公立はこだて未来大学 2020 年度 システム情報科学実習
グループ報告書

Future University-Hakodate 2020 System Information Science Practice

Group Report

プロジェクト名

AI するディープラーニング

Project Name

AI Love Deep Learning

グループ名

自然言語処理を用いた競馬予想

Group Name

Horse racing predictions using natural language processing

プロジェクト番号/Project No.

12

プロジェクトリーダ/Project Leader

松田顕　Ken Matsuda

グループリーダ/Group Leader

川崎景大　Keita Kawasaki

グループメンバ/Group Member

鈴木健斗　Kento Suzuki
新大実　Masamitsu Atarashi
川崎景大　Keita Kawasaki
藤原慎太郎　Shintaro Fujiwara
武田佑樹　Yuki Takeda

指導教員

竹之内高志　寺沢憲吾　香取勇一　佐々木博昭　片桐恭弘

Advisor

Takashi Takenouchi　Kengo Terasawa　Yuichi Katori　Hiroaki Sasaki　Yasuhiro Katagiri

提出日

2021 年 01 月 14 日

Date of Submission

January 14, 2021

概要

近年、AI の発達が進んでおり、その中でディープラーニングを用いた技術は様々な分野で応用されている。そこで我々のグループは地域性などの理由から競馬予想に注目した。競馬予想ではオッズと呼ばれる払戻金の倍率と勝利確率の積が回収率となる。我々はその期待値を算出し購入する馬券を決定する。ここで、オッズは集合知によって形成されており、集合知を上回らなければ儲けを出すことができないという問題がある。集合知は多数の人が予想を行うことで生じるため、集合知を上回るためにはそこに含まれない要素を組み込む必要がある。我々はそのひとつとして調教師のコメントに注目し、自然言語処理を用いて分析を行うことになった。

手法としては、調教師のコメントを形態素解析にかけて LSTM を用いて予測を行う。形態素解析には Mecab[1] という形態素解析エンジンを用い単語分割を行うこととした。また、単語埋め込みには Word2Vec を使用し単語を実数値であらかじめ決められた次元のベクトルで表現を行う。

データセットには調教師のコメントと着順のデータが必要になる。そのため、競馬新聞からデータの収集をすることを考えた。競馬新聞には多数の種類があり、紙面のものや電子的なものがある。そこで、紙面よりも電子的なものの方がデータ量が多い点からデータセットには電子版の新聞を使用することとした。いくつか候補が挙げられたが、我々は馬三郎 [2] という競馬新聞に注目した。ここで、著作権の問題を考慮しメールで問い合わせることでデータの使用の許可をとった。また、同じく権利の問題で電子版の新聞はコピー＆ペーストができないという問題が発生した。そこで、文字認識のアプリケーションを用いデータの収集を行うこととした。

前期ではグループ内でデータ班と実装班の 2 つに分かれて活動を行った。後期は当時の到達度からアプリケーションの実装ができないと判断したため改良したモデルを最終成果物として活動を行った。まず、モデルの実装とデータ収集を行った。データは 2019 年 1 月〜5 月分のものを使用した。最初に、着順と人気順から好走、普通、凡走のラベルを作成し多クラス分類を行うモデルの作成を行った。しかし、人気を入力に含めた場合、判断の際に人気の影響が強くなってしまった。最終的に、競馬では 1〜3 着が重要となることに着目し 3 着ごとのラベルを作成し着順のクラスを予測するモデルを作成した。

キーワード　ディープラーニング，自然言語処理，競馬

（文責: 武田佑樹）

Abstract

In recent years, the development of AI has been progressing, and in this context, technology using deep learning has been applied in various fields. Therefore, our group focused on horse racing predictions for regional and other reasons. Expected payout in prediction of horse racing is the product of the payout which called odds and the probability of winning. We calculate expected value and decide tickets to buy. The problem is that odds are formed by collective intelligence, and you cannot make a profit unless the predictions are higher than the collective intelligence. Collective intelligence is formed when many people make predictions, so it is necessary to incorporate elements that are not included in the collective intelligence in order to surpass it. We focused on trainer's comments as one of them, and we used natural language processing to analyze them. As a method, we applied morphological analysis to the trainer's comments and used LSTM to make predictions. The morphological analysis was performed using Mecab[1], a morphological analysis engine, to perform word segmentation. Word2Vec was used for word embedding to represent the words as vectors of predetermined dimensions with real number.

In the dataset, we need the trainer's comment and the data of order of finish. Therefore, we considered to collect data from horse racing newspapers. There are many types of horse racing newspapers, printed and some of which are electronic. Among them, we decided to use the electronic version of the newspaper because there are more data in the electronic version than in the paper. Among several candidates, we decided to use a horse racing newspaper called Umasaburo[2]. Here, we asked the publishers for permission to use the data by emailing them due to copyright issues. Also, due to the same rights issue, the electronic version of the newspaper could not be copied and pasted. Therefore, we decided to collect text data by using a character recognition application.

In the first semester, we were divided into two groups, the data group and the implementation group. In the second semester, we decided that we could not implement the application based on the progress at that time, so we improved the model. We implemented the model and collected data. Data from the period January-May 2019 was used. First step was to create a model to predict the order of arrival with the data of trainer's comments. Next, we created a model for classification by setting good, normal, and bad labels using order of arrival and popularity order. However, when the popularity order was used, the results were strongly influenced by the popularity order. Finally, focusing on the fact that the first three places are important in horse racing, we created a model to predict the order of arrival for each of the three places.

Keyword Deep Learning, Natural Language Processing, Horse Racing

（文責: 鈴木健斗）

目次

第 1 章　はじめに

1.1　背景

近年、機械学習技術が急速な発展を見せている。特にディープラーニングは目覚ましい成果をあげており、大きな注目を集めている。発展の背景としてはハードウェアやソフトウェアの進歩、インターネットの普及等により、大規模なデータセットを扱えるようになったことが考えられる。現在、競馬予想分野では勾配ブースティングを採用したモデルが主流となっている [3]。しかし、ディープラーニングが多岐に渡り応用される中で、競馬予想においてもその活用が期待される。

1.2　目的

中央競馬では、掛け金の総額から一定割合を控除した額を的中者全員に配分する仕組みを採用している。したがって、払戻金の倍率（オッズ）は (1 – 控除率)/得票率 となる。これより、オッズが集合知によって形成されることがわかる。利益を目的とした競馬予想では、回収率の期待値（オッズと勝利確率の積）が 100% を超える馬券を予想することになる。そのためには、全馬券の勝利確率をできる限り正確に算出する必要がある。単勝における統計的性質に、オッズ別の勝率が得票率に収束するというものがある [4]。これは、長期的な回収率が 還元率 $= 1 -$ 控除率 に収束することを意味している。利益を目的とする場合、回収率は少なくとも還元率という統計的なベースラインを超える必要がある。したがって、回収率が 100% を超える AI は、オッズを形成する集合知を上回っていなければならない。

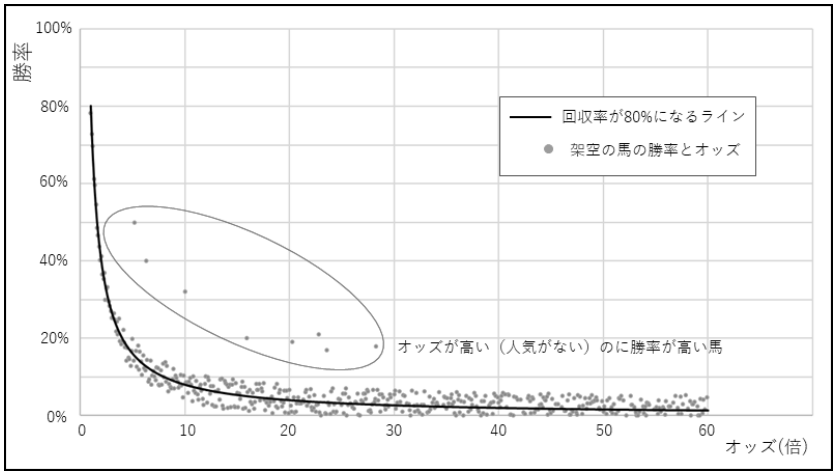

図 1.1　勝利確率とオッズの関係のイメージ

図 1.1 は、勝利確率とオッズの関係のイメージである。中央競馬の単勝及び複勝における控除率は 20% であるため、長期的な回収率は 80% に収束することになる。点の多くは回収率が 80% になるラインの付近に集中しているが、赤線で囲まれた点はラインから逸脱している。これらの点は、オッズと乖離して勝率が高い馬を意味している。このような馬の勝利確率を正確に見積もり、購入行動に反映することができれば、回収率は向上すると考えられる。集合知を上回る方法として、寺沢（2019）は、集合知を形成する集団があまり注目していないデータに可能性があると述べ

ている [4]。そこで、我々は競馬新聞に提供される厩舎コメントに注目した。厩舎コメントが機械学習で利用された先例はあまり見られない。また、競馬初心者にとって、厩舎コメントから厩舎関係者の真意を汲み取ることは難しいと考えられる。本グループでは、厩舎コメントと着順の関係性の発見を目指した。

1.3 現状における問題点

厩舎コメントでは、レースに際して、競走馬の最新の体調や厩舎関係者の意気込みが二文ほどの長さで紹介される。調教師を始めとする厩舎関係者は、立場上、競走馬の不調をある程度の建て前で取り繕うと推察される。人間の目による経験則では、競走馬が真に好調かどうかを常に見極めるのは困難であるという問題がある。しかしながら、ディープラーニングを用いた自然言語処理によって、人間の目では判断できない特徴を判断し、競馬予想における厩舎コメントの有用性を発見できる可能性がある。

