1　イントロダクション

この読み物は、”An Introduction to Statistical Learning”（http://www-bcf.usc.edu/ gareth/ISL/ から入手可能。）の解説として大和（以降、筆者とする）が記したものである。はじめに断っておくが、これは単なる日本語に訳したものではない。重要な部分をピックアップし、要約・加筆・修正したものである。ただし、図などは（今のところ）貼付ける予定はないので、Figure とかこの文章中に出くわせば、適宜英語の図の所を参照してほしい。（本当はこのテキストに貼付ければいいのだろうが、めんどくさい・・・）

機械学習は、ざっと言えば何らかのデータを理解するための一つのツール（もしくは複数の機械学習を用いるならばツール群と言った方がいいかもしれない）である。機械学習は、大別すると以下の２つに分類される。

- 教師付き学習 (supervised)：１つ、あるいは複数入力。アウトプットは１つ。
- 教師なし学習 (unsupervised)：１つ、あるいは複数入力。アウトプットはない。データ構造や関係性を探ることを目的とする。

以下、この本で使用されるデータについて述べる。ただし、いずれのデータおよび分析手法は後のチャプターで詳しく述べられるはずなので、今は概観をつかんでもらえればいいと思う。

1.1　給与データ

想像できるように、労働者の給与は社会人経験年数に関係している。また、教育にも依存するだろう（大学卒業、大学院卒業など)。後の章で、回帰分析の際このデータを使用する（データに関しては、Figure 1.1 参照)。

1.2　株式市場データ

株価自体は、ある正の値を持っていて、将来の株価予想などが行えれば便利である。ただ現実は予想が難しい（予想できれば、こんな訳してる場合じゃなく株取引を積極的にしてるはず)。一歩条件を緩くして、価格を正確に予測する代わりに、前日から今日の価格が上がるか、もしくは下がるかだけでも予測できないだろうか？ (ということで、そんなデータの説明が Figure 1.2。）ただ、こういう分析は教師付き学習ではあるけれでも、離散的なアウトプットをするので（前日に比べて価格が上がる、もしくは下がるの２通り)、統計では分類（Classification）と呼ばれる（回帰に対して)。

株価予想モデルは、チャプター４で行うが、60% の確率で上下変動を予想できるみたい。

1.3 遺伝子情報データ

教師なし学習として、遺伝子情報データを最後に紹介する。インプットデータは遺伝子の並び方であるが、アウトプットとして何か値を出すのではなく、インプットデータの特徴を捉えることを目的としている。(Figure 1.4 に遺伝子情報を何らかのスコア２つ Z_1, Z_2 に焼き直してクラスターした結果が示されている。左の図は機械学習による結果で、右はガンの種類１４種類を表したもの。この場合、ガンの種類の情報はデータとして持っているので、分類と捉えることも出来そう。)

1.4 機械学習の歴史

割愛。興味がある人はどうぞ。ただし、この本"An Introduction to Stastical Learning"のベースとなる本があって"The Elements of Statistical Learning" (http://statweb.stanford.edu/ tibs/ElemStatLearn/ から入手可能)、この本が終われば読んでみるのもいいかも。(筆者は大学院のときに Boosted Decision Trees の部分だけ読みました)

1.5 記法

割愛しますが、この本は極力行列表現をなくしているみたいです。

1.6 データセットなど R に関して

R でのコードもついています。使用するデータセットは下のライブラリ２つについています　b 7。

- ISLR
- MASS

なので、install.packages() でインストールしておいてください。