3.2 - Support Vector Machines

Giovanni Della Lunga
giovanni.dellalunga@unibo.it

Introduction to Machine Learning for Finance

Bologna - February, 2022

Support Vector Machines

- In this section we consider another popular category of supervised learning models known as **support vector machines**;
- Like **decision trees**, SVMs can be used for either classification or for the prediction of a continuous variable;
- We first consider **linear classification** where a linear function of the feature values is used to separate observations and in particular we will focus on **binary classification** where the separation is into only two categories;

Linear Separation

- **Linearly Separable Data points**: Data points can be said to be linearly separable if a separating boundary/hyperplane can easily be drawn showing distinctively the different class groups.
- Linear separable data points mostly require linear machine learning classifiers such as Logistic regression for example.

../5-pictures/chapter-4-4_pic_

Linear Separation

../5-pictures/chapter-4-4_pic_1.pn

Linear Separation

- Binary classification can be viewed as the task of separating feature space into two halves;
- A simple situation is that in which we attempt to classify loans into good loans and defaulting loans;

../5-pictures/chapter-4-4_pic_2.png

Loans Classification Example

- Consider two features: **credit score** and **income** of the borrower;
- We carry out an approximate scaling by subtracting 620 from the credit score (normalization);
- See Table 5.1 Hull

../5-pictures/chapter-4-4_pic_3.png

Loans Classification Example

- This is a **balanced data set** in that there are five good loans that defaulted;
- SVM does not work well for a seriously imbalanced data set and, if this is your condition, you need to use procedures to correct for this.

../5-pictures/chapter-4-4_pic_4.png

Linear Separation

circles are defaulting loans, squares are good loans

../5-pictures/chapter-4-4_pic_5.png

Linear Separation

- Which of the linear separators is optimal?

../5-pictures/chapter-4-4_pic_6.png

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

Example

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

../5-pictures/chapter-4-4_pic_7.png

SVM Approach

../5-pictures/chapter-4-4_pic_8.png

SVM Approach: Notation

../5-pictures/chapter-4-4_pic_9.png

The Math

../5-pictures/chapter-4-4_pic_10.png

The Math

- 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

The Math

- For the example in table 5.1 we can set x_1 equal to income and x_2 equal to credit score;
- All good loans must be to the north-east of the pathway while all defaulting loans must be to the south west of the pathway;
- This means that, if a loan is good, the income and credit score must satisfy:

$$w_1 x_1 + w_2 x_2 \geq b + 1$$

- While if the loan defaults it must satisfy:

$$w_1 x_1 + w_2 x_2 \leq b - 1$$

Example

- Specification of hard margin problem for our example
- In our example the task is to find b, w_1, and w_2 to minimize subject to

../5-pictures/chapter-4-4_pic_11.p

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.:

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

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

Hard Margin Vs Soft Margin

../5-pictures/chapter-4-4_pic_12.png

The soft margin problem

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

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

- 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:

../5-pictures/chapter-4-4_pic_13.p

Example

$C = 0.001$ results

../5-pictures/chapter-4-4_pic_14.p

Impact of C for Example

../5-pictures/chapter-4-4_pic_15.p

Non Linear Separability

- **Non-Linearly Separable data points**: This is the exact opposite of Linearly separable data points.
- View the image below, notice that no matter how one tries to draw a straight line, some data points will one way or the other get misclassified.

../5-pictures/chapter-4-4_pic_16.p

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 $(\nu - m)^2$ where ν is the feature value and m is its mean

The Kernel Trick

../5-pictures/chapter-4-4_pic_17.pn

The Kernel Trick

../5-pictures/chapter-4-4_pic_18.pn

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 $l_1, l_2, \ldots, l_m$, the new feature values are calculated as

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

- 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

../5-pictures/chapter-4-4_pic_19.p

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 ws

Example

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

../5-pictures/chapter-4-4_pic_20.p

Example

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

../5-pictures/chapter-4-4_pic_21.p