（文責: 新大実）

第 2 章　課題設定

2.1　課題の設定

本プロジェクトで競馬に関する活動をすることになった理由として、函館には競馬場があり地域との関連性もあることや担当の教員が詳しいことなどから、競馬がテーマとなった。我々のグループでは、競馬新聞の調教師のコメントを読み取り、自然言語処理を用いて競馬予測との関係性の分析を行う。ここで必要なのがコメント分析を行うプログラムとデータセットである。プログラムはディープラーニングを用い、自然言語処理でコメントと着順のデータを分析し、予測を行う。開発言語は Python を用いることとした。データセットには、過去の競馬新聞約 1 年分の調教師のコメントと着順のデータが必要となり、バックナンバーを閲覧できる競馬新聞と新聞の文字を認識するソフトを用意する必要がある。したがって、本プロジェクトの課題はコメントの分析を行うプログラムの作成と調教師のコメントのデータセットの作成である。

2.2　前期目標

プロジェクトの活動にあたって、最初にグループ全体でディープラーニングと競馬に関するメンバー全員の基本的な知識習得と共有を目標とした。その後、どのような形で競馬予測をするのかを定めていなかったため、習得した知識をもとに研究に関する具体的なテーマの決定を目標とした。テーマの決定後、グループ内で実装班とデータ収集班に分かれて活動を行うこととなった。実装班はまず、自然言語処理に関する知識習得を目標とし、その後プロトタイプの作成を目標とした。データ班では、著作権や入手のしやすさの問題からデータ作成に使える新聞がなかなか見つからなかったため、データに用いるバックナンバーの閲覧ができる競馬新聞の決定を目標とした。その後、新聞の約一年分のデータからデータセットの作成を目標とした。

2.3　後期目標

後期の目標としてまず、実装したプロトタイプと作成したデータセットから分析を行うことが挙がった。その後、有意な結果が得られた場合、アプリケーションとして実装することが目標となる。有意な結果が得られなかった場合、アプローチやデータの変更などの分析方法の変更が必要になるため、目標が変わる場合がある。

2.4　担当割り当て

まず、グループ内でプログラムの実装を行う班とデータ収集とデータセットの作成を行う班の 2 つに分け、各人の希望する担当を調べた。その結果、人数が均等になったため割り当てを行った結果、川崎、鈴木、新が実装班となり、武田と藤原がデータ班となった。

（文責: 武田佑樹）

第 3 章　活動内容

3.1　前期の活動

3.1.1　勉強会

前期の前半で行われた勉強会では、参考書 (「ゼロから作る Deep Learning」[5]) を用いて Deep Learning の仕組み・原理を実際に実装しながら学んだ。事前に参考書の章を 1 つずつ各自自習し、プロジェクト学習の時間内に担当している章の主要な部分の解説を行って理解を深めた。また、競馬について知識を深めるためオンラインでの観戦やスマートフォン用の競馬予想ゲームアプリをプレイするなどを行った。

3.1.2　テーマの決定

最初に、どのような要件の人工知能を作成するのか検討するためブレインストーミングを用いて案を出し合った。その結果、画像から馬の体調を予測してレースの結果にどのような影響を与えているかの分析を実行しようと考えた。しかし先行研究の調査として過去の卒業研究である 2 つの論文（パドック画像を用いた着差に基づく競走馬分類 [6]・馬体画像を用いた競走馬の着順分類 [7]）を読んだ結果、先行研究の成果が芳しくなかった事と我々が先行研究の結果を基にそれ以上の成果を出す見通しが立たなかった。また新型コロナウイルスの影響があり競馬場実地での動作確認が出来るか不透明だったため、画像から分析を行う事を諦めた。その次に、調教師のコメントを分析してみることを検討した。調教師は競走馬の調教が上手く行かなかったときに、立場上馬主の目があり良いことしか書けないので、本当に上手く行ったときと実際は上手く行っていない時の微妙なコメントの変化をディープラーニングで分析してみると面白いのではないかと考えた。そこで最終的なテーマを「調教師のコメントを自然言語処理で分析する」に決定した。

（文責: 鈴木健斗）

3.1.3　実装班の活動

3.1.3.1　勉強会

実装班では自然言語処理の実装を行うための技術的な習熟度が十分ではなかったので、最初に自然言語処理と Python の機械学習フレームワークを扱った参考書を用いて輪読を行うこととした。「Python と Keras によるディープラーニング」[8] を用いて 6 章の自然言語処理の仕組みと実装例を学び、実装班内で専門用語の解釈の確認を行った。そして参考書には含まれていなかった日本語の形態素解析の手法などを Web などから情報収集し、単語埋め込みへの変換も含めて実装した。

形態素解析器は Mecab[*1]、単語埋め込みは学習済みの Word2Vec[*2]を利用した。それらのプログラムをメンバー間で共有し、それぞれの環境で動かせることを確認した。

3.1.3.2 形態素解析器を用いた分かち書きの実装

「Python と Keras によるディープラーニング」では英文の自然言語処理を行っていたため、単語間の空白で分かち書きを行うことができる。しかし日本語は単語に分ける際に辞書などを用いて分割する処理が必要になる。実装班では形態素解析器 Mecab を用いて単語の分かち書きを実装した。Mecab は高速で分かち書きを行うことができ、比較的容易に利用することができるので利用することとした。

3.1.3.3 学習済みの Embedding 層の利用

分かち書きを行った単語をニューラルネットワークに入力するには、数値表現に変換する必要がある。数値表現には単語埋め込みと呼ばれる実数で 300 次元ほどのベクトルで単語を表現する方法を用いた。数値表現は実際のテキストから学習する必要があるが、学習は計算コストが大きく計算能力が高い計算機が必要であるため、実装班では学習済みの数値表現を利用することとした。日本語 Wikipedia で学習させた Word2Vec を利用した。

3.1.3.4 プロトタイプの実装

モデルの構築には Keras[9] を利用した。実際に我々が行う自然言語処理のタスクと類似した、ライブドアのニュース記事の分類を行った。データセットは NHN Japan 株式会社が運営する「livedoor ニュース」から HTML タグを取り除いて作成された、9 つのニュースカテゴリに分けられたテキストファイルが提供されている [10]。それを利用し、LSTM モデルを用いて「独女通信」、「IT ライフハック」、「家電チャンネル」の 3 種類のテキスト分類を行った。モデルは入力層に Embedding 層、隠れ層に LSTM、25 層の全結合層、出力層に活性化関数が sigmoid 関数の 3 層の全結合層を利用した。ニュース記事ごとのデータを 500 個用意し、Word2Vec が対応していない未知語はカットしている。最大入力単語数を 500 単語でバッチ化しバッチサイズ 32 の 20 エポック、スプリットを 0.2 で学習させた。学習データの正解率は 1.0、検証データでの正解率は 0.9267 となった。ランダムに文章分類を行った場合、正解率は 1/3 程度になるので、高い精度で予測を行うことができた。作成したニューラルネットワークのプロトタイプはデータがどの記事であるかの確率を出力する。そして最も高い確率のクラスをニューラルネットワークが判断したクラスとする。ここで正解率は正しく分類できたデータ数を全データ数で割った値である。

（文責: 川崎景大）

[*1] 京都大学情報学研究科－日本電信電話株式会社コミュニケーション科学基礎研究所 共同研究ユニットプロジェクトを通じて開発されたオープンソース 形態素解析エンジンである。日本語の文章を単語ごとに分割することができる。

[*2] 単語を入力するとその単語埋め込み表現 (分散表現) が得られる手法である。単語を実数値であらかじめ決められた次元のベクトルで表現できる。

3.1.4 データ班の活動

3.1.4.1 データ班の目的

データ班の活動の主な目的は調教師コメントと着順のデータ収集である。はじめに調教師コメントをどこから収集するか検討するところから始めた。

3.1.4.2 データ収集の方法

調教師コメントを収集する候補として競馬新聞や競馬新聞と提携しているアプリケーションなどが挙げられる。また、調教師コメントには競馬新聞ごとに特徴があり、どれから調教師コメントを収集するかで結果が変わってくるのではないかと考えた。そこでデータ班としては、コメントのわかりやすさとデータ収集のしやすさの 2 点に重きを置いて候補を決めることにした。ここでコメントのわかりやすさとは自分たちが実際に競馬新聞のコメントを読んでみて感じた感想である。競馬ブック web、競友、競馬ニホンなどが候補として挙げられたが、最終的には電子版の馬三郎 [2] からデータを収集することにした。理由として 3 つ挙げられる。1 つ目は、2007 年までの競馬新聞のバックナンバーが存在することである。他の競馬新聞ではバックナンバーがあまり存在せず、バックナンバーがあっても空白の期間が存在することから、バックナンバーが豊富にあることはデータ収集のしやすさの観点から見ても重要であった。2 つ目がコストを低額にできることである。電子版の馬三郎では月額ごとに利用料を支払う必要がある。しかし、それを支払ってしまえば、2007 年までのコメント、着順データを収集できることから競馬新聞を購入するよりコストを抑えることができると考えている。3 つ目が文字認識のしやすさである。競馬新聞を選択する段階では、実際に競馬新聞を購入してコメントデータを作成する方法と電子化されている情報から自分たちの必要な部分を抜き出して作成する 2 つの方法があった。本来は考えるまでもなく、電子化されているものからデータを作成する方が労力的にも楽である。しかし、今回自分たちの作成しようとしているコメントデータは新聞社特有のデータであり、著作権などの観点からプログラムでの自動収集やコピー&ペーストなどができないようになっていることが多い。その結果、一文字一文字を手打ちで作成する必要が出てくる。そこでデータ作成を少しでも簡単にするために文字認識を利用しようというアイディアが出た。文字認識を利用することを考えると、現実で新聞のコメントごとに文字認識をするよりは電子化された文字に対して文字認識をした方が精度が良くなると考えている。

