPM: maximize ϵ such that there exists (a, b) such that

$$\begin{aligned} \inf_{x \sim (\hat{x}_+, \Gamma_+)} \mathbf{Prob}\{a^T x \leq b\} &\geq 1 - \epsilon \\ \inf_{x \sim (\hat{x}_-, \Gamma_-)} \mathbf{Prob}\{a^T x \geq b\} &\geq 1 - \epsilon \end{aligned}$$

for *every* $(\hat{x}_\pm, \Gamma_\pm)$ in $\mathcal{X}_\pm$

here, $\mathcal{X}_\pm$ describe our uncertainty about the moments

a specific uncertainty model

we assume that $\mathcal{X}_\pm$ have the form

$$\mathcal{X} \quad = \quad \left\{(\hat{x}, \Gamma) \ : \ (\hat{x} - \hat{x}_0)^T \Gamma^{-1} (\hat{x} - \hat{x}^0) \leq \nu^2, \ \ \|\Gamma - \Gamma^0\| \leq \rho \right\},$$

where $\hat{x}_0, \Sigma_0$ and ν, ρ are given

- model inspired by maximum-likelihood approaches to moment estimation (under Gaussian assumptions)
- refined models possibles (cf. Iyengar & Goldfarb, 2001)

robust MPM

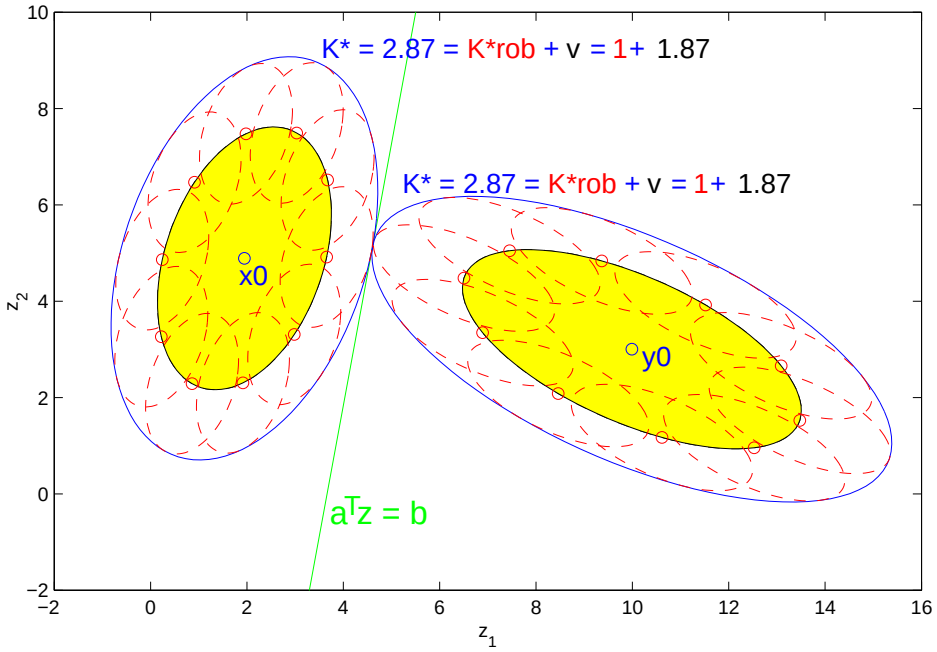
main result:

- the optimal robust minimax probability classifier with uncertainty sets$\mathcal{X}_\pm$ can be obtained by solving the **original** MPM problem, with $\Gamma_\pm = \Gamma_\pm^0 + \rho_\pm I_n$, and $\hat{x}_\pm = \hat{x}_\pm^0$
- if κ_*^{-1} is the optimal value of that problem, the corresponding upper bound on the worst-case misclassification error is

$$1 - \alpha_*^{\mathrm{rob}} = \frac{1}{1 + \max(0, (\kappa_* - \nu))^2}.$$

bottom line: errors in covariance matrices are handled by regularization, while errors in the mean affect the misclassification probability

robustness: geometric interpretation



experimental results

α and test-set accuracy (TSA) compared to BPB (best performance in Breiman, 1996) and to the performance of an SVM with linear kernel (SVML) and an SVM with Gaussian kernel (SVMG):

Dataset	*Linear kernel*		*Gaussian kernel*		*BPB*	*SVML*	*SVMG*
	α	TSA	α	TSA			
Twonorm	80.2 %	96.0 %	83.6 %	97.2 %	96.3 %	95.6 %	97.4 %
Breast cancer	84.4 %	97.2 %	92.7 %	97.3 %	96.8 %	92.6 %	98.5 %
Ionosphere	63.3 %	85.4 %	89.9 %	93.0 %	93.7 %	87.8 %	91.5 %
Pima diabetes	31.2 %	73.8 %	33.0 %	74.6 %	76.1 %	70.1 %	75.3 %
Sonar	62.4 %	75.1 %	87.1 %	89.8 %	-	75.9 %	86.7 %

variations

- minimize weighted sum of misclassification probabilities
- quadratic separation: find a quadratic set such that

$$\begin{aligned} \inf_{x\sim(\hat{x}_+,\Gamma_+)} \mathbf{Prob}\{x \in \mathcal{Q}\} &\geq 1-\epsilon \\ \inf_{x\sim(\hat{x}_-,\Gamma_-)} \mathbf{Prob}\{x \notin \mathcal{Q}\} &\geq 1-\epsilon \end{aligned}$$

  $\rightarrow$ leads to a semidefinite programming problem (see Vandenberghe, 2002)

