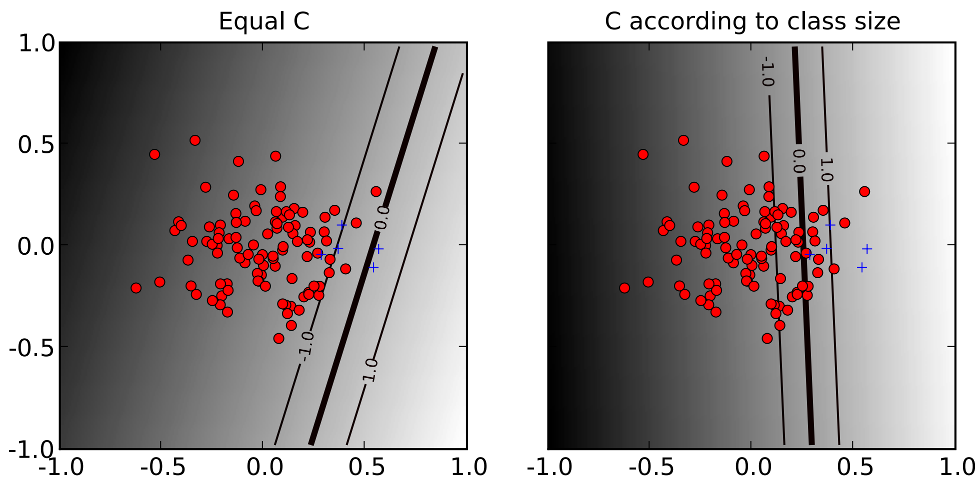
SVMs: error, regularization and unbalanced data

Chapter e-8



SVM: error +regularization?

Recall that most classifiers are based on a cost function that has the form

error term + regularization term

Let's express the SVM optimization problem in this form.

The hinge loss

The primal form of the SVM:

$$\underset{\mathbf{w},b}{\text{minimize}} \frac{1}{2}||\mathbf{w}||^2 + C\sum_{i=1}^{n} \xi_i$$

$$\text{subject to: } \; y_i(\mathbf{w}^\intercal \mathbf{x}_i + b) \geq 1 - \xi_i, \; \xi_i \geq 0, \quad i = 1, \ldots, n\,.$$

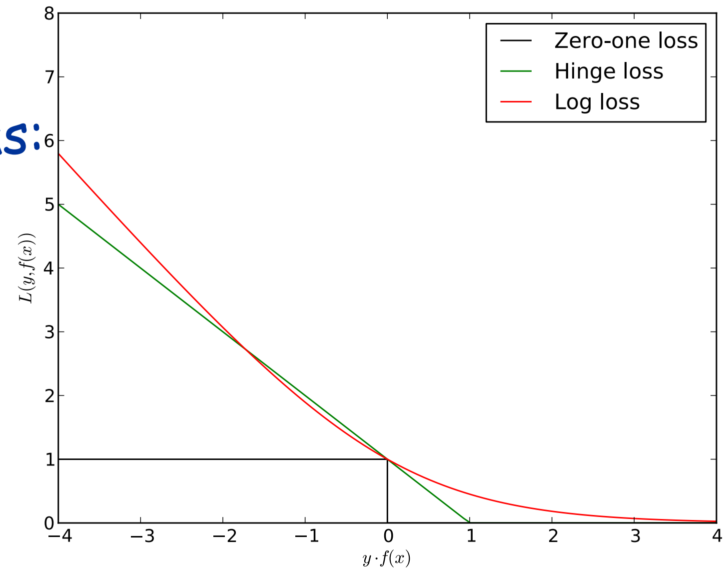
Let's define:

Hinge loss

$$E_{\text{svm}}(\mathbf{w}, b) = \frac{1}{n}\sum_{i=1}^{n} \max(1 - y_i(\mathbf{w}^\intercal \mathbf{x}_i + b), 0)$$

