MINIMAX SUPPORT VECTOR MACHINES

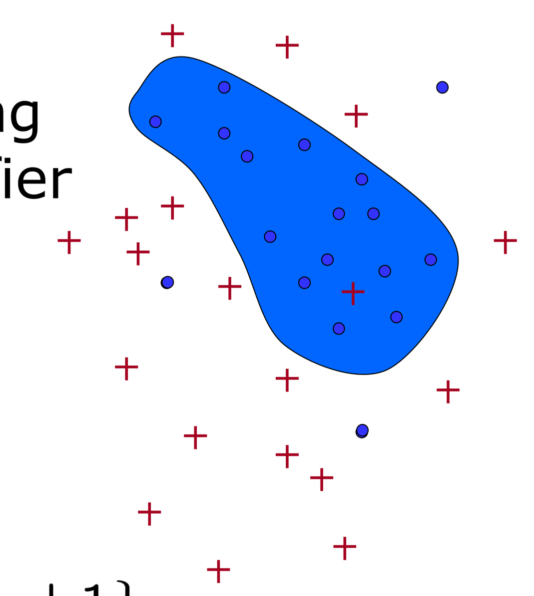
Mark A. Davenport, Richard G. Baraniuk
Rice University
Electrical and Computer Engineering

Clayton D. Scott
University of Michigan
Electrical Engineering and Computer Science

Overview

Classification

given some training data, find a classifier that generalizes



Notation

pattern: $\mathbf{x} \in R^d$
label: $y \in \{-1, +1\}$
classifier: $f : R^d \to \{-1, +1\}$

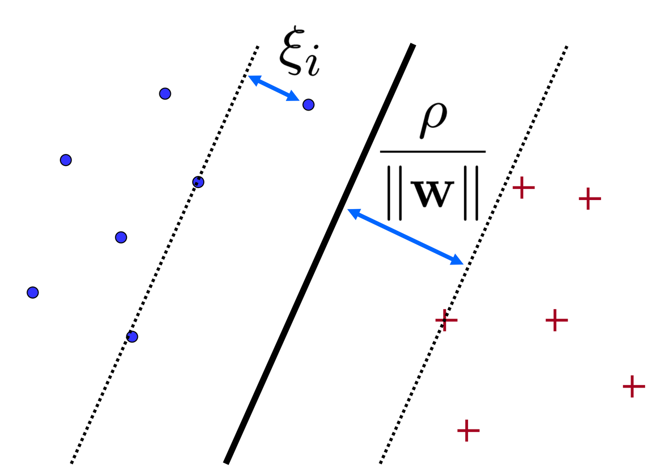
$$P_E(f) := \Pr(f(\mathbf{x}) \neq y)$$

Goal: minimize $P_E(f)$ by minimizing the *misclassification rate* using *support vector machines (SVMs)*

Support Vector Machines

Method for learning from training data
- Use "kernel-trick"
- Maximize the "margin"

$$\min_{\mathbf{w},b,\xi,\rho} \quad \tfrac{1}{2}\|\mathbf{w}\|^2 - \nu\rho + \frac{1}{n}\sum_{i=1}^{n} \xi_i \qquad \nu \in [0,1]$$
$$\text{s.t.} \quad (\langle \mathbf{w}, \mathbf{x_i} \rangle + b) y_i \geq \rho - \xi_i$$



Minimax Learning

False alarm: $P_F(f) := \Pr(f(\mathbf{x}) = +1 | y = -1)$
Miss: $P_M(f) := \Pr(f(\mathbf{x}) = -1 | y = +1)$

$$P_E(f) := \pi_- P_F(f) + \pi_+ P_M$$

$\Pr(y = -1)$ $\qquad$ $\Pr(y = +1)$

True class frequencies are often not represented by the data, resulting in too much/little emphasis on one class

- 100 training samples
- 50 have cancer
- 50 do not

50% of population has cancer

$$f^*_{mm} = \arg\min_f \max\left(P_M(f), P_F(f)\right)$$

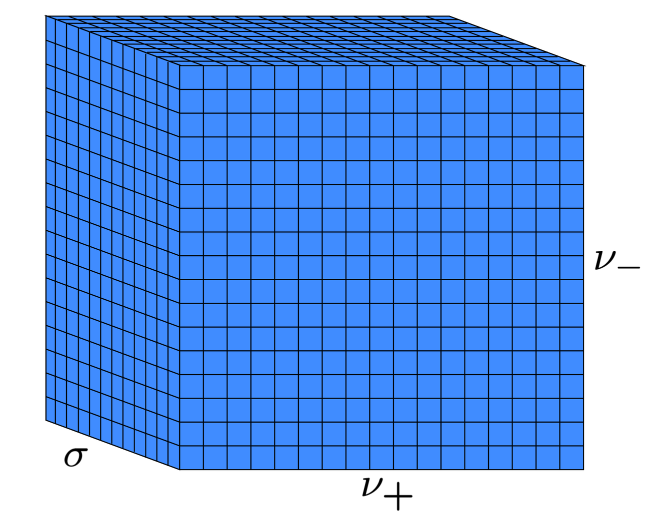
Minimax SVMs

Consider *cost-sensitive* SVMs
- Introduce class-specific weights
- Adjust weights to achieve desired error rates
- Cross-validation (grid search)
  - expensive, high-variance

