**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: \( x \in \mathbb{R}^d \)
  - label: \( y \in \{-1, +1\} \)
  - classifier: \( f : \mathbb{R}^d \rightarrow \{-1, +1\} \)
- \( P_E(f) := \Pr(f(x) \neq y) \)

**Goal:**
- Minimize \( P_E(f) \) by minimizing the misclassification rate using support vector machines (SVMs).

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

\[
\begin{align*}
\min_{w, \xi, \nu, \rho} & \quad \frac{1}{2} \|w\|^2 - \nu - \frac{1}{n} \sum_{i=1}^{n} \xi_i \quad \nu \in [0, 1] \\
\text{s.t.} & \quad (w^T x_i + b) y_i \geq \rho - \xi_i
\end{align*}
\]

**Minimax Learning**
- False alarm: \( P_F(f) := \Pr(f(x) = +1 | y = -1) \)
- Miss: \( P_M(f) := \Pr(f(x) = -1 | y = +1) \)
- \( P_E(f) := \pi_P P_F(f) + \pi_M P_M(f) \)

\[
\Pr(y = -1) \quad \Pr(y = +1)
\]

- 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

\[
\min f \quad \text{arg} \min \max (P_M(f), P_F(f))
\]

**Parameter Selection**
- Possible strategies:
  - Cross-validation (grid search) on a grid of parameters
    - slow
    - guaranteed to find "optimal" parameters
  - Coordinate descent
    - fast
    - potentially prone to errors
  - Many variants possible

**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
  - Balanced datasets
  - Unbalanced datasets

**Support Vector Machines**
- Method for learning from training data
  - Use "kernel-trick"
  - Maximize the "margin"

**Smoothing**
- Cross-validation
- True error rate
- Smoothing error estimates
- Bias reduction

**Key observations**
- Accurate error estimation is critical
  - smoothing always helps
- Coordinate descent 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