Soft Margin Support Vector Classification as Buffered Probability Minimization

Matthew Norton, Alexander Mafusalov, Stan Uryasev

Risk Management and Financial Engineering Lab,
University of Florida

July 29, 2015

Outline

Outline

VaR and POE

Some Notation...

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

Quantile (VaR) & Probability of Exceedance (POE)

- $q_\alpha(Y)$:= $\min\{z : P(Y \leq z) \geq \alpha\}$ =quantile of Y at $\alpha \in [0, 1]$

- *POE at z*: probability level where $q_\alpha(Y) = z$
  or $\{1 - \alpha : q_\alpha(Y) = z\} = P(Y > z) = p_z(Y)$

CVaR and bPOE

Some Notation...

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

Superquantile (CVaR) & Buffered Probability of Exceedance (bPOE)

- $\bar{q}_\alpha(Y) := E[Y|Y > q_\alpha(Y)]$ = *superquantile* of Y at $\alpha \in [0, 1]$

- *bPOE at z*: probability level where $\bar{q}_\alpha(Y) = z$
  or $\{1 - \alpha : \bar{q}_\alpha(Y) = z\} = \bar{p}_z(Y)$

POE vs bPOE?

Probability of Exeedance (POE)

- POE is concerned with the proportion of events exceeding a threshold $z \in \mathbb{R}$.
- DOES NOT consider the magnitude of these events.
- Considers only a count of events exceeding a threshold $z \in \mathbb{R}$.

Buffered Probability of Exceedance (bPOE)

- bPOE is concerned with the proportion of events, that when considered together, have average magnitude equal to some threshold $z \in \mathbb{R}$.
- bPOE is a probability measurement that accounts for the magnitude of tail events.

Why is bPOE important?

POE can hide critical information

- When the distribution of Y is heavy tailed, the magnitude of tail events is important.

Example: Hurricane Damage Data

Threshold (Damage in $ billions)	POE (%)	bPOE (%)
100	1	3
50	4	10
10	15	69
1	48	100
0.1	79	100

- Damage Data = Heavy-tailed distribution
- Notice: bPOE reflects heavy-tail (e.g. at threshold $10 billion)

Outline

Calculation of bPOE & Superquantile

Calculation of superquantile

Let $E[\cdot]^+ = E[\max(\cdot, 0)]$. The superquantile of Y at probability level $\alpha \in [0, 1]$ equals:

$$\bar{q}_\alpha(Y) = \min_\gamma \gamma + \frac{E[Y - \gamma]^+}{1 - \alpha}$$

Calculation of bPOE

bPOE of Y at threshold $z \in \mathbb{R}$ equals:

$$\bar{p}_z(Y) = \inf_{\gamma < z} \frac{E[Y - \gamma]^+}{z - \gamma} \tag{1}$$

Outline

Easy Probability Minimization

Consider a linear loss function $L(w, Y) = w^T Y$ with parameters $w \in \mathbb{R}^n$ and real valued random vector $Y \in \mathbb{R}^m$.

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

$$\min_w P(L(w, Y) > 0)$$

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

$$\min_w \bar{p}_0(L(w, Y)) = \min_w E[L(w, Y) + 1]^+$$

The Binary Classification Problem

THE DATA:

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

- $X_i \in \mathbb{R}^n$: vector of features for sample i
- $y_i \in \{-1, +1\}$: class label of sample i

THE TASK:

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

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

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

Application Examples:

- Credit scoring: $X =$ financial indicators; $y =$ creditworthy or not?
- Medical: $X =$ patient health indicators; $y =$ has disease or not?
- E-mail filtering: $X =$ e-mail text; $y =$ spam or not?
- Recommendation: $X =$ past purchases; $y =$ beer or wine?