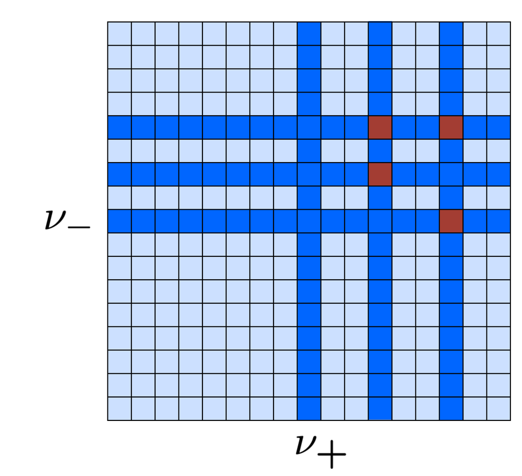
$$\min_{\mathbf{w},b,\xi,\rho} \quad \tfrac{1}{2}\|\mathbf{w}\|^2 - 2\nu_-\nu_+\rho + \frac{\nu_-}{n_+}\sum_{i \in I_+} \xi_i + \frac{\nu_+}{n_-}\sum_{i \in I_-} \xi_i$$
$$\text{s.t.} \quad (\langle \mathbf{w}, \mathbf{X_i} \rangle + b) Y_i \geq \rho - \xi_i$$
$$(\nu_+, \nu_-) \in [0,1]^2$$

Parameter Selection

Possible strategies:



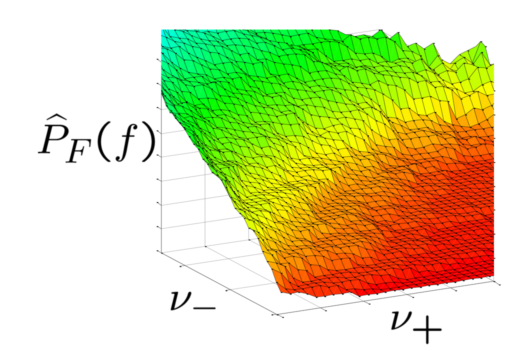
- CV estimates on a grid of parameters
  - slow
  - guaranteed to find "optimal" parameters



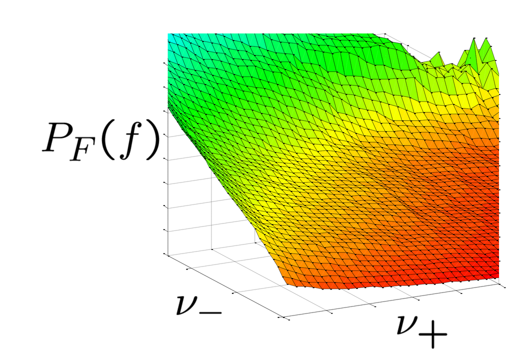
- Coordinate descent
  - fast
  - potentially prone to errors
- Many variants possible

