Machine Learning in Business
John C. Hull

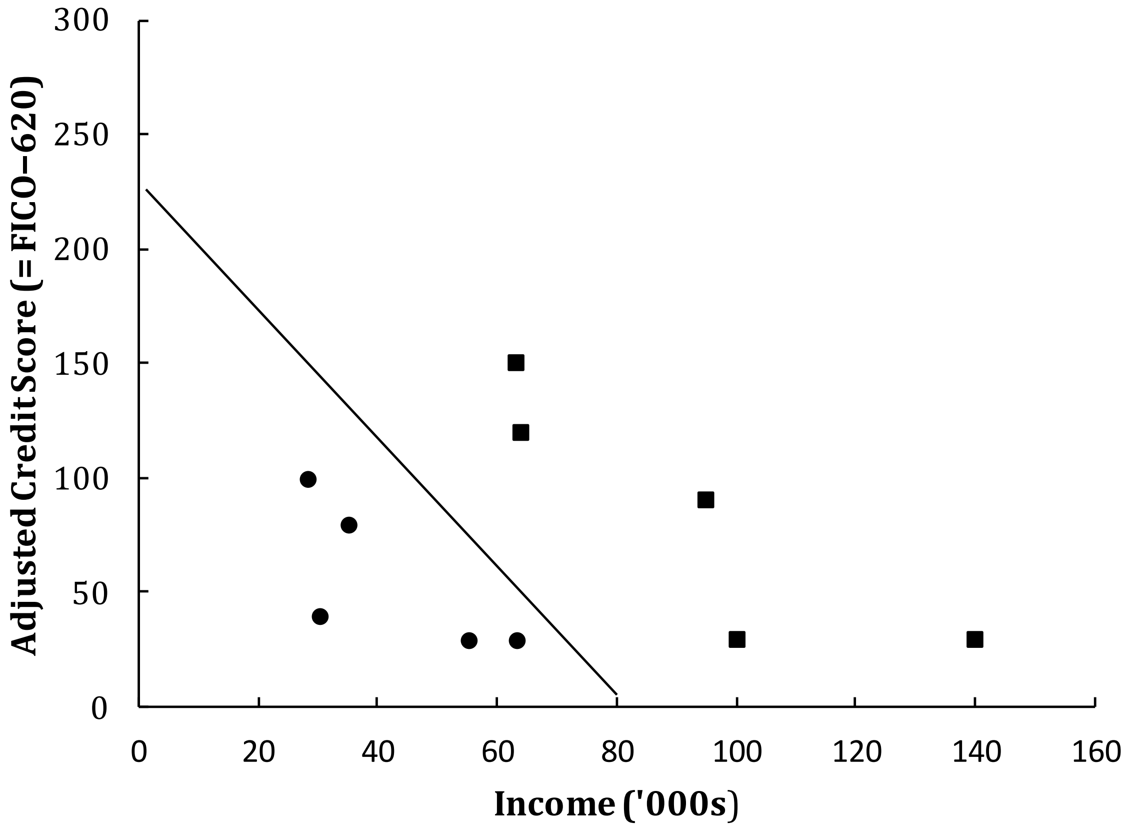
Chapter 5

Supervised Learning: SVMs

A Baby Example *(We carry out an approximate scaling by subtracting 620 from the credit score) Table 5.1*

Credit score	Adjusted credit score	Income ('000s)	Default =0; good loan=1
660	40	30	0
650	30	55	0
650	30	63	0
700	80	35	0
720	100	28	0
650	30	140	1
650	30	100	1
710	90	95	1
740	120	64	1
770	150	63	1

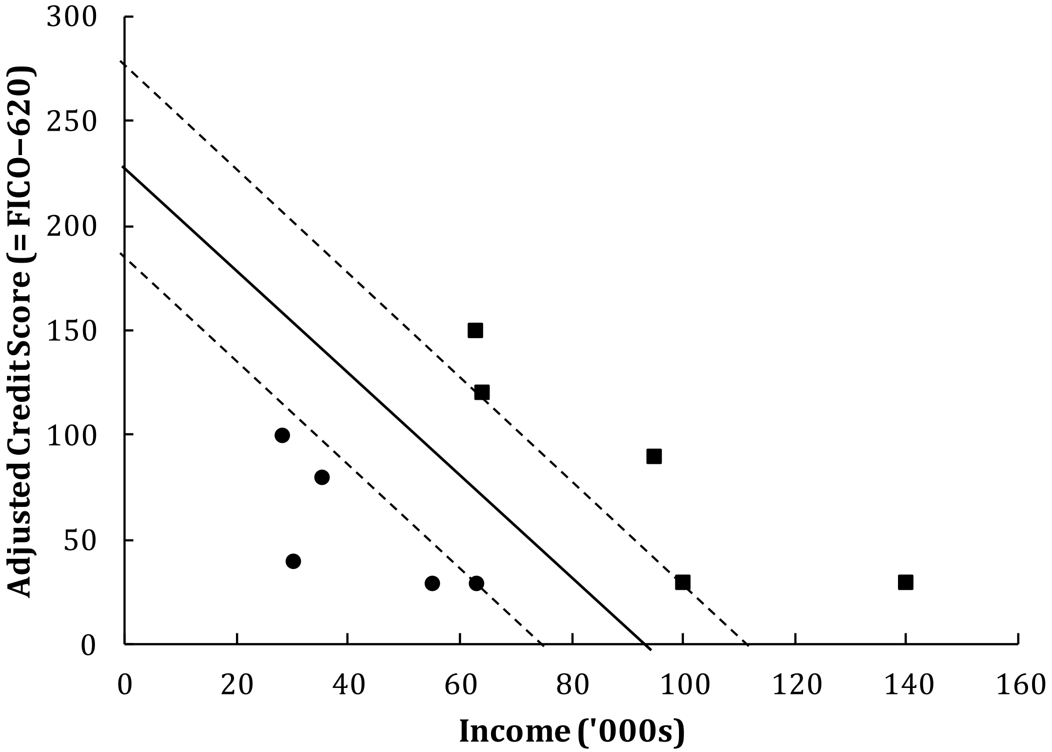
Linear Separation (circles are defaulting loans, squares are good loans) Figure 5.1



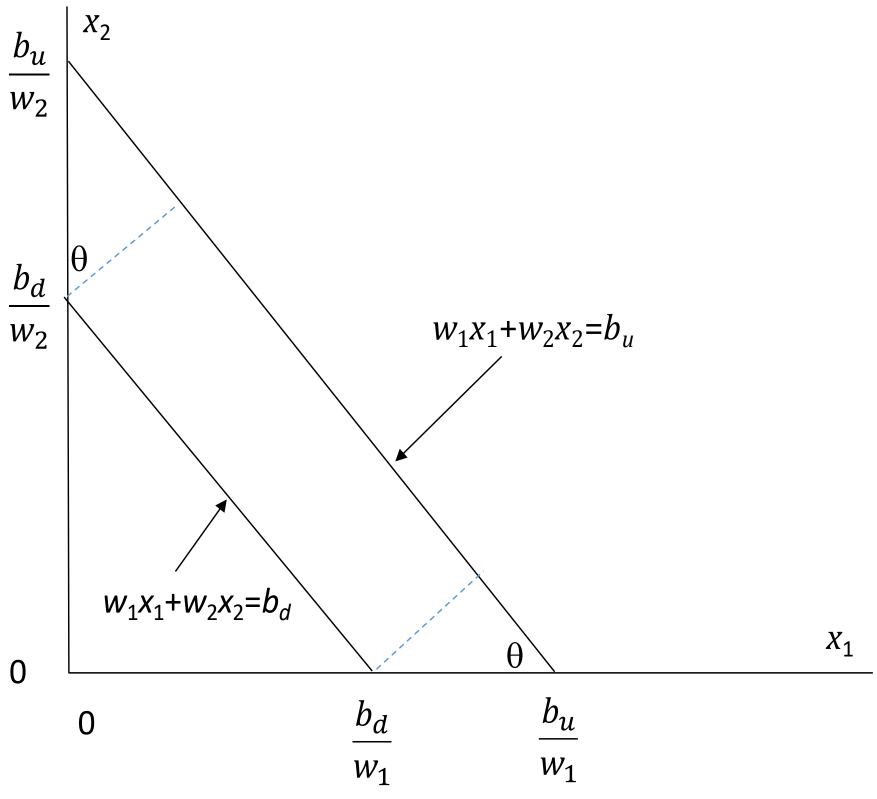
SVM Approach

- In the support vector machine (SVM) approach we find a pathway that separates the data into two classes as far as possible
- In the "hard margin" case perfect separation is possible (as in our example)
- The algorithm finds the widest path possible
- Data must be normalized. (We carry out approximate normalization by subtracting 620 from credit score)
- The support vectors are the observations at the edge of the pathway

Best pathway for example. Solid line would be used to distinguish good and bad loans



Notation (Figure 5.3)



The Math

If P is width of pathway

$$\sin\theta = \frac{Pw_1}{b_u - b_d} \qquad \cos\theta = \frac{Pw_2}{b_u - b_d} \qquad P = \frac{b_u - b_d}{\sqrt{w_1^2 + w_2^2}}$$

We can scale w_1, w_2, b_u, and b_d by the same constant without changing the model. We can therefore set $b_u = b+1$ and $b_d = b-1$ so that the width of the pathway is

$$P = \frac{2}{\sqrt{w_1^2 + w_2^2}}$$

In the hard margin case the algorithm minimizes $w_1^2 + w_2^2$ subject to perfect separation being achieved

Specification of hard margin problem for baby data

In our example the task is to find b, w_1, and w_2 to minimize $w_1^2 + w_2^2$ subject to

$$
\begin{aligned}
30w_1 + 40w_2 &\le b - 1 \\
55w_1 + 30w_2 &\le b - 1 \\
63w_1 + 30w_2 &\le b - 1 \\
35w_1 + 80w_2 &\le b - 1 \\
28w_1 + 100w_2 &\le b - 1 \\
140w_1 + 30w_2 &\ge b + 1 \\
100w_1 + 30w_2 &\ge b + 1 \\
95w_1 + 90w_2 &\ge b + 1 \\
64w_1 + 120w_2 &\ge b + 1 \\
63w_1 + 150w_2 &\ge b + 1
\end{aligned}
$$

The general hard margin problem

- The objective function is

$$\sqrt{w_1^2 + w_2^2 + \cdots w_n^2}$$

- We minimize this for values of w_i and b subject to the condition that there are no violations, i.e.:

$$\sum_i w_i x_i - b > 1 \text{ if loan good}$$

$$\sum_i w_i x_i - b < -1 \text{ if loan bad}$$

The Soft Margin Problem

We measure the violation of an observation as the extent to which the hard margin condition is violated

we minimize

$$C \times \text{sum of violations} + \sqrt{\sum_i w_i^2}$$

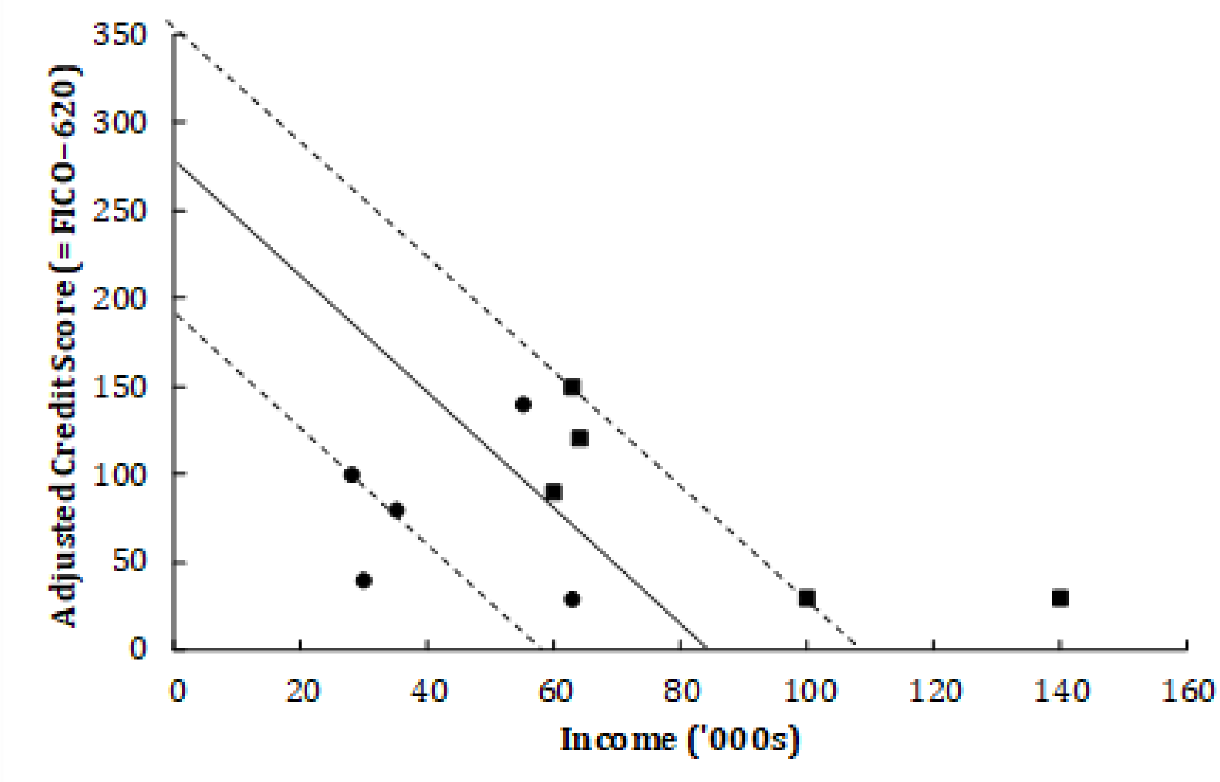
Changing C changes the trade-off between the width of the path and the violations

As C becomes smaller the pathway becomes wider with more violations

Changed example:

Credit score	Adjusted credit score	Income ('000s)	Default =0; good loan=1
660	40	30	0
650	**140**	55	0
650	30	63	0
700	80	35	0
720	100	28	0
650	30	140	1
650	30	100	1
710	90	**60**	1
740	120	64	1
770	150	63	1

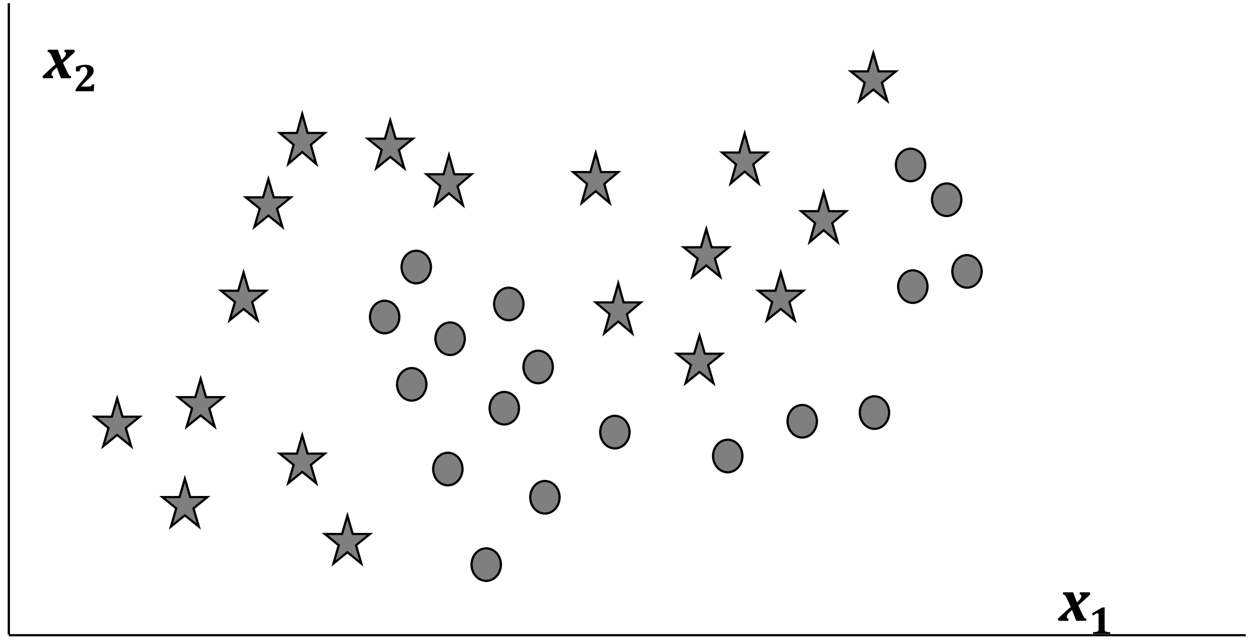
C=0.001 Results



Impact of C for Example

C	w_1	w_2	b	Loans mis-classified	Width of pathway
0.01	0.054	0.022	5.05	10%	34.4
0.001	0.040	0.012	3.33	10%	48.2
0.0005	0.026	0.010	2.46	10%	70.6
0.0003	0.019	0.006	1.79	20%	102.2
0.0002	0.018	0.003	1.69	30%	106.6

Non-linear separation (Figure 5.5)



Non-linear classification

- The objective is to create new features so that the boundary becomes linear
- Suppose there is a single feature (age?) and we find the low and high values of the feature tend to give one outcome while intermediate values give another outcome
- We could form a new feature as $(v\text{-}m)^2$ where v is the feature value and m is its mean

Forming new features

- We can add powers of each feature as a new feature.
- Alternatively, we can choose particular landmarks and create new features using the Gaussian Radial Basis Function (a similarity function). If values of features at a landmark are ℓ_1, ℓ_2, ……, ℓ_m, the new feature values are calculated as

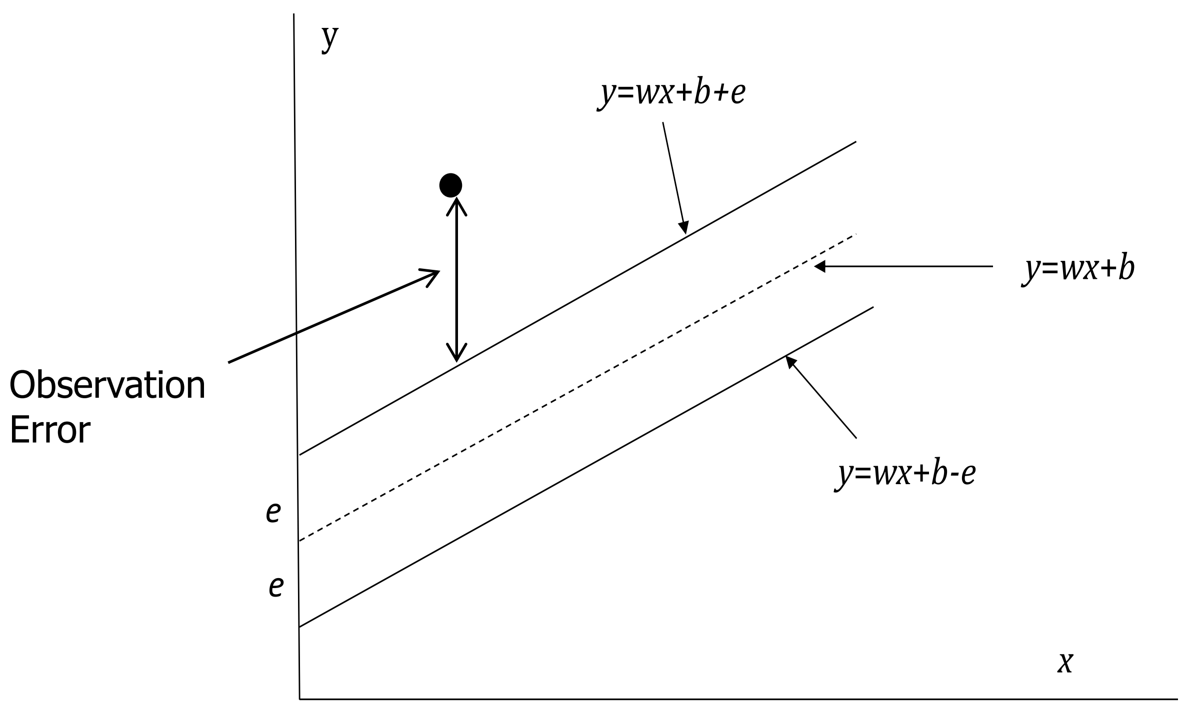
$$\exp\left(-\gamma \sum_{j=1}^{m} (x_j - \ell_j)^2\right)$$

- As the parameter γ increases the span of influence of a landmark decreases and the boundary becomes less smooth

SVM Regression: using SVM to predict a continuous variable

- We search for a pathway with a certain width that includes as many target values as possible
- If a target value lies within the pathway there is assumed to be no error
- If it lies outside the pathway the error is the difference between the actual value and the value predicted by the outer edge of the pathway

The Single Feature Case



General Case

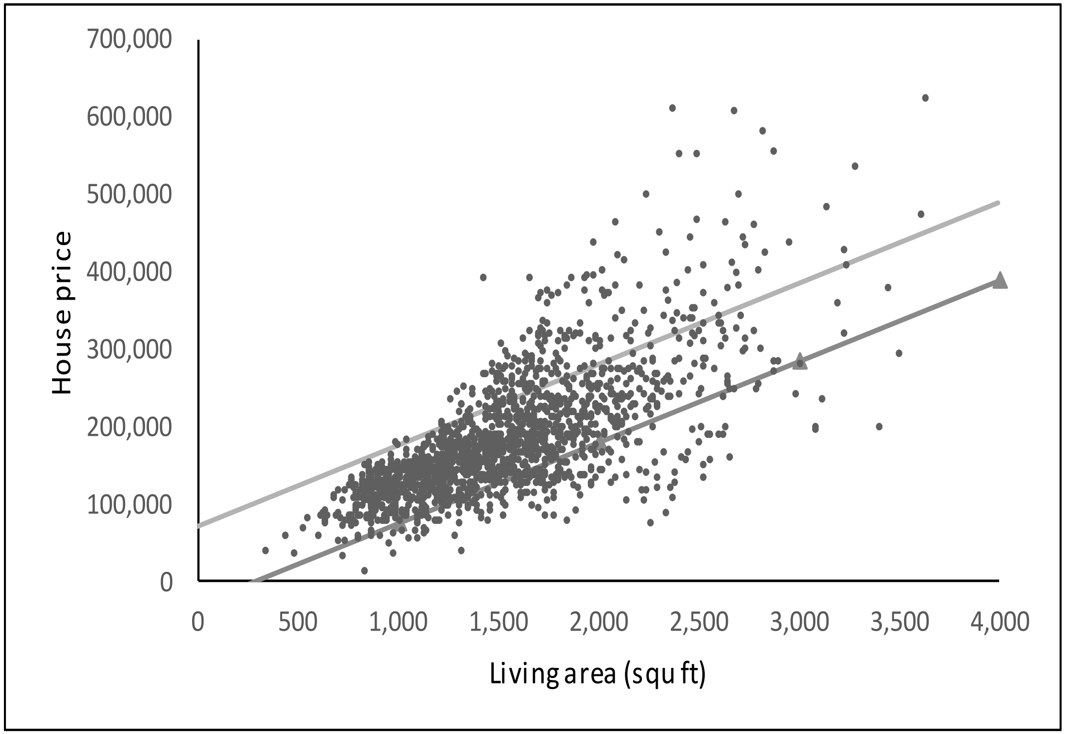
- We minimize

$$C \sum_{i=1}^{n} z_i + \sum_{j=1}^{m} w_j^2$$

where C is a hyperparameter

- z_i is the error (zero if observation lies within the pathway)
- The first term is concerned with reducing errors for observations outside the pathway
- The second term provides some regularization. It avoids large positive and negative w's

Predicting Iowa House Prices from Living Area when e=50,000 and C=0.01 (Figure 5.7)



Predicting Iowa House Prices from Living Area when e=100,000 and C=0.1 (Figure 5.8)