Outline

Margin Maximization: Support Vector Machines

Let's Think Geometrically

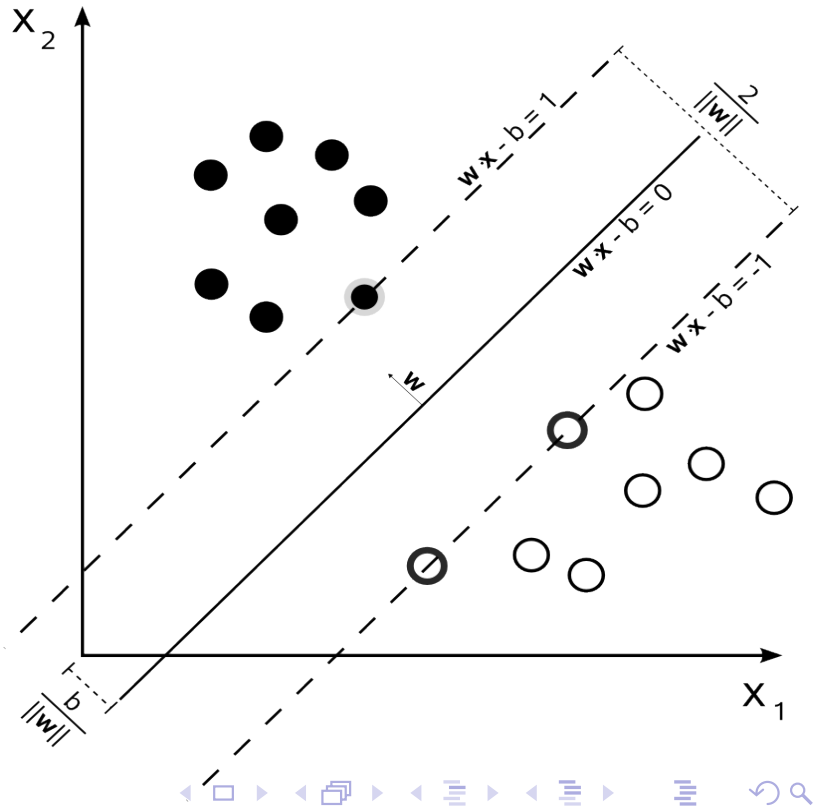
Find a 'separating' hyperplane

What is our function (i.e. hyperplane)?

$$f(X) = w^T X - b$$

What makes a good hyperplane? One with *maximum margin*

$$\max_{w,b} \frac{1}{\|w\|_2} \equiv \min_{w,b} \|w\|_2$$



Finding the Maximum Margin Hyperplane: SVM formulation

Step 1: Define our random loss function:

- $L(w, b, X_i, y_i) = -y_i(w^T X_i - b)$
- If prediction is wrong, $-y_i(w^T X_i - b) > 0$
- If prediction is right, $-y_i(w^T X_i - b) < 0$

Step 2: Define our formulation to solve for max margin hyperplane:

Hard Margin Support Vector Classifier

$$\begin{aligned} \min_{w,b} \quad & \|w\|_2 \\ s.t. \quad & 0 \geq -y_i(w^T X_i - b) + 1 \quad , \forall\, i \in \{1, ..., N\} \end{aligned} \tag{2}$$

Q: What if data not linearly separable?
A : Introduce tradeoff between in-sample accuracy and margin size

Soft Margin Support Vector Classifier: The C-SVM

$$\begin{aligned} \min_{w,b,\xi} \quad & C\|w\|_2 + \sum_{i=1}^{N} \xi_i \\ s.t. \quad & \xi_i \geq -y_i(w^T X_i - b) + 1 \quad , \forall\, i \in \{1, ..., N\} \\ & \xi \geq 0 \end{aligned} \tag{3}$$

Hard Margin Support Vector Classifier

$$\begin{aligned} \min_{w,b} \quad & \|w\|_2 \\ s.t. \quad & 0 \geq -y_i(w^T X_i - b) + 1 \quad , \forall\, i \in \{1, ..., N\} \end{aligned} \tag{4}$$

Outline

Let's Think Probabilistically

Find hyperplane that minimizes probability of losses

Step 1: Define our random loss function:

