Upload your answer report between 16:30-17:30, July 30.

Read the review https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5721660/ .

Is there any improvement in your skills to grab the content of the review?

If any, describe what they are and how they worked?

If no, describe how the lectures should be improved.

　まず、本講義を受講する前にこの論文を読んでも、何が書いてあるか全くわからなかっただろう。機械学習の勉強を始めるにあたってどこから勉強し始めれば良いのかもわからず、その糸口を得るために本講義を受講させていただいた。

　機械学習とは何かということや、pythonとjupyter notebookのダウンロードという非常に初歩的なところから始めて頂いたので、とても入りやすかった。また、全てにおいて根本的な概念に重点を置いて説明してくださったので、全体としてのフレームワークを押さえることができた。具体的なコードやデータマイニングを行うには詳細な知識がまだまだ足りないが、まずは概念的なところをつかんだ上で、詳細は教科書に書いてあることを自分で勉強するという方法が良いのだと思う。

　本論文に沿って、講義を受けたことで理解が深まった点、あまり理解できなかった点を考察する。

Tip 1: Check and arrange your input dataset properly

　ここでは、データセットのチェックとアレンジについて述べられている。最初の講義で、まず一般的な統計解析と機械学習の類似点と相違点、特徴などについて説明していただいたため、機械学習ではより大きいサイズのデータを扱い、より複雑なアルゴリズムを用いて予測をするということが理解できた。また、アルゴリズムそのものよりデータセットに注意を払うこと、外れ値に注意することなどは一般的な統計解析と同様である。

Tip 2: Split your input dataset into three independent subsets (training set, validation set, test set), and use the test set only once you complete training and optimization phases

　まずtraining setで学習させ、validation setで妥当性を検証し、test setに適応する。講義ではバイアスとバリダンスの問題も含めて解説して頂いたので理解しやすかった。

Tip 3: Frame your biological problem into the right algorithm category

　Target label（答）があるデータセットを利用できる場合は教師あり学習、target labelがない場合は教師なし学習のアルゴリズムを適応する。教師あり学習と教師なし学習、また回帰や分類、クラスタリングとそれらの関係についてはあまり理解できていなかった。分類や学習の話はそれも念頭におかれていたのだと思うが、自分ではそこまで考えが及んでいなかった。特に教師なし学習とはどういうものかというのがまだ理解できていないので、その辺りをさらに勉強したい。

Tip 4: Which algorithm should you choose to start? The simplest one!

　複雑なアルゴリズムではなく、最もシンプルなアルゴリズムを選択する。講義では、非常に複雑であろう機械学習に関して、毎回非常に根本的なところを可能な限りシンプルに説明して頂いたので、複雑な方が良い方法であるという誤った認識を矯正できた。シンプルな方法ほど学習しやすく制御しやすい。また、複雑であるほどオーバーフィッティングが起こりやすいということも教えて頂いたので、こうした概念は理解しやすかった。

Tip 5: Take care of the imbalanced data problem

　Imbalanced datasetを扱う場合は、training setを作成する場合に、その割合に気をつけなければならない。より多くのデータを集める必要があるが、不可能な場合は重み付けも有用な方法である。Imbalanced datasetや、F1スコアやMCCに関しては知らなかったため、あまり理解できなかった。本論文内に記載されていることは何となくわかったが、今後具体例などを使ってさらに勉強したい。

Tip 6: Optimize each hyper-parameter

　hyper-parameterという言葉は初めて聞いたが、パラメーターの数や、次元を増やすとどのようになっていくかということは講義で説明していただいたため、何となく概念は理解できた。ただ、具体的な手法などに関してはわからないので、今後は実際にデータを使って動かしてみたい。

Tip 7: Minimize overfitting

　これに関しては、講義で詳しく説明していただいたので、理解しやすかった。これまで、普通の統計解析ではデータにフィットするモデルほど良いと思っていたが、機械学習においては、training setにフィットしすぎると、test setに対する予測性能が悪くなるオーバーフィッティングという問題があることを学んだ。また、それを制御する方法として正則化やクロスバリデーションがあり、特に正則化についてはsubset selectionのところでも説明していただいたので、理解しやすかった。

Tip 8: Evaluate your algorithm performance with the Matthews correlation coefficient (MCC) or the Precision-Recall curve

　概念的なところは講義でも触れていただいたように思うが、Tip 5でも述べたようにMCCやF1スコア、Precision-Recall curveに関しては知らなかった。ここで挙げられている具体例は非常にわかりやすく、MCCやPR curveの概念や有用性について理解できた。自身の研究にも活かせそうなので、さらに掘り下げて勉強したいと思う。

Tip 9: Program your software with open source code and platforms

　もちろん独自のソフトウェアを開発する能力はないが、講義ではRとPythonを主に使用したので、RとPythonはユーザーが多くユーザーコミュニティも充実しているということが実感できた。コードを共有するという文化が根付いていて、非常に良いと思う。

Tip 10: Ask for feedback and help to computer science experts, or to collaborative Q&A online communities

　講義を通じて、いかにデータマイニングが専門知識を必要とするかを痛感した。もちろん自身でもその概念を理解してプログラムを動かせるようになりたいが、専門家のフィードバックやアドバイスは必要不可欠である。紹介してあるQ&Aサイトを見ると日本語でのサイトもあったので活用したい。また、より広く情報を得るためにも、英語でスムーズに理解できるよう精進したい。

　まとめると、本講義を受講したことで、これまで全く使ったことのなかったRやPythonに触れることができた。また、機械学習に関しての根本的な概念、考え方、注意すべき事項などについて学ぶことができた。しかし、まだ実際に自分でデータを使ってコードを作成したり動かしたりできる自信はない。今後は学んだことを念頭に置きながら、教科書やユーザーコミュニティを活用して、実装するための知識を身につけていきたい。