- nonlinear classification via kernels

  (using plug-in estimates of mean and covariance matrix)

outline

- convex optimization
- SVMs and robust linear programming
- minimax probability machine

▷ learning the kernel matrix

transduction

the data contains both

- labeled points (training set)
- unlabeled points (test set)

transduction: given labeled training set and unlabeled test set, predict the labels on the test set

kernel methods

main goal: separate using a nonlinear classifier

$$a^T \phi(x) = b$$

where ϕ is a nonlinear operator

define the **kernel matrix**

$$K_{ij} = \phi(x_i)^T \phi(x_j)$$

(involves both labeled and unlabeled data)

fact: for transduction, all we need to know to predicts the labels is the kernel matrix (and not $\phi(\cdot)$ itself!)

kernel methods: idea of proof

at the optimum, a is in the range of the labeled data:

$$a = \sum_i \lambda_i x_i$$

$\Longrightarrow$ solution of classification problem depends only on the values of kernel matrix K_{ij} for labeled points x_i, x_j

in a transductive setting, the prediction of labels also involves K_{ij} only, since for an unlabeled data point x_j,

$$a^T \phi(x_j) = \sum_i \lambda_i \phi(x_i)^T \phi(x_j)$$

involves only K_{ij}'s

fact: all previous algorithms can be "kernelized"

partition training/test

in a transductive setting, we can partition the kernel matrix as follows:

$$K = \begin{bmatrix} K_{tr,tr} & K_{tr,t} \\ K_{tr,t}^T & K_{t,t} \end{bmatrix}$$

where subscripts tr and t stand for "training" and "test", respectively

kernel optimization

what is a "good" kernel?

- *margin:* kernel "behaves well" on the training data,
  $\longrightarrow$ condition on the matrix $K_{tr,tr}$
- *test error:* kernel yields low predicted error
  $\longrightarrow$ condition on the full matrix K
- also, to prevent overfitting, the blocks in K should be "entangled"
  $\longrightarrow$ will restrict the search space with affine constraint

kernel optimization and semidefinite programming

main idea: kernel can be described via the Gram matrix of data points, hence is a positive semidefinite matrix

$\longrightarrow$ semidefinite programming plays a role in kernel optimization

margin of SVM classifier

kernel-based SVM problem for labeled data points:

$$\min_{a,b} \|a\|_2 \text{ subject to } y_i(a^T\phi(x_i) + b) \geq 1, \;\; i = 1, \ldots, N$$

(classifier depends only on training set block of kernel matrix K_{tr})

margin of optimal classifier is $\gamma = 1/\|a^*\|_2$

geometrically:

$\gamma^{-1} =$ distance between the convex hulls of the two classes

(can work with "soft" margin when data is not linearly separable)

generalization error

how well the SVM classifier will work on the test set?

from learning theory (Bartlett, Rademacher), generalization error is bounded above by

$$\frac{\sqrt{\mathbf{Tr}\, K}}{\gamma(K_{tr})}$$

where $\gamma(K_{tr})$ is the margin of the SVM classifier with training set block kernel matrix K_{tr}

hence, the constraints

$$\mathbf{Tr}\, K = c, \;\; \gamma(K_{tr})^{-1} \leq w$$

ensure an upper bound on the generalization error

margin constraint

using a dual expression for the SVM problem, margin constraint

$$\gamma(K_{tr}) \geq \gamma$$

writes as LMI (linear matrix inequality) in K

$$\begin{bmatrix} G(K_{tr}) & e+\nu+\lambda\cdot y \\ (e+\nu+\lambda\cdot y)^T & \gamma^{-1} \end{bmatrix} \succeq 0$$

where

- $G(K_{tr})$ is linear in K_{tr}
- $e =$ vector of ones
- λ, ν are new variables

avoiding overfitting

the trace constraint $\mathbf{Tr}\, K = c$ is not enough to "entangle" the matrix K

we impose an affine constraint on K of the form

$$K = \sum_i \mu_i K_i$$

where K_i's correspond to given different, *known* kernels and μ_i's will be our new variables

optimizing kernels: example problem

goal: find a kernel matrix that

- is positive semidefinite and has a given trace

$$K \succeq 0, \quad \mathbf{Tr}\, K = c$$

- belongs to an affine space (here K_i's are known)

$$K = \sum_i \mu_i K_i$$

- satisfies a lower bound γ on the margin on the training set, $\gamma(K_{tr})$

the problem reduces to a semidefinite programming feasibility problem

experimental results

	K_1	K_2	K_3	K^*
Breast cancer	$d = 2$	$\sigma = 0.5$		
margin	0.010	0.136	-	0.300
TSE	19.7	28.8		11.4
Sonar	$d = 2$	$\sigma = 0.1$		
margin	0.035	0.198	0.006	0.352
TSE	15.5	19.4	21.9	13.8
Heart	$d = 2$	$\sigma = 0.5$		
margin	-	0.159	-	0.285
TSE		49.2		36.6

wrap-up

- convex optimization has much to offer and gain from interaction with classification
- described variations on linear classification
- many robust optimization interpretations
- all these methods can be kernelized
- kernel optimization has high potential

see also

- Learning the kernel matrix with semidefinite programming
  Lanckiert, Cristianini, Bartlett, El Ghaoui, Jordan (ICML 2002)
- Minimax Probability Machine
  (Lanckiert, Bhattacharrya, El Ghaoui, Jordan) (NIPS 2001)
- Robust Novelty Detection with Single-Class MPM
  Lanckriet, El Ghaoui, Jordan, (NIPS 2002)