教学データからのモチベーション向上・キャリアパス支援のための情報推薦機構

平松　楓也

富山県立大学　情報基盤工学講座

June 24, 2020

はじめに

背景

近年の就職活動は売り手市場と言われていたが、コロナウイルスの影響で世界的に経済状況が悪化しており、買い手市場に推移していく可能性が考えられる。また、大手企業へ就職を考えた場合、応募人数が多く狭き門であることが多いため、企業は企業がより求めている人材を採用すると思われる。そのため、学生の間に、より効率的に企業が求める人材になるための勉強が必要になると思われる。

目的

過去の卒業生の就職先や、学業成績、野外活動のデータをクラスタリングし、在校生がより効率的に就職活動を行えるよう対話型の情報推薦機構の基礎技術を開発する。

完成目標

入学したばかりの大学1年生、卒業後就職したい企業が決まっている学生

その企業に就職するためにはどうすれば効率的に動けるかわからない

協調フィルタリングで過去の卒業生のデータからやるべきことを推薦

やるべき勉強、活動が明確になりモチベーション向上

就職したかった企業への内定

説明的データ分析

説明

事実を説明する
見つける

例

どんな人が何を買っているか？
ある広告がどれだけ売り上げに貢献しているか？

手法

BI、クラスタリング、アソシエーション分析

予測的データ分析

説明

未来や欠測値を予想する

例

ある商品群を閲覧した人の性別は？
広告を出稿したらどれだけ売り上げが上がるのか？

手法

分類・回帰、統計的機械学習、協調フィルタリング

指示的データ分析

説明

最適解を探す

例

利益を最大化するための、最適な仕入れ量は？
売上を最大化するには、どこに広告を出稿するべきか？

手法

最適化、実験結果

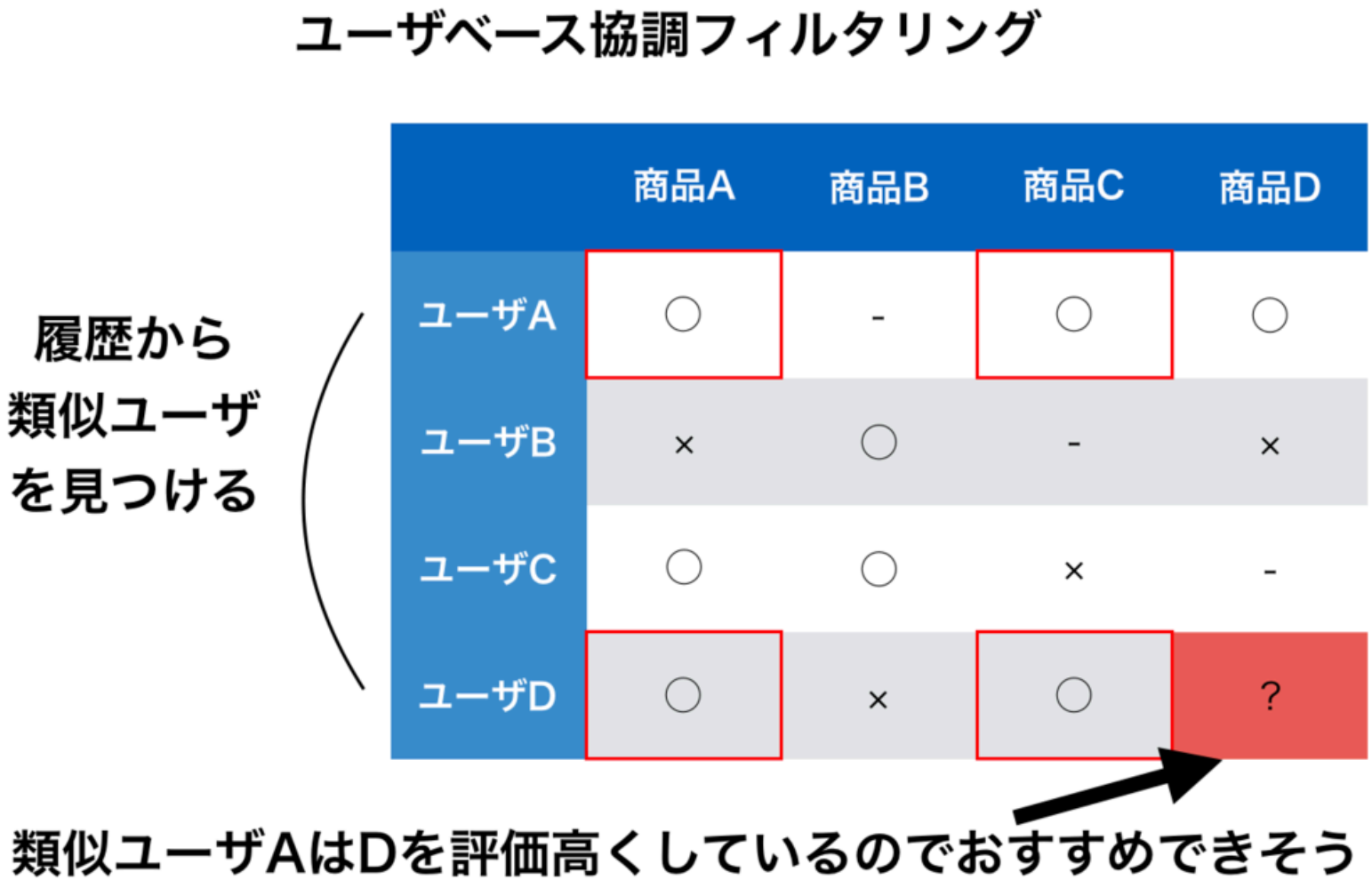
協調フィルタリング

協調フィルタリングとは、Amazon が開発したレコメンドエンジンで、多くのユーザの嗜好情報を蓄積し、あるユーザと嗜好の類似した他のユーザの情報を用いて自動的に推論を行う方法論である。

