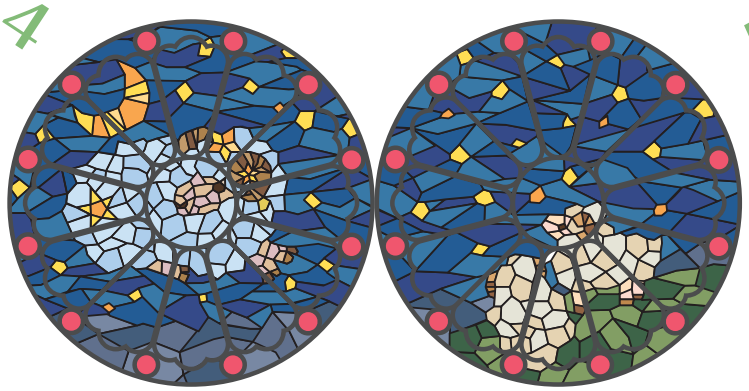

数理の窓

どうやって機械を信じますか？

最近の人工知能ブームの原因の一つに神経回路網モデルの有効性が確認されたことがある。特にそのようなモデルの一つであるディープ・ラーニング（以下DL）を応用したプログラムが囲碁界のトッププロを負かした2016年3月のニュースは衝撃的であった。さて、画像のパターン認識、自然言語処理、信用リスク判断など、DLの応用は急速に広がっているが、ビジネスに活用しようとすると必ず問題となるのが説明可能性（interpretability）である。この説明可能性は特に金融等の規制業界においては重要で、単に社内のモデルがリスク量はこの程度だと言っているので、というような説明だけでは許されない。過去のデータに対するフィッティング度合を数量的に示した上で、モデルの特徴を把握できるような分かりやすい説明が求められるのである。では、どのようなモデルなら分かりやすくて、どのようなモデルは分かりにくいのであろうか。

まず、説明変数の数が多すぎると分かりにくくなる。信用審査のモデルでも説明変数がかなり多いが、集約すると影響度の大きな数個の説明変数に限定できる。また説明変数に関して被説明変数が単調増加か単調減少であれば分かりやすいが、途中まで増加するけれども他の説明変数との関係次第では減少に転じたりする挙動のモデルは分かりにくい。さらに、大局的な動作が把握できず、局所的な動作しか説明されていないと、分かりやすいとは言いにくくなる。このような観点から見たときに分かりやすいモデルというのは線形で（即ち一次式で書けて）、重要な説明変数が少ないものということになるだろう。

そこで、DLの結果を分かりやすく説明するための工夫がいろいろと検討されている。DLを幾つかの線形モデルのつなぎ合わせとして説明を試みるもの、影響度の大きな説明変数に絞って神経回路網のつながり具合を精査するものなど様々である。恐らくこのような努力の結果としてDLが更に受け入れやすくなっていくものと想像できる。

しかし分かりやすさの追求には自ずと限界があるため、全く違う観点の評価方法が必要になってくるのではないか。例えば、優秀な人物を人が評価する際には、脳の中を調べるわけにはいかないので何らかのテストを実施して、一定以上の能力があることを見極めている。それと同じようにDLをテスト問題の正答率で評価するという考え方もあっていいだろう。場合によっては、評価用の別のプログラムが評価対象のDLに直接接続して質問を繰り出し、正答率で能力を見極めるということだって起こってくるだろう。このDLの説明可能性は、一見専門的で特殊な問題のように映るが、今後、人が高度な機械を如何に信用し受け入れていくのかという、より大きな問いを包含していて面白い。　　（小粥 泰樹）