The SVM problem now can be written as:

$$\underset{\mathbf{w},b}{\text{minimize}} \; E_{\text{svm}}(\mathbf{w}, b) + \lambda||\mathbf{w}||^2$$



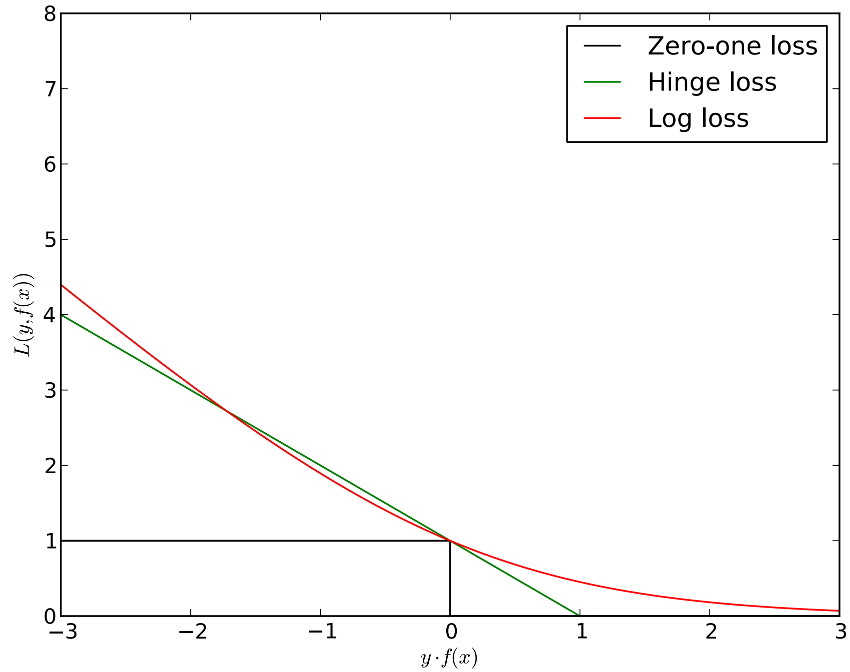
See page e-8-45

SVM: error + regularization

$$\underset{\mathbf{w},b}{\text{minimize}}\ E_{\text{svm}}(\mathbf{w},b)+\lambda||\mathbf{w}||^2$$

$$E_{\text{svm}}(\mathbf{w},b)=\frac{1}{n}\sum_{i=1}^{n}\max(1-y_i(\mathbf{w}^\intercal\mathbf{x}_i+b),0)$$

E_{svm} is an upper bound on E_{in} and is a margin-maximizing error function



misclassified

correctly classified with small confidence

correctly classified

L_1 Regularization

Regular SVM uses $||\mathbf{w}||^2$ as the regularizer

Another option: $$||\mathbf{w}||_1 = \sum_i |w_i|$$

This is the L_1 regularizer (aka Lasso), which is known to lead to very sparse solutions.

L_1 Regularization

The L_1 regularizer tends to generate much sparser solutions than a quadratic regularizer.