Smoothing

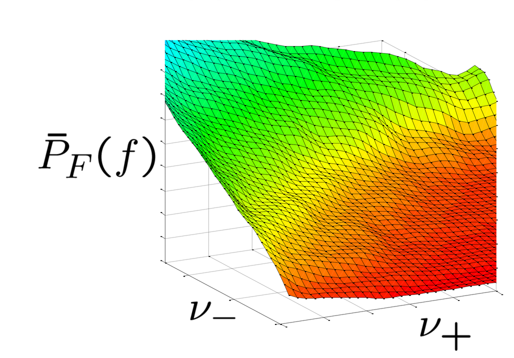
Cross-validation



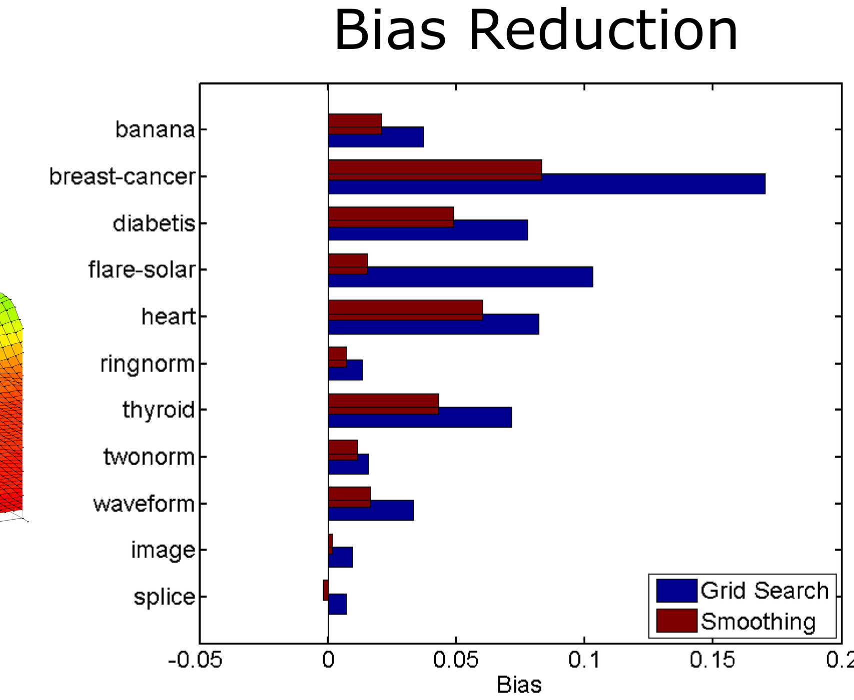
True error rate



Smoothed error estimates



Bias Reduction

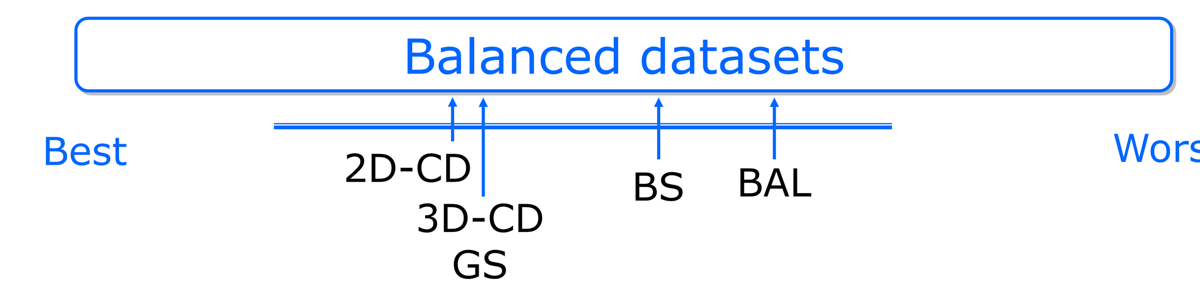


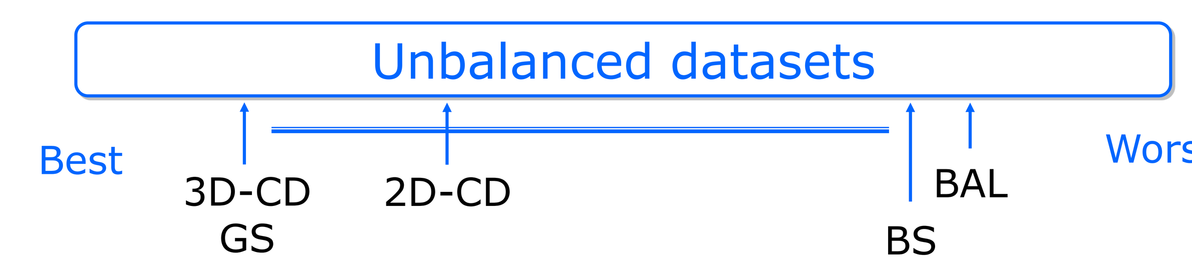
Experiments

11 datasets (100 permutations)
- Full grid search (GS)
- Coordinate descent (2D-CD, 3D-CD)
- Bias-shifting (BS)
- Balanced SVM (BAL)
- Minimax Probability Machine (MPM)

Results

Nemenyi test





Unbalanced Datasets

Dataset	GS	2D-CD	3D-CD	BS	BAL
banana	.193	.194	.189	.218	.226
breast-cancer	.451	.460	.477	.737	.564
diabetes	.340	.340	.338	.449	.455
flare-solar	.410	.412	.425	.548	.595
waveform	.168	.171	.168	.181	.210
image	.134	.133	.157	.097	.151
splice	.195	.196	.200	.335	.379

MPM Comparison

Dataset	SVM	MPM
banana	.619	.517
breast-cancer	.453	.421
diabetes	.332	.319
flare-solar	.392	.399
waveform	.175	.218
image	.320	.362
splice	.239	.296

Key observations

- Accurate error estimation is critical
  - smoothing always helps
- CD is surprisingly effective
- BS and BAL are significantly worse
- The minimax SVM outperforms the MPM
  - even when the MPM parameters are set by an oracle