Appendix A

The training of SVM models for prediction and classification is traditionally formulated as the goal of learning a maximum-margin hyperplane (i.e., a decision boundary) that uses information in features to separate observations belonging to two different labeled classes. In the context of post-PCI risk stratification, these features may correspond to demographic, comorbidity and laboratory variables, with the labeled observations corresponding to patients who experience adverse outcomes (positive examples) or remain event free (negative examples). As shown in Figure A1, the maximum margin hyperplane corresponds to the decision boundary with the maximal distance from any of the training examples. The choice of a maximum-margin hyperplane is supported by theoretical results in statistical learning that this approach maximizes the ability to correctly classify previously unseen examples [4].

Denoting the training data for model development as n training examples of the form $(\mathbf{x}_1,y_1),...,(\mathbf{x}_n,y_n)$, where each $(\mathbf{x}_i,y_i)$ tuple denotes the risk variables $\mathbf{x}_i$ for patient i and the corresponding label $y_i \in \{+1,-1\}$ signifies whether the patient experienced an event (+1) or remained event free (-1), the traditional formulation of SVM training learns a maximum-margin linear boundary of the form $\hat{y}_i = \text{sgn}(\mathbf{w}^T\mathbf{x}_i)$. In this formulation $\hat{y}_i$ is the predicted label for observation $\mathbf{x}_i$ and sgn(.) represents the signum function (+1 if input is greater than zero and -1 otherwise). The weights $\mathbf{w}$, which parameterize the linear boundary, are learned during training by solving the following convex optimization problem:

$$\min_{\mathbf{w},\xi\geq 0} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n}\xi_i$$
$$\text{subject to } \forall_i : y_i\left(\mathbf{w}^T\mathbf{x}_i\right) \geq 1-\xi_i$$

where slack variables ξ_i accounts for training examples that are not linearly separable.

Details on solving the SVM optimization problem are presented elsewhere [5]. Essentially, the formulation above corresponds to finding the maximum margin hyperplane (i.e., the $\|\mathbf{w}\|^2$ term in the minimization objective) subject to classifying the training examples correctly or using some minimal slack to account for cases that are not linearly separable. To further obtain probabilistic estimate for patients, SVM classification is supplemented with Platt scaling

of the outputs (Appendix C).

Appendix B

Similar to the notation used in Appendix A, the training data for developing an SVM model optimized for cohort-level performance can be represented as n training examples of the form $(\mathbf{x}_1,y_1),...,(\mathbf{x}_n,y_n)$, where each $(\mathbf{x}_i,y_i)$ tuple denotes the risk variables $\mathbf{x}_i$ for patient i and the corresponding label $y_i \in \{+1,-1\}$ signifying whether the patient experienced an event (+1) or remained event free (-1). The AUROC for a model parameterized by the weight vector $\mathbf{w}$ (i.e., $\hat{y}_i = \mathbf{w^T x}_i$ is the label predicted by the model) can then be computed from the number of swapped pairs Φ corresponding to the number of pairs of examples that are ranked in the incorrect order as:

$$\text{AUROC} = 1 - \Phi / n^+ n^-$$

where:

$$\Phi = \left| \{(i, j) : (y_i > y_j) \text{ and } (\mathbf{w}^T \mathbf{x}_i < \mathbf{w}^T \mathbf{x}_j)\} \right|$$

and n^+ and n^- represent the number of positive and negative examples respectively.

An alternate formulation of measuring the AUROC can be obtained by re-expressing the training data in terms of comparable positive-negative pairs. In this case, the data are represented as tuples of the form $(\mathbf{x}_{ij},y_{ij})$ where $y_{ij} = 1$ and $\mathbf{x}_{ij} = \mathbf{x}_i - \mathbf{x}_j$ for a given pair of positive $(\mathbf{x}_i)$ and negative $(\mathbf{x}_j)$ training examples. The error between the predicted $\hat{y}_{ij} = \mathbf{w^T x}_{ij} = (\mathbf{w^T x}_i - \mathbf{w^T x}_j)$ and y_{ij} is then proportional to 1 - AUROC. Defining this quantity as the AUROC loss function:

$$\Delta_{\text{AUROC}}(\mathbf{y}, \hat{\mathbf{y}}) = \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{1}{2} (1 - \hat{y}_{ij}) = \Phi$$

where $\mathbf{y} = (1,..,1)^T$ and $\hat{\mathbf{y}}$ is a vector denoting the $\hat{y}_{ij}$ predicted by the model stacked together, SVM training can be framed as finding a solution to the following optimization problem:

$$\min_{\mathbf{w},\xi\geq 0} \frac{1}{2}\|\mathbf{w}\|^2 + C\xi$$

$$\text{subject to } \forall_{\hat{\mathbf{y}}\in\mathbf{Y}\setminus\mathbf{y}} : \mathbf{w}^T[\sum_{i=1}^{n} y_i\mathbf{x}_i - \sum_{i=1}^{n} \hat{y}_i\mathbf{x}_i] \geq \Delta_{\text{AUROC}}(\mathbf{y},\hat{\mathbf{y}}) - \xi$$

where ξ corresponds to the slack variable, C represents the cost variable, and $\mathbf{Y}$ is the universe of possible vectors $\hat{\mathbf{y}}$. Due to the exponential size of $\mathbf{Y}$, the problem is solved efficiently using a sparse approximation-based approach [1,6].

