1年生実習 第3週

B5 研究室

2024 年 7 月 3 日

1　SVM(サポートベクターマシン)

本日使用するモデルは，SVM(サポートベクターマシン) です．SVM は，分類問題において高い性能を発揮することが知られています．

それでは，SVM の基本的な原理について見てみましょう．分類対象のデータとして，図 1 のような 2 クラスの 2 次元データを考えます．

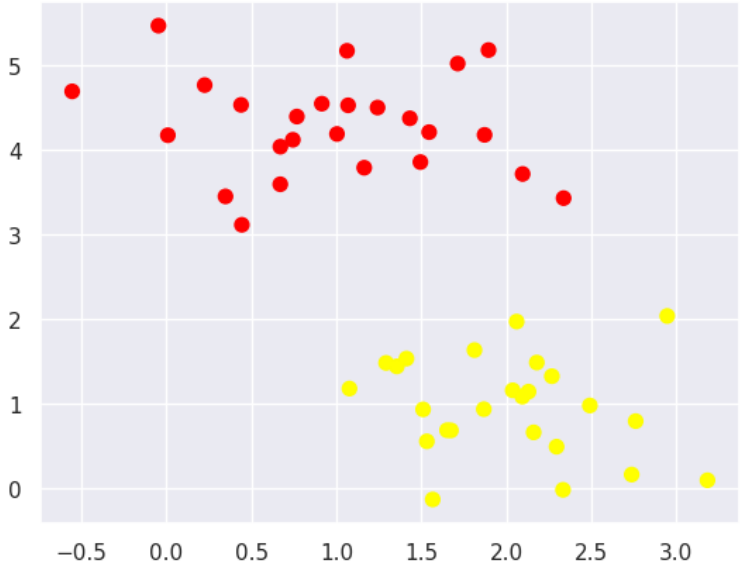

図 1: SVM のデータ

赤色と黄色の 2 クラスのデータを分類するために，2 つのクラスを分離する直線を見つけることが目標です．このとき，どのようにして直線を引くことができるでしょうか? 次のページに進む前に**定規で試しに線を引いてみてください．**

クラスを分割する線の例として，図 2 のような線が考えられます．

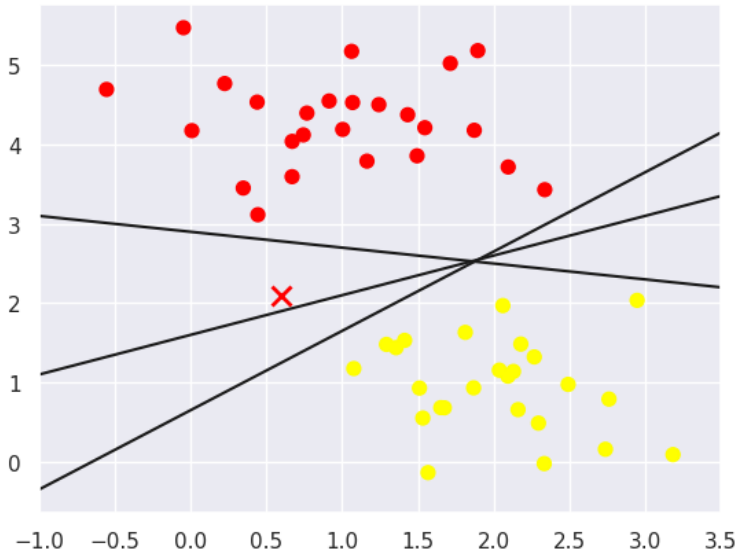

図 2: SVM の直線

これらの線は，赤色と黄色の線をうまく分割できています．しかしながら，どの線を境界とするかによって，図 2 の赤い × で示したデータのクラスが変わってしまいます．「クラス間に線を引く」という考え方は直感的で簡単ですが，実際の問題を解くためにはまだ不十分であることがわかります．

そこで登場するのが，マージン最大化という考え方です．マージンとは，クラス間の最も近いデータ点と境界線の距離のことです．クラス間の最も近いデータ点のことをサポートベクターと呼びます．図 2 の線におけるマージンを可視化した図を図 3 に示します．

