# データサイエンス基礎講座 超初級　第4限

フューチャーブリッジパートナーズ株式会社

長橋　賢吾

# 第4時限　データって、どうまとめるの……＜その２＞

- ドクター：第3限では、これまでの統計解析のアプローチから一歩進んで、データマイニングの一種のアソシエーション分析、階層化クラスタリング、重回帰分析を取り上げました。
- あゆみ：はい、難しかったですが、主成分分析よりはわかった気がします。
- ドクター：はい、主成分分析、因子分析は直感的に理解しにくい一方で、アソシエーション分析、階層化クラスタリング、重回帰分析は直感的にわかるのでハードルは低いですね。今回は、前回ちょっと学習した機械学習をもう少し深めていきます。

# カーネル法・サポートベクターマシンとは？

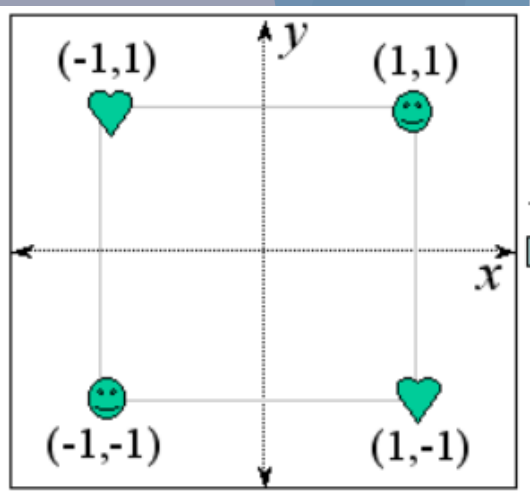
- ドクター：今回、最初に取り上げるのは、カーネル法・サポートベクターマシンです

- あゆみ：う～ん、まったくわからないです。。
- ドクター：はい、名前は難しいですが、やっていることは意外とシンプルです。

- あゆみ：カーネル法とサポートベクターマシンって何ですか？
- ドクター：はい、簡単にいうと、カーネル法という方法を使って機械学習をする方法（サポートベクターマシン）です。

- あゆみ：カーネル法。。。

# カーネル法・サポートベクターマシンとは？

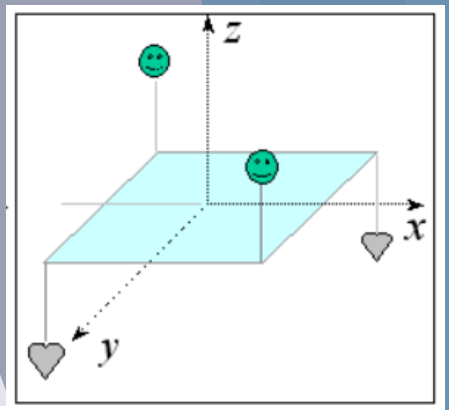



- ドクター：まず、最初にカーネル法から始めます。上図は、笑顔とハートを線引きをしたいのですが、どう線引き（線形分離）すればいいかわかりますか？

- あゆみ：えーと、わからないです。
- ドクター：笑顔は、(1,1),(-1,-1)、ハートは(-1,1),(1,-1)、単純に線をひいたら、交差してしまい線引きはできません

- あゆみ：たしかに、そうですね～

# カーネル法・サポートベクターマシンとは？




- ドクター：線引できないとき、どうするか、その答えを提供するのが、カーネル法です。
- あゆみ：というのは？
- ドクター：この4つの位置は、x,yという2次元の座標です。そして、その2次元の座標に加えてzという3次元の座標に変換します。
- あゆみ：うーん、なぜ、変換が必要ですか？
- ドクター：はい、先ほどの例では、x,yの2次元では線引きすることはできませんでしたが、それを3次元にすることで線引します。

# ロジスティクス回帰とは？

- ドクター：ロジスティクス回帰と回帰分析のちがいはわかりますか？

- まゆみ：わからないです～
- ドクター：はい、ロジスティクス回帰は、“回帰”と名前があるように、y=ax+bであらわされる回帰分析の一種です。

- まゆみ：はい、これはわかります。

- ドクター：ただし、決定的に違うのは、目的変数が二項分布、すなわち、目的変数が0～1の値しかとらないこと、であることです。

# ロジスティクス回帰とは？

- ドクター：二項分布であることは、機械学習のアプローチに近いです。たとえば、あるロジスティクス回帰式で、迷惑メールに関するロジスティクス回帰のモデル（回帰式）を作成した場合、説明変数はたくさんあっても、目的変数は、0(迷惑メール）、もしくは、1(正常メール）です。

- まゆみ：それと機械学習との関係は？
- ドクター：はい、先ほどのサポートベクターマシンは、あるデータを１と－１に分類することです。そして、その分類した学習データをもとに、テストデータを入れて、分類することが目的です。
- まゆみ：はい、分類することですよね。

# コラム4　ビックデータ時代の統計解析

ビックデータという名前を聞かない日がないくらい、あちこちでビックデータという単語を目にします。それは、かつてクラウドが新鮮な響きをもたらしていた2000年代、そこから、クラウドが当たり前になって、日常に欠かせなくなってきた2014年の現在、10年経過することで、すそ野が広がってきた感があります。ビックデータももう少ししたら、クラウドのような存在になるのかもしれません。

そのなかで、ビックデータ的統計解析について考えてみたいと思います。ビックデータは言うまでもなく膨大なデータであり、その膨大なデータを解析することで付加価値を得る考え方です。

一方、統計解析には大きく2つの方法があります。一つは、小さなデータから、母集団を推定する方法、世の中の多くの事象は正規分布に従っているので、少ないサンプルから、正規分布にそって母集団を推定する方法です。

そして、もう一つのアプローチは、膨大なデータから学習し、新しいデータに対して、何かしらの判断基準を提供するアプローチです。今回取り上げた教師あり学習であるサポートベクターマシン、ロジスティクス回帰は、この考え方です。

ビックデータ時代の統計解析という点では、やはり、後者の機械学習的なアプローチが役立ちやすい局面と思われます。そして、サポートベクターマシンのように、計算量が少ないために、膨大なデータを学習しても、計算量は指数関数的に増えない手法も、この機械学習的なアプローチに一役買っています。

だからこそ、重要なこと、やはり、データです。どれだけデータを集めるのか、データを征するものはビジネスを征する、そんな時代がやってきそうです。