Similar to the discussion for the traditional SVM in Appendix A, probabilistic estimates for SVM classification optimized for cohort-level performance metrics are obtained through Platt scaling of the model outputs (Appendix C).

Appendix C

To produce probabilistic risk estimates for each patient, the output produced by traditional SVM classification (Appendix A) or the use of SVM classification optimized for cohort-level performance (Appendix B) are transformed using the improved method for Platt scaling [2] proposed by Lin et al. [3]. In this case, the probability of an event is estimated as:

$$\Pr(y_i = 1 \mid \mathbf{x}_i) \equiv \frac{1}{1+\exp(A\mathbf{w}^T\mathbf{x}_i + B)}$$

where the parameters A and B of Platt scaling are determined by regularized maximum likelihood estimation on data from patients undergoing PCI in 2004 [3].

Appendix Figure:

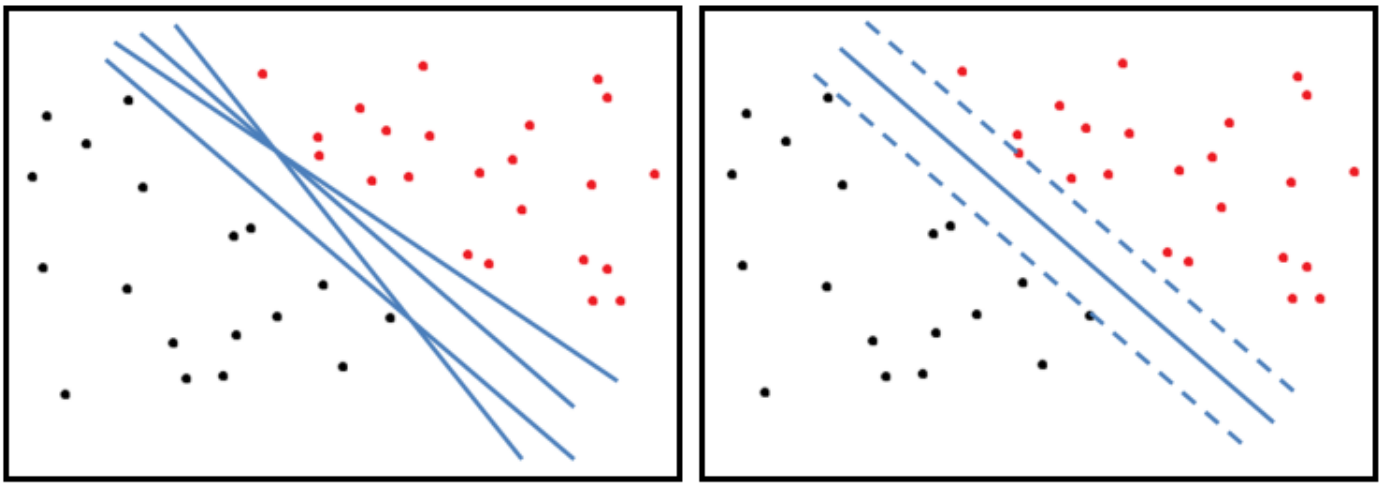

Figure A1: Given training data from two different classes (colored red and black) there are a potentially infinite number of decision boundaries that can separate the data (left panel). The maximum-margin hyperplane, roughly speaking, chooses the boundary that is 'in the middle' and provides a separation from the closest examples in an attempt to maximize the margin of error (right panel).

Supplemental References

1. Tsochantaridis I, Hofmann T, Joachims T, et al. Support vector machine learning for interdependent and structured output spaces; 2004. pp. 104-111.

2. Platt J Probabilistic outputs for support vector machines. Bartlett P Schoelkopf B Schurmans D Smola, AJ, editor, Advances in Large Margin Classifiers: 61–74.

3. Lin HT, Lin CJ, Weng RC (2007) A note on Platt's probabilistic outputs for support vector machines. Machine Learning 68: 267-276.

4. Vapnik VN, Lerner AY, Chervone AY (1965) Learning Systems for Pattern Recognition Using Generalized Portraits. Engineering Cybernetics, (1) 63.

5. Cristianini N, Shawe-Taylor J (2000) An introduction to support vector machines: and other kernel-based learning methods. Cambridge; New York: Cambridge University Press.

6. Joachims T. A support vector method for multivariate performance measures; 2005. pp. 377-384.