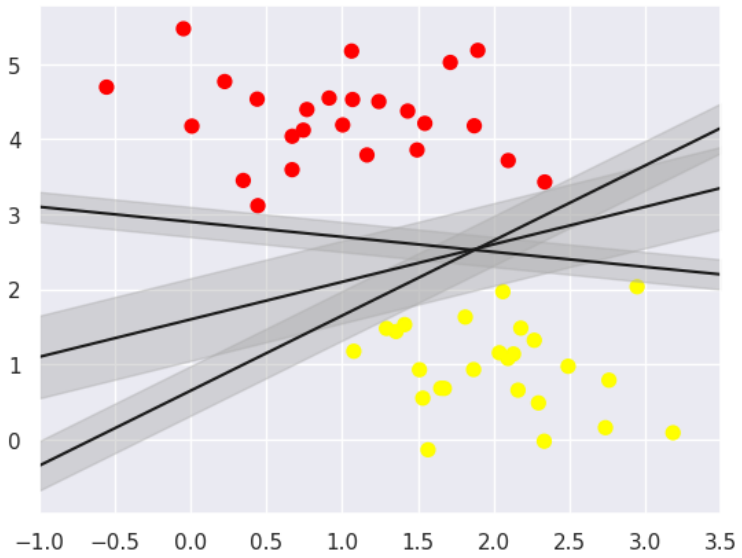

図 3: SVM のマージン

図 3 の例では，中央の線が最も大きなマージンを持っています．しかしながら，この線は最大のマージンを持っているわけではありません．SVM では，このようなマージンを最大化する直線を見つけることができます．SVM によって見つけられた直線は，図 4 のようになります．

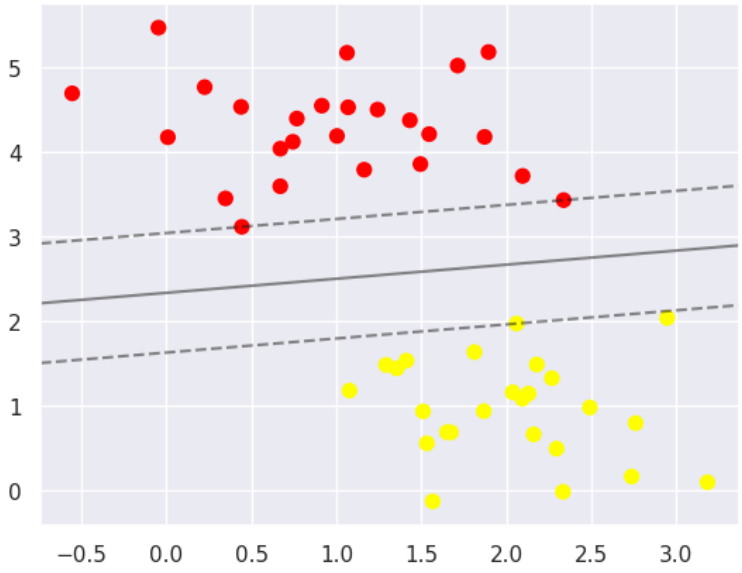

図 4: SVM のサポートベクターとマージン

図 4 の直線は，サポートベクターによって定義されるマージンを最大化する直線です．

では，データ数が変わった場合はどうなるでしょうか? 図 5 に，データ数が変化した場合の SVM の分割例を示します．

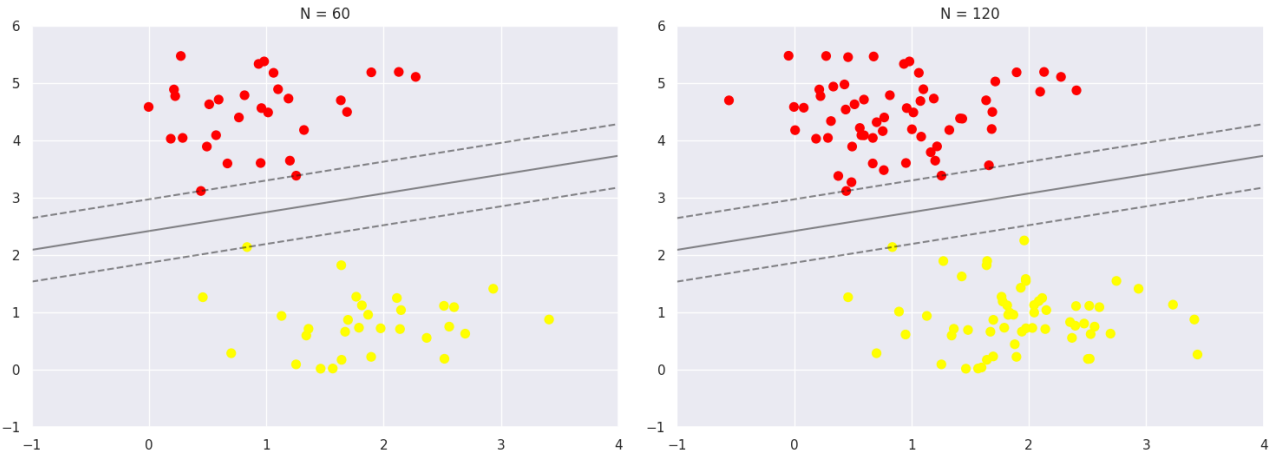

図 5: データ数が変化した場合の決定境界 (右：60 点，左：120 点)

