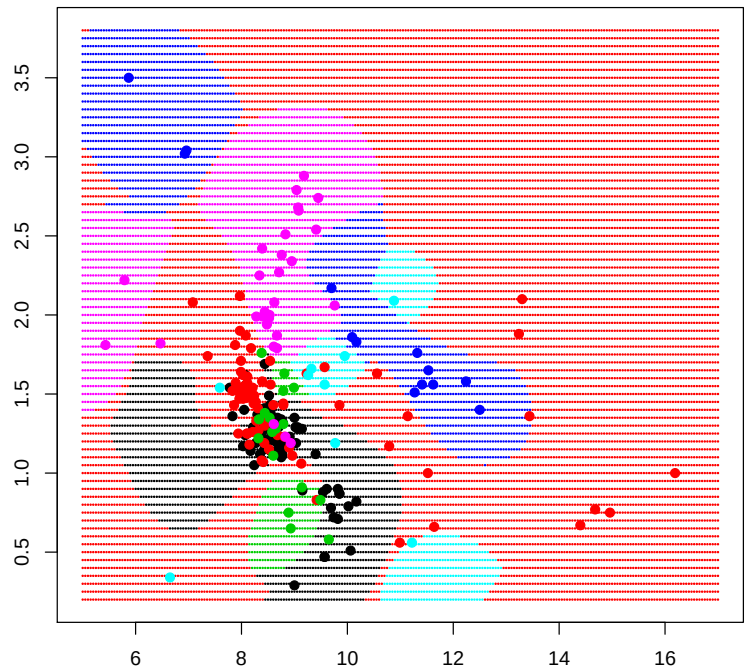
Multiclass SVM

Statistical Machine Learning

Ryan Haunfelder



Multiclass SVMs

- SVMs were originally designed for binary classification.
- There are numerous methods for extending them for use with multiclass problems where $\mathcal{G} = \{1, 2, \ldots G\}$.
  - One-against-all
  - One-against-one
  - Directed acyclic graph SVM (DAGSVM)
  - All-together methods

One-against-all

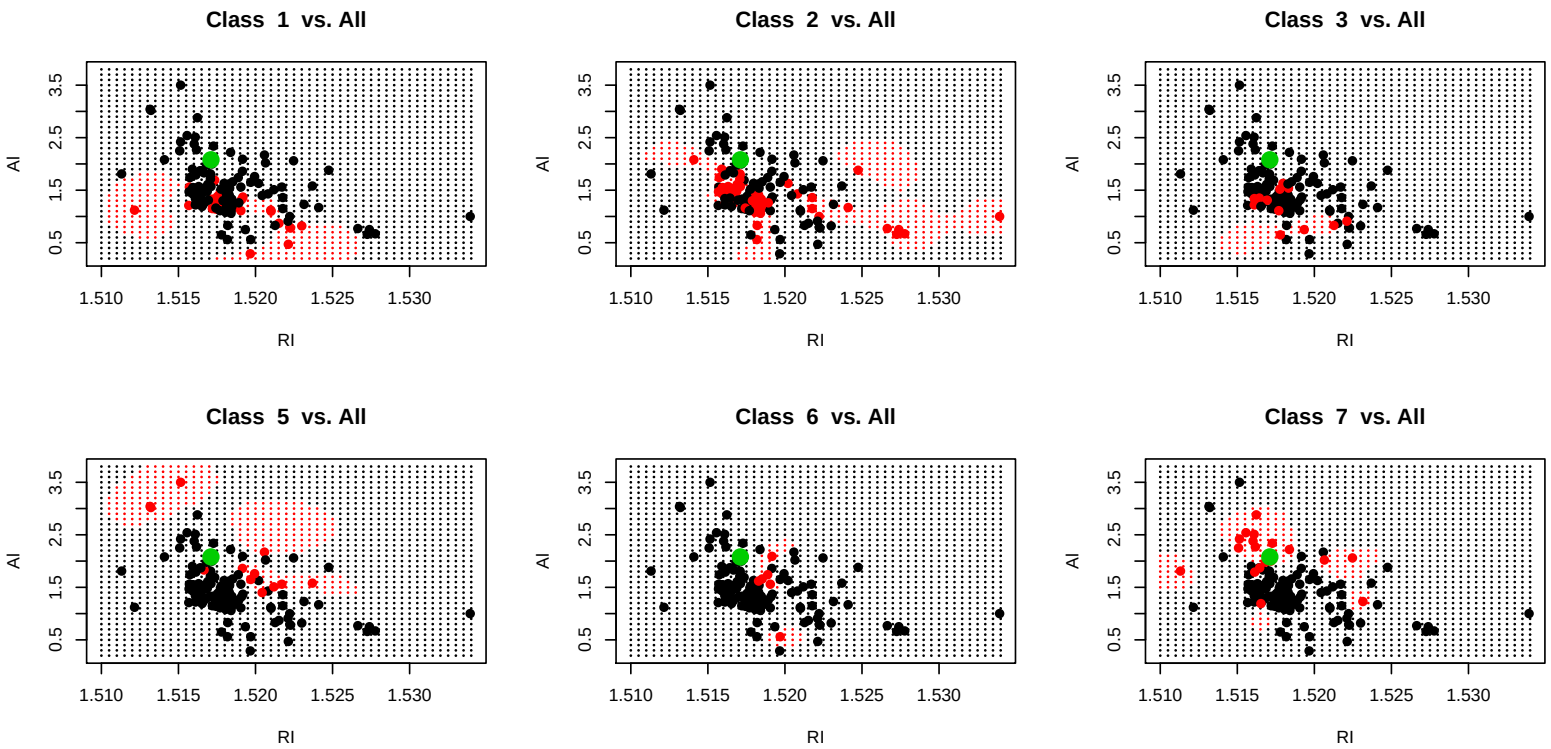
- The one-against method will fit G separate SVMs classifying each group vs the rest of the data. Thus, G programs are fit to the data ($j \in \mathcal{G}$),

