課題Review“Ten quick tips for machine learning in computational biology”を読んで

医学専攻　博士課程２年　新川神奈

このレビューを読んで、受講した前後の変化は、用語の理解度が上がったこと、また概念全体へのイメージがついたこと、また未知の内容も比較的スムーズに読み進められた点である。

専門用語に対する詳細な理解は、まだまだ足りないが、当初のレベルからすると前進した。当初は、例えば「教師あり・なし」と聞いても、全く検討がつかなかった。論文中には、Tip1、Tip2でデータの独自性を把握し、前処理をすることの大切さや、テストセットは最終的に決定したアルゴリズムで最後に行う（lock box approach）ことの大切さが述べられている。授業を受けた今は、感覚的ではあるが、「データの詳細をまず知ること、そして整理すること。ベストな方法で解析を行うために試行錯誤して最終的な結果へ導くこと。」の流れを想像できる。

Tip3には自分のデータセットを正しいアルゴリズムカテゴリーに落としこみ、解析する重要性が述べられている。特に、この章は授業を受けたことで、引っかかることなく読めるようになったと感じる。授業では、教師あり学習について詳しく学んだ。そのため、SVMsやｋ-NNという概念について、また最もシンプルなアルゴリズムから始めることという内容についても納得がいった。シンプルなモデルから始めることで、間違っても修正すべきところが見つけやすく、初心者にとっては過程も結果も把握しやすいものになる、という考え方は、非常に勉強になった。

　Tip7ではoverfittingについて述べられている。授業でもトレーニングセットで性能を上げて過学習した場合、実際のデータセットにはフィットしない問題について学んだため、興味深かった。また、“Overfitting is the bugbear of machine learning”という言葉は特に印象深かった。これに対処するためにcross-validaitonおよびregularizaionという手法があるとのことだが、こちらについては理解が追いついていない。正則化については授業で詳しい説明を受け、過学習を防ぐための手法としてRidgeやLassoなどの手法について学んだ。

Tip8では、不均衡なデータセットにも対応できるよう、accuracyやF1 scoreではなく、MCCやPrecision-Recall 曲線を用いることを勧めており、興味深く読み進めることができた。Tip9では、オープンソースのソフトウェアを使用することが勧められていたが、他の研究者との共有や将来どの団体に所属しても使用できるようにするために、これらのソフトを習熟することの必要性を、改めて感じた。またTip10には、専門家に相談する必要性や、そのツールについて書いてあり、将来的に役立ちそうだと感じた。

　理解が困難な箇所もあったが、以前に比較すると、確実に前進したことを感じる。特に、授業で扱って頂いた箇所については、親しみを感じた。

独学では、何年かけてもたどりつけないような、概念への理解を促して頂き、ありがとうございました。