Introduction to Machine Learning

Brown University CSCI 1950-F, Spring 2012
Prof. Erik Sudderth

Lecture 17:
Kernels & Support Vector Machines
Preview: Unsupervised Learning & Clustering

Many figures courtesy Kevin Murphy's textbook,
Machine Learning: A Probabilistic Perspective

Kernels or Features?

$N \longrightarrow$ number of training examples

$M \longrightarrow$ number of features

$L \longrightarrow$ cost of kernel function evaluation, at worst $\mathcal{O}(M)$

$\Phi \longrightarrow$ *NxM* matrix evaluating each feature for all training data

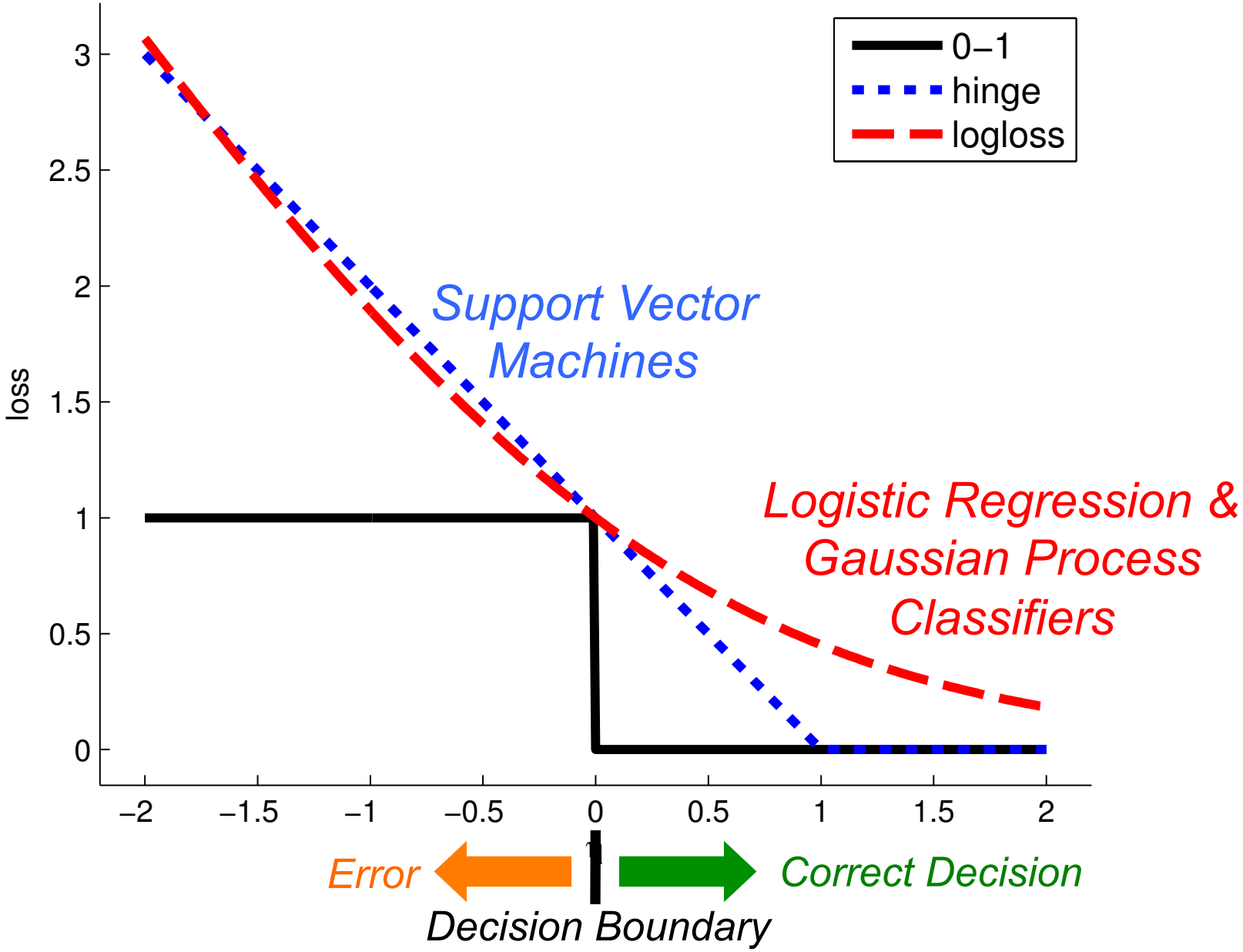
- Feature-based linear regression: $\mathcal{O}(NM^2 + M^3)$
- Kernel-based GP regression: $\mathcal{O}(LN^2 + N^3)$
- Roughly, the difference corresponds to using either