$$\min_{\beta_0,\beta} \frac{\|\beta\|}{2} + C \sum \xi_i^j$$
$$f_j(X_i) \geq 1 - \xi_i^j \quad \text{if } Y_i = j$$
$$f_j(X_i) \leq -(1 - \xi_i^j) \quad \text{if } Y_i \neq j$$

- A point is classified to the group which has the largest decision function. I.e. class$=\underset{j}{argmax} \hat{f}_j(X_i)$.

Glass Example: One-against-all

- The glass data contains chemical analysis of 7 different types of glass. Potential features include refractive index (RI) and elemental composition (here we will investigate Aluminum-Al). Correct classification is helpful in crime scene analysis.



One-against-one

- Another approach is to define machines for all $\frac{k(k-1)}{2}$ pairs of classes.

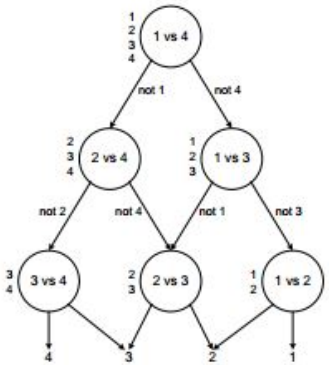
$$\min_{\beta_0,\beta} \frac{\|\beta\|}{2} + C \sum \xi_i^{jk}$$
$$f_{jk}(X_i) \geq 1 - \xi_i^{jk} \quad \text{if } Y_i = j$$
$$f_{jk}(X_i) \leq -(1 - \xi_i^{jk}) \quad \text{if } Y_i = k$$

- A new point is classified by obtaining class vote from each classifier and choosing the class with the most votes. This strategy leads to issues with potential ties.
- This is the default method of the svm function in the e1071 package.

Glass Example: One-against-one

Directed acyclic graph SVM (DAGSVM)

- A directed acyclic decision graph is created with nodes representing one-vs-one SVMs. The path taken through the DAG is called the evaluation path.