また、協調フィルタリングには二種類あり、ユーザベース協調フィルタリングとアイテムベース協調フィルタリングがある。

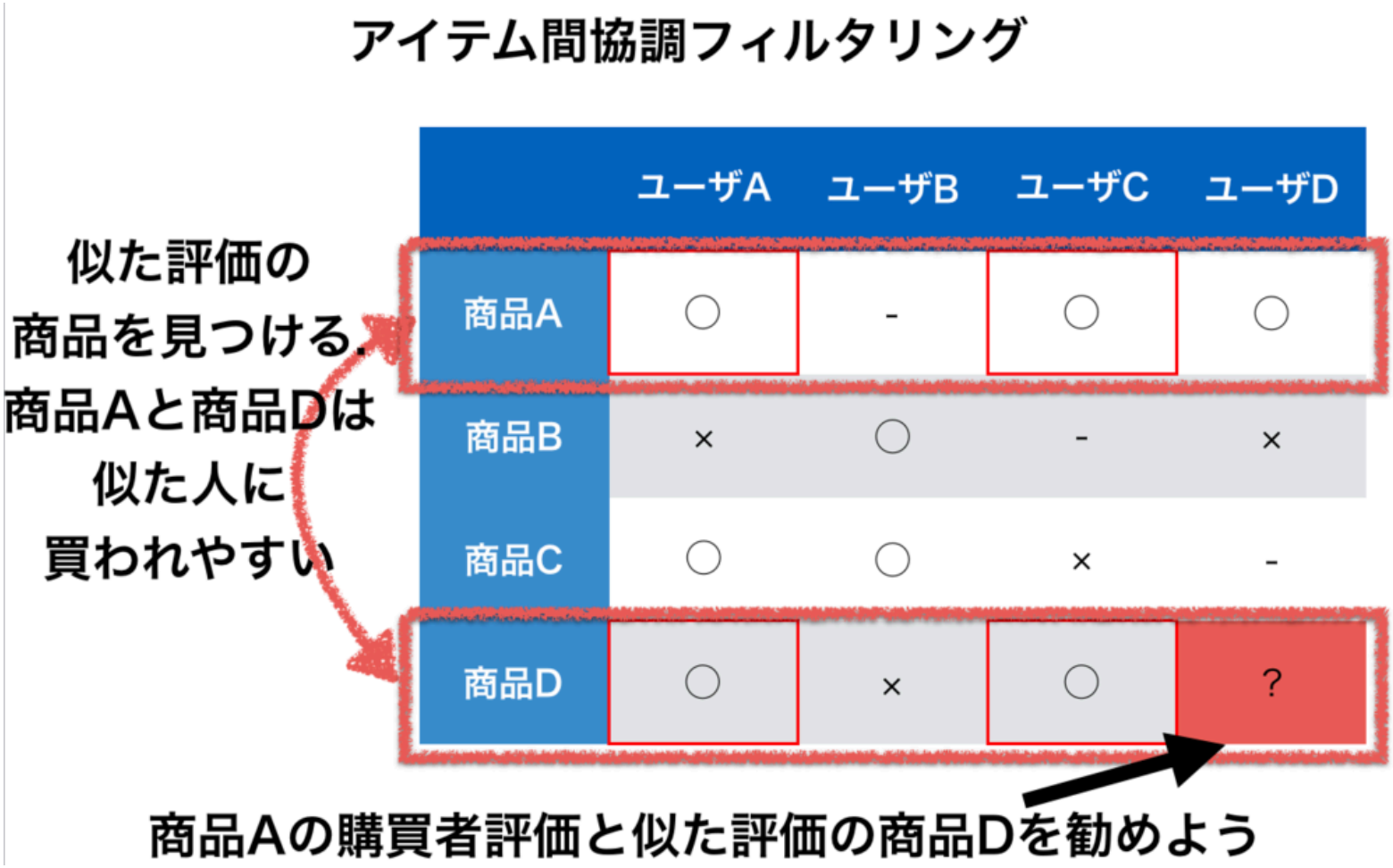
ユーザーベース協調フィルタリング

ユーザベース協調フィルタリングでは「ユーザ A は未評価アイテム I に対して、当該ユーザと似たような嗜好をしている他ユーザと同じような評価をするだろう」という仮定に基づいている。