$$(\Phi^T \Phi)^{-1} \qquad\qquad (\Phi \Phi^T)^{-1}$$

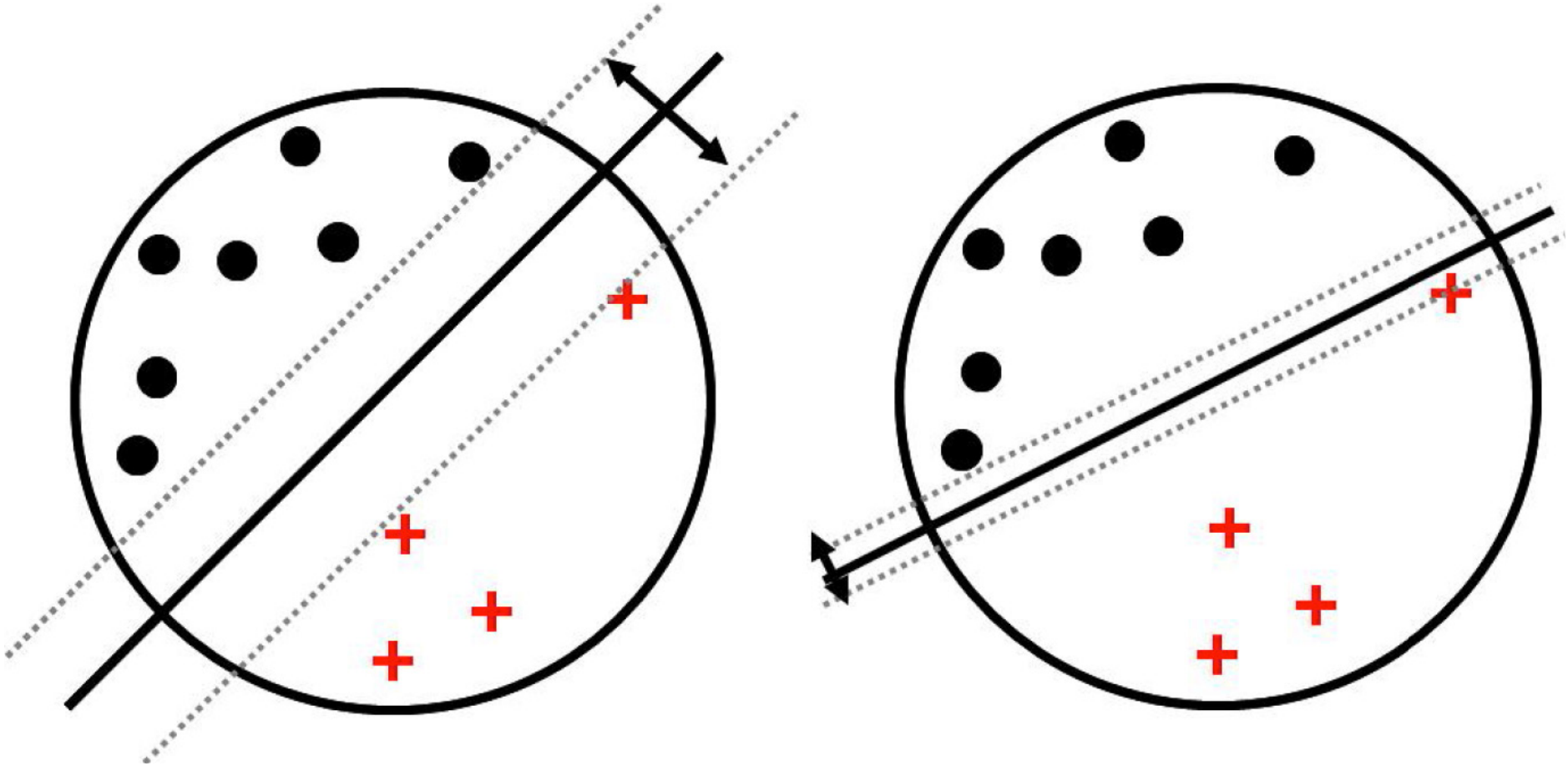
- Relative costs of logistic regression and GP classification are similar, per iteration of optimization-based learning
- What if *N* and *M* are both large???

