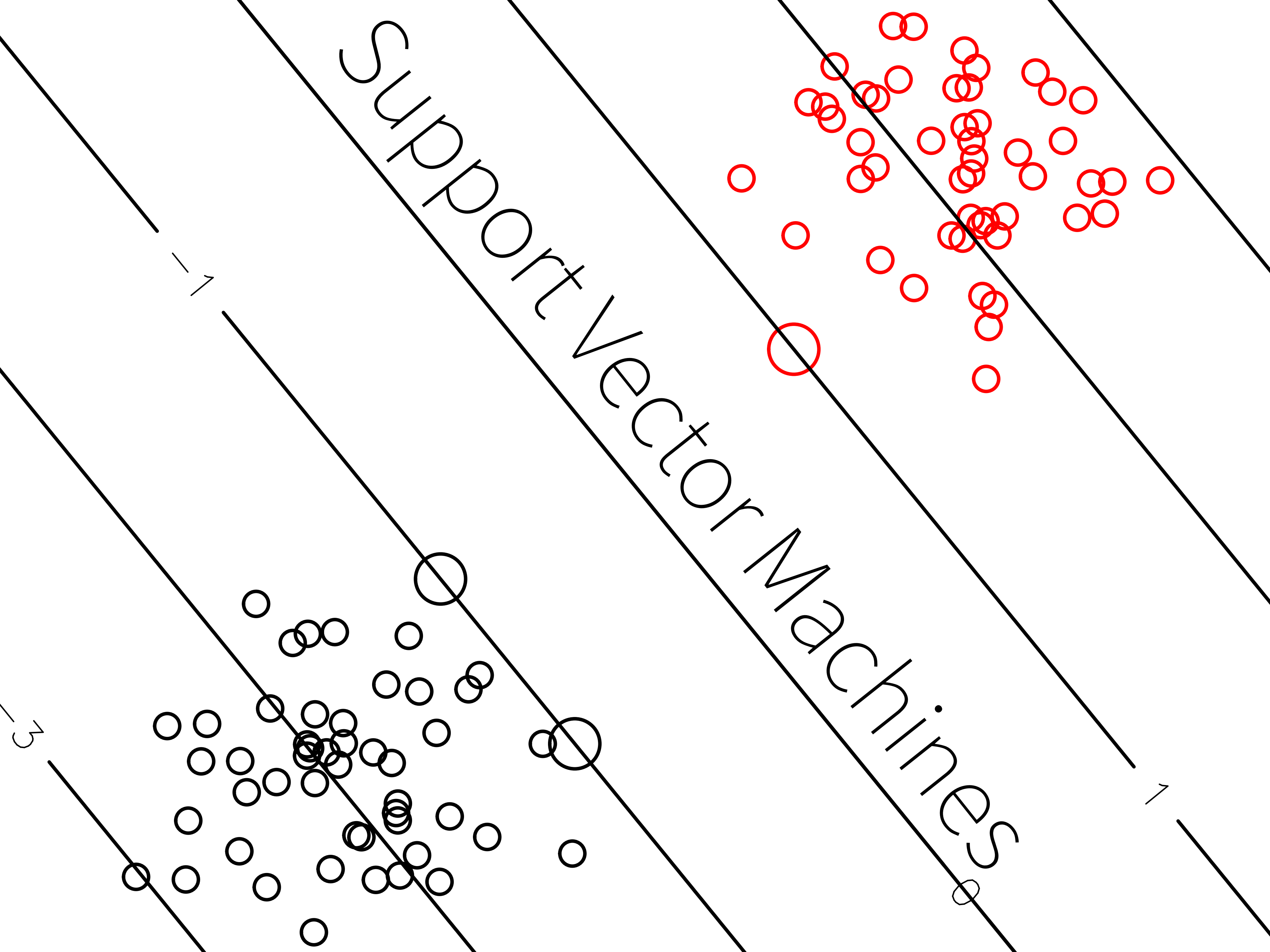

Support Vector Machines
−1
−3
1
0

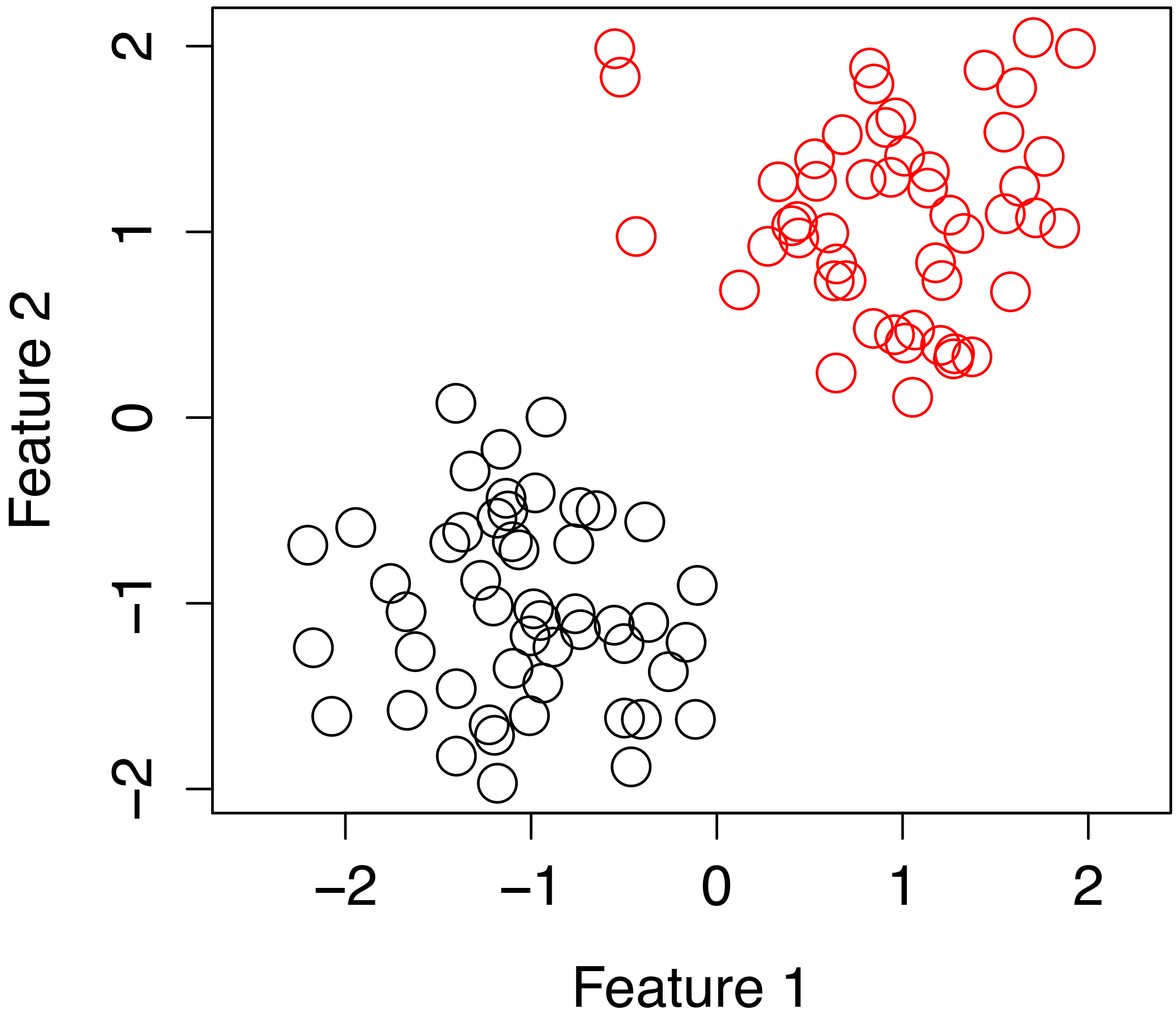
More on classifiers

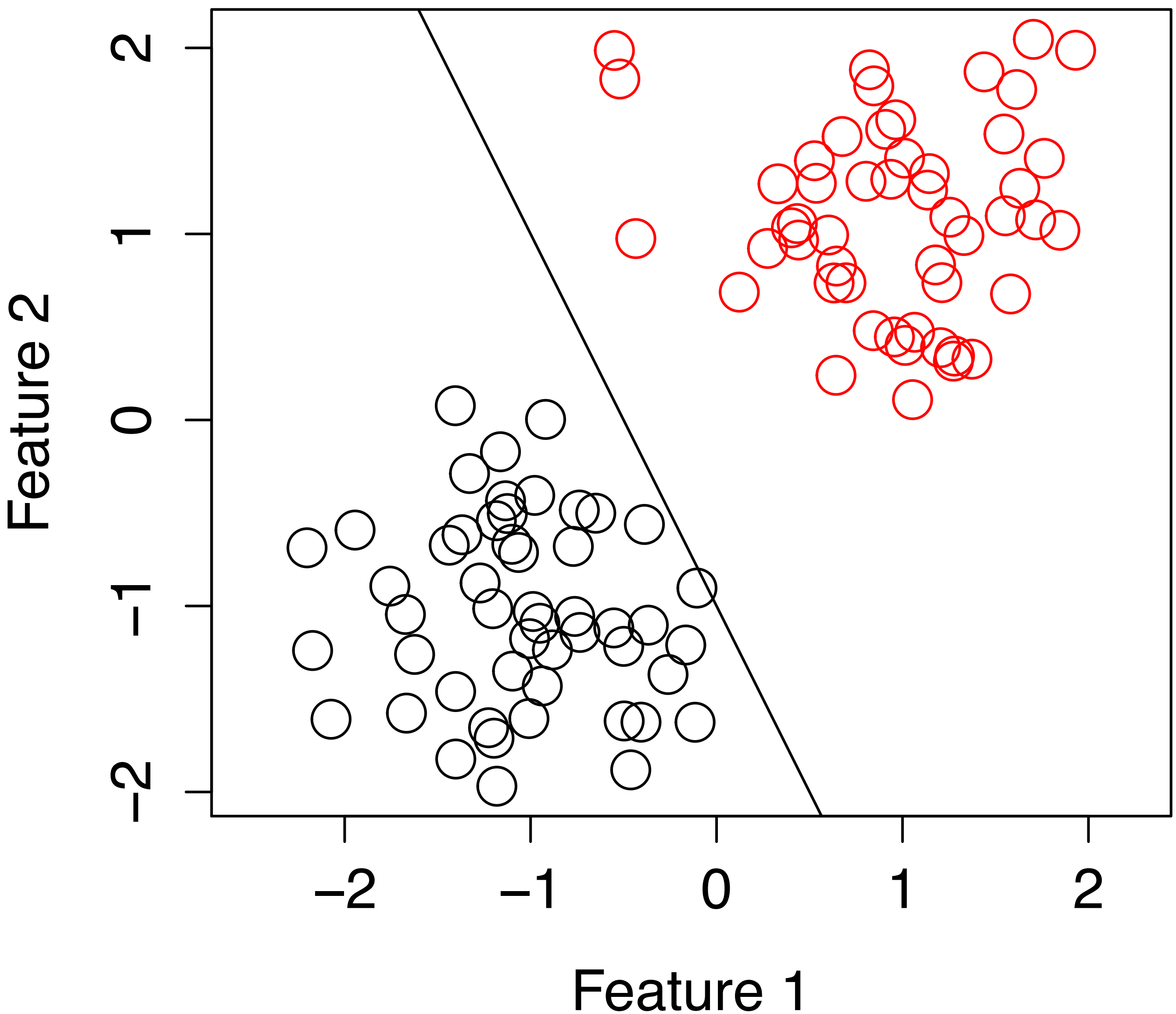
- Density based classifiers
  - Linear normal density based classifiers
  - Parzen classifier
- Distance based classifiers
  - Nearest neighbor rule
  - Nearest Mean
  - Support vector machine

Why is there more than one classifier?

