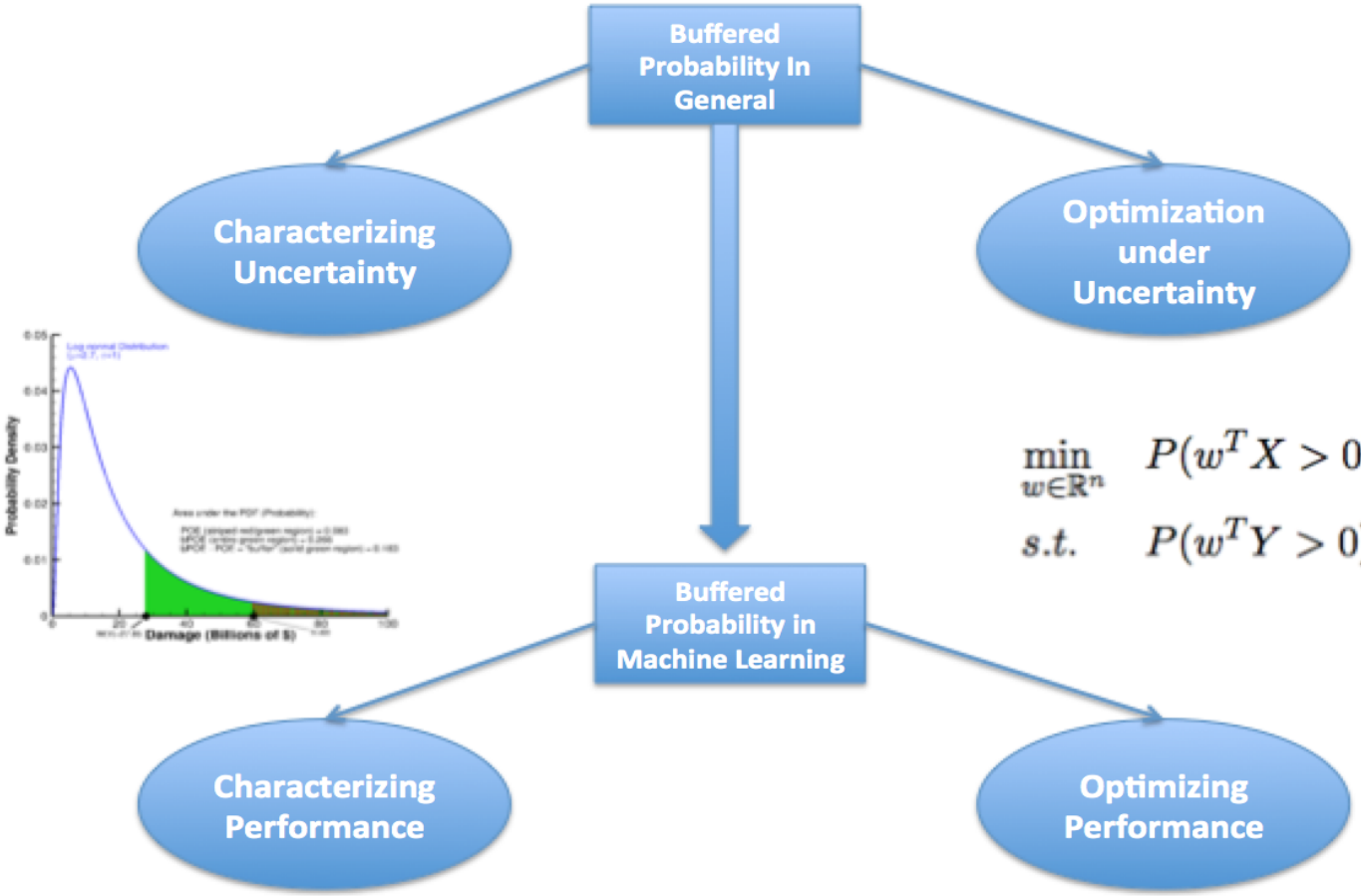
Measuring Performance of Machine Learning Algorithms with Buffered Probability of Exceedance