SVM は境界線から遠いデータに影響されず，サポートベクターのみを用いて境界線を引いていることがわかります．このような，境界から離れたデータに対する影響が少ない性質が，SVM の特徴の一つです．

ここまで見てきたデータは，決定境界が直線の場合のデータです．直線で分割できるデータのことを，線形[1]分離可能なデータといいます．では，境界線が直線ではない場合はどうなるでしょうか? 図 6 に，直線で分割できないデータを示します．

[1]”線形”という用語は，様々な場面で登場しますが，基本的には，直線で表せる関係を意味します．

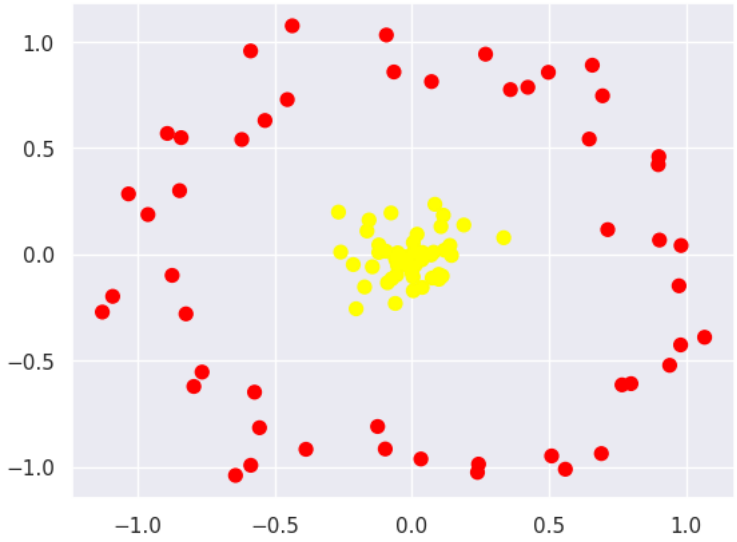

図 6: 直線で分割できないデータ

図 6 のデータを分割するために，図 6 に高さ方向の次元を追加した 3 次元空間を考えてみましょう．このとき，図 6 のデータを 3 次元空間にプロットするための式として，次のような式が考えられます．

$$r = e^{-(x^2+y^2)} \tag{1}$$

式 (1) は，2 次元データ (x, y) を 3 次元データ (x, y, r) に変換する式です．この関数は，原点が最も高い値を持ち，原点から離れるほど値が小さくなるような関数です．(図 7) このような，距離に基づいて値が決まる関数のことを，放射基底関数 (Radial Basis Function: RBF) といいます．

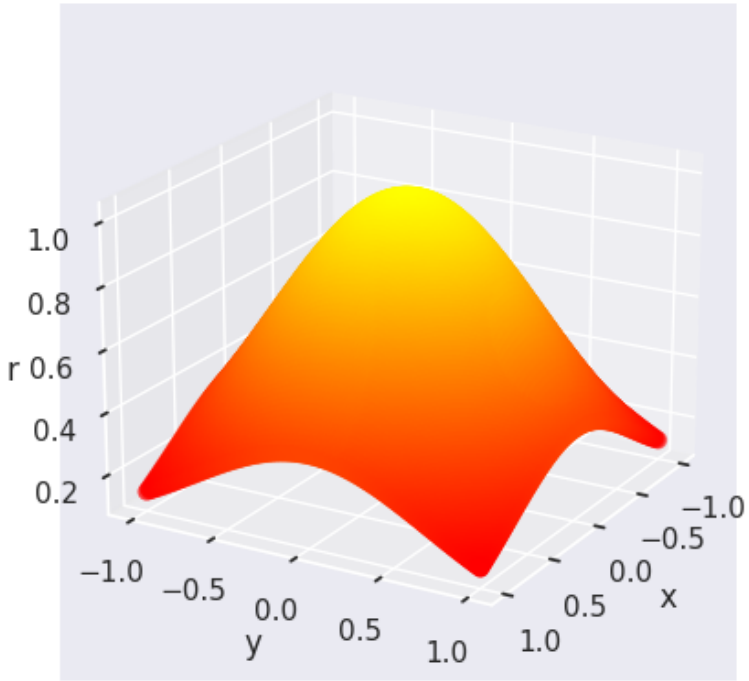

図 7: RBF 関数

では，RBF 関数を用いて，図 6 のデータを 3 次元空間にプロットしてみましょう．図 8 に，RBF 関数を用いて 3 次元空間にプロットしたデータを示します．