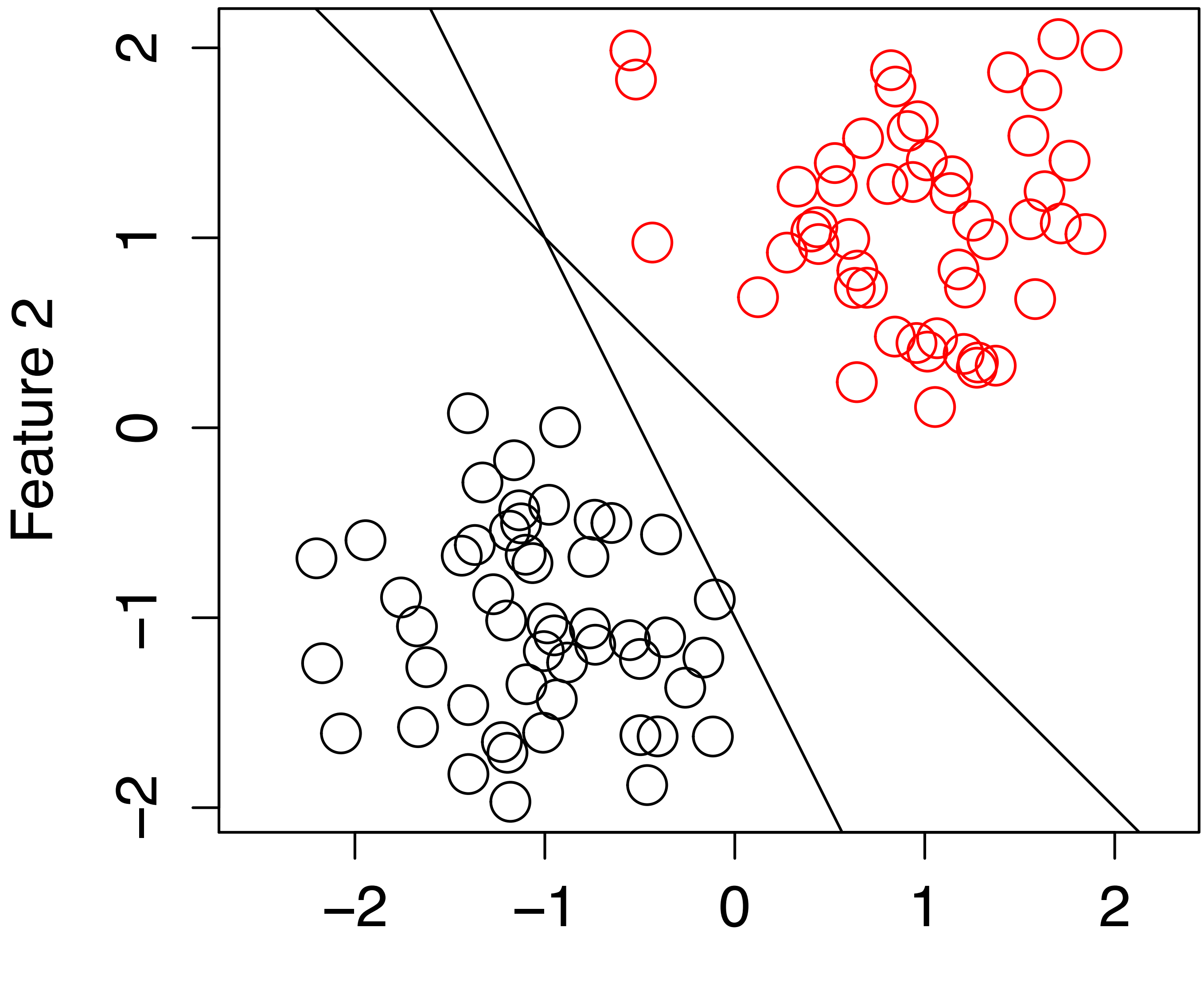
- Different problem/assumptions
- Different strengths/weaknesses
  - Needs a large/small sample
  - Take a short/long time to train
  - Has a clear interpretation
- Different goals (generative/discriminative)
- NFL: there is no single best classifier

Classification problem (Separable)