Matthew Norton
joint work with Stan Uryasev and Alexander Mafusalov

Assistant Professor, Naval Postgraduate School, Operations Research Department

October 1, 2018

General Methodology



$$\min_{w \in \mathbb{R}^n} \quad P(w^T X > 0)$$
$$s.t. \quad P(w^T Y > 0) \leq \alpha$$

The Binary Classification Problem

THE DATA:

We have samples $\{(X_1, y_1), \ldots, (X_N, y_N)\}$ of (X, y) where

- $X_i \in \mathbb{R}^n$: vector of features for sample i
- $y_i \in \{-1, +1\}$: class label of sample i
- $X^+ = (X|y = +1), X^- = (X|y = -1), m^+ + m^- = N$

THE TASK:

Using the labeled samples, construct a function $f_\theta : \mathbb{R}^n \to \mathbb{R}$ that predicts labels.

$$X \to \boxed{f_\theta(X) \leq 0} \to y' = -1$$

$$X \to \boxed{f_\theta(X) > 0} \to y' = +1$$

What would we like to solve?

Common Performance Metrics

- Misclassification Rate $:= P(-yf_\theta(X) > 0)$
- AUC $:= 1 - P(f_\theta(X^+) < f_\theta(X^-))$
- False Positive Rate (FPR) $:= P(f_\theta(X^-) > 0)$
- False Negative Rate (FNR) $:= P(-f_\theta(X^+) > 0)$

Ideal Problem: $\xi(f_\theta(X), X) :=$ Error in prediction $f_\theta(X)$ given input X

$$\min_\theta P(\xi_0(f_\theta(X), X) > 0)$$
$$s.t. \;\; P(\xi_1(f_\theta(X), X) > 0) \leq \alpha$$

What is usually solved?

Popular in ML: Empirical Risk

$$\min_{\theta} E[\ell_0(f_\theta(X), X)]$$
$$s.t. \;\; E[\ell_1(f_\theta(X), X))] \leq R$$

- What ℓ do we choose? Does it sufficiently encode the true objective /requirements/ performance metrics?
- How do we set R? Particularly if $E[\ell(f_\theta(X), X)]$ only approximates our *real* performance metric.

Example: Error type control in classification

$$\begin{aligned} \min_{\theta} \quad & P(-f_\theta(X^+) > 0) \\ s.t. \quad & P(f_\theta(X^-) > 0) \leq \alpha_{fp} \end{aligned} \tag{1}$$

- Scanning for explosives in the airport (DHS 1.5$ million Kaggle competition)
- Imbalanced Data Sets
- Early warning detection systems