- $L(w, b, X_i, y_i) = -y_i(w^T X_i - b)$
- If prediction is wrong, $-y_i(w^T X_i - b) > 0$
- If prediction is right, $-y_i(w^T X_i - b) < 0$

Step 2: Define formulation: Minimize probability of losses exceeding some threshold $-C \leq 0$:

$$\min_{w,b} \quad p_{-C}\left(\frac{-y(w^T X - b)}{\|w\|}\right) \equiv \min_{w,b} \quad P\left(\frac{-y(w^T X - b)}{\|w\|} \geq -C\right)$$

Problem! With empirical observations, probability minimization is non-convex, discontinuous.
Idea: Minimize buffered probability of exceedance instead!

$$\min_{w,b} \quad \bar{p}_{-C}\left(\frac{-y(w^T X - b)}{\|w\|}\right) \tag{5}$$

$$\downarrow$$

Using bPOE formula, (5) becomes the EC-SVM

$$\begin{aligned} \min_{w,b,\xi} \quad & \sum_{i=1}^{N} \xi_i \\ s.t. \quad & \xi_i \geq C\|w\| - y_i(w^T X_i - b) + 1 \\ & \xi \geq 0 \end{aligned} \tag{6}$$

Outline

bPOE minimization and C-SVM: Equivalent problems!

Soft Margin Maximization: The C-SVM

$$\begin{aligned} \min_{w,b,\xi} \quad & C\|w\|_2 + \sum_{i=1}^{N} \xi_i \\ s.t. \quad & \xi_i \geq -y_i(w^T X_i - b) + 1 \\ & \xi \geq 0 \end{aligned} \tag{7}$$

bPOE minimization: The EC-SVM

$$\begin{aligned} \min_{w,b,\xi} \quad & \sum_{i=1}^{N} \xi_i \\ s.t. \quad & \xi_i \geq C\|w\| - y_i(w^T X_i - b) + 1 \\ & \xi \geq 0 \end{aligned} \tag{8}$$

What did we prove?

- C-SVM and EC-SVM are equivalent for $C \geq 0$.
- Equivalence holds for ANY general norm.

Outline

Why is this interesting? What else have we proved?

C-SVM

- Derived with geometric intuition
- Previously, no interpretation for C-parameter
- Previously, no interpretation for optimal objective value

EC-SVM

- Derived from probabilistic intuition
- Free parameter, C, has statistical interpretation (bPOE threshold)
- Optimal objective value is a probability level

"..the parameter C has no intuitive meaning."
Shawe-Taylor, Cristianini (2004)

Outline

A Larger Framework

A Larger Framework: Superquantiles and Soft Margins

$\hat{C} \in [0, +\infty)$		$C \in (-\infty, +\infty)$	$=$	$-z$
$\uparrow$		$\uparrow$		$\uparrow$
C-SVM	$\subset$	EC-SVM	$\equiv$	$\min_{w,b} \bar{p}_z\left(-y(w^T X - b)\right)$
$\Updownarrow$		$\Updownarrow$		$\Updownarrow$
ν-SVM	$\subset$	Eν-SVM	$\equiv$	$\min_{w,b} \bar{q}_\alpha\left(-y(w^T X - b)\right)$
$\downarrow$		$\downarrow$		$\downarrow$
$\hat{\nu} \in (\nu_{min}, \nu_{max}]$		$\nu \in [0, 1]$	$=$	$1 - \alpha$

Key:

- $\Updownarrow$ — Formulations generate same set of optimal hyperplanes.
- $\subset$ — The right hand side formulation is an "extension" of the left hand formulation.
- $\equiv$ — Formulations are objectively equivalent.
- $\uparrow$ or $\downarrow$ — Arrow points to parameter values for the formulation

Summary

- SVM's are an extremely popular tool for classification
- We showed that SVM's can also be viewed as a simple bPOE minimization problem
- We derived an equivalent SVM that is more interpretable than the classical formulation
- We showed that it fits into a larger framework, fully connecting soft margin classification with superquantile concepts