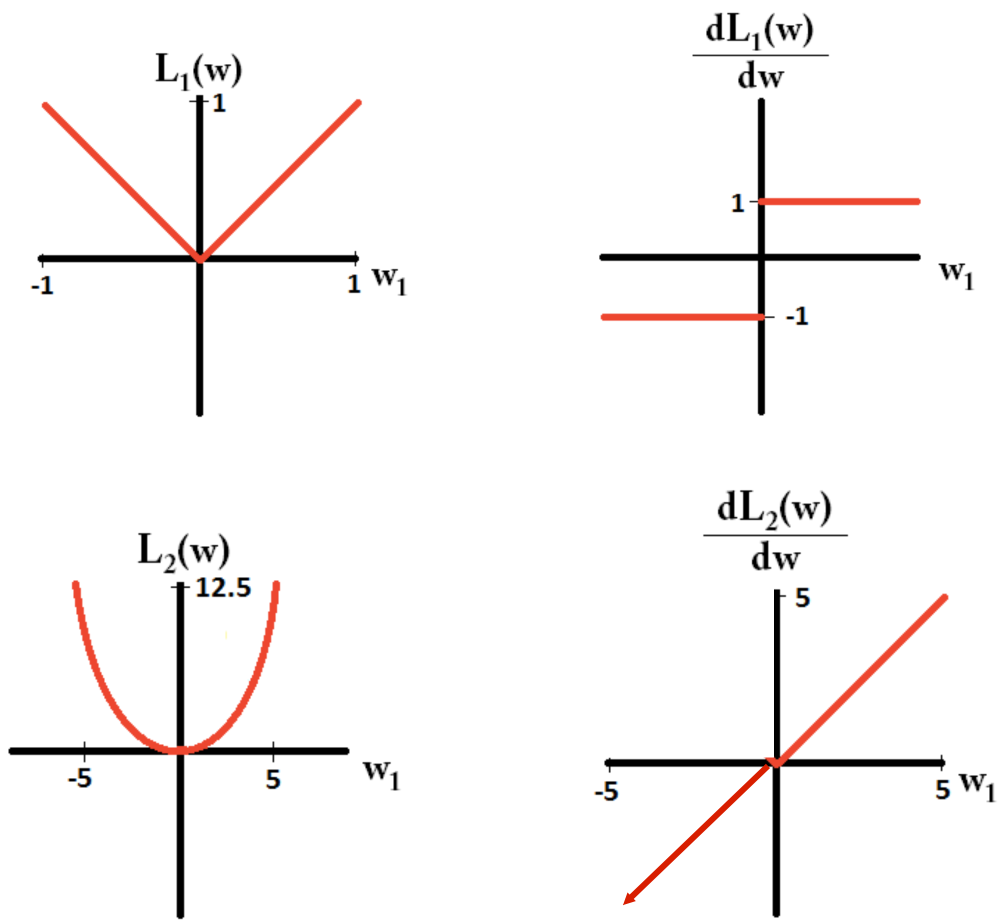
