データサイエンス基礎講座 超初級　第５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５時限　データを入力して、結果が出た！ 感激！！

- ドクター：第4限では、機械学習を深く理解するということで、カーネル法・サポートベクターマシン、ロジスティクス回帰、そして、決定木を扱いました。

- あゆみ：難しかったです～
- ドクター：はい、やはり、機械学習は簡単に理解できるものではないので、繰り返し学習することが大事です。

k平均法とは？

- ドクター：今回、最初に取り上げるのは、k平均法(kmeans)です。

- あゆみ：平均ですか？
- ドクター：はい、平均ですが、単に平均をとるってことではありません。

- あゆみ：kもあるんですよね？
- ドクター：k平均法は、非階層クラスタリングの手法です。

- あゆみ：非階層クラスタリング？

k平均法とは？

- ドクター：非階層クラスタリングは、第3時限で取り上げた階層化クラスタリングと同じクラスタリングです。
- あゆみ：クラスタリングはまとめることですよね。

- ドクター：はい、その通り、クラスタリングは似た者同士をまとめる手法で、正解データがない上でまとめることから教師無し学習とも言います。

- あゆみ：これは、わかります～

コラム５　アンサンブル学習と集合知

ジェームス・スロウィッキー著「みんなの意見は案外正しい」（角川文庫）は、今回取り上げたアンサンブル学習を考える上でいろいろ示唆があります。

この本の冒頭では、ある論文の内容が紹介されています。それは、19世紀の英国の科学者フランシス・ゴルトンが、カウンティフェアに参加したときに、雄牛の体重を予測するという大会で、参加者が提出した787の予測値の平均が、実際の体重とほぼ一致し、どの個人、専門家による予想値よりも誤差が少ないという結果が得られました。

すなわち、みんなの意見は案外正しい、というわけです。スロウィッキーは、こうした「みんなの意見は案外正しい」という状況が成立する条件として、１．意見の多様性、２．独立性、３．分散化、４．集約をあげました。その一方で、「みんなの意見が案外正しい」が成立しない状況として、１．均一化、２．中央集中、３．分裂、４．模倣、５．情動を上げます。

ひるがえって、アンサンブル学習もたぶんにこの「みんなの意見は案外正しい」の要素が含まれていそうです。アンサンブル学習の前提は、独立です。弱学習器は、それぞれ独立しており、その独立した結果をもとに集約し、多数決・平均をとります。

というわけで、コンピュータで正しい答えを得る、という考え方が変わりつつあるのかもしれません。それは、一台のコンピュータで、完成度の高いアルゴリズムによって、分類・回帰するのではなく、大量のコンピュータ・クラスタが独立した手法・アルゴリズムによって、分類・回帰し、それを集約することで、正しい答えを得る。こうした時代になりつつあると「みんなの意見は案外正しい」から思いました。