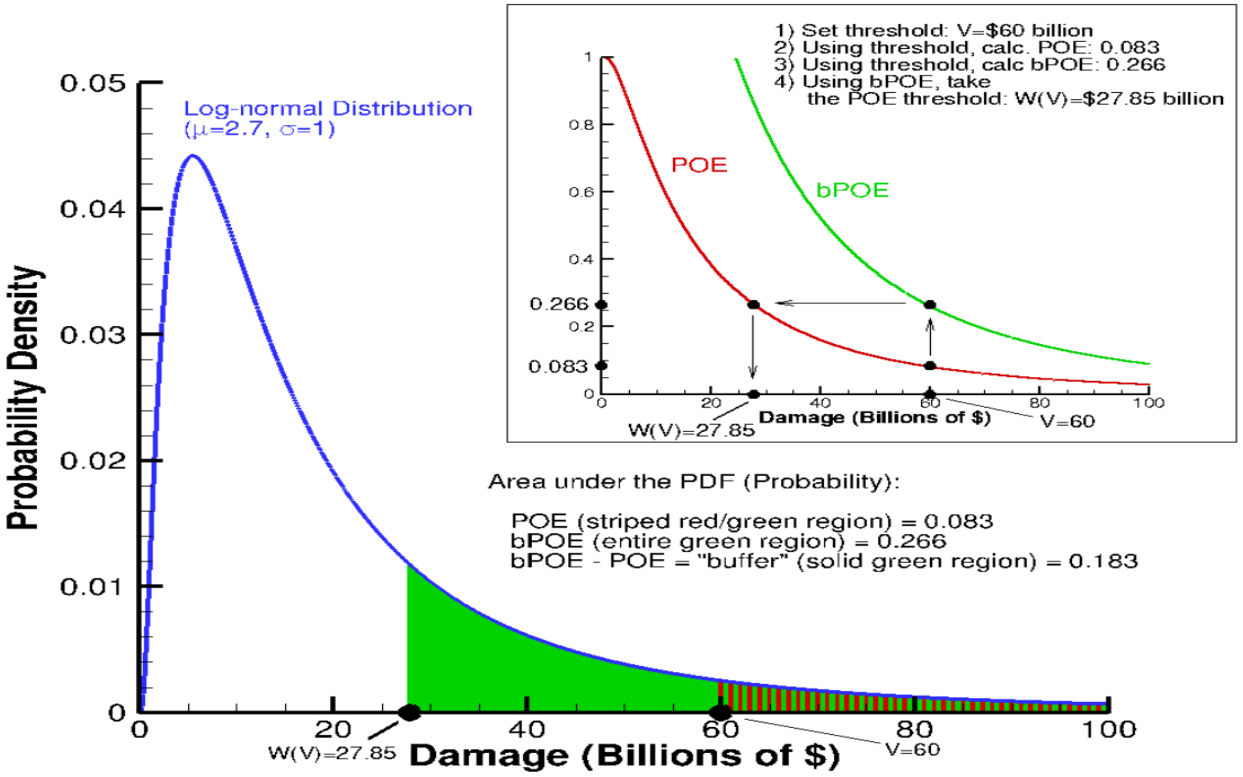
What is Buffered Probability and why is it important?

Probability of Exceedance (POE)

Proportion of largest outcomes (losses) exceeding threshold $z \in \mathbb{R}$.

Buffered Probability of Exceedance (bPOE)

Proportion of largest outcomes (losses) that average to threshold $z \in \mathbb{R}$.



Probability of Exceedance

Some Notation...

- X:= random variable
- $z \in \mathbb{R}$:= threshold level
- $\alpha \in [0, 1]$:= probability level
- $F_z(X) = P(X \leq z)$:= CDF

Quantile (VaR) & Probability of Exceedance (POE), Continuous Case

- $q_\alpha(X)$:= $\{z : P(X \leq z) = \alpha\}$ =quantile of X at $\alpha \in [0, 1]$
- $p_z(X)$:= $\{1 - \alpha : q_\alpha(X) = z\}$ = POE of X at level z
  or equivalently, $\quad p_z(X) = 1 - F_z(X) = P(X > z)$

Buffered Probability of Exceedance (bPOE)

Some Notation...

- X:= random variable
- $z \in \mathbb{R}$:= threshold level
- $\alpha \in [0, 1]$:=probability level

Superquantile (CVaR) & Buffered POE (bPOE), Continuous Case

- $\bar{q}_\alpha(X) := E[X|X > q_\alpha(X)]$ = *superquantile* of X at $\alpha \in [0, 1]$
- $\bar{p}_z(X) := \{1 - \alpha : \bar{q}_\alpha(X) = z\}$ = *bPOE* of X at level z

- *bPOE* is a generalization of *Buffered Probability of Failure* Royset & Rockafellar (2010) and can be defined by *superdistribution* studied by Royset & Rockafellar (2013)
- *bPOE* studied in general context also by Mafuslalov & Uryasev (2014)

bPOE Formula: Norton, Mafusalov, Uryasev (2014)

Let $[\cdot]^+ = \max\{0, \cdot\}$. bPOE of X at threshold $z \in \mathbb{R}$ equals:

$$\bar{p}_z(X) = \inf_{a \geq 0} E[a(X - z) + 1]^+ \tag{2}$$

Calculation of POE

Let $I\{\cdot\}$ be 0-1 indicator. POE of X at threshold $z \in \mathbb{R}$ equals:

$$P(X > z) = E[I\{X > z\}] \tag{3}$$

Easy bPOE Optimization

Linear loss function $L(w, X) = w^T X$ with parameters $w \in \mathbb{R}^n$ and random vector $X \in \mathbb{R}^n$.

Minimize POE of loss at threshold $z = 0$ ... Hard Problem!

$$\min_w P(L(w, X) \geq 0)$$

Minimize bPOE of loss at threshold $z = 0$ ... Easy Problem!

$$\begin{aligned}
\min_w \bar{p}_0(L(w, X)) &= \min_w \min_{a \geq 0} E[a(L(w, X) - 0) + 1]^+ \\
&= \min_{w, a \geq 0} E[L(aw, X) + 1]^+ \\
&= \min_y E[L(y, X) + 1]^+
\end{aligned}$$

Measuring Classifier Performance with POE

$$\underline{\text{Performance} = P(\xi(f_\theta, X) > 0)}$$

- Misclassification Rate $:= P(-yf_\theta(X) > 0)$
- AUC $:= 1 - P(f_\theta(X^+) < f_\theta(X^-))$
- False Positive Rate (FPR) $:= P(f_\theta(X^-) > 0)$
- False Negative Rate (FNR) $:= P(-f_\theta(X^+) > 0)$

Measuring Classifier Performance with bPOE

Define metrics with bPOE and an error function:

$$\underline{\text{Performance} = \bar{p}_0(\xi(f_\theta, X))}$$

- *Buffered* Misclassification Rate $:= \bar{p}_0(-yf_\theta(X))$
  - Equivalent to SVM when minimized (Norton, Mafusalov, Uryasev 2017).
- *Buffered* AUC $:= 1 - \bar{p}_0(f_\theta(X^-) - f_\theta(X^+))$
  - Equivalent to RankSVM when maximized (Norton, Uryasev 2018).
- Works for Regression: *Buffered* Residuals $:= \bar{p}_\epsilon(|f_\theta(X) - y|)$
  - Equivalent to SV Regression when minimized (Norton 2018).

Neyman-Pearson Classification

Ideal Approach: Minimize False Negative Rate such that False Positive Rate is less than $\alpha_{fp} \in [0, 1)$.

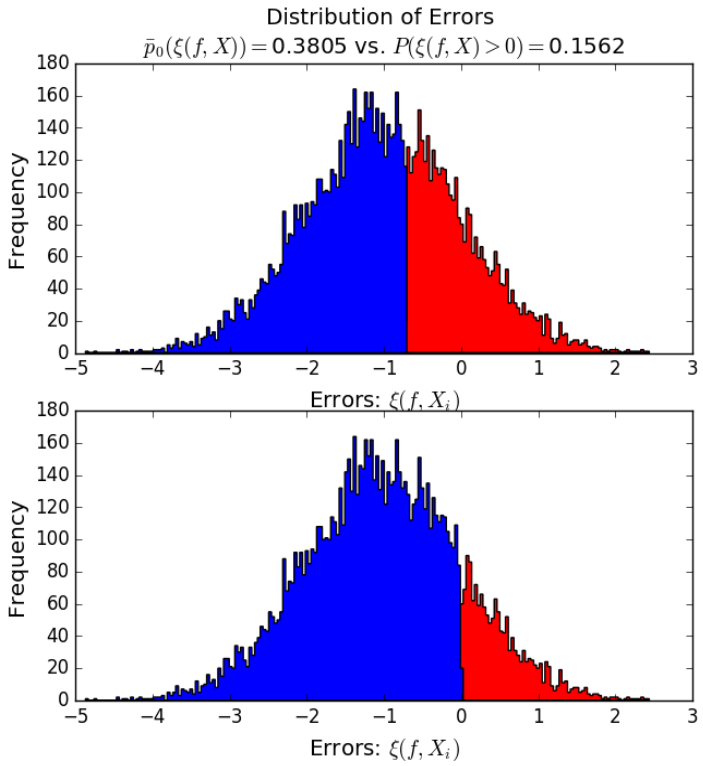
$$\begin{aligned} \min_{\theta} \quad & P(-f_\theta(X^+) > 0) \\ s.t. \quad & P(f_\theta(X^-) > 0) \leq \alpha_{fp} \end{aligned} \tag{4}$$

bNP classification: Primal Formulation

New Approach: Minimize Buffered False Negative Rate with Buffered False Positive Rate constraint

- *Buffered* False Positive Rate (bFPR) $:= \bar{p}_0(f_\theta(X^-))$
- *Buffered* False Negative Rate (bFNR) $:= \bar{p}_0(-f_\theta(X^+))$

$$\begin{aligned} \min_\theta \quad & \bar{p}_0(-f_\theta(X^+)) \\ s.t. \quad & \bar{p}_0(f_\theta(X^-)) \leq \alpha_{fp} \end{aligned}$$



bNP classification: The Linear Case

Assume linear f and modify error function to be class-specific *margin error* and consider non-zero threshold $z \in \mathbb{R}$.

$$\xi(f, X) \rightarrow \frac{-y(w^T X + b))}{\|w\|_2}$$

$$\begin{aligned} \min_{w,b} \quad & \bar{p}_z \left(\frac{-w^T X^+ - b}{\|w\|_2} \right) \\ s.t. \quad & \bar{p}_0 \left(\frac{w^T X^- + b}{\|w\|_2} \right) \leq \alpha_{fp} \end{aligned} \quad \equiv \quad \begin{aligned} & \min_{w,b,\gamma} E[-w^T X^+ - b + 1]^+ - z\|w\|_2^2 \\ & \gamma + \frac{E[w^T X^- + b - \gamma]^+}{\alpha_{fp}} \leq 0 \end{aligned}$$

bNP classification: Dual Formulation and Connection to SVM's

$$\max_{\beta} \quad \sum_{i|y_i=+1} \beta_i + \frac{1}{4z} \sum_{i=1}^{N} \sum_{j=1}^{N} y_i y_j \beta_i \beta_j X_i^T X_j$$

$$s.t. \quad 0 \leq \beta_i \leq \frac{1}{m^+} \quad , \text{ if } y_i = +1$$

$$0 \leq \beta_i \leq \frac{1}{m^- \alpha_{fp}} \sum_{i|y_i=-1} \beta_i \quad , \text{ if } y_i = -1$$

$$\sum_i y_i \beta_i = 0$$

- QP with linear constraints
- Optimal solution has support vector expansion.

bNP classification: Dual Formulation and Connection to SVM's

$$\max_{\beta} \quad \sum_{i|y_i=+1} \beta_i + \frac{1}{4z} \sum_{i=1}^{N} \sum_{j=1}^{N} y_i y_j \beta_i \beta_j k(X_i, X_j)$$

$$s.t. \quad 0 \leq \beta_i \leq \frac{1}{m^+} \quad , \text{ if } y_i = +1$$

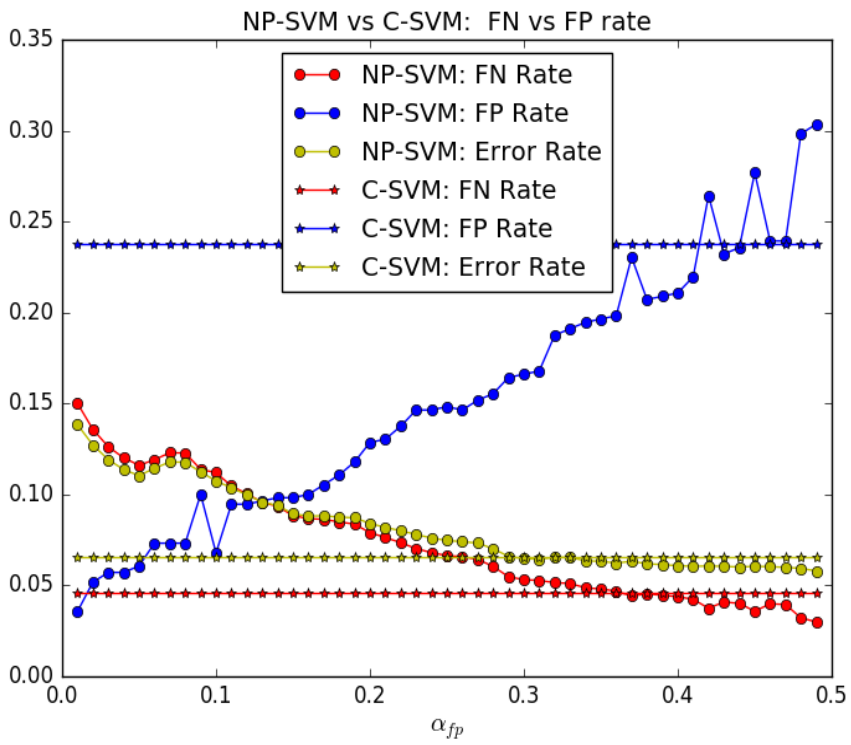
$$0 \leq \beta_i \leq \frac{1}{m^- \alpha_{fp}} \sum_{i|y_i=-1} \beta_i \quad , \text{ if } y_i = -1$$

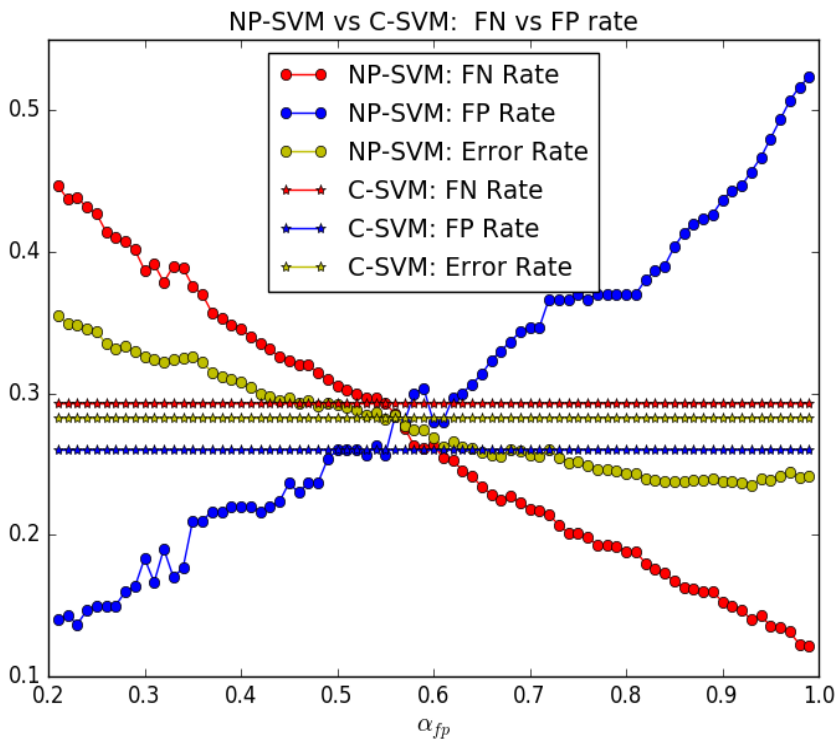
$$\sum_i y_i \beta_i = 0$$

- Can use Kernel Trick for non-linear classification.
- $k(X_i, X_j) = \phi(X_i)^T \phi(X_j)$
- $\phi : \mathbb{R}^n \to \mathbb{R}^d, d >> n.$
- e.g. $k(X_i, X_j) = e^{-c\|X_i - X_j\|_2}$

bNP classification: Experiments

Trade-off curves for many α_{fp} compared to baseline C-SVM for two datasets, Spam Detection (Left) and Credit Default Prediction (Right).





Thank You!

Thank You!
Questions?