3.1.4.3 著作権関連について

ここでコメントデータに対する著作権の問題が出てきたことがわかる。調査してみたところ、日本の法律では機械学習におけるデータ作成に関して、さまざまな条件は存在するが著作権を気にすることなく作成できることがわかった。機械学習について関わる部分としては著作権法の第十一条、第十二条、第十二条の二、第三十条の四第二号、第四十七条の七である。これらの中でデータセットの著作権とデータの著作権が別物であることや情報解析としてデータを使う分には著作権を意識しなくてもよいということが述べられている。今回の場合も、法律の適用範囲である。ただ、気をつけなければいけない点が存在する。自分たちは競馬新聞と提携しているアプリケーションからデータ収集をすることになるので、それを利用する際には利用規約に同意する必要がある。機械学習におけるデータセットの作成は著作権を気にしなくても良くなると書いたが、今回は利用規約の方に違反する恐れがある。そこで一度サービス提供元にメールで今回のプロジェクトの主旨を伝え、利用規約に抵触する恐れはあるか問い合わせることにした。その結果、許可を得ることができ

たので利用規約に違反しないようにデータ収集を行っていく。

3.1.4.4　後期の活動

実装したプロトタイプと作成したデータセットから分析を行っていく。有意な結果が得られた場合、アプリケーションとして実装することが目標となる。有意な結果が得られなかった場合はアプローチやデータの変更などの分析方法を変えていく必要があるため、目標が変わる場合がある。夏休みと後期を通して電子版の馬三郎からコメントと着順のデータセットを作成することを目的に活動していく。

（文責: 藤原慎太郎）

3.2　後期の活動

3.2.1　実装班の活動

厩舎コメントデータを入力として、着順の予測を行うモデルを作成した。このモデルは前期に作成したプロトタイプを用いた。前期のデータ班と並行して作業を行っていたため、一番初めのデータセットは 2019 年 1 月から 3 月の 3 ヶ月分のコメントと着順のデータで着順の予測を行った。このモデルは 1 着〜18 着までの分類を行っていたため、データ量が分散されていたことやコメントだけで細かく 1 着毎に判断出来る問題ではないため、正しく機能していなかった。

それから着順と人気順のデータを用い、馬の走りの評価として 1 着〜5 着かつ人気よりも高い順位でゴールした馬は好走、人気とあまり変わらない順位であれば普通、人気よりも著しく低い順位であれば凡走といった形で三つにラベル付けを行った。詳細なラベル付けは第 5 章にて示す。このラベル付けを行ったモデルではデータ数に偏りがあり、好走と凡走のデータ数が少なく普通のデータ数が多い状態であった。実装班のみの活動ではこの問題を解決することは出来なかったが、精度の向上のために JRA により提供されているうまやことば [11] の厩舎関係用語を Mecab のユーザ辞書に追加を行ったり、コメントデータの頻出単語の上位 500 単語のみを残す前処理をしてから学習を行った。この作業によってモデルの精度が劇的に向上する事はなかったが、プロトタイプのモデルを厩舎コメントに適用できるよう改良することが出来た。

（文責: 鈴木健斗）

3.2.2　データ班の活動

当初は馬三郎の中ですべての対象データを収集してデータセットを作成しようとしていたが、文字認識の精度があまり良くなかったので、作業時間が大幅に増加していた。そこで馬三郎でしか得ることができない情報である厩舎コメントのみを文字認識を用いて収集し、着順などのデータは netkeiba.com[12] のサイトをスクレイピングすることで収集することにした。収集したデータごとにデータ順が異なっていたが、特定の項目によって昇順、または降順を基準としたデータ順序になっていたので作業を簡略化させるためにエクセルマクロの作成も行った。エクセルマクロが完成した後、対応付けの作業やデータ修正の作業に入った。データセット作成に関する具体的な手法は 5 章で述べる。

（文責: 藤原慎太郎）

3.2.3 全体での活動

班ごとに行う活動が終わり、全体で最終的なまとめを行う活動に移った。

3.2.3.1 モデルの改良

まず、当時の到達度から目標であった厩舎コメントから馬の調子を予測するアプリの開発を行えないと判断したため、構築したモデルを最終成果物とした。しかし、実装班のみで作成していたモデルでは予測精度も低く上手く動作もしていなかったため、全体で話し合いアイディアを出し合った。中でも実現出来た内容は以下のとおりである。

- 表記ゆれ対策
- 不均衡データへの対処
- データセットから障害レースを抜いたデータで学習

表記ゆれへの対策というのは、「メンバー」と「メンバ」などの表記が違う同じ意味の単語を統一出来る同義語辞書を用いた。本来は競馬用語で同じ意味の単語を統一するためのアイディアだったが、競馬用語について同義であるかどうか判断しづらいものが多かったため競馬用語用に同義語辞書を作成することは断念した。不均衡データへの対処というのは、好走、普通、凡走での 3 分類を行ったとき普通のデータが著しく多く、モデルの予測でほとんどの分類を普通にしてしまっていた問題への対処である。この問題に対してはアンダーサンプリングというデータ数が少ない方のデータにデータ数を合わせて学習を行う手法を取った。アンダーサンプリングを適用するために Python に対応している imbalanced-learn というライブラリを用いた。また、障害レースに出場する馬のコメントは「飛越」などの通常のレースとは違う単語が使われていることが多く、通常のレースの予測にはあまり関係が無いと判断したため、作成済みのデータセットから障害レースのデータを除去する作業を行った。

（文責: 鈴木健斗）

第 4 章　課題解決のプロセスの詳細

4.1　前期の担当課題と他の課題の連携内容

4.1.1　川崎景大

参考書を輪読し、一部の章の解説を行った。並行して未来大学の過去の卒業論文や指導教員の論文を参考にテーマの検討も行った。実装班に異動し、自然言語処理と Python のフレームワークを解説している参考書を輪読した。そして実装班内で参考書内の用語の確認を行った。参考書の輪読を踏まえ、Mecab[1] による文章の分かち書きと Word2Vec による単語埋め込み表現への変換を行うプログラムを作成した。

（文責: 川崎景大）

4.1.2　藤原慎太郎

前期は主に事前調査やディープラーニングの学習、データ班での活動を行った。まず、事前調査ではテーマの検討を行う前に現在はどのようにディープラーニングが活用されているのかをインターネットや論文を通して調べた。他にも競馬というテーマを決めたときに今まで触れたことがなかったので知識習得として事前に調査をした。ディープラーニングの学習では、「ゼロから作る Deep Learning」[5] という本を用いて輪読を行ったので、担当の章を他の人に説明する必要がある。そこで資料作りを行った。担当する章では他の人の質問を受け付けるので、その章の内容を理解して説明の場に臨めるようにネットで同じ内容を勉強するなどを行った。データ班の活動では、さまざまな種類の競馬新聞をインターネットを通じてサンプルを入手することで、競馬新聞の候補を選択するのに役立てた。また、著作権関連の調査も行った。著作権に関連して馬三郎 [2] のサービス提供元に利用規約についてメールで質問を行うことや馬三郎と文字認識の購入のためのお金の申請を各先生方に伺った。提供元にメールをするときは寺沢先生に質問内容を確認していただいた。後期からは馬三郎と文字認識を用いてデータセットを作成していく。

（文責: 藤原慎太郎）

4.1.3　武田佑樹

前期ではディープラーニングについての知識習得、中間発表、データ班での活動を主に行った。ディープラーニングの知識習得では参考書の輪読を行い、2 章分を担当し、内容の解説を行った。中間発表ではプロジェクト紹介文の担当になったため、グループの紹介文作成、評価担当の割り振りを行った。データ班での活動では競馬新聞や文字認識アプリの検索をし、データセットの作成を進めている。

（文責: 武田佑樹）

4.1.4　鈴木健斗

前期で行ったことは準備段階となっており、ディープラーニングの学習やその他事前知識の習得が出来た。ディープラーニングの学習は参考書の「ゼロから作る Deep Learning」[5] の輪読という形で行い、各メンバーが章ごとに内容をまとめてきて説明することで学習した。また、実装班でフレームワークを用いた実践的なプログラムの書き方を参考書の「Python と Keras によるディープラーニング」[8] にて学習した。その他事前知識は主に競馬や AI による競馬予想についての知識であり、元々競馬について詳しいメンバーが居なかったため競馬観戦から始め、AI による競馬予想に関する論文を読み先行研究の調査を行った。後期では実装班でデータセットに合わせたプログラムの実装を行っていく。

（文責: 鈴木健斗）

4.1.5　新大実

前期は輪読、中間発表、プログラム作成が主な活動となった。輪読の担当箇所を読み込み、資料を作成、解説した。また、中間発表準備ではポスター制作班に参加し、プロジェクトの概要を執筆した。発表当日は発表者として事前に作成したスライドを用いて発表を行った。夏休み中には自然言語処理の勉強会に参加し、簡単なプログラムを Python で書いた。

（文責: 新大実）

4.2　後期の担当課題と他の課題の連携内容

4.2.1　川崎景大

JRA から提供されている厩舎関係用語 [11] をもとに Mecab でユーザー辞書の登録を行った。また好走、普通、凡走の予測モデルの学習で使用するアンダーサンプリング手法について調査をした。調査した手法の中で One-sided selection とランダムアンダーサンプリングを適用するプログラムをそれぞれ作成し、最終的にランダムアンダーサンプリングを適用することに決定した。このときの問題点として、Python のライブラリーを利用していたが、公式ドキュメントや論文での調査が足りなく、ランダムアンダーサンプリング以外の手法を正しく適用できなかったことが挙げられる。また他のメンバーが作成した好走、普通、凡走の予測モデルのプログラムを変更し、着順クラスの予測を行うプログラムの作成を行った。メンバーがラベルの作成関数やモデルのパラメータ設定などを一般化してくれていたため、スムーズに変更を行うことができた。この経験により、グループ活動でのコーディング作業で統一性をもって開発することの大切さを学ぶことができた。他に担当した課題として作成したモデルのテストデータでの再現率と適合率を計算するプログラムを作成した。しかし再現率と適合率を正しく理解していなかったため、メンバーや教員に指摘されるまで間違った指標でモデルの評価を行っていた。他に担当した課題として輪読した本の内容をもとに、LSTM のスタッキングで性能が改善するか調査した。成果発表の Web サイト作成では実装班で実施した手法とその結果について表や図を用いてまとめた。前期と後期を通して、自分は機械学習に関する勉強が不足している状態で進めていたことが問題点としてあったため、今後は理論的