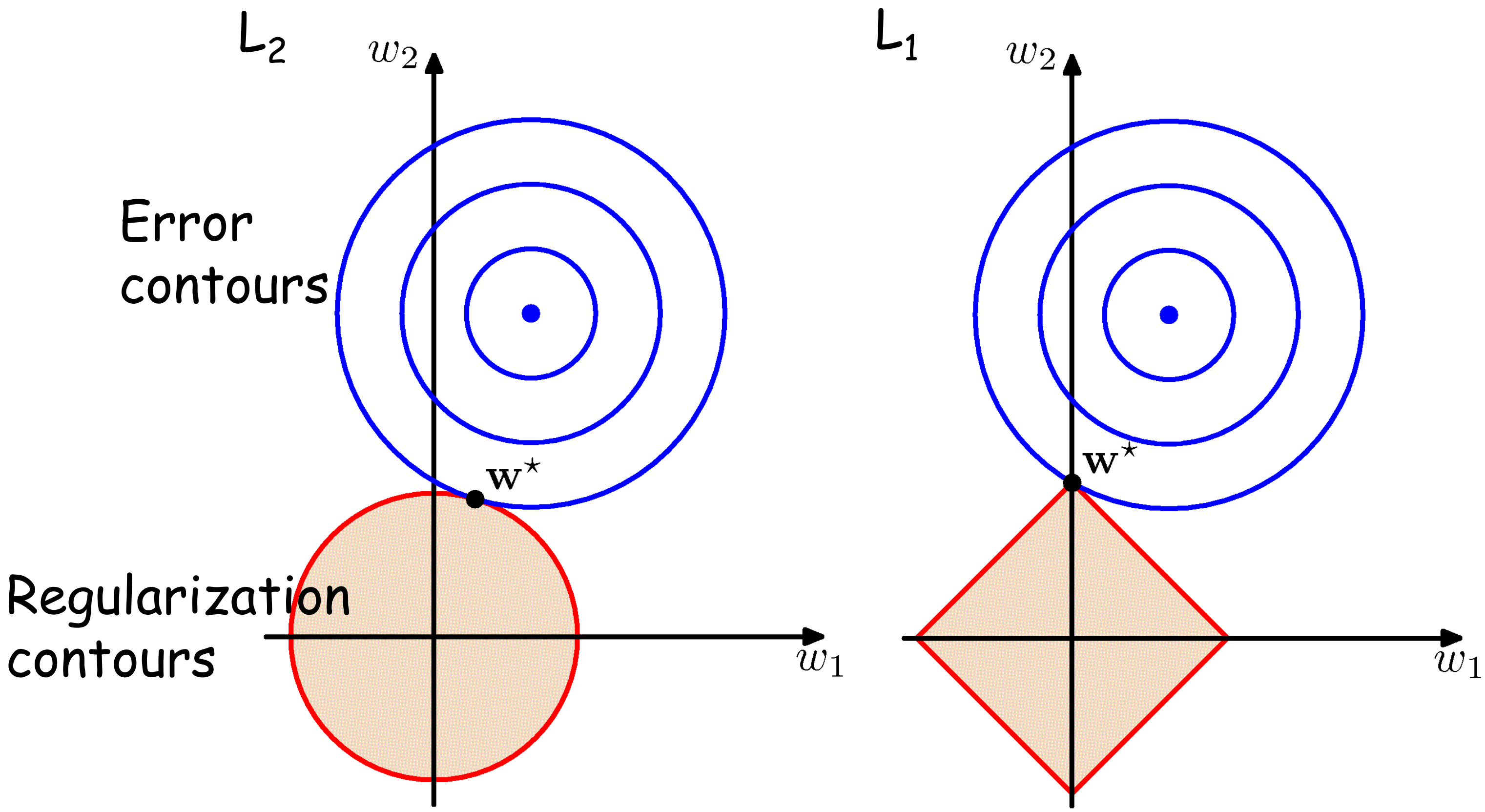