1. Barbero, A. , Takeda, A. , Lopez, J. Geometric Intuition and Algorithms for Ev-SVM Journal of Machine Learning Research 16 (2015) 323-369
2. Chang, C. C., and Lin, C. J. (2001) Training v-Support Vector Classifiers: Theory and Algorithms Neural computation, 13(9), 2119-2147
3. Cortes, C., Vapnik V. (1995) Support-vector networks Machine learning 20.3 (1995): 273-297.
4. Davis, J.R., Uryasev S. (2014). Analysis of Hurricane Damage using Buffered Probability of Exceedance. Research Report 2014-4, ISE Dept., University of Florida.
5. Mafusalov, A. and S. Uryasev. (2014) Buffered Probability of Exceedance: Mathematical Properties and Optimization Algorithms Research Report 2014-1, ISE Dept., University of Florida, October 2014

6. Norton M., Mafusalov, A, Uryasev S. (2015) Research Report 2014, ISE Dept., University of Florida
7. Norton M., Uryasev S. (2014) Maximization of AUC and Buffered AUC in Classification Research Report 2014-2, ISE Dept., University of Florida, October 2014
8. Perez-Cruz, F., Weston, J., Herrmann, D. J. L., and Scholkopf, B. (2003) Extension Of The ν-SVM Range For Classification NATO SCIENCE SERIES SUB SERIES III COMPUTER AND SYSTEMS SCIENCES, 190, 179-196.
9. Rockafellar, R. T., and Uryasev, S. (2002) Conditional value-at-risk for general loss distributions Journal of Banking and Finance (2002), 1443??1471.
10. Rockafellar, R.T. (2009). Safeguarding Strategies in Risky Optimization. Presentation at the International Workshop on Engineering Risk Control and Optimization, Gainesville, FL,

February, 2009.

11. Rockafellar R.T., Royset J.O. (2010). On Buffered Failure Probability in Design and Optimization of Structures. Reliability Engineering & System Safety, Vol. 95, 499-510.

12. Schoelkopf, B., Smola, A. J., Williamson, R. C., and Bartlett, P. L. (2000) New Support Vector Algorithms Neural computation, 12(5), 1207-1245

13. Takeda, A., and Sugiyama, M. (2008) v-Support Vector Machine as Conditional Value-At-Risk Minimization In Proceedings of the 25th international conference on Machine learning (pp. 1056-1063). ACM.

14. Uryasev, S. (2014) Buffered Probability of Exceedance and Buffered Service Level: Definitions and Properties. Report 2014-3, ISE Dept., University of Florida

$$\begin{aligned} \max_{\beta} \quad & \sum_{i=1}^{N} \beta_i \\ s.t. \quad & \| \sum_{i=1}^{N} \beta_i y_i x_i \|^* \leq C \\ & \sum_{i=1}^{N} \beta_i y_i = 0 \\ & 0 \leq \beta_i \leq 1 \ , \ \forall \ i \in \{1, \ldots, N\} \end{aligned}$$

$$\begin{aligned} \max_{\beta} \quad & \sum_{i=1}^{N} \beta_i \\ s.t. \quad & \| \sum_{i=1}^{N} \beta_i y_i x_i \|^* \leq C \sum_{i=1}^{N} \beta_i \\ & \sum_{i=1}^{N} \beta_i y_i = 0 \\ & 0 \leq \beta_i \leq 1 \ , \ \forall \ i \in \{1, \ldots, N\} \end{aligned}$$

$$\begin{aligned} \max_{\beta} \quad & \sum_{i=1}^{N} \beta_i \\ s.t. \quad & \sum_{i=1}^{N} \sum_{i=1}^{N} \beta_i \beta_j (y_i y_j \phi(x_i)^T \phi(x_j) - C^2) \leq 0 \\ & \sum_{i=1}^{N} \beta_i y_i = 0 \\ & 0 \leq \beta_i \leq 1 \ , \ \forall \ i \in \{1, \ldots, N\} \end{aligned}$$