な背景をしっかりと検討したうえで研究活動などを行う必要があると考えられる。以上より、プロジェクト学習では多くの貴重な経験と学びを得ることができた。これらの経験をこれからの活動で活かしていきたいと考えている。

（文責: 川崎景大）

4.2.2 藤原慎太郎

後期はデータ収集、マクロの作成、データの修正などのデータ班としての活動のほか、全体では作成したモデルに対する評価などを行った。主に時間を費やした作業はデータ収集であるが、予定通りの作業とはいかずに工夫しながら作業を進めていく必要があった。馬三郎を用いてのデータ収集であり、厩舎コメントなどのデータはコピー&ペーストができないことから文字認識を用いて収集する予定であった。しかし、実際に文字認識を用いてデータを収集してみると、文字認識の精度があまり良くなかったという結果であったため、文字修正にかかる時間や文字サイズを大きくして精度を上げなければいけなかったので、1 度に文字認識できる範囲を狭める必要もあり、作業時間は想定より大幅にかかることになった。この現状のまま、作業を進めていくとデータ収集できる量も減少していくと予想できるので作業を効率化できるところは効率化するという意識を持つ必要があった。そこで、エクセルマクロの活用、スクレイピングで収集できるデータは別の対象から収集してくるなどの対策をすることにした。エクセルマクロ、スクレイピングのどちらも初めて触れる分野であったが、便利なツールであることは知っていたため、この機会に勉強することにした。エクセルマクロでは収集した厩舎コメントの文字を一括で修正するマクロ、収集したそれぞれのデータを対応付けさせるためにある項目を基準にレースごとに昇順や降順に変更させるマクロ、特定の文字を消去させるマクロの 3 つを作成した。エクセルマクロの作成に時間がかかってしまったため、スクレイピングについて勉強する時間はあまりなかった。そこで既存の製品であり、スクレイピングする作業をプログラムを打つのではなく視覚的に作業することができる Octoparse[13] を用いることにした。これらの対策を提案したことで、作業にかかる時間は大幅に減少したと考える。データ班としての活動を通して、作業を効率化させる方法や班員同士の役割分担、コミュニケーションの仕方などさまざまなことを学ぶことができた。今回のプロジェクト学習で学んだことをこの機会で終わりにせず、卒業研究や将来携わるであろうプロジェクト、日頃の生活から有効に活用していきたい。

（文責: 藤原慎太郎）

4.2.3 武田佑樹

後期ではデータ班での活動、成果発表を行った。データ班での活動は主にデータセットの作成となった。担当区分は 2019 年 1 月〜5 月前半の担当となった。最初は 1 月の担当となり、文字認識ソフトを用いたコメントデータと着順、人気データ収集を手動で行っていたが、文字認識ソフトの誤字の多さから作業量が多いのに加えて時間がかかっていた。文字の表示を拡大すれば認識の精度が上がるため文章全体が映る程度に競馬新聞アプリのウィンドウサイズに合わせて拡大して文字認識することで誤字を抑えていたが、マクロを使用して文字置換を行うことになったため、作業時間の短縮ができた。置換する単語のリストはデータ班のメンバーがあらかじめメモにとっておいた単語を合わせて使用した。プログラムは変換する単語のリストに単語を追加していけばその分まで置

換するようにしたため新たな誤字を発見していくたびに追加しており、後半のコメント直しをする手間が少なくなった。誤字修正の他にも着順と人気のデータをとるのに時間がかかっていた。着順と人気のデータは文字認識でとろうとすると拡大がしにくく、張り付ける際に並び替えを行わなければならないため、スクレイピングツールを使うこととなった。着順、人気のデータは Web サイトで収集できるため、スクレイピングツールで抽出することで作業の簡略化ができた。しかし、スクレイピングツールで出力したデータの順番がコメントデータの記載されている順番と違うことがわかった。そのためスクレイピングツールが出力したデータの順番をコメントデータの順番に合わせるマクロを用いて対応付けが可能になった。データ収集が終わった後はデータセットの修正を行った。データセットの 1 月と 2 月分はすべての作業を手動で行っており、また量が多かったため作業に慣れておらず誤字が多くみられた。3 月分に関しても 5 月分で新たに見つかった誤字の修正が必要なため行った。成果発表会の準備では、Web サイトの記述、修正を行った。Web サイトではデータ班の活動内容の修正、手法の検討の図の修正を行った。成果発表会では中間発表と同じく質疑応答を担当した。今回のデータ班での活動では 2 人という少ない人数でより多くのデータを収集するということが必要だった。さらに、著作権やアプリの形式などの問題上、簡単に集める方法が見つからないため大量のデータセットを手作業で作成するために効率的な作業を行う工夫を行ってきた。より効率的な方法や精度の良い文字認識ツールがあれば 5 ヵ月分以上のデータを収集することができたかもしれないが、これらの作業で用いた技術を他の場面でも生かせるようにしていきたい。今回のプロジェクト学習を通じて作業の効率化をするための方法やグループ活動を計画的に進める方法が学べた。

（文責: 武田佑樹）

4.2.4 鈴木健斗

後期は実装班の中で主にモデルの精度の向上についての調査と話し合いの進行を行っていた。作業は手法についての検討を行う事と実装に取り組む事が多かったため、モデルに対しての理解には苦労したがディープラーニングのアプローチに関しては多くの学びがあった。また、他の実装班のメンバーが理解しやすいようにプログラムを書いてくれていたことがモデルの理解のしやすさの一因であった。モデルに多く変化を加えてからは作業量の偏りと効率を考えた上で、精度を向上するための手法を調べてきて班員に提案することを多く行い、タスク管理を担当することで円滑に作業をすることが出来た。しかし、不均衡なデータに対しての配慮不足によるミスや、実施した手法の中でも表記ゆれの対応等の効果の検証がしづらい実装を行っていた事によって時間を大きく使ってしまった。グループでのプログラムの実装に用いた GitHub をグループメンバーが使い慣れていなかった事や、オンラインの作業のため進捗の確認や教え合う事がしづらい環境であった事が要因であると考えている。そのため、グループで実際に集まる事は作業を大きく活性化させる事がわかった。反省を活かして今後オンラインで作業しなければならない時はビデオ会議で行う事や、より教え合いやすいツールを用いるべきである。全体の作業では主に実装班で纏まりきらなかった結果についての方針について話し合った。この時データ班の新しい意見に助けられたため、会議での第三者視点は重要であった。また、成果発表会に向けての調整や Web サイトの記述やグラフの作成を行った。Web サイトの作成は最終成果物の作成と同時進行であったため、グループメンバーのスケジュール管理が難しかった。また、この日までにこの作業を終わらせれば良い等の指標も立てないまま活動していたため、計画段階で終わりまでの見通しを立てて活動するべきであった。本プロ

ジェクトの活動でグループ毎に分かれて活動した時の連携の取り方や、小グループを作った時のコミュニケーションの取り方やスムーズに役割分担をする方法を学ぶことが出来たため、今後活かしていきたい。

（文責: 鈴木健斗）

4.2.5 新大実

前期に作成した livedoor ニュースコーパス多クラス分類のプログラムを、厩舎コメントに対応できるように変更した。データセットの読み込みやテキストクリーニング、分かち書きのメソッドの修正を行った。また、Sudachi 同義語辞書 [14] を用いた表記ゆれの検出及び統一、頻出語抽出のメソッドを作成した。全単語の出現率や重要度などの情報を xlsx ファイルとして出力した。その結果、文書数のばらつきを鑑みても、走るや距離といった語は全体として頻度に差が見られず、逆に上位やスムーズといった単語は上位に比較的集中しており、前身や悪いといった単語は上位では見られにくいという傾向がわかった。

プログラムの共有等を想定して Goolgle Colab 用の環境を整えた。Google Colab は使用しなかったが、GitHub を導入してソースコードを共有できるようした。しかし、GitHub の使用にもやや不慣れな点があり、今後の使用を考えても理解を深めるべきだと思われた。Mecab 辞書やアンダーサンプリング、再現率・適合率のプログラムについては、他のグループメンバーに一任してしまい、ほとんど触る機会がなかったのは反省点だと考えている。また、深層学習や自然言語処理、競馬に関してなど、全体的に知識の不足を実感し、学習の時間が足りていなかったことについても深く反省している。最終成果発表会に際して、当日の発表用スライド及び発表用原稿を作成した。当日は発表及び一部の質疑応答に携わった。

本科目を通じて得た経験は通常の授業では決して得ることのできない大変貴重なものだったと思う。遠隔の共同作業の機会はこれからさらに増えていくのではないだろうか。今後の研究活動でも役立てていきたい。

（文責: 新大実）

第 5 章　モデルの構築

本グループでは、プロジェクトが全体として掲げる「ディープラーニングなどの最新技術の活用によるシステムの構築」という目的に基づいて、ニューラルネットワーク（LSTM）を用いた競馬予想を行った。モデルの構築にあたって『ゼロから作る Deep Learning』[5]『Python と Keras によるディープラーニング』[8] の輪読を実施し、必要な知識を習得、共有した。本章では、本グループが構築したニューラルネットワークモデルについて、我々が利用した様々な関連技術と共に紹介する。

5.1　モデルの概要

ここでは我々が構築したモデルの概要を説明する。本モデルは厩舎コメントを入力として、競馬予想を行うニューラルネットワークモデルである。モデルは埋め込み層、LSTM 層、ドロップアウト層、全結合層から構成される隠れ層を持つ。モデルの詳細は以降の節を参照されたい。プログラムの実装は Python で行った。モデルの構築には Keras[9] を用いた。Keras は、Python で書かれた高水準のニューラルネットワークライブラリである。

（文責: 新大実）

5.2　データセットの作成

電子版の競馬新聞アプリである馬三郎 [2] のバックナンバー機能を用いてコメント収集を行った。データの範囲は 2019 年 1 月〜5 月であった。レースは土、日、祝日に行われ、1 日のレース数は 24 または 36 レース程、1 レース当たりの馬の数は 5〜18 頭となっており 1〜5 月全体のコメントデータ数は約 20,000 データとなった。データは Excel に馬ごとに対応するコメントデータと着順、人気をとっていった。コメント収集の際、アプリの形式としてプログラムでの自動収集やコピー＆ペーストができないようになっていたため、瞬間テキスト [15] という文字認識ソフトを用いて収集を行った。このソフトは起動して読み取りたい文字の書かれている範囲を選択し範囲内の文字を抽出して文章としてコピーできるものである。しかし、認識の精度があまり良いものではなく、アプリのコメントの部分を拡大して文字認識を行うと多少の改善が見られたが、ウィンドウサイズの関係で拡大できる範囲に限りがあり、誤字が多く見られた。したがって、文字修正の必要があったため収集したコメントを一度確認する必要があった。1 月〜2 月のコメント収集を行っていた際は手作業で誤字をした部分の修正を行っていたがその分手間がかかり速度が遅くなってしまう問題が出てきた。誤字が発生する部分はある特定の単語やパターンが多かった。そこで Excel で誤字をした部分の一括置換を行うマクロを組むことで文字修正の手間を省くことを考えた。そのため、誤字を手作業で直す際に修正した箇所のメモをとりマクロで置換する部分のリストの作成を行った。文字修正のマクロはコメントデータを読み取っていき、リストに書かれている誤字の部分を修正する部分に書き換えるものにした。文字修正のマクロを作成した後は文字認識したコメントをそのまま張り付けてマクロを起動し、修正した後の文章で新たな誤字が見つかった場合、その都度修正する単語のリストに追加していくことで手間を少なくしていった。しかし、似た単語同士の

間違いなど、置換すると他の文章に影響してしまう部分は手作業でそのまま行っていった。また、文字修正の部分はある程度自動化したが、着順と人気のデータは手動でとっていた。アプリの形式上、着順と人気のデータはコメントとは別のページにありその分の手間が多かった。ここで、コメントデータは競馬新聞でしか取れないが、人気と着順のデータは Web サイトで取れることを利用しスクレイピングで収集することとした。スクレイピングには Octoparse[13] というツールを使用し、収集するサイトには netkeiba.com[12] を利用した。Octoparse では 1 日ごとのページの URL を入力していき、その日のすべてのレースから着順、人気を抽出し Excel に出力した。しかし、馬三郎のコメントデータは馬番順で並んでおり、netkeiba.com のデータは着順ごとに並んでいるためそのまま対応付けることができない問題が出てきた。そのため、スクレイピングした人気と着順のデータを馬番ごとに並び替えるマクロを作成することでコメントデータを対応付けできるようにした。以上のように、コメントデータはアプリから文字認識ソフトを用いてコピーし、文字置換と手動で誤字修正を行ったあとスクレイピングで抽出した着順、人気データを並び替えコメントデータとの対応付けをするという方法でデータセットの作成を行った。最終的なデータセットの構成を図 5.1 に示す。

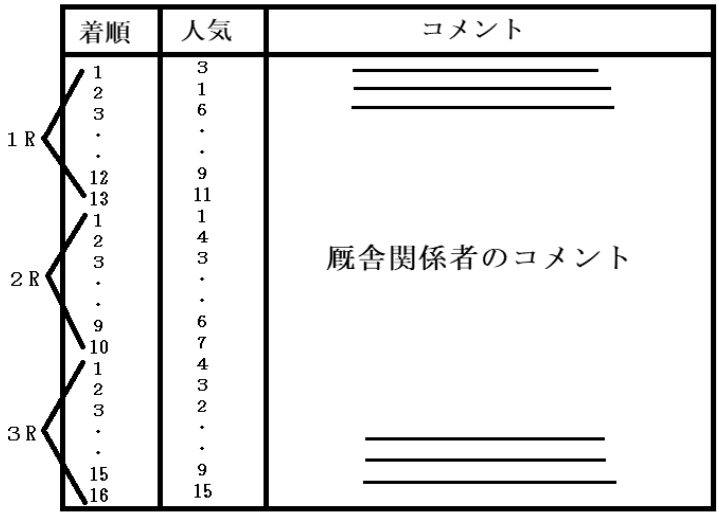

図 5.1　最終的なデータセットの構成

（文責: 武田佑樹）

5.3　モデルの実装

上述の通り、本グループはニューラルネットワークを用いた競馬予想を行った。ニューラルネットワークとは、複数の人工ニューロンとシナプス結合から構成されるネットワークモデルである。シナプス結合の強さを重みという。ニューラルネットワークは、入力層と複数の隠れ層、出力層を含む多層構造になっている。ニューラルネットワークの学習とは、出力とラベルデータの誤差（損失関数）が最小となるように重みを最適化することである。誤差逆伝播法によってすべての重みに対する損失関数の勾配を求め、勾配の逆方向に重みを更新する。

5.3.1 特徴量

入力層に入力されたデータは、シナプス結合を通じて以降の層へと伝播されていく。本グループでは、入力データの特徴量として、インターネット競馬新聞「競馬新聞 デイリー馬三郎」[2] によって提供されている中央競馬の厩舎コメントを用いた。収集におけるシステム的な限界もあり、2019 年 1 月から 2019 年 5 月までに開催されたレースを対象とした。データ収集の詳細については、3.1.4.2 節「データ収集」を参照されたい。

5.3.2 前処理

本グループで扱った厩舎コメントは「前回は途上の段階で、放牧を挟んだ。動きはいいし、今の未勝利戦なら能力は上位。〈牧師〉」[2] といった 2 文程のテキストデータである。テキストは単語あるいは文字を単位とするシーケンス（系列）データとして解釈される。この単位はトークンと呼ばれる。テキストをシーケンスデータとして扱うためには、テキストをトークンに分割（トークン化）し、数値テンソルに変換する必要がある。

5.3.2.1 Mecab によるトークン化

本グループでは、形態素解析器「Mecab」[1] を利用してトークン化を行った。Mecab とは、京都大学情報学研究科 – 日本電信電話株式会社コミュニケーション科学基礎研究所 共同研究ユニットプロジェクトを通じて開発されたオープンソース 形態素解析エンジンである。形態素解析とは、文を意味を有する最小の言語単位である形態素に分割し、形態素の品詞情報を解析する技術である。Mecab では形態素解析と共に分かち書きにも対応している。分かち書きとは、文を形態素に分割することである。Mecab はユーザー辞書にも対応している。Mecab 辞書に標準で記載されていない競馬用語を形態素として登録した。我々は Mecab 辞書およびユーザー辞書の形態素をトークンとして、コメントデータのトークン化を行った。

5.3.2.2 クリーニング

トークン化に際して、コメントデータのクリーニングを行った。クリーニングとは、文書に含まれるノイズを事前に除去する作業である。これは機械学習モデルの精度向上を目的としている。我々が処理したノイズを以下に列挙する。

- 文末に記載されていた担当調教師名の除去 例)〈牧師〉
- 句読点や括弧を始めとする記号の除去
- 数字の#への置換
- 全角文字の半角文字への置換

5.3.2.3 表記ゆれの統一

コメントデータに含まれる「いっぱい：一杯」「取消：取り消し」などの表記ゆれを修正した。表記ゆれの検出は、目視および Sudachi 同義語辞書 [14] の参照によって実施した。Sudachi 同義語辞書は、Sudachi 辞書に登録されている語に対して同義語情報を付与したものである。同義語辞書ソースのフォーマットは次の通りである。

- 0：グループ番号
- 1：体言/用言フラグ (省略可)
- 2：展開制御フラグ (省略可)
- 3：グループ内の語彙番号 (省略可)
- 4：同一語彙素内での語形種別 (省略可)
- 5：同じ語形の語の中での略語情報 (省略可)
- 6：同じ語形の語の中での表記ゆれ情報 (省略可)
- 7：分野情報 (省略可)
- 8：見出し
- 9：予約
- 10：予約

辞書の実体は次のようになっている。

- 000001,1,0,1,0,0,0,(), 曖昧,,
- 000001,1,0,1,0,0,2,(), あいまい,,
- 000001,1,0,2,0,0,0,(), 不明確,,
- 000001,1,0,3,0,0,0,(), あやふや,,
- 000001,1,0,4,0,0,0,(), 不明瞭,,
- 000001,1,0,5,0,0,0,(), 不確か,,

検出した表記ゆれをまとめた dictionary を以下に記載する。key が修正前の語、value が修正後の語である。

{' メンバ': ' メンバー', ' 割引': ' 割り引き', ' 取消': ' 取り消し', ' いっぱい': ' 一杯', ' かえる': ' 変える', ' だす': ' 出す', ' サッパリ': ' さっぱり', ' 入口': ' 入り口', ' 割引き': ' 割り引き', ' 乗り代わり': ' 乗り替わり'}

5.3.2.4　非頻出語の除去

指導教員のアドバイスを受けて、非頻出語を除去し、全コメントデータにおける頻出語上位 500 語のみを入力に用いた。

5.3.2.5　単語へのインデックスの割り当て

ニューラルネットワークはテキストを入力として受け取ることができないため、トークンを分散表現と呼ばれる数値ベクトルに関連付け、テキストをベクトル化する必要がある。本グループでは学習済みの分散表現として、「日本語 Wikipedia エンティティベクトル」[16] を利用した。これは、日本語版 Wikipedia の本文全文から学習した、単語、および Wikipedia で記事となっているエンティティの分散表現ベクトルである。Embedding 層は、トークンに対応した整数を入力として受け取り、整数に対応する分散表現を出力する隠れ層である。Embedding 層への対応のため、トークンのシーケンスデータをインデックスのシーケンスデータに変換した。

5.3.2.6　アンダーサンプリング

アンダーサンプリングとは、不均衡データを学習に用いる際、多数のクラスのデータ数を少数のクラスに合わせることで、学習データのクラス間の偏りを整える処理である。我々は、後述する好

走、普通、凡走分類モデルにおいて、データ数の多かった普通クラスに対してアンダーサンプリングを適用した。

5.3.3　入力層

本モデルは厩舎コメントを特徴量とする入力データから、競馬予想を行うニューラルネットワークモデルである。予想対象については模索があり、着順をラベルとした 18 クラス分類（着順予想）や、複勝を意識した 3 着毎の 5 クラス分類（17、 18 着は除外)、好走、普通、凡走の 3 クラス分類などを試みた。好走、普通、凡走分類モデルについては、図 5.2 に従ってラベル付けを行った。特徴量についても、人気順を含むモデルと含まないモデルの両方を試みた。

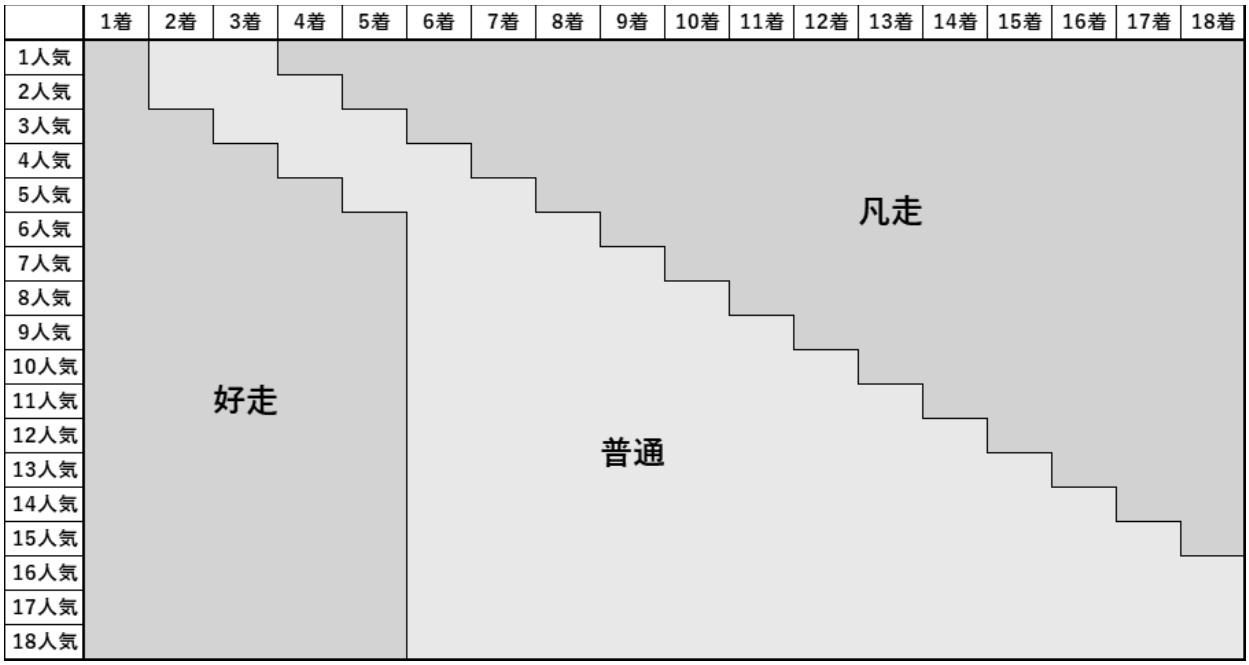

図 5.2　ラベルの割り当て

5.3.4 Embedding 層

前述したように、Embedding 層はトークンに対応した整数を入力として受け取り、整数に対応する分散表現を出力する隠れ層である。本層の重みベクトルとして「日本語 Wikipedia エンティティベクトル」の分散表現ベクトルを読み込んだ。この際、gensim ライブラリを利用した。

5.3.5 LSTM 層

シーケンスデータを処理するためのディープラーニングアルゴリズムの一つに、LSTM（Long Short-Term Memory）が知られている。LSTM はリカレントニューラルネットワーク（Recurrent Neural Network。以下、RNN) を基盤として発展させたものである。RNN は内部にループを含むニューラルネットワークである。シーケンスの一つ前の要素の出力が次の要素の入力に再利用される。これによって、シーケンスデータを一連の入力として扱うことを可能にしている。しかし、勾配消失問題（誤差逆伝播法による学習において、勾配が消失または発散してしまう問題）があるため、単純な RNN は長期間の依存関係を学習することができない。LSTM は、RNN の隠れ層を LSTM ブロックに置き換えたニューラルネットワークである。LSTM ブロックは LSTM セル、忘却ゲート、入力ゲート、出力ゲートから構成される。過去の情報を再入力することで、LSTM は勾配消失問題に対応している。Keras の LSTM 層を 2 層追加した。2 層の LSTM はいずれもユニット数 32、dropout=0.5 を指定した。

5.3.6 Dropout 層

Dropout は学習時にノードをランダムに消去する手法である。アンサンブル学習では、個別に学習させた複数のモデルの出力の平均、あるいは多数決を取る。Dropout はランダムなノードの削除によって、アンサンブル学習と同等の効果を実現していると考えられる。Dropout 層のパラメータには、dropout=0.5 を指定した。

5.3.7 Dense 層

Dense 層（全結合層）では、隣接 2 層のすべてのノードがシナプス結合されている。ユニット数は 16、活性化関数は未指定とした。

5.3.8 出力層

ユニット数にはクラス数、活性化関数にはソフトマックス関数を指定した。

5.3.9 コールバック

コールバックは訓練中に適用される関数の集合で、モデル内部の状態と統計量を可視化する [9]。

EarlyStopping と ReduceLROnPlateau を利用した [9]。EarlyStopping は、監視する値の変化が停止した時に訓練を終了させる関数である。ReduceLROnPlateau は、評価値の改善が止まった時に学習率を減らす関数である。いずれも検証データの精度を監視した。

（文責: 新大実）

第 6 章　中間発表

6.1　ポスター作成

中間発表の提出物として、プロジェクトのメインポスターを作成した。作業は他グループのポスター担当者と共同して行い、本グループからは鈴木、新が参加した。ポスターはバイリンガルで記述し、事前知識のない者でもグループの概要を理解できる内容を目指した。

（文責: 新大実）

6.2　Web サイト作成

6.2.1　Web サイト作成の目的

事前に参加者に提供する資料のひとつとして、Web サイト/ビデオのどちらかの作成がプロジェクト学習 WG により決定された。我々は Web サイトの作成を行うこととした。そして我々が取り組む課題は競馬に関係しているので、競馬になじみがない人でも率直に理解できるように作成する必要がある。さらに Web サイト作成とともにグループ内で用語の確認、今後の予定の見積もりや手法の検討も行う必要がある。

6.2.2　Web サイトの構成

Web サイトの構成として、「背景・目的」、「活動内容」、「手法の検討」、「最終成果物」、「今後の予定」で構成した。

6.2.3　背景・目的

この項では我々が競馬分析を行うにあたり、その動機となった問題点、課題点、それらを解決するためにどのようなことを目的としているかを記述した。分かりやすく説明するため、オッズと勝利確率の関係を説明する簡易的なイメージを作成した。また競馬には一般の人々になじみがない用語が用いられるので、それらの説明も記述した。

6.2.4　活動内容

この項では中間発表準備開始までの活動内容を記述した。

6.2.5　手法の検討

この項では中間発表時点で考えている課題の解決方法を記述した。データの収集、自然言語処理の方法など箇条書きで記述した。

6.2.6　最終成果物

この項では分析を行った後に成果物として何を作成するかを記述した。我々は自然言語処理を行って、調教師のコメントがレースの結果と関係があるかを調べ、それを利用したシステムを構築することを成果目標として記述した。

6.2.7　今後の予定

この項では課題解決までの流れを見積もり、月ごとにどのような活動を行うかを記述した。さらにデータ収集とプログラムの実装を別々に行うデータ班と実装班のメンバーを記述した。

（文責: 川崎景大）

6.3　中間発表の評価

プロジェクト学習の中間発表は例年は大学内で行われていたが、今年度は新型コロナウイルスの影響がありオンライン形式での開催となった。Zoom を用いたオンライン発表会が 2020 年 7 月 17 日に行われた。プロジェクト学習の中間発表が終了後、事前に作成していたアンケートフォームにおいて聴講者には発表評価をしていただいた。アンケートでは発表技術と発表内容について 1 から 10 の 10 段階の評価とコメントをしてもらう。記入していただいたものの一部を以下に示す。

発表技術:

- Web サイトが見やすく、効果的に使われている
- スライドに質問の方法を明記しており、オンラインならではのわかりやすさがあった
- マイクに吐息が当たって聞きづらい部分があった
- 動画などによる説明がないとせっかくの活動内容が伝わりにくい

発表内容:

- 着目点が面白い
- 発表がわかりやすかった
- ポスターの内容が項目ごとによくまとめられていてみやすかった
- 競馬に関する語句の説明が足りない
- 競馬の分析をしたいと思い立った経緯が知りたかった

表 6.1　中間発表

評価	発表技術	発表内容
1	0	0
2	0	0
3	0	0
4	0	1
5	4	2
6	2	3
7	6	8
8	11	12
9	12	10
10	5	5
合計	40	41
平均	8.0	7.9

コメントから反省する点がわかってきた。まず、競馬に関する語句の説明が足りないという意見がいくつかあった。中間発表を聞く中には競馬に関する知識がない人も多くいるということで、単語については気をつかってきたが妥協していた面もあったので、成果発表では単語説明の欄を設けるなどして対処していきたい。また、テーマに関しての理解を得ることはできたが、競馬自体を選んだ理由の説明をしておらず指摘を受けた。実際に背景・目的の中に記載していなかったので気をつけていきたい。しかし、着目点が面白いという意見や発表がわかりやすかったという意見もあり、全体的には本プロジェクトについて理解していただけたと感じている。これからの活動で参考になるコメントも多かったので有効に活用していきたい。

（文責: 藤原慎太郎）

第 7 章　成果発表

ポスターは中間発表で作成し、5.1 で記述されているため省略する。

7.1　Web サイト

Web サイトで被っている項目の説明は省略する.