- This is advantageous because it allows for analysis of generalization. Let,

$$\epsilon_j(\mathcal{G}) = P\{x : \text{ x in class j and x is misclassified by G}$$

or x is misclassified as class j by G $\}$

Directed acyclic graph SVM (DAGSVM)

- If we are able to correctly distinguish class j from the other classes in a random m-sample with a DDAG G over N classes containing K decision nodes with margins γ_i at node i, then with probability $1-\delta$,

$$\epsilon_j(G) \leq \frac{130R^2}{m}\left(D' \log 4em \log 4m + \log \frac{2(2m)^{N-1}}{\delta}\right)$$

where $D' = \sum \frac{1}{\gamma_i^2}$ and R is the radius of a ball containing the support of the distribution.

All-together Methods

- Cranmer and Singer and Weston and Watkins propose an approach with a single optimization problem.

	One-against-one		DAG		One-against-all		[25], [27]		C&S	
Problem	(C, γ)	rate	(C, γ)	rate	(C, γ)	rate	(C, γ)	rate	(C, γ)	rate
iris	$(2^{12}, 2^{-9})$	**97.333**	$(2^{12}, 2^{-8})$	96.667	$(2^{9}, 2^{-3})$	96.667	$(2^{12}, 2^{-8})$	**97.333**	$(2^{10}, 2^{-7})$	**97.333**
wine	$(2^{7}, 2^{-10})$	**99.438**	$(2^{6}, 2^{-9})$	98.876	$(2^{7}, 2^{-6})$	98.876	$(2^{0}, 2^{-2})$	98.876	$(2^{1}, 2^{-3})$	98.876
glass	$(2^{11}, 2^{-2})$	71.495	$(2^{12}, 2^{-3})$	**73.832**	$(2^{11}, 2^{-2})$	71.963	$(2^{9}, 2^{-4})$	71.028	$(2^{4}, 2^{1})$	71.963
vowel	$(2^{4}, 2^{0})$	**99.053**	$(2^{2}, 2^{2})$	98.674	$(2^{4}, 2^{1})$	98.485	$(2^{3}, 2^{0})$	98.485	$(2^{1}, 2^{3})$	98.674
vehicle	$(2^{9}, 2^{-3})$	86.643	$(2^{11}, 2^{-5})$	86.052	$(2^{11}, 2^{-4})$	**87.470**	$(2^{10}, 2^{-4})$	86.998	$(2^{9}, 2^{-4})$	86.761
segment	$(2^{6}, 2^{0})$	97.403	$(2^{11}, 2^{-3})$	97.359	$(2^{7}, 2^{0})$	97.532	$(2^{5}, 2^{0})$	**97.576**	$(2^{0}, 2^{3})$	97.316
dna	$(2^{3}, 2^{-6})$	95.447	$(2^{3}, 2^{-6})$	95.447	$(2^{2}, 2^{-6})$	95.784	$(2^{4}, 2^{-6})$	95.616	$(2^{1}, 2^{-6})$	**95.869**
satimage	$(2^{4}, 2^{0})$	91.3	$(2^{4}, 2^{0})$	91.25	$(2^{2}, 2^{1})$	91.7	$(2^{3}, 2^{0})$	91.25	$(2^{2}, 2^{2})$	**92.35**
letter	$(2^{4}, 2^{2})$	**97.98**	$(2^{4}, 2^{2})$	**97.98**	$(2^{2}, 2^{2})$	97.88	$(2^{1}, 2^{2})$	97.76	$(2^{3}, 2^{2})$	97.68
shuttle	$(2^{11}, 2^{3})$	99.924	$(2^{11}, 2^{3})$	99.924	$(2^{9}, 2^{4})$	99.910	$(2^{9}, 2^{4})$	99.910	$(2^{12}, 2^{4})$	**99.938**

References I

Koby Crammer and Yoram Singer.
On the learnability and design of output codes for multiclass problems.
Machine Learning, 47:201–233, 2002.

Chih-Wei Hsu and Chih-Jen Lin.
A comparison of methods for multiclass support vector machines.
IEEE Transactions on Neural Networks, 13:415–425, 2002.

JC Platt, N Cristianini, and J Shawe-Taylor.
Large Margin DAGs for Multiclass Classification.
nips, pages 547–553, 1999.