Approximate!!! Endless options, none perfect…

Losses for Binary Classification

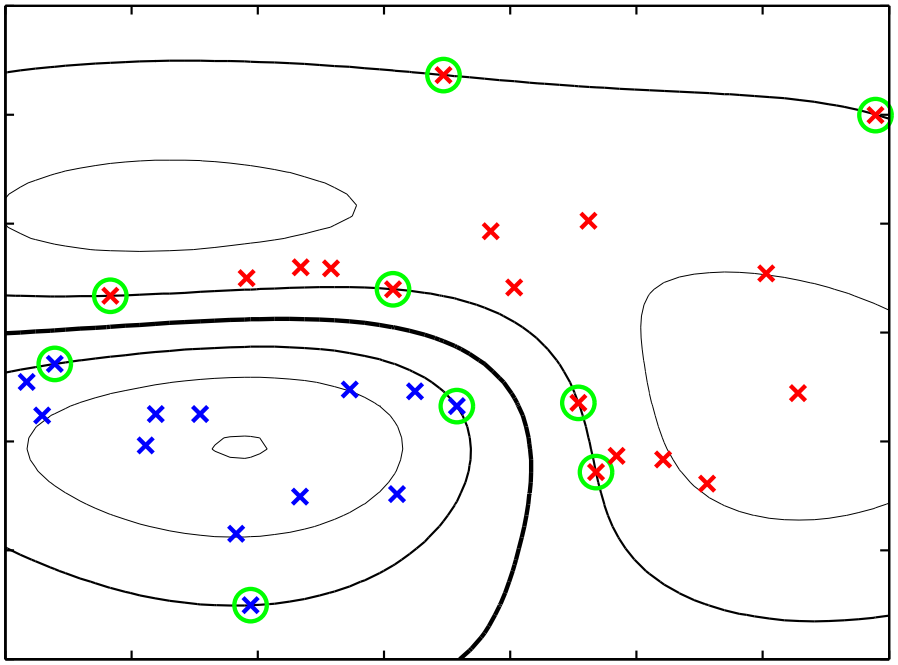


Maximum Margin Hyperplanes

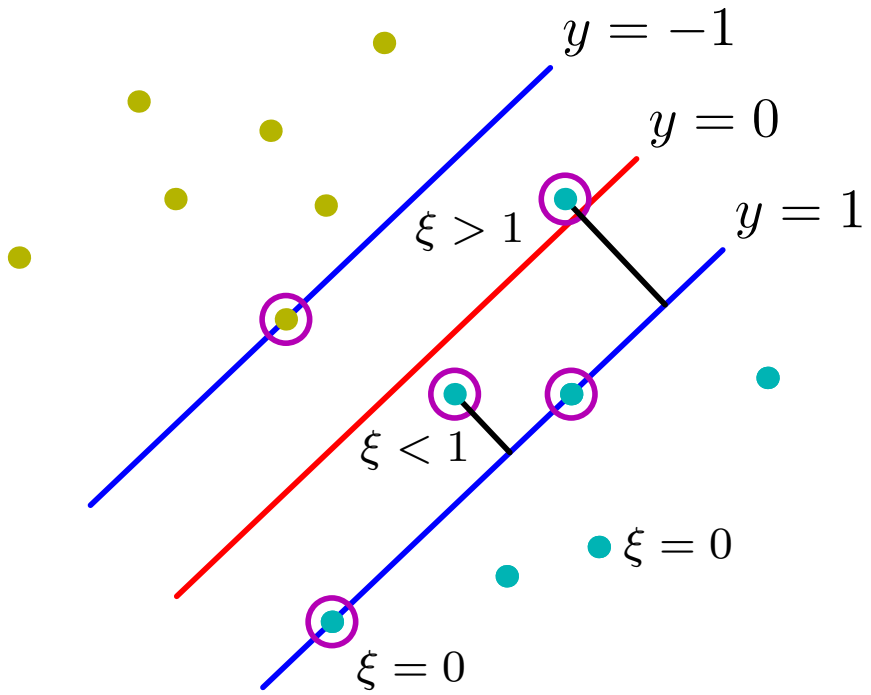


If multiple linear classifiers perfectly separate training data, which should I choose?