Figure adapted from http://stats.stackexchange.com/questions/45643/why-l1-norm-for-sparse-models

L_1 Regularization

The L_1 regularizer tends to generate much sparser solutions than a quadratic regularizer.

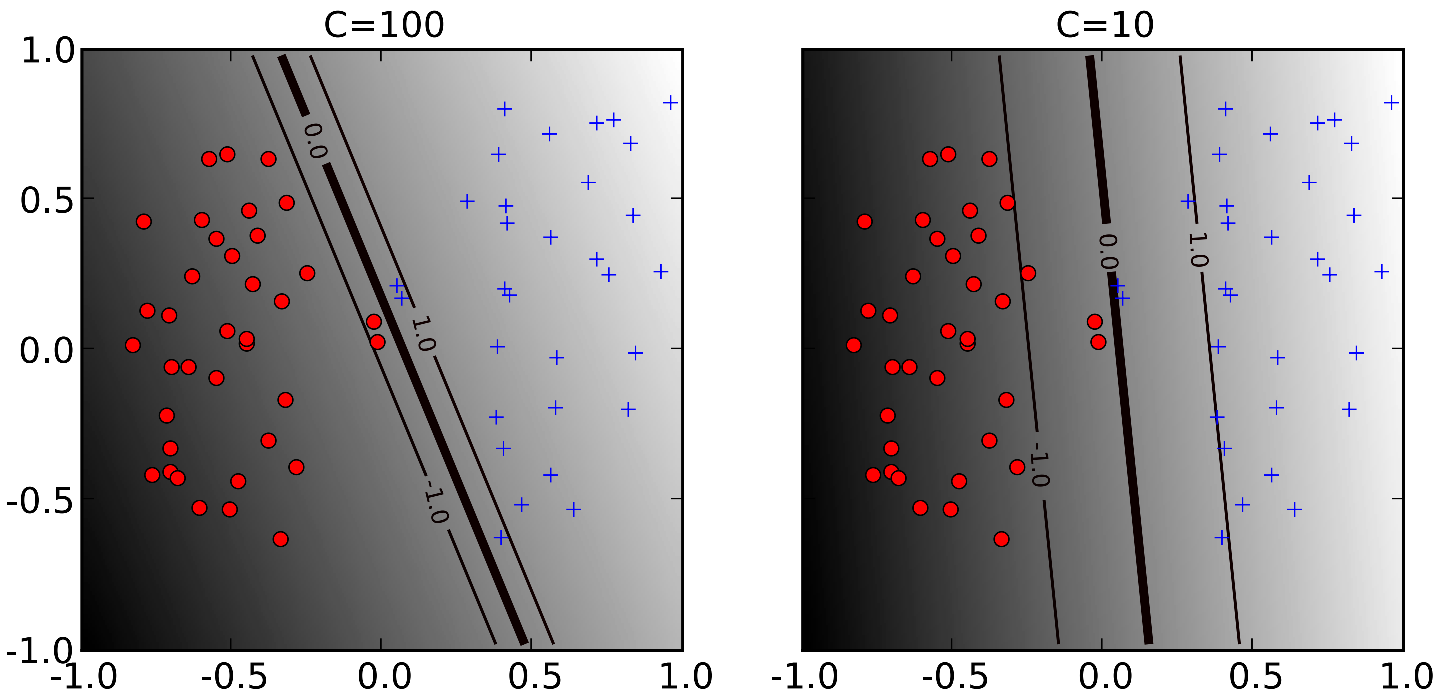


The role of the soft margin parameter

SVM for the non-separable case:

$$\underset{\mathbf{w},b}{\text{minimize}} \frac{1}{2}||\mathbf{w}||^2 + C\sum_{i=1}^{n}\xi_i$$

$$\text{subject to: } \; y_i(\mathbf{w}^\intercal \mathbf{x}_i + b) \geq 1 - \xi_i, \; \xi_i \geq 0, \quad i = 1, \ldots, n\,.$$