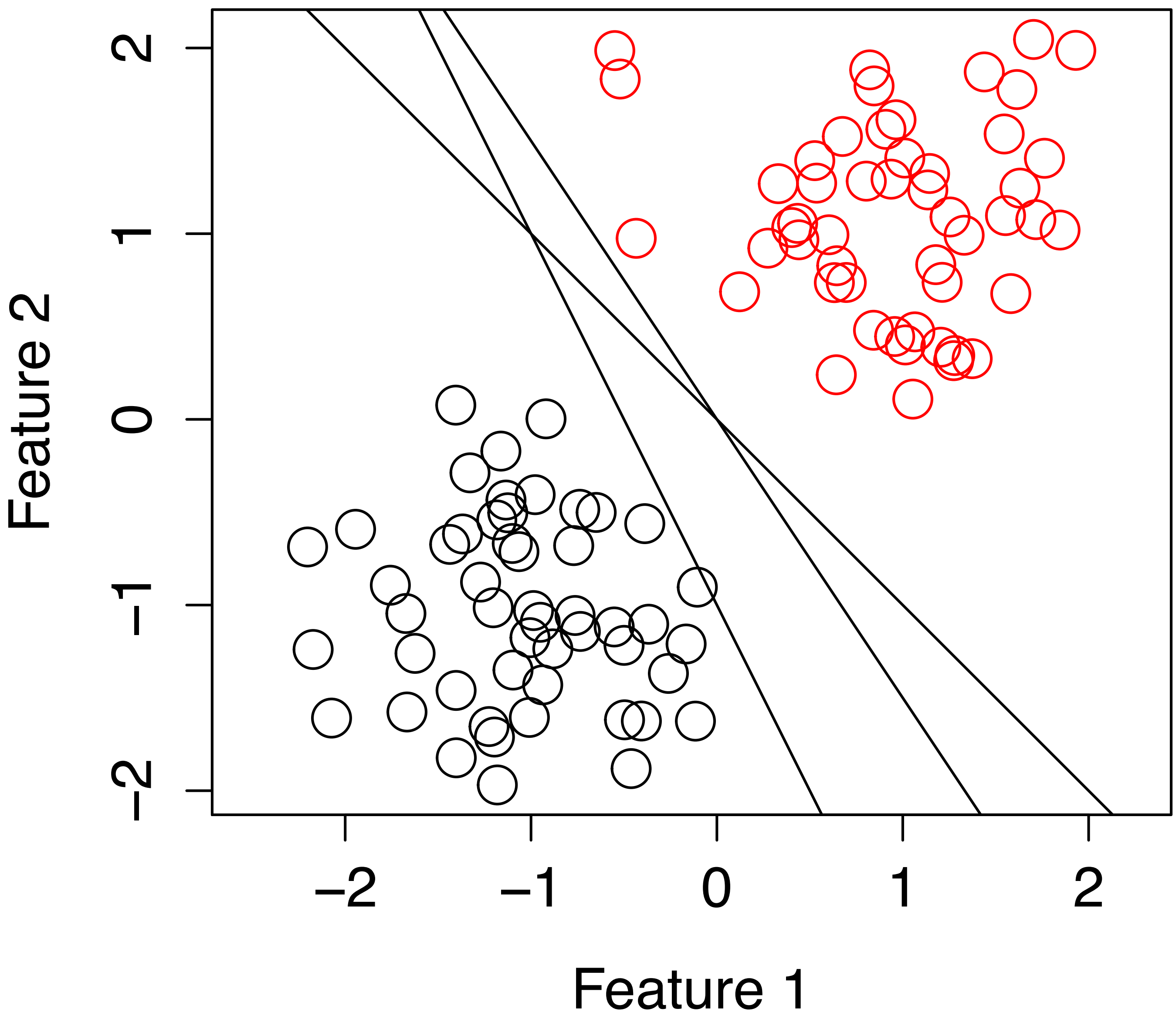




Feature 1
Feature 2
−2
−1
0
1
2

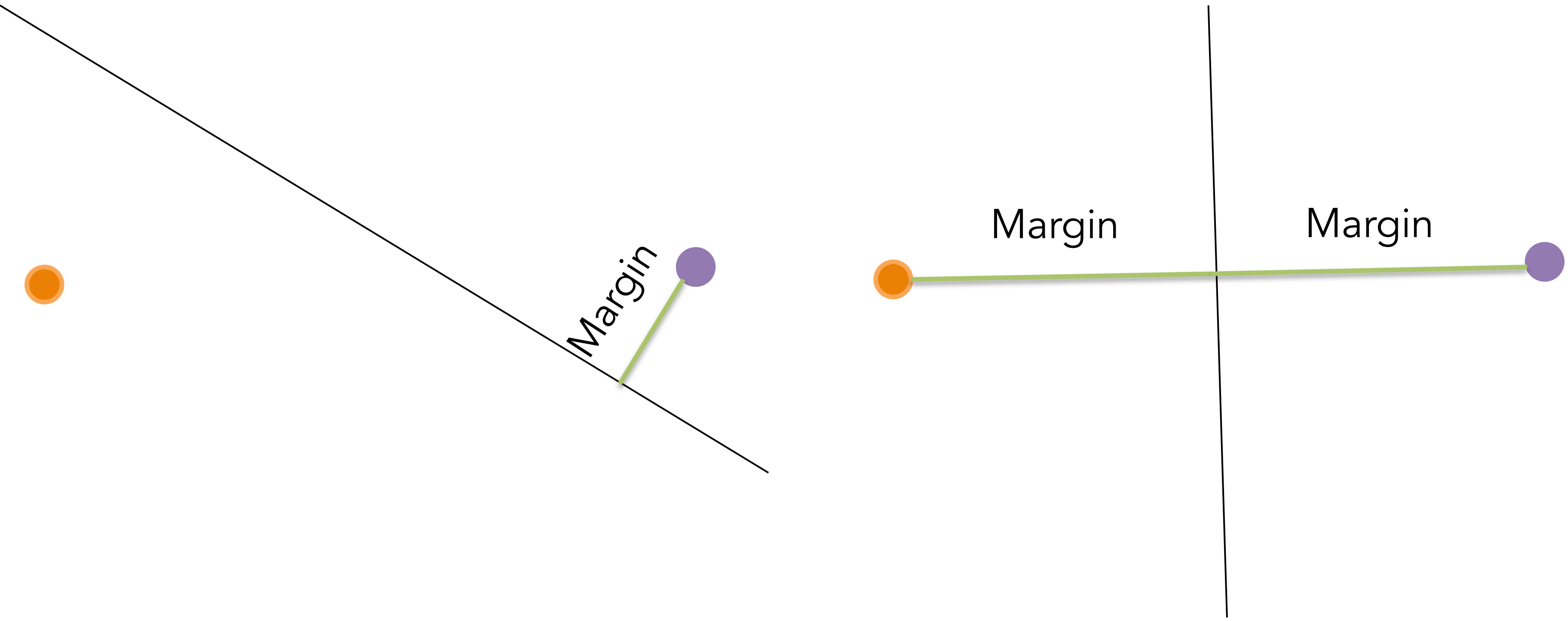


Feature 2
Feature 1
−2
−1
0
1
2



Feature 2
2
1
0
−1
−2
−2
−1
0
1
2
Feature 1

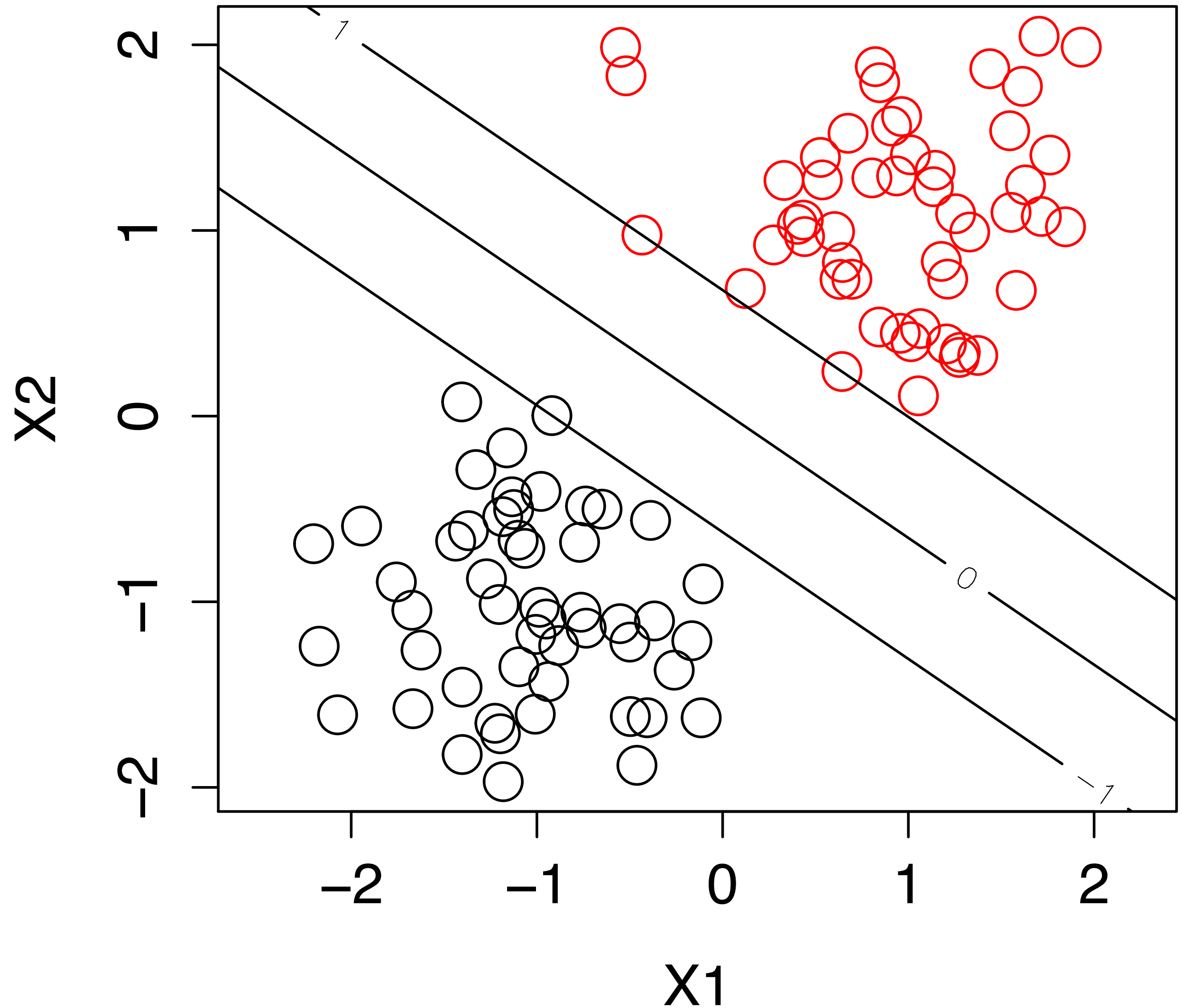
What is the margin



- Let’s assume a classifier that separates the training data data perfectly
- The margin is the minimum distance of any object to the decision boundary

Support Vector Classifier (Visually)

Idea: for separable data, maximize the margin



Support Vector Classifier (Formally)

How do we actually do this?

Max Margin

Inverse Margin

$$\min_{\mathbf{w},b} \frac{1}{2}\|\mathbf{w}\|^2$$

$$\text{subject to } y_i(\mathbf{w} \cdot \mathbf{x}_i + b) \geq 1, i \in [1, m].$$

Label of the classifier (+ or -)

Decision of the classifier (+ or -)

Decision needs to be correct, with a margin

(Vapnik & Chervonenkis, 1965)

(Mohri et. Al. – Foundations of Machine Learing, Ch. 4.2)

Why is it called a support vector?

- Suppose we remove a non-support vector from the datasets. Does the decision boundary change? (Demo!)
- Suppose we add a new data point inside the margin. Does the decision boundary change?
- One can show the decision function depends only on the support vectors.