7.1.1　実施した手法

この項ではデータ班、実装班がそれぞれの活動の詳細と実施した手法、結果について記述した。データ班の項目ではデータセットの作成についての具体的な手法、使用したソフトの使い方を図を用いて説明した。実装班の項目では作成したモデルの説明と実行結果、考察を記述した。多クラス分類の際のラベルの振り分け方を図示した。また、作成したモデルごとに出力されたモデルの精度のグラフとラベルごとの再現率、適合率をまとめた表を用いて結果を説明した。また、学習させたモデルを使用した複勝回収率の算出した結果を記述した。

7.1.2　まとめ

この項では本プロジェクトの活動と結果をまとめ、今後の展望について記述した。

（文責: 武田佑樹）

7.2　成果発表の評価

プロジェクト学習の成果発表会は前期と同じように新型コロナウイルスの影響により、オンライン形式での開催になった。Zoom を用いたオンライン発表会が 2020 年 12 月 4 日に行われた。評価方法は中間発表と同様であるが、グループごとの評価ではなくプロジェクトごとの評価に変更された。聴講者には発表技術と発表内容について評価していただいた。グループ B に関するものを抜粋して記入していただいたものの一部を以下に示す。

発表技術:

- 手短な発表はありがたかった。以前より細やかに情報が記述されていた
- あらかじめ用意した質問に答えるなど、かなり工夫して発表していた
- 「馬三郎」などの説明のない用語がでているが、最初に一言それは何かを書いておくべき
- もう少しサイトに載せる内容をまとめたほうがよい

発表内容:

- 試行錯誤して苦労した様子がなんとなく伝わってきた
- 今後の展望にある、データ数をより増やした場合の結果も気になる

- 学習不足が否めない結果となったが、とても将来への期待を感じさせるものでもあった
- プログラム時間が少なかったのか，完成度が少し低かったのが残念

表 7.1　成果発表

評価	発表技術	発表内容
1	0	0
2	0	0
3	0	0
4	0	0
5	1	1
6	4	3
7	11	9
8	9	11
9	8	9
10	3	3
合計	36	36
平均	7.8	7.9

表 7.1 より、発表技術の平均評価点数は 7.8 であり、発表内容の平均評価点数は 7.9 であった。発表技術の平均評価点数は中間発表のときと比べて下回ってしまった。今回の成果発表はポスターを簡潔な内容で作成したので、Web サイトは詳細に記載するといった方法で臨んだ。しかし、事前に閲覧できる時間が 1 時間しかないこともあり、聴講者のコメントには Web サイトの内容をもう少しまとめたほうが良いというものが多かった。また、中間発表のコメントに専門的な内容が多いので語句の説明を増やしたほうが良いというコメントがあった。そこで語句の説明には気をつかっていたがまだ不十分な点があったのは事実であり、反省していかなければいけない。発表内容に関しては中間発表と同様の点数であり、本プロジェクトで行った内容について理解していただけたと考える。

（文責: 藤原慎太郎）

第 8 章　結果

8.1　前期

8.1.1　ディープラーニングの知識習得

ディープラーニングについて同じ書籍を利用して学ぶことで、メンバー内で共通の認識を得ることができた。また CNN の実装を Python で実装することができるようになった。しかし、テーマは自然言語処理を用いた分析であるため CNN だけではなく RNN や形態素解析、単語埋め込みなどのディープラーニングの領域ではない自然言語処理に関する知識を勉強する必要があると考えられる。

（文責: 川崎景大）

8.1.2　テーマの決定

過去の未来大学の卒業研究の論文では、パドック画像などを解析する画像認識によるアプローチで着順の予測を行っていた。まだ試みられた例が少ないアプローチである自然言語処理による調教師のコメントからの着順予測を行うことで、集合知には含まれない関係性を発見できる可能性が考えられる。しかし、データセットの作成や自然言語処理の勉強などは過去の卒業論文などからノウハウを得ることができないので、勉強したディープラーニングの理論などを実践する際に問題が発生したり、時間がかかることなどが考えられる。

（文責: 川崎景大）

8.1.3　データの収集とモデルの構築

メンバーを分割したことで、グループで解決すべき問題が分割され、効率的な活動ができるようになったと考えられる。また、我々はディープラーニングについて基礎的なところから学習を開始したため、中間発表の反省として技術的な面で遅れていることが課題として挙げられていたが、メンバーを分割したことで技術的な勉強をしつつ、データの収集も行うことができるようになった。課題としてデータセットの作成に多大な時間がかかることが挙げられる。

（文責: 川崎景大）

8.2　後期

8.2.1　好走、普通、凡走 予測モデル (人気順の入力なし)

コメントデータを説明変数として 好走、普通、凡走 予測モデルを作成し、学習させた。学習過程が図 8.1 である。またテストデータでの再現率と適合率を図 8.2 に示す。ランダムに予測を行っ

た場合の再現率と適合率を図 8.3 に示す。

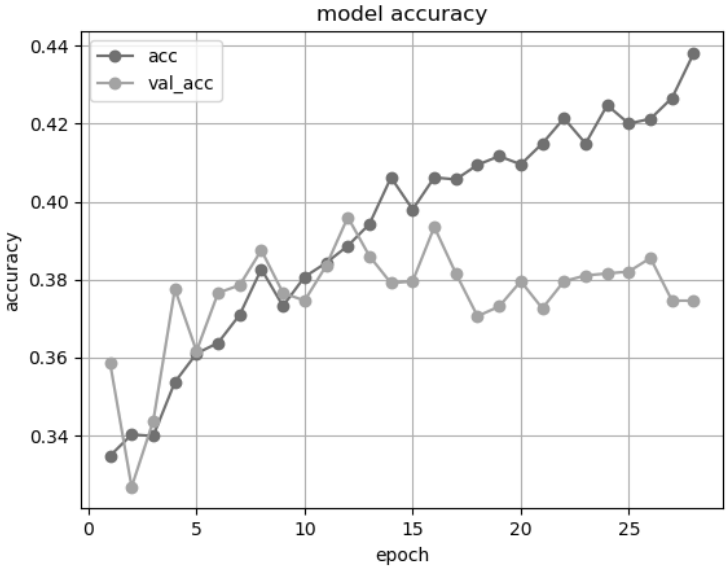

図 8.1　学習過程

	データ数	正解数	不正解数	再現率	適合率
好走	880	328	866	0.373	0.275
普通	1985	792	503	0.399	0.612
凡走	851	309	918	0.363	0.252

図 8.2　モデルの適合率と再現率

	データ数	正解数	不正解数	再現率	適合率
好走	880	293.3	945.3	0.333	0.237
普通	1985	661.7	577.0	0.333	0.534
凡走	851	283.7	955.0	0.333	0.229

図 8.3　ランダムに予測した場合の適合率と再現率

訓練データと検証データにアンダーサンプリングを適用し、テストデータに関してはアンダーサンプリングを適用していないので注意が必要である。

ランダムな予測を行った場合、各ラベルの再現率は 1/3 、各ラベルの適合率は ラベルのデータ数/データ数の合計 となる。再現率ではランダムな結果と比べて 3〜6% 良い性能を示した。適合率では普通のラベルがランダムな結果と比べて約 8% 高く、他の 2 つのラベルでは約 3% 高い結果となった。

ランダムな予測に比べて作成したモデルの各ラベル予測の再現率は少しだけ高かったが、モデルはコメントデータから好走、普通、凡走を判断できる要素を十分に検出できてない。またモデルを実際の競馬の予測に応用する場合、1〜3 着の予測を行うことが主な目的となり、適合率が低いため現状では実用的な段階には届いていない。

残された問題点として、ラベルごとのコメントの性質を調査し、共通するものを明らかにすることでより人間が扱いやすい判断材料を抽出する必要があると考えられる。

8.2.2　好走、普通、凡走 予測モデル (人気順の入力あり)

コメントデータと人気順を説明変数として 好走、普通、凡走 予測モデルを作成し、学習させた。学習過程が図 8.4 である。またテストデータでの再現率と適合率を図 8.5 に示す。ランダムに予測を行った場合の再現率と適合率を図 8.6 に示す。

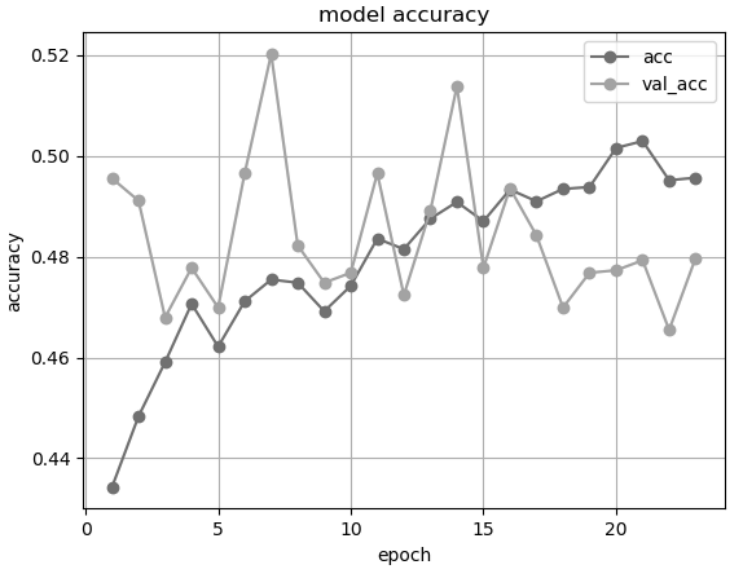

図 8.4　学習過程

	データ数	正解数	不正解数	再現率	適合率
好走	895	433	798	0.484	0.352
普通	2022	1309	457	0.647	0.741
凡走	799	261	458	0.327	0.363