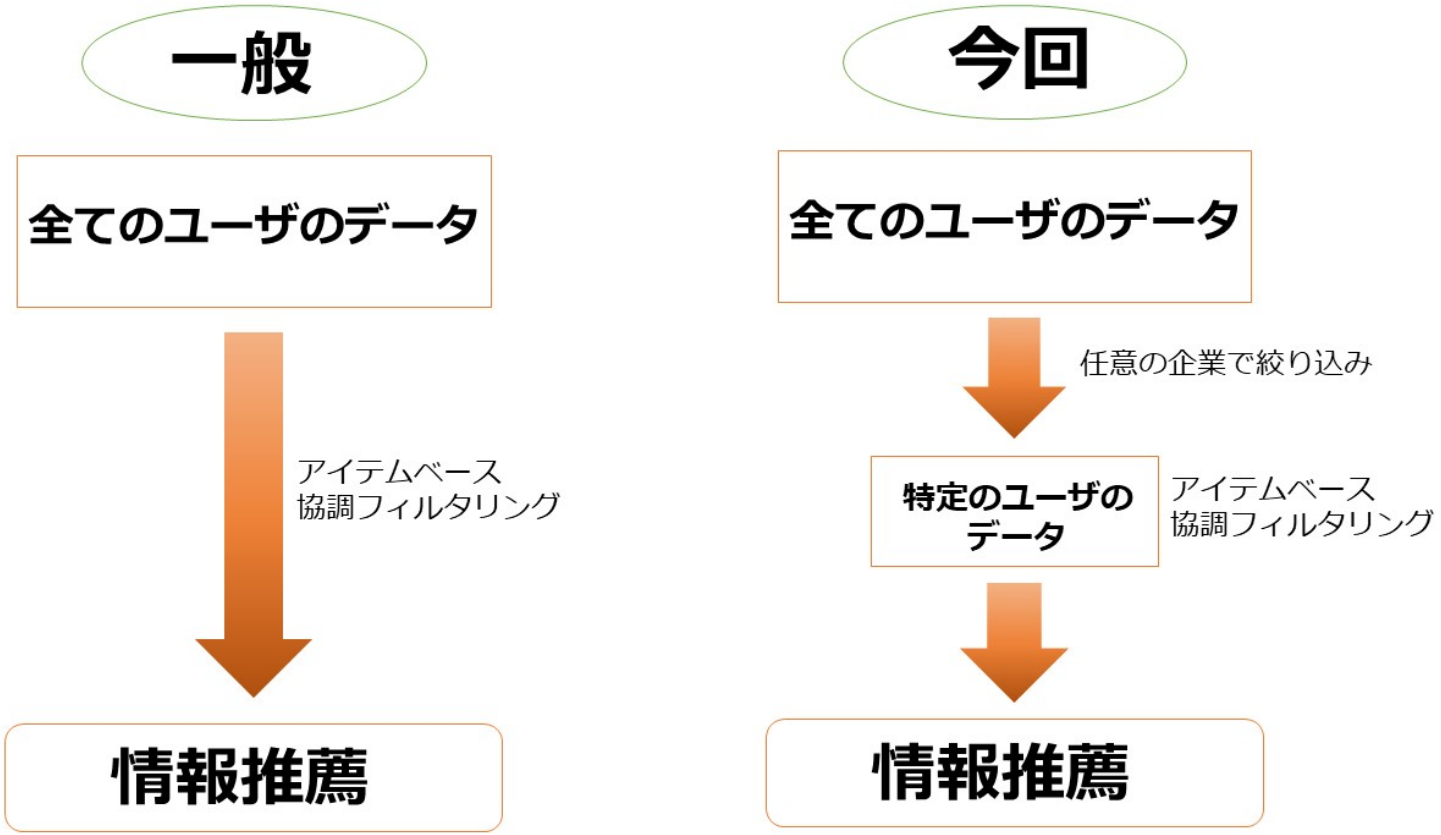
アイテムベース協調フィルタリング

アイテムベース協調フィルタリングでは「アイテム同士の類似度とあるユーザ A の過去に評価したアイテムの評価点を用いて未評価アイテム I の評価点を予測する」というアプローチである。



今後行うアイテムベース協調フィルタリングについて

一般に使われる協調フィルタリングは全ユーザのデータを基にフィルタリングを行うのに対し、今回では、ユーザ A が就職を希望している企業に就職したユーザのみでフィルタリングを行い情報推薦を行おうと考えている。



自分の卒論の構成

1 章 はじめに
2 章 教学データの活用（従来研究）
2.1、教学データの構成・内容（論文５本）
2.2、教学データから得られる知見（論文５本）
3 章 （従来研究）（数学的な手法）
3.1
3.2 情報推薦の仕組み
4 章 提案手法（オリジナリティ）
5 章 数値実験並びに考察
6 章 まとめと今後の課題

デモデータの進展

・csv ファイルを処理，管理しやすいように複数に分けることにした．

・アドミッション__出身高校・入試種別.csv
・アドミッション__受験科目・成績.csv
・カリキュラム__履修・評価.csv
・カリキュラム__科目情報.csv
・キャリア関連__インターンシップ.csv
・キャリア関連__就職情報.csv
・キャリア関連__資格・免許.csv
・課外活動__サークル.csv
・課外活動__アルバイト.csv
・卒業後__企業側.csv
・卒業後__学生.csv

アドミッション

	A	B	C	D	E
1	StudentN	出身高校	入試種別		
2	1713001	宇都宮中ヶ	前期		
3	1713002	四日市西	前期		
4	1713003	苾道	前期		
5	1713004	片山学園	後期		
6	1713005	恵那	前期		
7	1713006	藤枝明誠	推薦		
8	1713007	氷見	前期		
9	1713008	四日市西	後期		
10	1713009	松阪	前期		