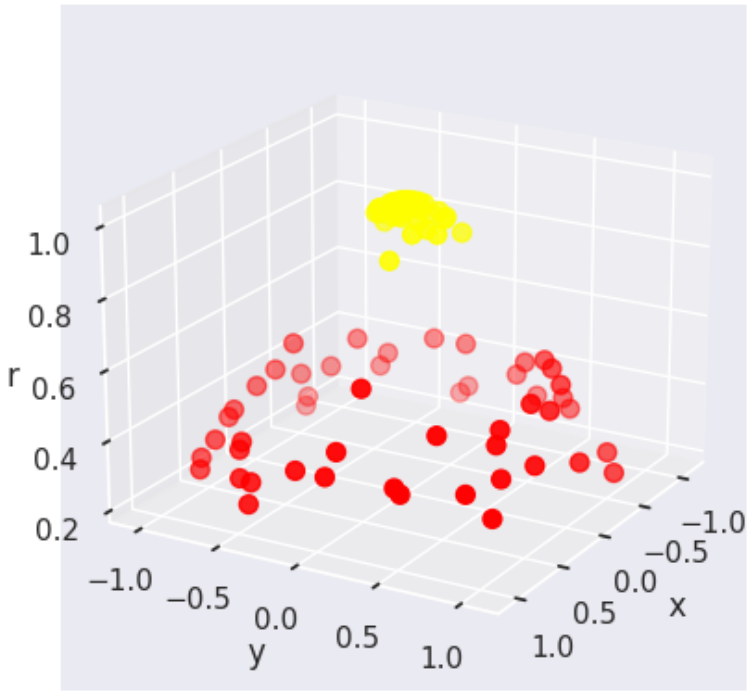

図 8: RBF 関数を用いて 3 次元空間にプロットしたデータ

図 8 を見れば，だいたい $r = 0.7$ となるあたりの平面で分割すればうまくいきそうだとわかります．しかしながら，式 1 はたまたまうまくいっただけであり，(当たり前ですが) すべての場合でうまくいくわけではありません．今回の例で考えると，RBF 関数の中心をうまく設定することができなければ，データをうまく分割することができない，ということです．詳細は割愛しますが，SVM では，カーネルトリックという素晴らしい手法によりこの問題を解決しています．SVM はカーネルトリックという手法の恩恵を受けている，ということだけ覚えておいてください．

話をデータの分類に戻します．図 9 に RBF 関数によって分類された結果を示します．

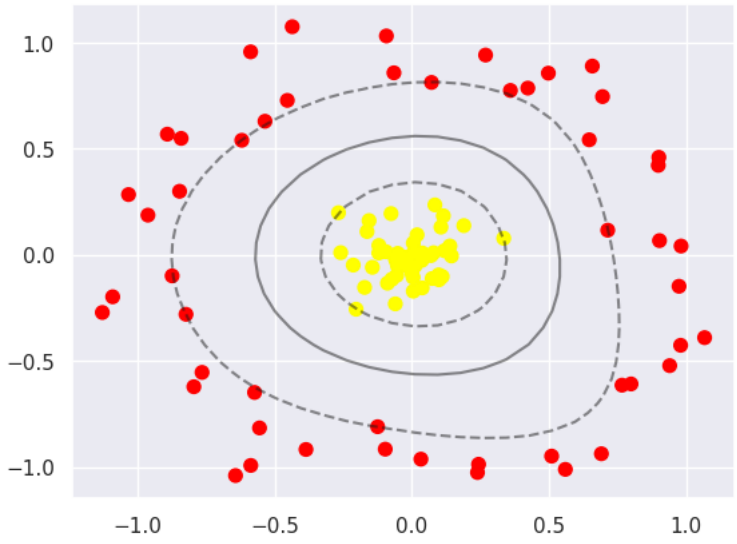

図 9: SVM の分類結果 (RBF 関数)

線形に分割できなかったデータを RBF 関数を用いて，マージンが最大となるようにうまく分割することができました．RBF 関数のような，データを高次元空間に写像する関数をカーネル関数といいます．RBF カーネル以外にも，多項式カーネルやシグモイドカーネルなど，様々なカーネル関数が存在します．カーネル関数を用いて，

データを高次元空間に写像する手法のことをカーネル法といいます．カーネル法を使用した SVM は，非線形なデータに対しても高い性能を発揮することが知られているため，頻繁に使用されます．

2　機械学習によるデータ分類

今週の実習では，先週の実習で皆さんから集めた手書き文字データを分類するための機械学習モデルを構築します．現状のデータセットの状態は以下の通りです．

- 取得文字: ○× の 2 文字
- 取得文字数: 3 セット × 10 人の 60 文字
- データ長：最長のデータに合わせて引き伸ばし (M1 の方で処理済)
- 座標軸：X, Y, Z の 3 軸

このデータセットを用いて，分類を行うための機械学習モデルを構築します．

2.1　データの準備

先週取得したデータを OneDrive からダウンロードし，Google Drive 上の任意の場所にアップロードします．

ここから後ろはまた後で書きます

3　課題

参考文献

[1] Jake VanderPlas. Python データサイエンスハンドブック 第 2 版: 菊池彰訳. オライリージャパン, 2021, 545p.