図 8.5　モデルの適合率と再現率

	データ数	正解数	不正解数	再現率	適合率
好走	895	298.3	940.3	0.333	0.241
普通	2022	674.0	564.7	0.333	0.544
凡走	799	266.3	972.3	0.333	0.215

図 8.6　ランダムに予測した場合の適合率と再現率

人気順を説明変数に加えた場合、加えていない場合と比べて普通ラベルの予測の再現率が他のラベルに比べて高くなることが確認された。

競馬のオッズは集合知により形成され、勝利確率を正確に反映している。オッズをもとに決定される人気順も勝利確率を求めるのに良い指標となる。また人気順と着順の組み合わせは同じ値であ

る場合が最も多く、ランダムアンダーサンプリングを適用しても人気順と着順が同じ値になる組み合わせが最も多くなる。それらには人気順 1 番と 1 着の組み合わせを除いて普通ラベルが割り当てられているため、作成したモデルは人気順のみでラベル予測を行うように学習している可能性が考えられる。学習を行う際は集合知が形成する指標を説明変数に用いると、その正確さから本来の目標とするコメントデータからの特徴の抽出ができなくなる可能性がある。

8.2.3　3 着ごとの予測モデル

コメントデータを説明変数として 3 着ごとの推定モデルを作成し、学習させた。学習過程が図 8.7 である。またテストデータでの再現率と適合率を図 8.8 に示す。ランダムに予測を行った場合の再現率と適合率を図 8.9 に示す。

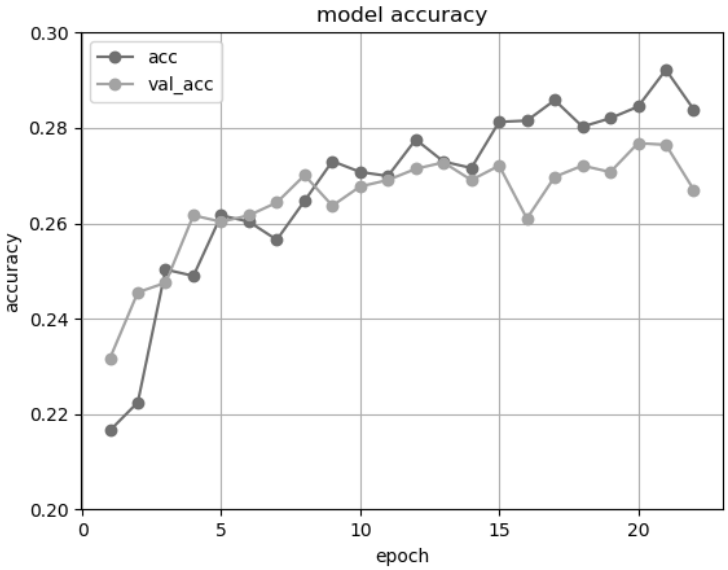

図 8.7　学習過程

	データ数	正解数	不正解数	再現率	適合率
1~3着	843	377	841	0.447	0.310
4~6着	793	195	611	0.246	0.242
7~9着	738	163	584	0.221	0.218
10~12着	680	5	18	0.007	0.217
13~16着	662	285	637	0.431	0.309

図 8.8　モデルの適合率と再現率

	データ数	正解数	不正解数	再現率	適合率
1~3着	843	168.6	574.6	0.200	0.227
4~6着	793	158.6	584.6	0.200	0.213
7~9着	738	147.6	595.6	0.200	0.199
10~12着	680	136.0	607.2	0.200	0.183
13~16着	662	132.4	610.8	0.200	0.178

図 8.9　ランダムに予測した場合の適合率と再現率

ランダムな予測を行った場合、各ラベルの再現率は 1/5 になるが、10-12 着の再現率が大きく悪くなる結果となった。また適合率はランダムな予測と比べ、1-3 着は約 8%、13-16 着は約 13% 良い結果となった。

残された問題点として、10-12 着の再現率が他のラベルと比べて低い原因を調査し、解決したうえでモデルの性能を評価する必要があると考えられる。

（文責: 川崎景大）

第 9 章　今後の課題と展望

9.1　前期のまとめ

私たちのグループでは「自然言語処理を用いた競馬分析」をテーマとして活動してきた。テーマを選ぶ上で、函館に競馬場がある点とディープラーニングを学ぶ際に競馬を用いた例が多く存在することが影響を与えた。しかし、競馬に精通しているメンバーは居らず、競馬という大きな枠組みの中の何に取り組むのかについては漠然としていたため、前期は主に競馬とディープラーニングについての基礎知識を勉強する時間を多く必要とした。まず、グループメンバー各自がテレビや競馬をテーマとしたゲームに触れることで競馬の知識や問題点を学んでいった。次に各自で競馬に触れて気づいたことをグループ内で共有することで取り組んでいくテーマを絞っていった。最終的に「調教師のコメントを自然言語処理で分析する」というテーマに決定した理由として、調教師は調教が上手くいかなかったときに、立場上、馬主の目があり良いことしか書けないので、その微妙な変化をディープラーニングで分析してみると面白いのではないかと考えた結果である。並行してディープラーニングの学習も行った。「ゼロから作る Deep Learning」[5] という本を用いて、メンバー 1 人ずつ担当の章を決めることで知識の共有と学習の効率化を目指した。テーマの決定とディープラーニングの学習が終了したのが 7 月だった。そこからはグループメンバーを実装班とデータ班の 2 つの班に分けて活動していった。実装班は主に自然言語処理の学習と実装を目的としている。また、データ班は自然言語処理で分析する際のデータセットの作成を目的としている。後期はデータ班が作成したデータセットを用いて、実装班が自然言語処理で分析していく。学習させたニューラルネットワークを用いて、新たにコメントを入力したときに予想される着順を出力するアプリケーションの作成を目的に活動していく。

（文責: 藤原慎太郎）

9.2　後期のまとめ

後期の最初はデータ班と実装班ごとに分かれた作業から始まった。データ班は競馬新聞と提携しているアプリケーションである馬三郎 [2] の厩舎コメントと netkeiba.com[12] のサイトを対象に 2019 年 1〜5 月の約 20000 データを収集し、本プロジェクトで用いるデータセットを作成した。実装班は、夏季休暇中に作成した Word2Vec のモデルと LSTM ネットワークを用いて、5 ヶ月分の着順とコメントデータを Mecab[1] により単語分割し、着順の予測を行うモデルを作成した。それぞれの班の作業が終了したのが 10 月半ばであり、その後は全体での作業となった。当初予定していたモデルでは、着順と人気を考慮した好走、普通、凡走の 3 分類モデルであった。しかし、実際に競馬予想をする上で必要となる項目が好走の適合率であるが、3.5 割付近という結果であった。また、人気順も入力データとして用いていたが、出力結果に大きく影響しており、本来の目的であるコメントから着順予想できるか判断しづらかった。そこで、新たにコメントのみを入力とする 3 着ごと推定モデルの作成をすることにした。この段階で 11 月に突入しており、当時の到達具合からアプリケーションの開発に着手することはできないと判断し、改良したモデルを最終成果物とすることにした。新たに作成したモデルの 1〜3 着の適合率は約 3 割であり、5 分類していることか

ら、ランダムで出力した結果と比較すると若干の精度の向上を見ることができたが、やはり実用的なものであるとは言えないモデルであった。これまでに作成したモデルを用いて複勝回収率の算出も行ったが、収束すると言われている 8 割付近という結果であったため、今回のモデルによってコメントが集合知に含まれない要素であるという結論には至らなかった。

しかし、通年の活動を通して、ディープラーニングの学習やプロジェクトの進め方など様々なことを学ぶことができ、有意義なプロジェクト学習にすることができたと言える。

（文責: 藤原慎太郎）

参考文献

[1] Mecab, 2020 年 12 月 23 日アクセス, [online] http://taku910.github.io/mecab/

[2] 馬三郎, 2020 年 9 月 22 日 アクセス, [online] https://uma36.com/

[3] 城崎哲, AI 競馬 人工知能は馬券を制することができるのか. ガイドワークス, 2020.

[4] 寺沢憲吾, 情報学者が競馬予想に踏み出す時に知っておくべきこと, 情報処理, Vol. 60, No. 2, pp. 154-158 (2019).

[5] 斎藤康毅. ゼロから作る Deep Learning. オライリージャパン, 2016.

[6] 大島洋明, パドック画像を用いた着差に基づく競走馬分類の検討, 平成 30 年度公立はこだて 未来大学卒業論文 (2019).

[7] 柳川大輝, 馬体情報を用いた競走馬の着差分類, 平成 31 年度公立はこだて 未来大学卒業論文 (2020).

[8] Francois Chollet. Python と Keras によるディープラーニング. マイナビ出版, 2018.

[9] Keras Documentation, 2020 年 12 月 23 日アクセス, [online] https://keras.io/ja/

[10] RONDHUIT, 「ダウンロード - 株式会社ロンウイット, livedoor ニュースコーパス」, 2020 年 1 月 8 日アクセス, [online] https://www.rondhuit.com/download.html#ldcc

[11] JRA 日本中央競馬会, 2020 年 12 月 23 日アクセス, [online] https://www.jra.go.jp/

[12] netkeiba.com, 2020 年 12 月 23 日アクセス, [online] https://www.netkeiba.com/

[13] Octoparse, 2020 年 12 月 23 日アクセス, [online] https://www.octoparse.jp/

[14] GitHub, Sudachi 同義語辞書, 2020 年 12 月 23 日アクセス, [online] https://github.com/WorksApplications/SudachiDict/blob/develop/docs/synonyms.md

[15] SOURCENEXT, 瞬間テキスト, 2020 年 12 月 23 日アクセス, [online]https://www.sourcenext.com/product/pc/use/pc_use_001516/

[16] 日本語 Wikipedia エンティティベクトル, 2020 年 12 月 23 日アクセス, [online] http://www.cl.ecei.tohoku.ac.jp/ m-suzuki/jawiki_vector/