図 1: 出身高校・入試種別

	A	B	C	D	E
1	StudentN	センター試	二次試験		
2	1713001	390	277		
3	1713002	389	292		
4	1713003	400	281		
5	1713004	874	0		
6	1713005	389	281		
7	1713006	0	0		
8	1713007	395	254		
9	1713008	885	0		
10	1713009	388	283		

図 2: 受験科目・成績

・出身高校のリストは県大の主な出身校を載せたサイトがあったのでそれを参考にした・点数は平均点を富山県立大学の H29 の平均点を参考に標準偏差 10 でランダムに発生

カリキュラム

	A	B	C	D	E	
1	StudentN	経済学１	富山と日オ	環境論１	環境論２	E
2	1713001	1	2	1	5	
3	1713002	3	5	2	3	
4	1713003	5	5	5	2	
5	1713004	2	3	4	5	
6	1713005	2	3	5	1	
7	1713006	4	3	4	4	
8	1713007	3		3	5	
9	1713008	2	5	2	2	
10	1713009	5	3	5	4	

図 3: 履修・評価

キャリア関連

	A	B	C	D	E
1	StudentN	学部卒業後	就職先		
2	1713001	就職	デンソーテクノ		
3	1713002	就職	スギノマシン		
4	1713003	就職	コマツNTC		
5	1713004	就職	富山村田製作所		
6	1713005	進学	YKK AP		
7	1713006	進学	スズキ		
8	1713007	就職	立山科学グループ		
9	1713008	進学	豊田合成		
10	1713009	進学	澁谷工業		

図 4: 就職情報

	A	B	C	D	E	F	G	H
1	StudentN	普通自動車	TOEIC公開	MOS Wor	MOS Exce	簿記2級以	基本情報技術者	
2	1713001	1	0	0	0	0	1	
3	1713002	1	0	0	0	0	0	
4	1713003	1	0	0	0	0	0	
5	1713004	1	0	0	0	0	0	
6	1713005	0	0	0	0	0	0	
7	1713006	1	0	0	0	0	0	
8	1713007	0	0	0	0	0	0	
9	1713008	1	0	0	0	0	0	
10	1713009	1	0	0	0	0	0	

図 5: 資格・免許

・就職先は富山県立大学の HP を参考にし、学科、学部、院ごとに別ランダム振り分け

・0 が未取得、 1 が取得済み

課外活動

	A	B	C	D	E
1	StudentN	サークル			
2	1713001	Chat Box			
3	1713002	無所属			
4	1713003	無所属			
5	1713004	アカペラサークル			
6	1713005	フットサルサークル			
7	1713006	PDC			
8	1713007	軟式野球部			
9	1713008	無所属			
10	1713009	TRPG・映画研究会			

図 6: サークル

	A	B	C	D	E
1	StudentN	アルバイト			
2	1713001	串網			
3	1713002	アルビス			
4	1713003	秋吉			
5	1713004	アルバイトなし			
6	1713005	個別指導塾スタンダード			
7	1713006	串網			
8	1713007	アルバイトなし			
9	1713008	アルバイトなし			
10	1713009	セブンイレブン			

図 7: アルバイト

・サークルも県大に実在するサークルを全て同じ確率でランダムに振り分け、所属率は 50 %

・アルバイト先は県大周辺にあるアルバイト募集している店舗をランダムで振り分け、アルバイトしている確率は 90 %

決定木分析とは

決定木分析とは，分類木 (雨の日と晴れの日などの条件で分ける) と回帰木 (○○円などの変わりうる値で分ける) を組み合わせたもので，ツリーによってデータを分析する

【メリット】

1. 分析結果が見やすい

【デメリット】

1. 過学習を起こしやすい

決定木が枝分かれする判定は, エントロピー (ジニ不純度) を計算しノードをどう分類すれば一番利得が大きくなるかで分類する

決定木分析：環境

【環境】

1. jupyter notebook
2. これは python 言語の延長線上にあるもので，データ分析により向いている．
3. anaconda からインストール可能で，分割して実行結果を表示できる点が一番のメリット

決定木分析：得たい知見

【得たい知見】

1. 入試種別，センター試験，二次試験から学部卒業後の進路にどのような関係があるのか
2. 参考にした論文では，大学での目的，入学時不安，高校での成績から一年次 GPA にどのような関係があるのかを調べていた

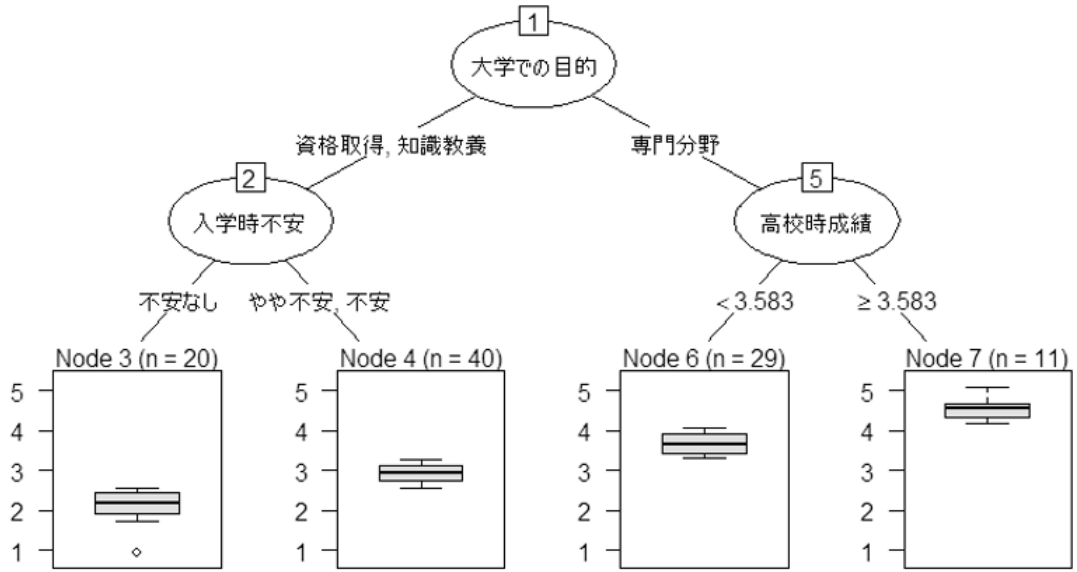


図 9: 学修成果に関わる入学時特徴についての決定木分析結果（模擬データ）

決定木分析：適したcsvの作成

【お膳立て】

1. センター試験と二次試験の点数が学科によって最大点数が違うので，100 点満点に正規化
2. 前回データを分割管理することに決めたので，アドミッション 2 つと就職情報を結合し，いらない項目を削除し，成績と進学.csv を新たに作成
3. 決定木分析では数値しか使えないので，入試種別と学部卒業後を前期 0，後期 1，推薦 2 と就職 0, 進学 1 に置換

	StudentNumber	入試種別	正センター試験	正二次試験	学部卒業後
0	1713001	1	73	0	0
1	1713002	0	58	58	1
2	1713003	0	61	62	0
3	1713004	0	58	58	0
4	1713005	0	63	60	0
..	...	...	...	...	...
352	1718031	0	69	62	1
353	1718032	2	0	0	1
354	1718033	0	68	62	0
355	1718034	2	0	0	0
356	1718035	1	72	0	1

[357 rows × 5 columns]

図 10: 成績と進学.csv

決定木分析：変数設定

【変数設定】

1. 今回は学部卒業後への影響を考えるので目的変数を学部卒業後にする
2. その他の入試種別とセンター試験点数，二次試験点数を説明変数にする
3. 木の深さの最大は 4 とする

【グラフの用語】

1. gini はジニ不純度のこと。0〜1 の値をとり、大きいほど不純度が高い
2. sample はその枝にある要素数
3. value はその枝にある要素数のうち、左が就職の数で右が進学の数

決定木分析：結果

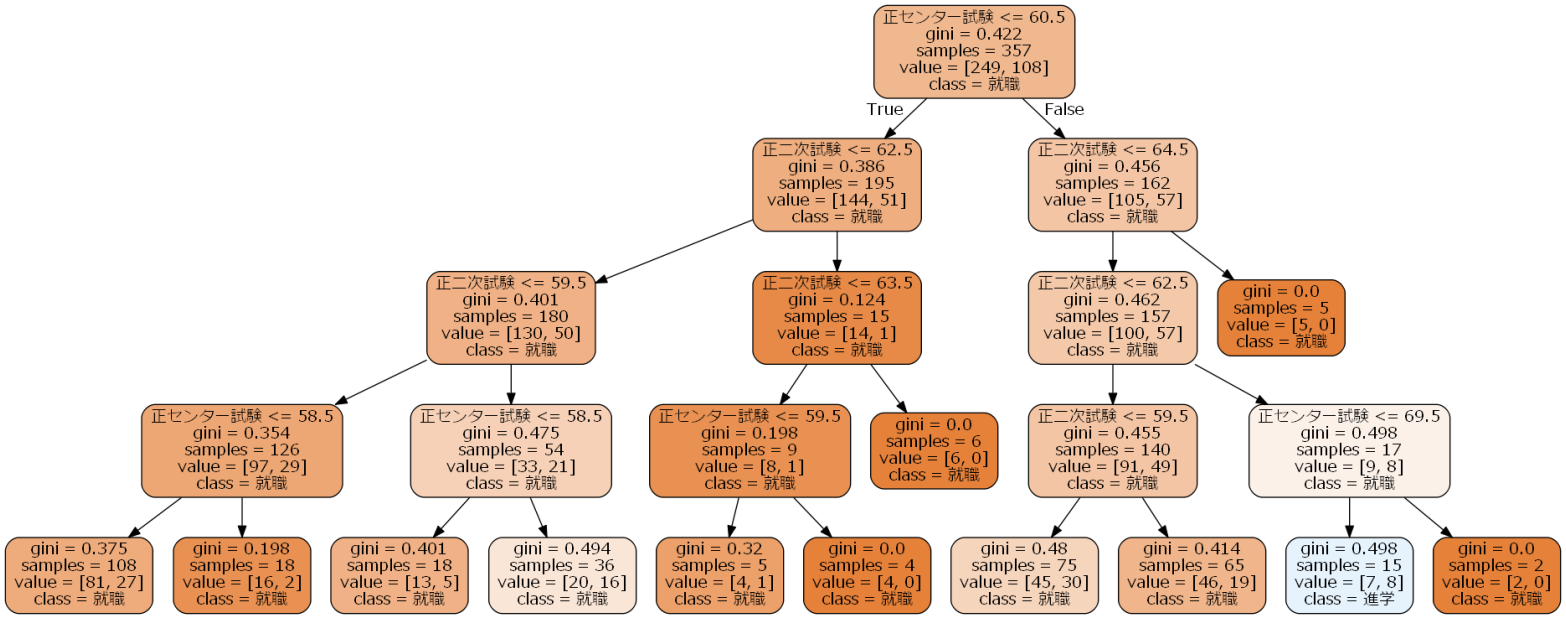


図 11: 決定木分析結果

おわりに

まとめ

1. 本当は 1 年次 GPA との関係性を決定木分析したかったが間に合わなかった。
2. 擬似データが乱数で発生させたものなので分析結果があまり良くなかったのでこれに関しては実際のデータを説明変数に入れるしかない気がする
3. しかし，実際のデータがもらえるのであれば決定木分析の基盤は完成しているので適用は楽に行けそう

今後の課題

1. 一度本当にこの決定木分析が正しく行われているかを検証するためにも進学する人の入試成績が高いようにするなどしてもう一度決定木分析をやってみたほうがいいかもしれない
2. 卒論 2.2 ではもっと他の手法で得られる知見もやっていきたいので次はクラスター分析などの分析を IR で行っている論文を探し，自分でできるか進めていきたい