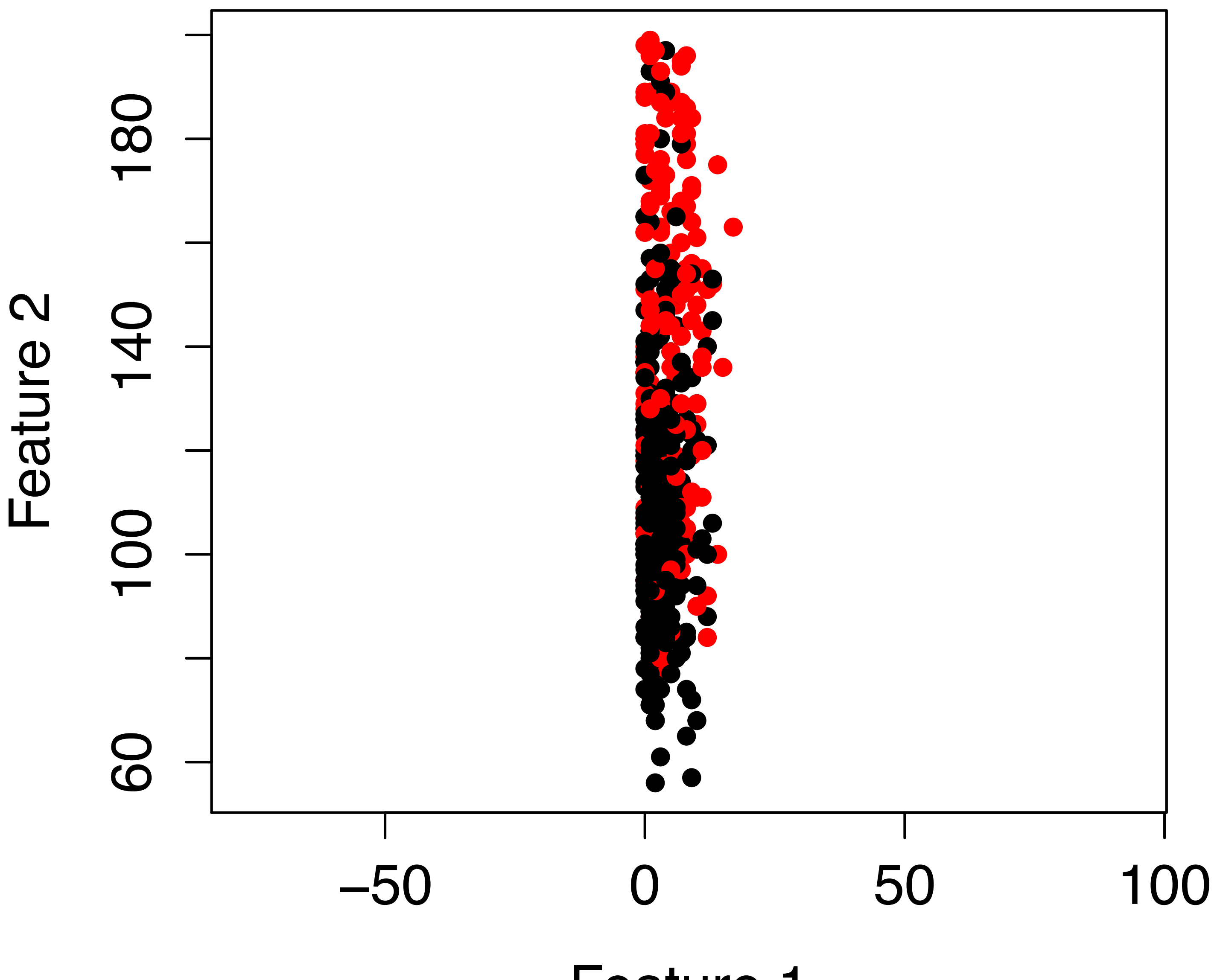
$$f(x) = \sum_{i=1}^{n} \alpha_i (x_i^\top x)$$

- (Note the alpha's come from the optimization)

Why maximize margin?

- Intuitively, if the the margin is wide, it is unlikely that a new object will ‘jump over’ the margin.
- More formally, one can prove that for the worst possible dataset, the larger the margin, the smaller the generalization error (with high probability)
- In practice, we generally do not have ‘the worst possible dataset’, but maximizing the margin works well in many cases

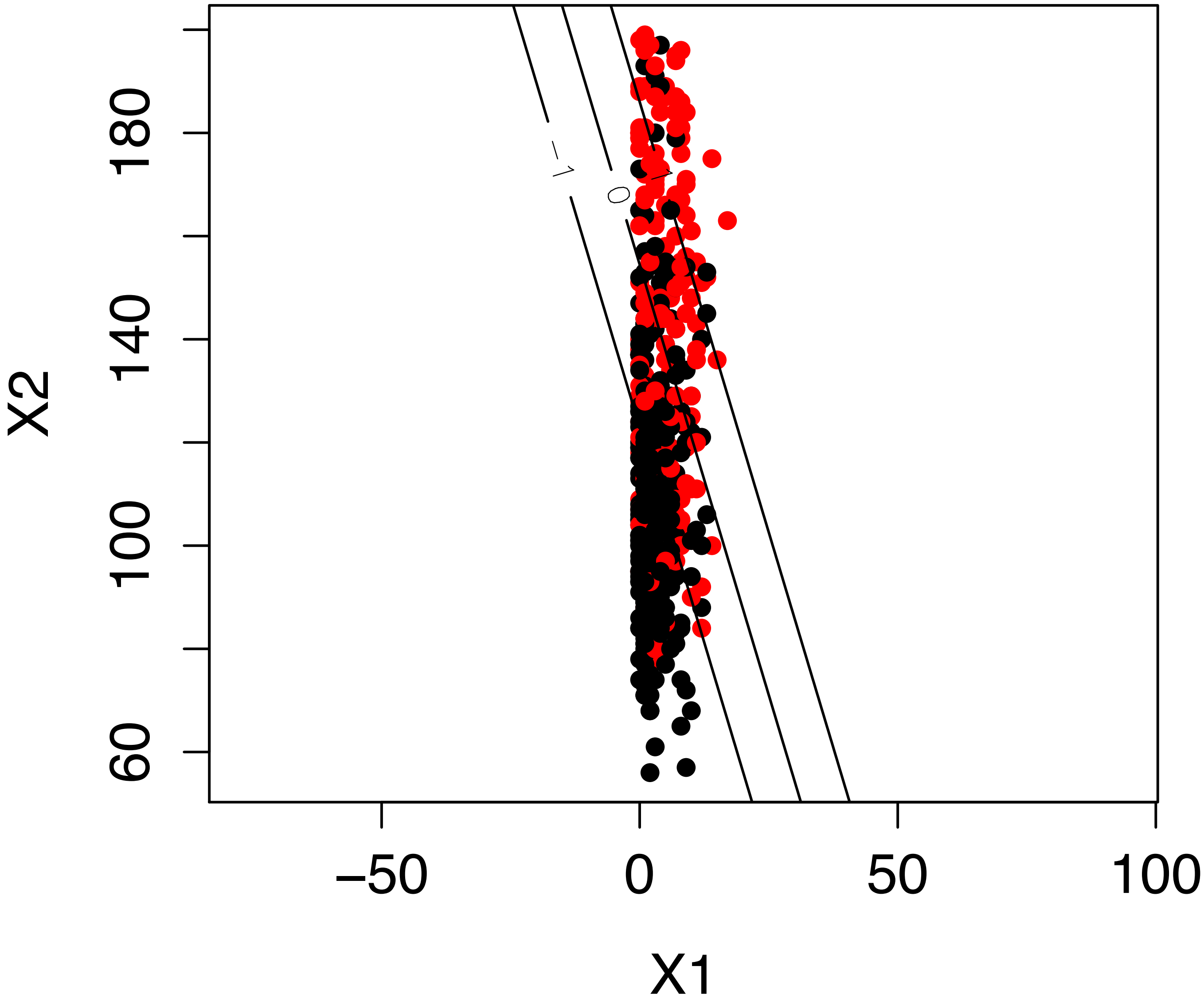
Not linearly separable (1)



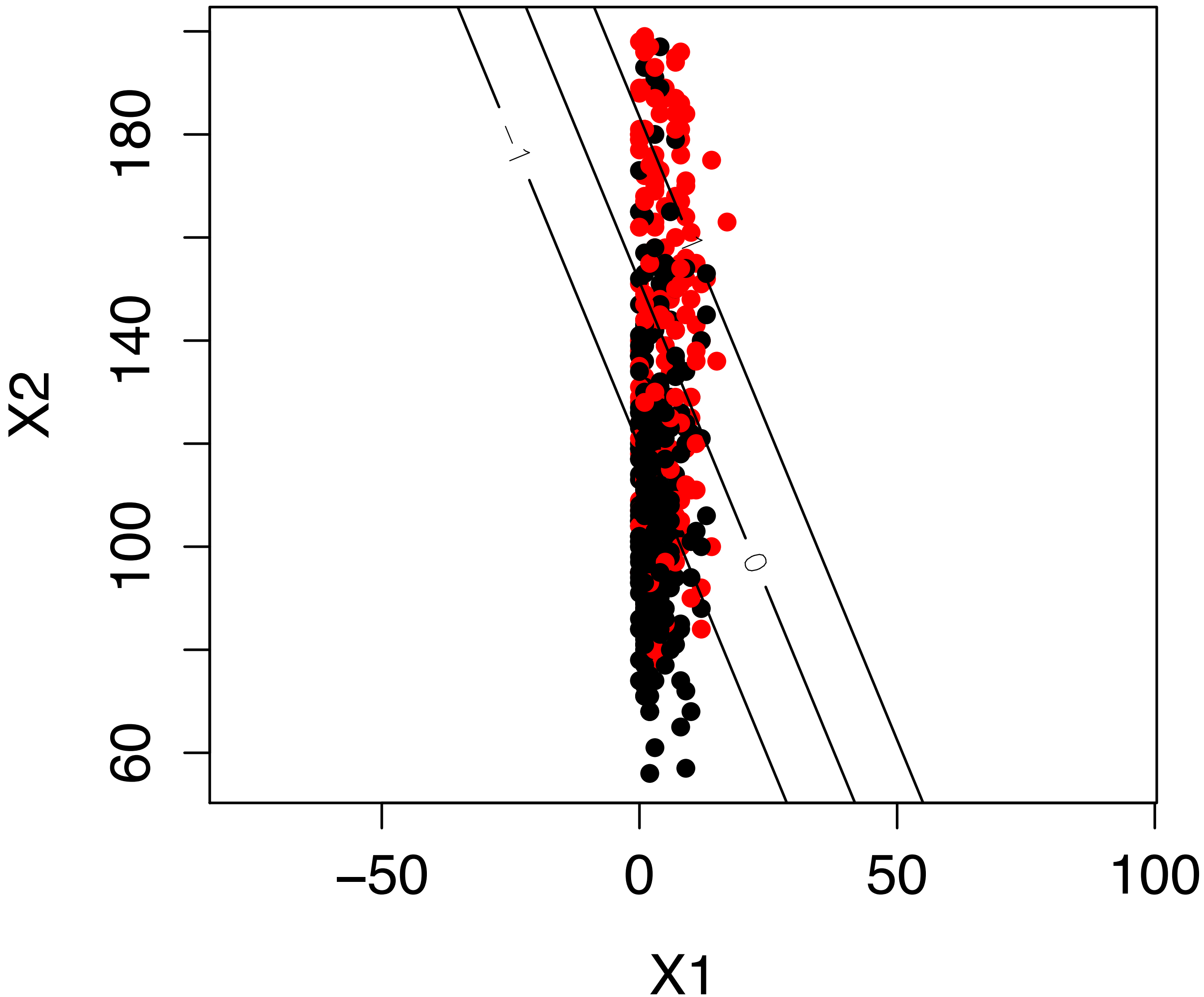
Not Linearly Separable (2)

- We have to allow some objects to be on the wrong side of the boundary.
- But how many are allowed to be on the wrong side?
  - How ‘expensive’ is it to be inside the margin, or on the wrong side of the boundary
- There is a trade-off between a wider margin and more objects on the wrong side of the margins.

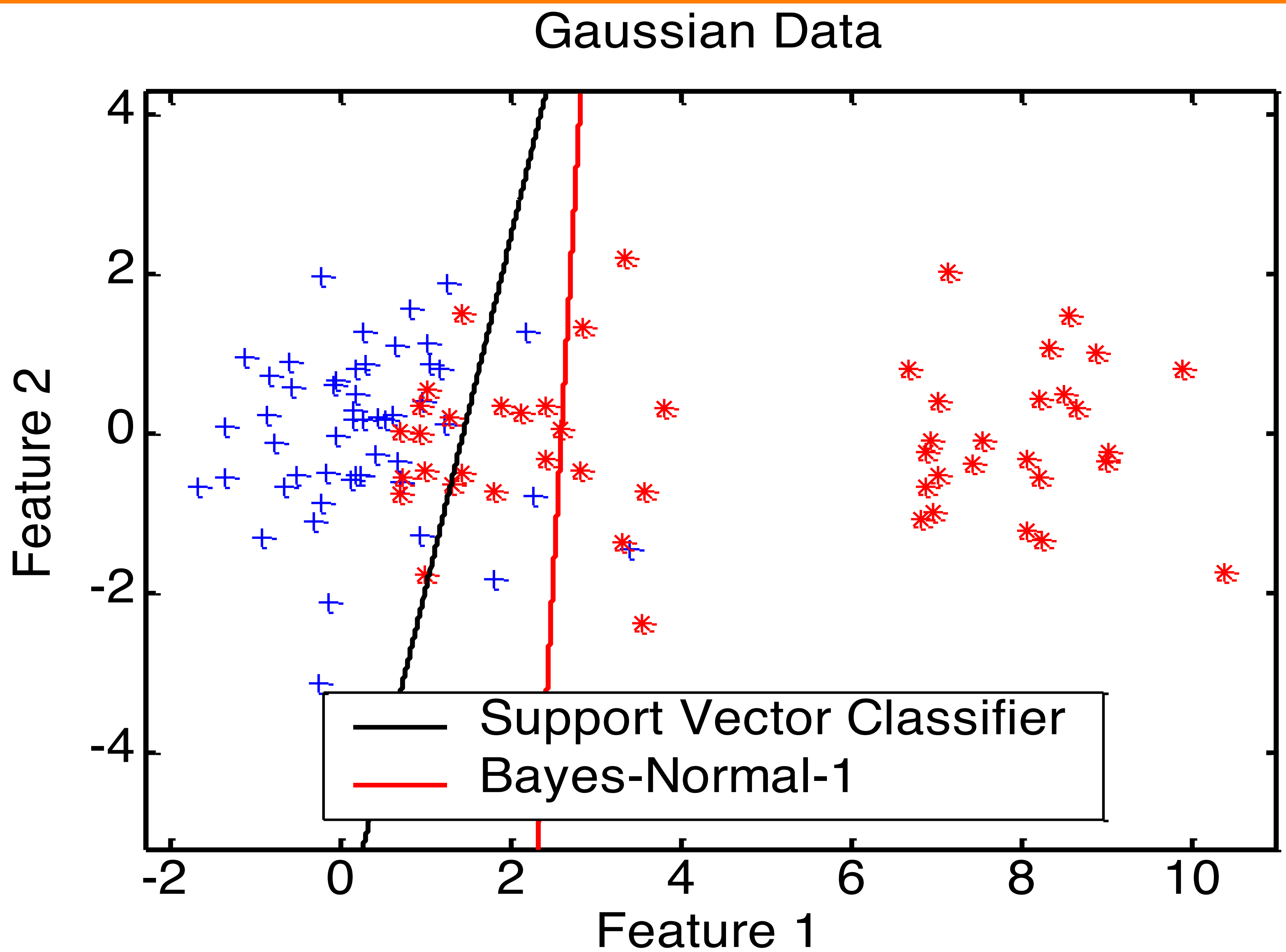
Diabetes revisited (Large C)



Diabetes revisited (Small C)



SVM: Difference with LDA



Conclusions

- One way of choosing the decision boundary is to maximize the margin
- The canonical example is the support vector classifier
- We can use an adaptation of this technique to allow for overlapping classes
  - This requires us to make a trade-off between a smaller margin and more object not outside the margin

References

- James et. al. - Introduction to Statistical Learning Section 9.1
- More complicated alternative (less recommended):
  - Hastie et. al. - The Elements of Statistical Learning, 2nd Edition, Section 4.5 Separating Hyperplanes (skip the mathematical details), especially 4.5.2 and Figure 12.1James et. al. - Introduction to Statistical Learning Section 9.1