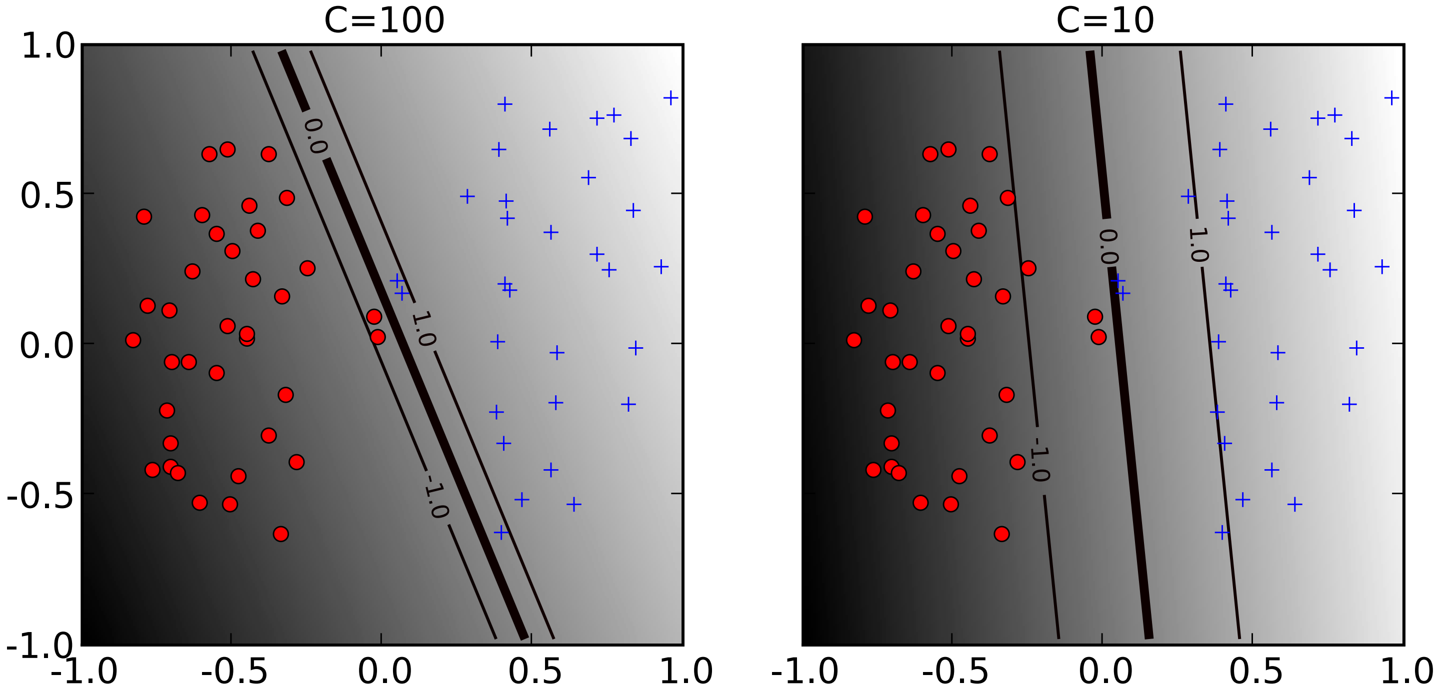


The role of the soft margin parameter

SVM for the non-separable case:

$$\underset{\mathbf{w},b}{\text{minimize}} \frac{1}{2}||\mathbf{w}||^2 + C\sum_{i=1}^{n}\xi_i$$

$$\text{subject to:} \quad y_i(\mathbf{w}^\intercal \mathbf{x}_i + b) \geq 1 - \xi_i,\ \xi_i \geq 0, \quad i = 1, \ldots, n\,.$$



Soft margin is useful even if the data is linearly separable!

A potential problem for unbalanced data

SVM for the non-separable case:

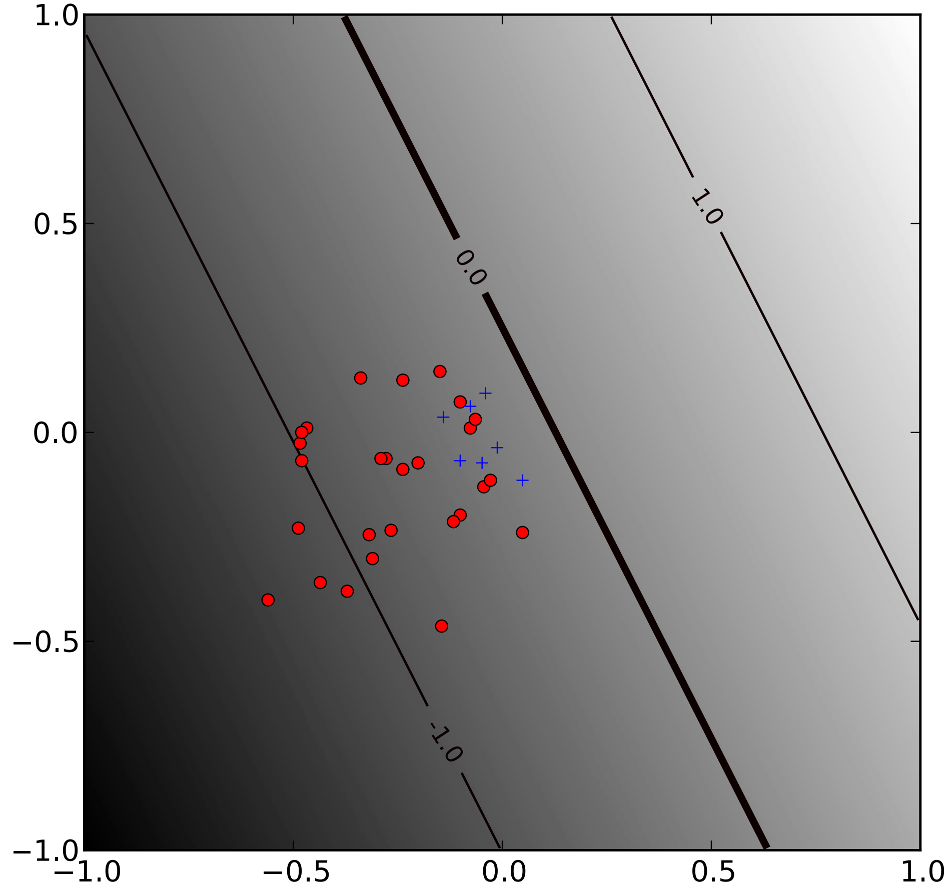
$$\underset{\mathbf{w},b}{\text{minimize}} \frac{1}{2}||\mathbf{w}||^2 + C\sum_{i=1}^{n} \xi_i$$