Support Vectors & Slack Variables



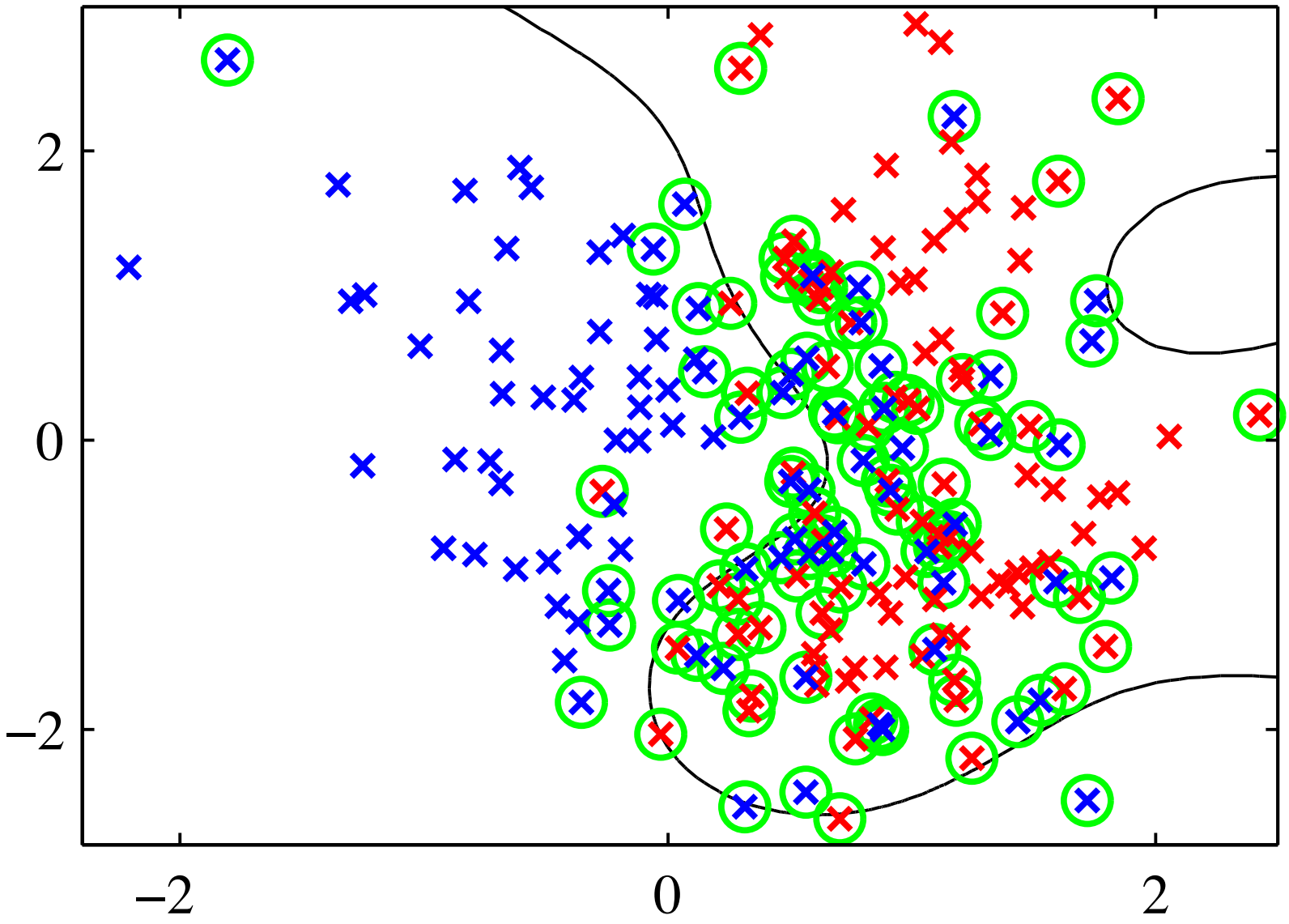
Support vectors (green) for data separable by radial basis function kernels, and non-linear margin boundaries



Linear decision boundary in feature space, where data violating margin have nonzero "slack variables"

C. Bishop, Pattern Recognition & Machine Learning

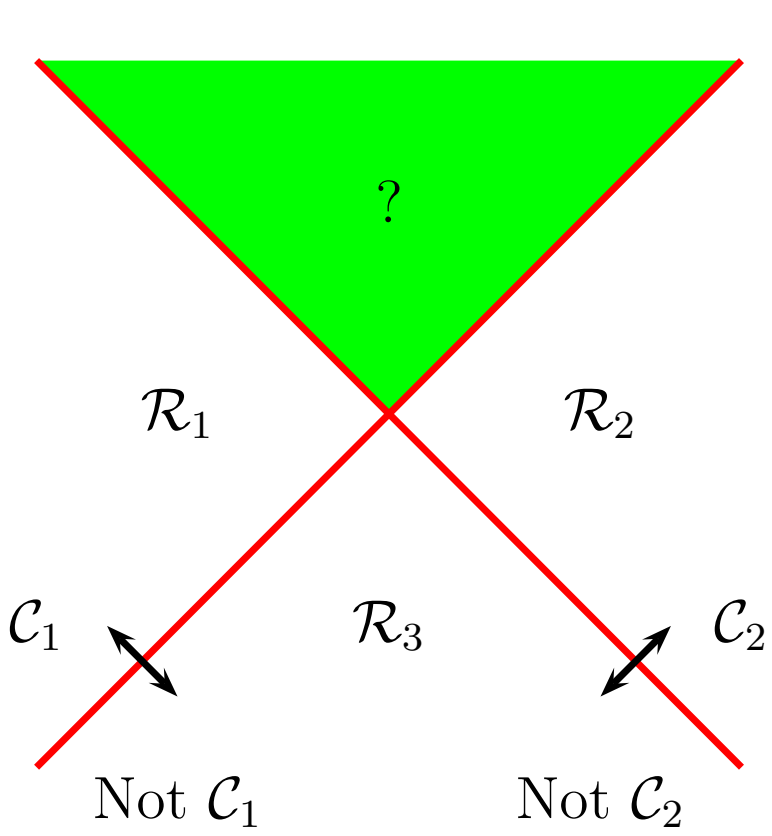
How Many Support Vectors?



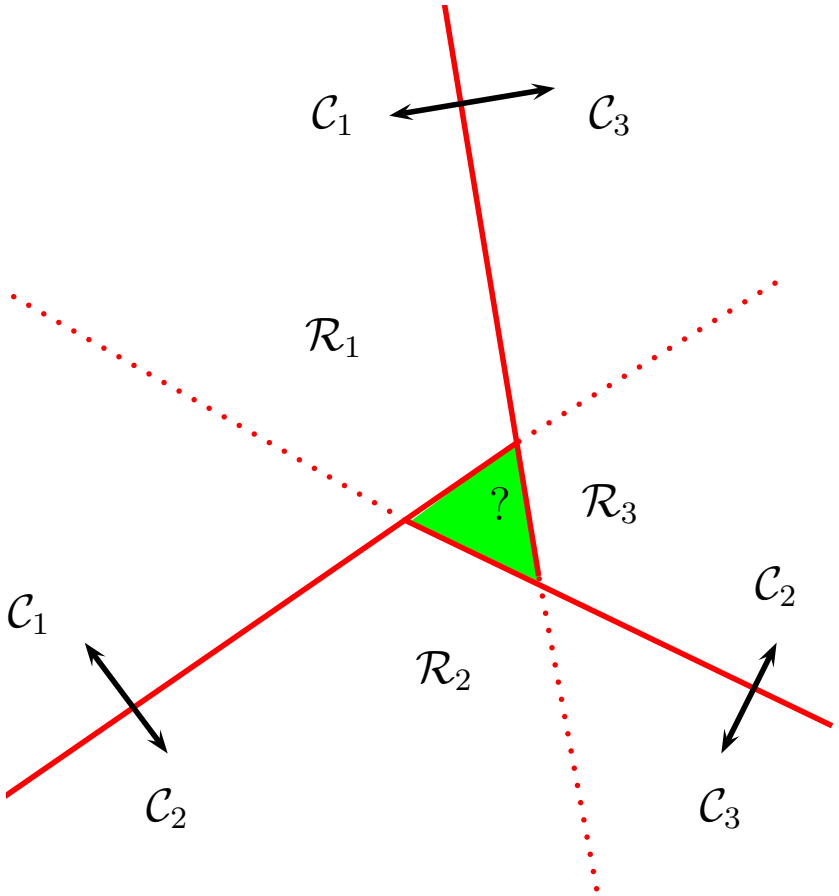
C. Bishop, Pattern Recognition & Machine Learning

Multiclass Support Vector Machines

*Complicated by the fact that binary SVM classifiers are **not** calibrated probabilistic models*



One versus Rest
(One versus All)



One versus One
(separate each pair of classes)

On to Unsupervised Learning

	Supervised Learning	*Unsupervised Learning*
Discrete	classification or categorization	clustering
Continuous	regression	dimensionality reduction

- **Goal:** Infer label/response y given only features x
- **Classical:** Find latent variables y good for *compression* of x
- **Probabilistic learning:** Estimate parameters of joint distribution p(x,y) which *maximize marginal probability* p(x)

Clustering can be Ambiguous