$$\text{subject to: } \; y_i(\mathbf{w}^\intercal \mathbf{x}_i + b) \geq 1 - \xi_i, \; \xi_i \geq 0, \quad i = 1, \ldots, n\,.$$

$C\sum_{i=1}^{n} \xi_i$ is the penalty for misclassification

If there are only a few positive examples, the penalty for misclassifying them will be small.

What happens when data is unbalanced



The SVM is essentially ignoring the minority class!

Solving the problem

Replace $$C\sum_{i=1}^{n}\xi_i$$

With: $$C_{\oplus}\sum_{i\in pos\_class}^{n}\xi_i + C_{\ominus}\sum_{i\in neg\_class}^{n}\xi_i$$

Choosing the parameters such that:

$$C_{\oplus}Pos \approx C_{\ominus}Neg$$

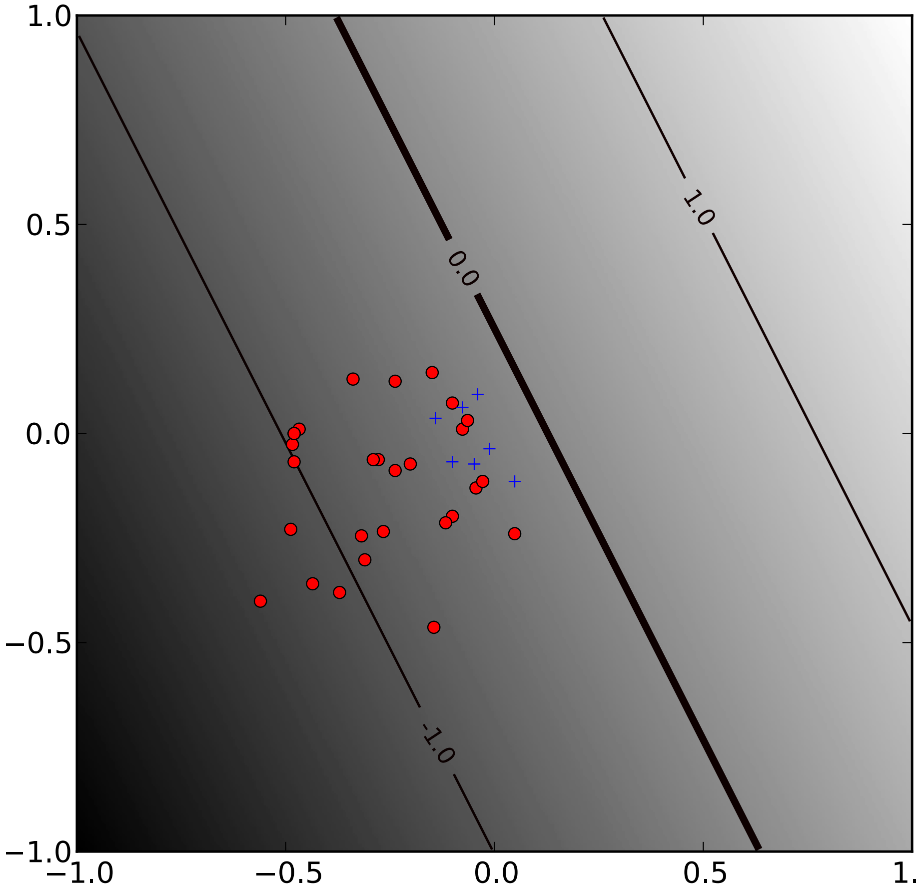
A choice that achieves this:

$$C_{\oplus} = C\frac{n}{Pos}, \quad C_{\ominus} = C\frac{n}{Neg}$$

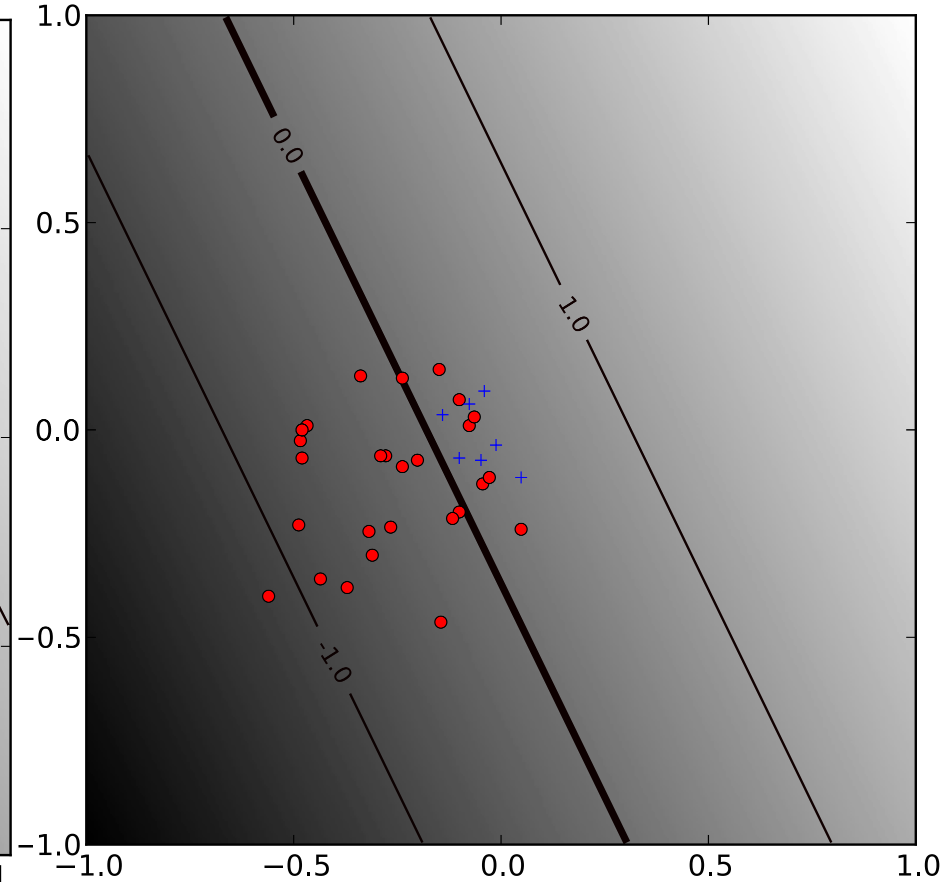
Essentially optimizes balanced error rather than regular error rate.

Effect of unequal soft-margin constants

Comparing the two ways of choosing the soft-margin constant:



Equal C



C inversely proportional to class size

Interim conclusions

SVMs:

- Deliver a large-margin hyperplane, and in so doing can control the effective model complexity.
- Express the hyperplane using only a few support vectors
- Control the sensitivity to outliers and regularize the solution through setting C appropriately.

Coming next:

- Nonlinearity.

These properties make SVMs one of the most useful classification approaches

SVMs for regression

SVR - SVM Regression

Based on the epsilon-insensitive loss: