卒業論文

教育データに基づくキャリアパス推薦機能を搭載した学習支援システムの開発と評価

Development and Evaluation of a Learning Support System with Career Path Recommendation Based on Educational Data

富山県立大学 工学部 情報システム科
2220067 山本 藤也
指導教員 Antonio Oliveira Nzinga Rene 准教授
提出年月: 2026年2月

目 次

図一覧

表一覧

記号一覧

以下に本論文において用いられる用語と記号の対応表を示す.

用語	記号	用語	記号
学生（サンプル）数	m	科目数（特徴量数）	n
学生 ID（添字）	i	科目 ID（添字）	j
成績行列	$\mathbf{X}$	成績行列の要素（成績値）	x_{ij}
学生 i の特徴ベクトル	$\mathbf{x}_i$	想定入力ベクトル（成績を仮定した入力）	$\tilde{\mathbf{x}}$
就職ラベル（二値）	y_i	業界（クラス）	c
業界集合	$\mathcal{C}$	業界 c に対する二値ラベル	$y_i^{(c)}$
業界就職確率	π_i	線形予測子	z_i
シグモイド関数	$\sigma(\cdot)$	重み係数ベクトル	$\mathbf{w}$
科目 j の係数	w_j	バイアス項	b
オッズ	odds_i	対数オッズ（ロジット）	$\mathrm{logit}(\pi_i)$
正則化強度	λ	目的関数	$J(\mathbf{w}, b)$
未履修科目集合	$\mathcal{U}$	確率上昇幅	$\Delta\pi_j$

第1章

はじめに

§ 1.1　本研究の背景

近年，大学における教育環境は急速に変化しており，学生の学習履歴や成績，行動ログといった多様な情報が ICT 技術の発展によって日常的に蓄積されるようになった [1,5,7]．これらのデータは一般に「教育ビッグデータ」と総称され，その分析を通じて教育改善や学修支援を行う教学 IR（Institutional Research）やラーニングアナリティクスが，多くの高等教育機関で導入されつつある [1,5,6]．こうした活動は，学生一人ひとりの学習行動を可視化し，学習意欲や理解度の把握に寄与してきた [6]．

一方で，大学教育の高度化・多様化に伴い，学生の履修選択にかかる負担はこれまで以上に大きくなっている．選択科目や専門領域が拡大する中，学生は膨大な科目群から自身の興味や将来のキャリア志向に沿った履修計画を主体的に構築する必要がある．しかし，どの科目がどの業界・職種と関連しているかといった因果的・構造的な情報は十分に提示されておらず，実際には断片的な情報や周囲の経験に基づいて科目選択が行われているのが現状である．

加えて，教学データ自体にも多くの課題が存在する．学生の履修には興味本位の受講や気まぐれな選択が含まれやすく，キャリアと直接関係しないノイズが大量に混入する．また，業界ごとの進路人数にも大きな偏りがあり，履修履歴データは本質的に疎で一貫性に欠ける．このような不完全なデータ環境において，高精度かつ安定した推薦を行うことは容易ではない [3,5]．

本研究室でもこれまで，協調フィルタリング（Collaborative Filtering）を用いた科目推薦手法の研究が進められてきた [10–12]．学生間の類似度に基づく推薦は一定の効果を示したものの，協調フィルタリングはデータの疎性やノイズに弱く，履修のばらつきが大きい場合には推薦結果が安定しない [10]．また,「似ている先輩が履修していたから推薦する」という相関ベースの仕組みであるため，特定の科目がキャリアにどのように影響するかを説明できず，推薦の根拠を利用者に提示することが困難であった [13]．

このように，教学データの蓄積が進む一方，学生のキャリア形成を支援するためには「どの科目がキャリアにとって本質的に重要か」を可視化し，なおかつノイズを含む現実のデータ環境でも安定して動作する推薦手法が必要である [5,6]．こうした背景から，本研究では従来手法の限界を克服する新たな分析モデルの構築を目指す．

§ 1.2　本研究の目的

本研究の目的は，教学データに基づいて学生のキャリアパスを精度高く予測し，その形成に寄与する科目を統計的に特定したうえで，推薦理由を論理的かつ説明可能な形で提示するアルゴリズムとシステムを構築することである [5,6]．これは単なる推薦機能の高度化にとどまらず，教学 IR やラーニングアナリティクスの成果をキャリア支援という実務領域へ橋渡しする，新たなデータ活用基盤の確立を意味する [1,6]．

先行研究では，協調フィルタリングに基づく科目推薦手法が開発され，履修履歴の類似度を用いることで一定の効果が示されてきた [10–12]．しかし，協調フィルタリングは相関ベースのパターン抽出に依存するため，学生ごとの履修行動のばらつきやノイズの多さにより精度が安定せず，結果が大きく変動するという課題があった [3,10]．特に，業界ごとの進路人数が少ない大学環境ではデータが疎になりやすく，推薦の根拠を十分に説明できない点が実運用上の大きな障壁となっていた [13]．

そこで本研究では，従来手法の構造的な限界を克服するため，L1 正則化ロジスティック回帰（Lasso Logistic Regression）を新たな中心技術として採用し，キャリアパス推薦への適用を試みる [14,15]．L1 正則化はスパース推定により，数多くの履修科目の中からキャリアに関連する因子のみを自動的に抽出し，関係の薄い科目の重みをゼロに縮約する [15]．この性質により，ノイズを含む教学データ環境でも頑健な推定が可能となり，各科目の影響度を符号と大きさで直感的に解釈できる [14,16]．これは,「なぜその科目がキャリア形成に有効なのか」を説明する上で極めて重要であり，データ駆動型キャリア支援における透明性と納得性の向上に寄与する [5,13]．

また本研究では，提案手法の有効性を厳密に検証するため，キャリアと科目の関連強度や履修確率を制御可能な合成データセットを独自に生成する．実データは個人情報保護などの理由で利用が難しく，そもそも「どの科目がキャリアに影響したか」という真の正解（Ground Truth）が観測できない [13]．そのため，合成データによりモデルが重要科目を適切に識別できているかを純粋に評価することが可能となる．さらに，データ規模やノイズ率を段階的に変化させることで，提案手法が現実の教学データに近い多様な状況でどの程度の安定性・頑健性を示すかを多角的に分析する．

加えて，本研究はアルゴリズム開発にとどまらず，大学のキャリアセンターや教務担当者が実際に利用できるよう，科目の重要度係数を視覚的に出力するキャリアパス分析ツールの実装も目的に含める．Python 環境を必要としない GUI アプリケーションとして提供することで，専門知識を持たない職員でも容易に扱うことができ，教学データの利活用を実務レベルに引き上げることが可能となる [1,6]．

以上の取り組みにより，本研究は「解釈性」「安定性」「頑健性」を兼ね備え，教学データの制約に左右されずに運用可能なキャリアパス推薦モデルの確立を目指す [3,5]．これは従来の学修支援にとどまらず，学生の将来設計やキャリアデザインに寄与する新たな学習支援の枠組みを提示するものであり，高等教育機関におけるデータ利活用の新しい方向性を示すことを最終的な目的とする [5,6]．

§ 1.3　本論文の概要

本論文は次のように構成される．

第 1 章 本研究の背景と目的を述べる．背景では教学データ分析の動向と従来手法の課題について説明し，目的ではキャリアパス推薦における本研究の位置付けを示す．

第 2 章 教学 IR やラーニングアナリティクスなどの教育データ活用の枠組みを整理し，さらに情報推薦システムや機械学習アルゴリズムなど，本研究に関連する基礎技術について述べる．

第 3 章 本研究で採用するロジスティック回帰分析および L1 正則化による特徴選択の理論的基盤を示し，キャリアパス予測モデルとして適用する際の考え方について述べる．

第 4 章 提案手法の詳細について説明する．合成データの生成方法，キャリアパス推薦アルゴリズム，および分析結果を出力するためのシステム実装について述べる．

第 5 章 合成データを用いた数値実験の設定と評価指標を示し，予測精度・推薦内容・安定性の観点から提案手法の有効性を検証する．

第 6 章 本研究の成果をまとめ，今後の課題と展望について述べる．

第2章

教育データの利活用と関連技術

§ 2.1　教学 IR とラーニングアナリティクス

近年，大学教育を取り巻く環境は急速に変化し，教育の質保証や学習成果の可視化が強く求められている [1,5,7]．このような状況では，経験則や個人の勘のみに依存した教育改善には限界があり，大学に蓄積されたデータに基づいて現状を把握し，改善策を検討するデータ駆動型の意思決定が重要になる [5]．こうした流れを背景として，大学に蓄積されたさまざまな教学データを収集・分析し，教育改善や大学経営に活用する教学 IR（Institutional Research）や，Learning Management System (LMS) 等に蓄積される学習ログを用いて学習過程を分析するラーニングアナリティクスの重要性が高まっている [1,5,6]．

教学 IR は，成績や履修情報といった比較的マクロなデータに加え，入試情報や進路情報なども含めて統合的に扱い，カリキュラム設計や学生支援，大学経営の意思決定に役立てることを目的としている [1,5]．例えば，特定科目の不合格率の推移や履修者数の変化を把握し，カリキュラムの再設計や開講計画の最適化に反映するなど，組織レベルの改善に結び付けやすい点に特徴がある [1]．一方，ラーニングアナリティクスは，LMS へのアクセスログ，課題提出履歴，小テスト結果，教材閲覧履歴などのミクロな学習データを対象とし，学習状況の可視化やリスクの早期発見，学習方略の改善支援などに利用される [6,7]．たとえば，学習行動が停滞している学生群を早期に検出し，学修相談や学習支援へつなげるといった運用が想定される [6]．図 2.1 に教学ビッグデータの構成例を示す．

教学データの例を表 2.1 に示す．入学前の出身高校や入試成績から，在学中の履修登録，授業への出欠状況，課題提出状況，GPA，さらには卒業後の満足度調査に至るまで，学生一人ひとりに関する多様な情報が時系列に沿って蓄積されていることがわかる．このように，教学データは従来の「成績表」にとどまらず，学習過程全体を記述する豊富な情報源となっている [5]．また，教学 IR が扱うデータは部局や担当部署ごとに分散している場合が多く，データ統合（ID 連携や欠損処理，定義の統一）を含めた運用設計が分析の成否に直結する [1,4]．

一方で，教学データは一般的なビッグデータとは異なる特徴を持つことが指摘されている [1,3,5]．主な特徴を以下に整理する．

1．対象人数は比較的少ないが，データの種類は多様である

一つの大学に限れば在籍者数は数千〜数万人規模であり，通信事業者や大型 EC サイトなどと比べると対象人数は多くない．その一方で，一人の学生に関して収集されるデータの種類は年々増加しており，成績情報，授業出欠，LMS ログ，図書館利用履歴，

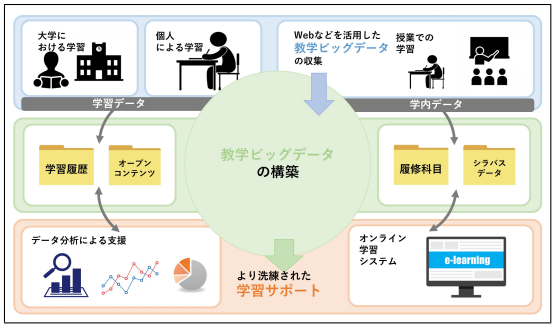

図 2.1: 教学ビッグデータ・アナリティクス

表 2.1: 教学データの例

取得時期	教学データ	内容
入学前	出身高校	課程差別，判定値，etc.
	入試情報	入試区分，成績
	入学前学習	取組状況，提出物
入学時	導入教育	オリエン欠席，テスト結果，etc.
各セメスター	履修登録	履修科目
	授業	出欠状況，課題提出，etc.
	学生生活	部活，アルバイト，etc.
	成績	科目成績，GPA，etc.
4 年次	就職活動	活動履歴，内定状況
卒業後	卒業後	満足度，アンケート

各種アンケートなど，多様な性質のデータが混在している [1]．このため，分析ではデータの意味（教育的解釈）と統計的処理（前処理やモデル化）を両立させる必要がある [3]．

2．履修データは本質的に疎である

学生数に対して提供される科目数が多いため，履修履歴を行列として表現すると，多くの要素が「未履修（0）」となる疎行列となる．この性質は，類似度に基づく推薦手法などにとって不利に働く [3]．また，科目は年度や担当教員により内容が変化し得るため，同一名称であっても学習成果が一様でない可能性がある [5]．

3．学習行動にはノイズが多く含まれる

興味本位の履修や友人に誘われた授業の受講など，キャリア形成と直接関係しない行動が多数含まれる．そのため，単純に履修パターンを比較しても，意味のある因果関係を見つけ出すことは容易ではない [3,5]．教育データ分析では「相関が見えること」と「教育的に意味があること」が一致しない場合がある点が重要である [3]．

4．個人情報性がきわめて高い

教学データは学生個人の能力や関心，行動傾向が詳細に反映されたデータであり，匿名性が低い．このため，データの収集・分析・共有に際しては，個人情報と同等以上の慎重な取り扱いが必要となる [13]．特に，分析結果を学生支援へ還元する際には，過度な監視や不利益なラベリングにつながらない設計・説明責任が求められる [13]．

さらに，日本の大学においては，専門人材の不足や情報公開文化の未成熟，財政的制約などにより，教学データ分析の成果を十分に教育現場へ還元できていないという課題も指摘されている [4]．したがって，現実的な制約（少人数，疎行列，ノイズ，運用体制）を前提としつつ，教育改善やキャリア支援に直結する具体的な分析・システム設計を行うことが求められている [5,6]．本研究は，教学データを用いてキャリアパスに関連する科目を同定し，学生や教職員にとって解釈しやすい形で提示するモデルを構築することで，教学 IR／ラーニングアナリティクスの成果を「意思決定に使える形」に変換する一例を提示することを目的とする [5,13]．

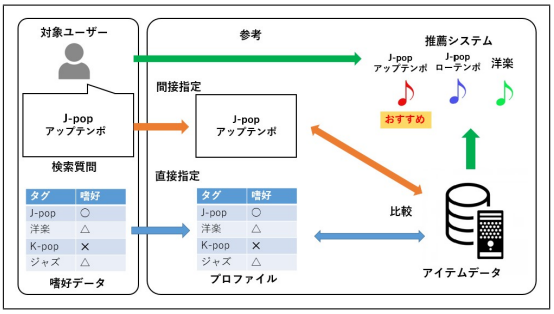

図 2.2: 内容ベースフィルタリング

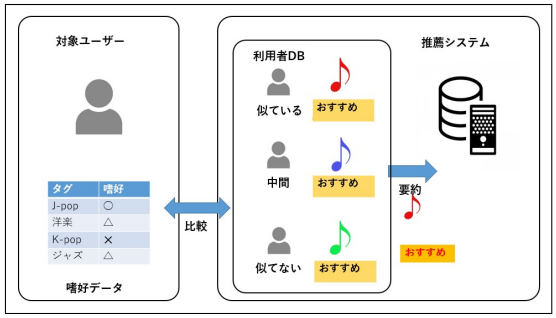

図 2.3: 協調フィルタリング

§ 2.2　情報推薦システム

現代の情報社会においては，インターネット上に膨大な情報が流通しており，利用者はその中から自分にとって有用な情報を取捨選択しなければならない．この問題を解決する技術として，情報推薦システム（Recommender System）が広く利用されている [10]．情報推薦システムは，利用者の嗜好や行動履歴をもとに，利用者の関心に合致すると考えられるアイテム（商品，記事，科目など）を自動的に提示することを目的とする [10,11]．教育分野においては，科目推薦や学習コンテンツ推薦に加え，履修計画支援や学修到達のガイダンスなどへの応用が期待される [10]．

情報推薦の代表的な手法として，内容ベースフィルタリングと協調フィルタリングの 2 種類が知られている [10,12]．

内容ベースフィルタリング（Content-based Filtering）

アイテムの内容に基づいて推薦を行う手法であり，アイテムごとの特徴量（ジャンル，キーワード，説明文など）と，利用者が過去に高く評価したアイテムの特徴量を比較することで，類似性の高いアイテムを推薦する [10]．図 2.2 に概要を示す．教育分野においては，科目のシラバス情報やキーワード，担当教員などを特徴量として用いることで，内容的に類似した科目を推薦することが可能である．一方で，科目内容を定量的なベクトルとして表現するには高度な自然言語処理や特徴設計が必要であり，特に実習科目や学際科目などの扱いが難しいという課題がある [10]．また，内容ベースは「既に好んだ内容に近いもの」を推薦しやすく，多様性が低下し得る点にも注意が必要である [10]．

協調フィルタリング（Collaborative Filtering）

利用者間またはアイテム間の類似性に基づいて推薦を行う手法である [10,11]．「似た嗜好を持つ利用者は似たアイテムを好む」という仮定に立ち，新たな利用者に対しては，過去の類似利用者が高く評価したアイテムを推薦する [12]．図 2.3 に概要を示す．教育分野では，履修履歴をもとに「似た履修パターンを持つ先輩」が受講していた科目を推薦する形で応用されている．協調フィルタリングはアイテム側の特徴設計が不要で汎用性が高い一方，データの疎性やノイズの影響を強く受ける [3,10]．

協調フィルタリングには，ユーザ・アイテム行列を直接参照して類似度計算を行うメモリベース法と，行列分解や潜在因子モデルなどを事前学習して用いるモデルベース法が存

在する [10]．メモリベース法は実装が容易でデータ追加に柔軟である一方，大規模・疎なデータでは類似度が安定せず精度低下を招きやすい [3,10]．モデルベース法は高速で高精度な推薦が可能であるが，モデルの学習コストが高く，データ更新のたびに再学習が必要となる [10]．教育データのように運用上の制約が大きい領域では，推薦精度だけでなく「更新容易性」「説明可能性」「運用負担」も重要な評価軸となる [13]．

さらに，教育分野において協調フィルタリングを用いる場合には，次のような課題が顕著である [3,13]．

- 履修履歴行列が疎であり，学生間・科目間の類似度を安定して算出することが難しい．
- 興味本位の履修や偶発的な履修が多く，キャリア形成と無関係なノイズが推薦結果に影響する．
- 「なぜその科目が推薦されたのか」という説明が,「似た先輩が履修していたから」という相関ベースの根拠にとどまり，因果的な解釈ができない．
- 新入生や転学部生などデータが少ない学生に対しては適切な推薦を行いにくい（コールドスタート問題）．

本研究では，これら協調フィルタリングの構造的な限界を踏まえ，履修科目とキャリアとの関係性を明示的なモデルとして表現できる線形モデル，とくにL1 正則化ロジスティック回帰を用いたアプローチを採用する [14,15]．このような「係数に基づく説明」は，教育現場で要求される説明責任や合意形成（学生・教職員の納得）に資する [13]．また，L1 正則化により不要な科目の係数がゼロに縮約されるため，ノイズの多い履修データに対しても頑健な推定が期待できる [15]．本研究は，推薦精度の最大化よりも，現実的なデータ制約下で安定して動作し，かつ推薦根拠を提示できるシステム設計を重視する点に特徴がある [3,5]．

§ 2.3　機械学習アルゴリズム

本節では，本研究における比較対象となる代表的な機械学習アルゴリズムの特徴を整理し，学習モデルの性質と教学データとの適合性について考察する．特に，本研究が目指す「解釈性の高い推薦」「ノイズや疎データに対して安定した推定」に対して，どの手法が適しているかを明確化することを目的とする [3,5,13]．

一般に機械学習モデルは，入力特徴量（本研究では履修科目や成績など）から出力（本研究では業界ラベル）を予測する写像として捉えられる．この写像の形が単純で，人間が追跡できる構造を持つものが線形モデルであり，複雑な条件分岐や相互作用を内部に多数含むものが非線形モデルである．教育支援の文脈では，予測精度と同程度に「なぜその結果になったのか」を説明できることが重視されるため，モデルの解釈性は重要な評価軸となる [13]．

線形モデルと非線形モデルの基本的な考え方

線形モデルは，入力特徴量の重み付き和によって出力を決めるモデルである．例えば，履修科目をベクトル $\mathbf{x}$（履修有無，単位数，成績などで表現）とし，ある業界への就職確率を p とすると，ロジスティック回帰では $\mathbf{x}$ の各成分（各科目）が出力に与える影響が係数 $\mathbf{w}$ として表現される [14]．このとき,「どの科目がどの程度プラス（あるいはマイナス）に働いたか」を係数の符号と大きさとして説明できるため，推薦理由の提示と相性が良い．さらに，L1 正則化を組み合わせると，係数の一部が厳密に 0 になり，多数の科目の中からキャリアに関係する科目のみを自動的に抽出できる [15]．これは，科目数が多く履修行動にノイズが混入しやすい教学データにおいて特に有効である．

一方，非線形モデルは，特徴量の組合せや相互作用を複雑に表現できるモデルであり，代表例として決定木系（ランダムフォレスト，勾配ブースティング，LightGBM 等）や，カーネル SVM，ニューラルネットワークなどが挙げられる．これらは複雑な境界を学習できるため高い予測性能を示す場合があるが，内部で多数の分岐や木の集合・段階的更新が行われるため,「この学生に対してなぜこの科目を薦めるのか」を係数のように一意に示すことが難しい [10, 13]．また，データ規模が小さい・疎である場合には，学習が不安定になり過学習を招く可能性も指摘されている [3]．

ここで本研究における「解釈性」とは，推薦（あるいは予測）結果に対して，どの特徴量（科目）がどの程度寄与したかを人間が追跡し，説明できる程度を指す [13]．教学データはノイズ・欠損・利用者数の偏りなどの特徴を持つため，高度な表現力よりも，むしろデータ品質の低さに対して頑健に動作するか，そして推薦根拠を人間に説明可能かといった観点の方が重要となる [3, 5]．

本研究の設定（履修科目→業界予測）に沿った各モデルの位置づけ

本研究では，学生の履修科目情報から就職先の業界を予測し，その予測に寄与した科目を根拠として推薦理由を提示することを目的とする．この目的に照らすと，次の点がモデル選定上の重要条件となる．

- 科目数が多く，履修行列が疎であっても学習が破綻しにくいこと（疎性・小規模性への頑健性）．
- ノイズ（興味本位の履修など）を含んでも推定が極端にぶれないこと（安定性）．
- 推薦理由を「どの科目が効いたか」として説明できること（解釈性・透明性）．

線形モデルのうち，ロジスティック回帰は二値分類に適した標準的手法であり,「当該業界に進む確率」を確率値として出力できる点がキャリア支援の文脈と整合する [14]．さらに，L1 正則化ロジスティック回帰は，重要でない科目の係数を 0 にすることで特徴選択を同時に実現し,「重要科目の抽出」と「説明可能な予測」を同一枠組みで行える [15]．このため，本研究の中心モデルとして位置付けられる．

一方，ランダムフォレストは多数の決定木の多数決・平均によるモデルであり，非線形な関係を表現できるが，木の集合として意思決定が形成されるため，単一の係数のような

形で推薦根拠を提示することが難しい．LightGBM は勾配ブースティングに基づき高い性能を示すことが多いが，逐次的に木を追加して誤差を補正する構造上，判断の根拠が分散しやすい．SVM（RBF）は高次元空間での分離境界を学習できる一方，境界の形状が直感的に解釈しにくく，教育現場での説明には工夫を要する．ナイーブベイズは確率モデルとして概略の説明は可能であるが，特徴量独立の仮定が強く，履修科目間の関係を十分に反映できない場合がある [10].

以上を踏まえ，本研究で扱う教学データの特性（疎性・ノイズ・少量）と，キャリア支援システムに要求される要件（解釈性・根拠の提示）を総合すると，L1 正則化ロジスティック回帰は最も適したモデルであると位置づけられる．

これらを踏まえ，各モデルの特徴を「解釈性」「線形／非線形」の 2 軸で整理したものを表 2.2 に示す．表 2.2 より，本研究が求める「説明可能性」を満たすのは線形モデルであるロジスティック回帰であり，中でも L1 正則化を適用することで，多数の科目の中からキャリアに関連する特徴量のみを自動選択できる点が優れている [15]．これにより，どの科目が就職先の業界選択に強く寄与しているかを重み係数の大小として直接的に提示することができ，推薦根拠の可視化に大きく貢献する [14].

一方，ランダムフォレストや Light Gradient Boosting Machine (LightGBM) といった非線形モデルは高精度であるものの，個々の科目が結果に与える影響度を明確に示すことが難しく，教育現場での説明責任（アカウンタビリティ）の観点で課題が残る [13]．また，データ量が少ない場合には過学習しやすく，履修データのような疎なデータに対してはモデルの安定性が低下する傾向がある [3].

表 2.2: 比較対象モデルの解釈性と線形性

モデル	解釈性	線形性
L1 ロジスティック回帰	高（係数で説明）	線形
ランダムフォレスト	低（木の集合で複雑）	非線形
LightGBM	低（内部が複雑）	非線形
SVM（RBF）	低（境界が非直感的）	非線形
ナイーブベイズ	中（確率で概略）	近似線形

第3章

キャリアパス予測モデルの基礎理論

本章では，本研究の中核となるキャリアパス予測モデルの理論的基盤を整理する．第 2 章で述べたように，教学データは「科目数に比べて履修が少ない（疎である）」「興味本位の履修などノイズが混入する」「個人情報性が高い」といった特徴を持つ．本研究は，このような制約の強いデータ環境でも安定して動作し，かつ推薦理由を説明可能な形で提示できることを重視する．そのため，本章ではロジスティック回帰を基礎として，L1 正則化による特徴選択（スパース化）を組み合わせたモデルの考え方を，段階的に詳しく述べる．

§ 3.1　ロジスティック回帰によるキャリアパス予測

本研究で扱う「キャリアパス予測」は，学生の成績（履修）情報から，特定の職種カテゴリに進む確率を推定する問題として定式化できる．たとえば「IT・通信系エンジニアに進むか否か」「モノづくり系エンジニアに進むか否か」といった形で，あるカテゴリに属するかどうかを二値で表すことができる．このような二値分類問題に対して，ロジスティック回帰は確率的なモデルとして自然であり，さらに係数が解釈しやすいという利点を持つ [14].

A. 入力データ（成績ベクトル）とラベルの定義

まず，学生 i の成績データを M 科目のベクトルとして

$$\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{iM})^{\mathrm{T}} \tag{3.1}$$

と表す．式 3.1 の各要素 x_{ij} は，学生 i が科目 j で得た成績（あるいは未履修）を表す．本研究では成績値を次の尺度で統一する：

表 3.1: 成績値の尺度

値	意味
5	秀
4	優
3	良
2	可
1	不可
0	未履修

表 3.1 に示すように，0 は欠損ではなく「履修していない」という意味を持つ値である．したがって，本研究では欠損補完として 0 を入れるのではなく，入力値として 0 を保持することで，履修データの疎性をモデルへ正しく反映させる．

次に，職種カテゴリを目的変数（ラベル）として定義する．二値分類の形で扱うため，ある職種カテゴリ（以降，対象カテゴリと呼ぶ）に進んだかどうかを

$$y_i \in \{0, 1\} \tag{3.2}$$

で表す．式 3.2 において $y_i = 1$ は「対象カテゴリに進んだ」，$y_i = 0$ は「それ以外」を意味する．

実際の入力データは，学籍番号をキーとして「成績データ」と「職種ラベル」を対応づけて扱う．以下に，形式を理解するための小さな例を示す（科目名は実データの一部を簡略化している）．

表 3.2: 学籍番号と成績データの対応例（抜粋）

学籍番号	数学 I	プログラミング 1	線形代数	卒業研究
1714001	4	5	3	5
1714002	3	4	5	5
1714007	3	4	0	5

表 3.2 では，たとえば学籍番号 1714007 の「線形代数」が 0 となっているが，これは「未履修」を意味する．この 0 は「データが欠けている」ではなく「履修していない」という行動そのものを表しており，本研究ではこの性質を重要視する．

表 3.3: 学籍番号と職種カテゴリラベルの対応例（抜粋）

学籍番号	職種カテゴリ
1714001	モノづくり系エンジニア
1714002	モノづくり系エンジニア
1714007	IT・通信系エンジニア

表 3.3 のように職種カテゴリが与えられるとき，特定のカテゴリを対象にした二値分類では，そのカテゴリに該当する学生を $y_i = 1$，それ以外を $y_i = 0$ として学習データを構成する．この対応づけにより,「どの科目の成績が，対象カテゴリへの進路と関連が強いか」を統計的に推定できるようになる．

B. 線形スコアと確率への変換（シグモイド関数）

ロジスティック回帰は，まず入力ベクトル $\mathbf{x}_i$ を重み付きに足し合わせた線形スコア

$$z_i = \mathbf{w}^{\mathrm{T}}\mathbf{x}_i + b \tag{3.3}$$

を計算する．式 3.3 で，$\mathbf{w} = (w_1, \ldots, w_M)^{\mathrm{T}}$ は各科目に対応する係数ベクトル，b はバイアス項である．ここで z_i は実数全体を取り得るため，このままでは確率として解釈できない．

そこでロジスティック回帰では，線形スコア z_i をシグモイド関数で $[0,1]$ に写像し，対象カテゴリへ進む確率 $\hat{p}_i$ を

$$\hat{p}_i = P(y_i = 1 \mid \mathbf{x}_i) = \sigma(z_i) = \frac{1}{1 + \exp(-z_i)} \tag{3.4}$$

として定義する．式 3.4 により，どのような z_i に対しても $\hat{p}_i$ は 0 から 1 の範囲に収まり，「対象カテゴリに進む確率」として扱える．

さらに，意思決定（分類）としては，確率がある閾値を超えたら対象カテゴリと判定する．典型的には 0.5 を用い，

$$\hat{y}_i = \begin{cases} 1 & (\hat{p}_i \geq 0.5) \\ 0 & (\hat{p}_i < 0.5) \end{cases} \tag{3.5}$$

とする．式 3.5 は分類規則を表すが，本研究では「確率」そのものも重要である．なぜなら，確率が高い／低いという情報は，推薦や説明（納得性）において役立つからである．

C. 係数の意味（解釈性）

ロジスティック回帰が本研究に適している大きな理由は,「係数 $\mathbf{w}$ を通じて科目の寄与を説明できる」点にある．式 3.3 の w_j が正であれば，科目 j の成績が高いほど z_i が増え，式 3.4 により確率 $\hat{p}_i$ も上がりやすい．逆に w_j が負であれば，科目 j の成績が高いほど対象カテゴリの確率が下がる方向に働く．したがって，係数の符号は「寄与の方向」，大きさは「寄与の強さ」を表すと解釈できる．

さらにロジスティック回帰は対数オッズの線形モデルとしても説明できる．式 3.4 を変形すると，

$$\log \frac{\hat{p}_i}{1 - \hat{p}_i} = \mathbf{w}^{\mathrm{T}} \mathbf{x}_i + b \tag{3.6}$$

が得られる．式 3.6 は「対数オッズ（logit）」が入力の線形結合で表されることを意味する．この形は，係数 w_j を「科目 j が 1 単位増えたときに対数オッズがどれだけ増えるか」として理解できるため，教育現場での説明（なぜその科目が重要なのか）に結びつけやすい．

D. 学習（最尤推定）と損失関数

次に，係数 $\mathbf{w}$ と b をどのように決めるかを述べる．ロジスティック回帰では，各データ点 $(\mathbf{x}_i, y_i)$ に対して，

$$P(y_i \mid \mathbf{x}_i) = \hat{p}_i^{\,y_i} (1 - \hat{p}_i)^{1 - y_i} \tag{3.7}$$

と表せる（ベルヌーイ分布）．式 3.7 に式 3.4 の $\hat{p}_i$ を代入することで，パラメータ $(\mathbf{w}, b)$ を含む確率モデルが得られる．

全データ m 件に対して独立と仮定すると，尤度（likelihood）は

$$L(\mathbf{w}, b) = \prod_{i=1}^{m} \hat{p}_i^{\,y_i} (1 - \hat{p}_i)^{1 - y_i} \tag{3.8}$$

となる．式 3.8 を最大化することが最尤推定であるが，積の形は扱いにくいため，通常は対数尤度

$$\ell(\mathbf{w}, b) = \sum_{i=1}^{m} \Big[y_i \log \hat{p}_i + (1 - y_i) \log(1 - \hat{p}_i) \Big] \tag{3.9}$$

を最大化する．式 3.9 は加算形になるため最適化に都合がよい．

さらに，最小化問題として扱うため，負の対数尤度（交差エントロピー損失）を

$$J(\mathbf{w}, b) = -\frac{1}{m} \sum_{i=1}^{m} \Big[y_i \log \hat{p}_i + (1 - y_i) \log(1 - \hat{p}_i) \Big] \tag{3.10}$$

と定義し，$J(\mathbf{w}, b)$ を最小化することで学習を行う．式 3.10 は「正解ラベルに高い確率を与えるほど損失が小さくなる」ように設計された関数であり，分類問題における標準的な目的関数として広く用いられる [14]．

ここまでで，ロジスティック回帰が「入力→線形スコア→確率→損失最小化」という流れで学習されることが明確になる．しかし教学データでは，科目数 M が多い一方で，各学生が履修する科目は限られ，さらにノイズも混入するため，式 3.10 のみを最小化すると過学習が起こりやすい．この問題への対処が，次節の正則化である．

§ 3.2　L1 正則化による特徴選択と頑健化

教学データにおける重要な特徴は,「科目が多い（高次元）」「履修は少ない（疎）」「ノイズが多い」という三点である．このとき，ロジスティック回帰をそのまま学習すると，多数の科目に小さな係数が割り当てられ，モデルが複雑化しやすい．結果として，特定のデータの癖（偶然の履修・偶然の高得点）を拾ってしまい，新しい学生に対して予測が不安定になる可能性がある．

A. L1 正則化付き目的関数

この問題に対して本研究では，L1 正則化（Lasso）を導入する [15]．L1 正則化では，式 3.10 に係数の絶対値和を罰則として加え，

$$J_{\mathrm{L1}}(\mathbf{w}, b) = -\frac{1}{m} \sum_{i=1}^{m} \Big[y_i \log \hat{p}_i + (1 - y_i) \log(1 - \hat{p}_i) \Big] + \lambda \sum_{j=1}^{M} |w_j| \tag{3.11}$$

を最小化する．式 3.11 において $\lambda \geq 0$ は正則化の強さを表すハイパーパラメータである．λ を大きくすると，係数がより強く 0 へ縮められ，モデルが単純化する．

式 3.11 が重要なのは，L1 罰則が「不要な係数をちょうど 0 にする」性質（スパース解）を持つ点である．つまり，科目数が多くても，本質的に必要な科目だけが係数として残り，関係の薄い科目はモデルから実質的に排除される．

B.「特徴選択」と「説明可能性」の関係

教育データでは，モデルが高精度であっても「なぜその推薦が出たのか」が説明できないと，学生や教職員が納得して利用しづらい．L1 正則化によって係数が 0 になる科目が明確に生じると,「この職種カテゴリに対して，どの科目が関係あり（係数が残る）で，どの科目が関係薄い（係数が 0）か」をはっきりと示せる．

さらに，係数が残った科目についても，式 3.3 と式 3.4 の関係から,「その科目の成績が高いと確率が上がるのか（$w_j > 0$）／下がるのか（$w_j < 0$）」という方向性まで説明できる．このように，L1 正則化は単に過学習を抑えるだけでなく,「説明のために必要な形」へモデルを整える役割も持つ．

C. 疎な入力（未履修=0）との整合

本研究の入力では未履修を 0 として保持する（表 3.1）．この設計は，線形モデルであるロジスティック回帰と相性がよい．なぜなら，式 3.3 において $x_{ij} = 0$ であれば，その科目 j の寄与は $w_j x_{ij} = 0$ となり,「履修していない科目はモデルに影響しない」という直感的な意味づけが成立するからである．

一方で，非線形モデルでは「0 という値そのもの」を複雑に解釈してしまう場合があり，データ量が少ない条件では予測が不安定になることがある．本研究では疎性を前提にしているため,「0 は未履修」を素直に反映しやすい線形モデルに優位性があると考える．

§ 3.3　本研究における適用アプローチ

本節では，これまで述べた理論を「キャリアパス推薦」という具体的な目的へ接続するための考え方を整理する．本研究は，最終的に学生へ推薦科目を提示するが，その基盤となるのは「職種カテゴリごとの重要科目を統計的に同定する」ことである．この同定結果（係数）は，推薦理由としても，教職員向けの分析材料としても利用できる．

A. 職種カテゴリごとのモデル化（One-vs-Rest の考え方）

実データでは職種カテゴリは複数存在する（表 3.3）．これを扱う方法の一つは多クラス分類であるが，本研究では「カテゴリごとの重要科目を明確に得る」ことを優先し，カテゴリ k ごとに二値分類を作る One-vs-Rest（OvR）の考え方を用いる．すなわち，元のラベルを $y_i \in \{1, \ldots, K\}$ としたとき，カテゴリ k に対する二値ラベルを

$$y_i^{(k)} = \begin{cases} 1 & (y_i = k) \\ 0 & (y_i \neq k) \end{cases} \tag{3.12}$$

で定義する．式 3.12 により，カテゴリ数 K 個の二値分類問題に分解できる．

この分解の利点は,「カテゴリ k に対する係数 $\mathbf{w}_k$」が得られ，カテゴリごとに重要科目の集合が直接読み取れる点にある．たとえば $w_{kj} > 0$ で大きい科目 j は,「カテゴリ k に進

む確率を押し上げる傾向が強い科目」として解釈できる．逆に $w_{kj} = 0$ の科目は，L1 正則化の結果として「カテゴリ k との関連が弱い」とみなされた科目である．

B. 「重要科目の抽出」と「推薦提示」のつながり

本研究では，モデルの係数をそのまま「重要科目」として扱えることがポイントである．ロジスティック回帰は式 3.3 と式 3.4 により，係数の方向と強さが確率へどのように反映されるかが明確である．そのため，推薦の際にも，単に科目名を列挙するのではなく,「なぜその科目が推薦されるか」を係数（寄与の方向・強さ）を根拠として提示できる．

また，成績尺度が 0–5（表 3.1）であるため，最大評価（5: 秀）を想定したときの影響を考えることもできる．たとえば，ある科目 j の係数が w_{kj} であるとき，その科目で高評価を得ることが確率へ与える影響を「スコア」として整理し，推薦の優先順位付けへ接続することが可能となる．このような「係数→説明→推薦」という流れが，協調フィルタリングのような相関ベース手法では難しかった点である．

本章では，(i) 成績データをベクトル $\mathbf{x}_i$ として表し（式 3.1），(ii) ロジスティック回帰により確率を推定し（式 3.3〜式 3.4），(iii) 尤度にもとづく学習を損失最小化として整理し（式 3.8〜式 3.10），(iv) L1 正則化により重要科目をスパースに抽出する（式 3.11），という一連の理論を段階的に述べた．さらに，職種カテゴリが複数ある場合には OvR による二値化（式 3.12）でカテゴリごとの重要科目を得る方針を示した．

次章では，これらの理論を実際の処理手順（学習・係数抽出・推薦提示・評価）へ落とし込む際の設計と実装方針，および提案手法を検証するための評価環境（合成データ生成を含む）について詳述する．

第4章

提案手法の詳細

本章では，本研究で提案するキャリアパス推薦手法について，システムとしての位置づけ，データ設計，学習・推薦アルゴリズム，および実装方針をまとめて述べる．第1章・第2章で述べたように，教学データは多様である一方で，疎性やノイズ，個人情報性などの制約が強い．そのため本研究では，単に高精度な予測を目指すだけでなく，(i) 推薦根拠を説明可能な形で提示できること，(ii) データ品質のばらつき（疎性・ノイズ・人数偏り）に対して安定して動作すること，の両立を重視して設計を行う．

§ 4.1　システム構成と合成データの生成

本研究の枠組みは，役割の異なる二つの系から構成される．第一は，学生および教職員が利用する「運用システム」である．運用システムは，成績（履修）情報と進路情報にもとづいてモデルを学習し，業界（職種カテゴリ）ごとの重要科目を抽出し，学生へ推薦を提示することで履修計画・キャリア形成を支援する．第二は，提案手法の動作検証・比較評価を目的とする「評価環境」である．

特に本節で述べる合成データ生成は，運用システムの内部機能ではなく，提案手法の妥当性を客観的に検証するために運用系とは独立して用意した「評価専用のデータ生成器」により実現する．この分離によって，運用側の実装を変更することなく，データ規模や疎性，ノイズ率といった条件を制御した再現可能な実験が可能になる．

A. システム全体像（運用系と評価系の分離）

運用システムは，学生・教職員の意思決定を支援する実務的な機能を担う．学生に対しては，志望する業界（職種カテゴリ）に対して推定された重要科目を提示し，履修の候補を推薦する．教職員に対しては，業界ごとの重要科目を整理して出力することで，カリキュラム改善やキャリア支援施策の検討を支援する．

一方で評価環境は，研究としての検証基盤であり，運用システムとは独立に動作する．ここでは研究者が正解構造（本当に重要な科目集合）を設計した上で成績データと進路ラベルを生成し，提案手法がその正解構造をどの程度回復できるか（重要科目を適切に抽出できるか）を検証する．

表 4.1: 運用システムと評価環境の役割分担

区分	目的・役割
運用システム	実データ（成績・進路）に基づきモデルを学習し，重要科目の抽出と推薦提示を行う．学生・教職員に対する実務的な支援を目的とする．
評価環境（評価専用）	正解構造を含む合成データを生成し，疎性・ノイズ率・規模などの条件を制御して，提案手法の動作検証・比較評価を可能にする．

【ここに図表：運用系と評価系の分離を示す構成図（概念図）】

B. データフローと入力形式

運用システムが扱う入力は，学生を単位とした成績（履修）情報と進路（職種カテゴリ）情報である．成績は科目ごとの評価値を並べたベクトルとして表し，進路は学生ごとに一つのラベルとして表す．両者は学籍番号などの識別子で対応づけられ，教師あり学習のデータセットが構成される．

本研究では，成績値を 0 から 5 の整数尺度で統一し，5: 秀，4: 優，3: 良，2: 可，1: 不可，0: 未履修として扱う．0（未履修）は欠損ではなく「履修していない」という意味を持つ入力値であるため，欠損補完ではなく値として保持する．この設計により，教学データが持つ疎性をそのまま入力に反映できる．

成績値は表 3.1 に示す尺度（0：未履修〜5：秀）に従う．

また，評価環境は運用系と独立に動作し，運用系が受け取る入力と同等の形式（成績ベクトルと進路ラベル）を外部から生成して供給する．すなわち評価環境が生成する合成データは,「運用システムに投入するテスト入力」として機能し，同一の学習・推薦処理に対して条件を変えた検証を繰り返せるようにする．

【ここに図表：入力データの流れ（成績・進路の結合〜学習〜出力）】

C. 合成データ生成（評価専用）

実データのみの評価では,「どの科目が本当にキャリアに寄与したか」という正解（Ground Truth）が観測できず，重要科目抽出の妥当性を定量的に検証しにくい．そこで評価環境では，研究者が正解構造を設計した上で合成データを生成し，提案手法が正解構造をどの程度回復できるかを検証可能にする．

学生数を N, 科目数を M, 職種カテゴリ数を K とし, 成績行列を $\mathbf{G} \in \{0,1,2,3,4,5\}^{N\times M}$, ラベルを $\mathbf{y} \in \{1,\dots,K\}^N$ と定義する．各カテゴリ k に対して，重要度の異なる科目集合（コア科目，準コア科目，無関係科目）を割り当て，集合ごとに成績分布を変えることでカテゴリの特徴が埋め込まれるよう設計する．

さらに，疎性を再現するため無関係科目では未履修（0）が多数となる分布を設定する．加えて現実の教学データに見られる「気まぐれな履修」や「ばらつき」を再現するため，成

績行列の一部をランダムに置換するノイズ注入を行う．ノイズ率を制御することで，提案手法の頑健性（ノイズ条件下での重要科目抽出の安定性）を評価できる．

$$G_{ij} \leftarrow \mathrm{Unif}\{0,1,2,3,4,5\} \quad \text{for } (i,j) \in \mathcal{I}_{\text{noise}} \tag{4.1}$$

式 4.1 はノイズ注入の操作を表す．ここで $\mathcal{I}_{\text{noise}}$ はノイズを付与する要素集合であり，その大きさがノイズ率に対応する．この操作により，コア科目にも低成績が混入する一方，無関係科目にも高成績が混入するため，単純な相関にもとづく方法では重要科目が見えにくい条件が生成される．その結果，提案手法が「本質的な重要科目」をどの程度安定して抽出できるかを，条件統制された形で検証できる．

【ここに図表：合成データの検証可視化（ヒートマップや分布の例)】

§ 4.2　キャリアパス推薦アルゴリズム

本節では，運用システムが行う学習・推定・推薦提示の手順を述べる．提案手法は，L1 正則化ロジスティック回帰を中核とし，(1) 職種カテゴリごとの分類モデルを学習する段階，(2) 学習済み係数にもとづいて推薦候補科目を抽出・順位付けする段階，の二段階からなる．本研究では推薦理由を説明可能な形で提示することを重視するため，モデル内部の寄与を係数として取り出せる設計を採用する．

A. 問題定式化

学生 i の成績（履修）ベクトルを $\mathbf{x}_i \in \{0,1,2,3,4,5\}^M$ とし，職種カテゴリラベルを $y_i \in \{1,\ldots,K\}$ とする．多クラス分類を直接扱うことも可能であるが，本研究では職種カテゴリごとに「そのカテゴリに進むか否か」を判定する形に分解することで，カテゴリごとの重要科目をより明示的に抽出できるようにする．具体的には One-vs-Rest 形式として，カテゴリ k に対する二値ラベル

$$y_i^{(k)} = \begin{cases} 1 & (y_i = k) \\ 0 & (y_i \neq k) \end{cases} \tag{4.2}$$

を定義する．式 4.2 により，カテゴリ数 K 個の二値分類問題に分解される．この分解は，カテゴリごとに「寄与の大きい科目」を係数として解釈する上で都合がよい．

B. 学習手順と前処理方針

成績情報と進路情報は，学籍番号などの識別子で対応づけて学習データを構成する．未履修は 0 として明示的に保持し，欠損値は原則として扱わない設計とする．また，本研究で扱う成績尺度はすでに 0–5 の共通スケールであり，0（未履修）が「効果なし」に対応する明確な意味を持つため，標準化のような変換は必須ではない．この点は，疎性を保持したまま線形モデルに入力できる利点でもある．

カテゴリ k のモデルにおける係数ベクトルを $\mathbf{w}_k$，バイアスを b_k とする．予測確率は第3章で述べたロジスティック関数により与えられ，係数 $\mathbf{w}_k$ の符号と大きさがカテゴリ k への寄与を表す．さらに L1 正則化を用いることで，寄与が小さい科目の係数が 0 に縮約され，重要科目が自動的に選択される（スパース化）．この性質は，科目数が多く疎でノイズを含む教学データに対して，過学習を抑えつつ解釈可能性を高める上で重要である．

また，職種カテゴリごとの人数が偏る場合，少数カテゴリの学習が不利になることがある．この問題に対しては，学習時にカテゴリ比率を考慮する重み付けを行うなど，不均衡データに配慮した設定を行う．さらに評価では，カテゴリ比率を維持した分割にもとづく交差検証を採用し，条件に依存しない安定した比較を可能にする．

C. 推薦スコアと推薦候補の抽出

学習後，カテゴリ k における係数 $\mathbf{w}_k = (w_{k1}, \ldots, w_{kM})$ が得られる．本研究では，係数が正である科目を「当該カテゴリへの進路に正の寄与を持つ科目」とみなし，推薦候補として抽出する．学生 i に対しては，未履修（成績 0）の科目に限定して推薦することで，すでに履修済みの科目を重複して提示しない設計とする．

推薦候補科目 j の優先度を決めるため，係数に基づくスコアを定義する．成績の最大値が 5（秀）であることを踏まえ,「もしその科目で最高評価を得た場合の影響」を想定した単純スコアとして,

$$S_{ij}^{(k)} = w_{kj} \times 5.0 \tag{4.3}$$

を用いる．式 4.3 の $S_{ij}^{(k)}$ は，学生 i がカテゴリ k を志望する場合に，科目 j を履修し高評価を得ることの理論上の寄与の大きさを表す．なお，係数そのもの w_{kj} による順位付けと本質的に同等であり，運用上は提示の分かりやすさのためにスコア化して扱う．

さらに実運用では，推薦スコアの大小だけで科目を提示すると，卒業要件や必修条件と整合しない科目が上位に現れることがある．そこで本研究では，卒業要件（必修・選択必修・自由科目など）に配慮したフィルタリングと優先順位付けを行い，学生が実際に履修計画へ落とし込みやすい提示となるように設計する．具体的には，(i) 必修科目で未履修のもの，(ii) 単位不足のカテゴリに属する科目，(iii) スコア上位の科目，の順に優先して提示する方針とする．

【ここに図表：推薦処理のフローチャート（志望カテゴリ入力→未履修抽出→スコア計算→要件フィルタ→提示)】

D. 比較手法

提案手法の有効性を検証するため，代表的な機械学習手法を比較対象とする．比較においては，入力形式や評価手順を可能な限り統一し，モデル構造の違いが結果へ与える影響を観察できるようにする．比較対象には，非線形モデル（例：決定木ベースのアンサンブル，カーネル法）や，確率モデル（例：単純ベイズ）などを含め，解釈性・精度・安定性の観点から総合的に評価する．特に，非線形モデルは高精度が期待される一方で，どの科目がどの程度寄与したかを学生や教職員に説明することが難しい場合がある．したがって本研究では，精度のみならず説明可能性を含めた観点で比較を行う．

§ 4.3　関連資料の提供とシステム実装方針

本節では，運用システムとしての実装方針と，成果物の構成を述べる．本研究の成果は，(i) 学生向けの推薦提示機構，(ii) 教職員向けの分析結果提示機構，(iii) 研究としての評価環境（合成データ生成と実験の再現基盤）の三つを中心に整理できる．ここで評価環境は研究用の検証基盤であり，運用系とは分離される点に注意する．

A. 学生向け提示機構（Web インターフェース）

学生向け提示機構は，学生が自身の志望カテゴリを入力し，推薦科目とその根拠を確認できるように設計する．提示においては，単に科目名を列挙するのではなく,「なぜその科目が推薦されるのか」を理解しやすい情報を併記することが重要である．本研究では，係数（寄与の方向と強さ）を推薦根拠の中核とし，必要に応じてスコア（式 4.3）やカテゴリ上位の重要科目一覧を併せて提示できる形を想定する．

また，学生が推薦を履修計画へ落とし込むためには，科目の概要（授業内容，キーワード，履修条件など）にアクセスできることが有用である．そのため，推薦科目からシラバス情報へ到達できる導線を用意し，推薦と科目理解を連続した体験として提供する方針とする．

【ここに図表：学生向け画面（推薦リスト表示）のイメージ】

B. 教職員向け提示機構（分析結果の出力）

教職員向け提示機構では，業界（職種カテゴリ）ごとの重要科目を整理して出力する．ここで重要科目とは，L1 正則化により係数が 0 へ縮約されずに残った科目，および係数が大きい科目を指す．これにより，カテゴリごとの特徴（どの科目が進路と関連しているか）を可視化でき，カリキュラム設計やキャリア支援の検討材料として利用できる可能性がある．

また，教職員が専門的なプログラミング環境を前提とせずに利用できるよう，入力データを指定して分析を実行し，結果を一覧として出力できる形が望ましい．本研究では，操作手順を簡素化し，導入コストを抑えた形での提供を想定する．

【ここに図表：教職員向け画面（分析実行と結果一覧）のイメージ】

C. 再現性と実験条件の管理（評価環境）

研究としての妥当性を担保するため，評価環境では再現性を重視する．具体的には，合成データ生成における乱数シードを固定し，同一条件であれば同一データが得られるようにする．また，データ規模（学生数，科目数，カテゴリ数）やノイズ率などの条件を明示的に管理し，提案手法と比較手法の差が，条件の違いに起因していないことを確認できるようにする．

評価環境は運用系とは独立したプログラムとして設計されるため，運用系の入力形式と同等のデータ（成績・進路）を生成して供給する形となる．この構造により，運用系の学習・推薦処理をそのまま用いて，テスト入力のみを差し替えた比較評価が可能となる．

【ここに図表：実験条件（規模・ノイズ率）を整理した表】

本章では，提案手法を「運用システム」と「評価環境」に分離して設計する方針を示し，運用系が扱う入力形式（成績 0–5 の尺度と進路ラベル）と学習・推薦の流れ，および評価専用の合成データ生成の位置づけを述べた．次章では，本章で述べた設計に基づき，評価指標と実験設定を整理し，提案手法が精度・安定性・説明可能性の観点でどのような挙動を示すかを検証する．

第5章

数値実験と考察

§ 5.1　実験の概要と評価指標

§ 5.2　実験結果

第6章

おわりに

謝辞

本研究を遂行するにあたり，多大なご指導と終始懇切丁寧なご鞭撻を賜った富山県立大学工学部電子・情報工学科情報基盤工学講座の奥原浩之教授，António Oliveira Nzinga René 講師に深甚な謝意を表します．また，数値実験の実施にあたりご助力下さった, 富山県立大学電子・情報工学科 3 年生の堀田遥斗氏に感謝いたします．最後になりましたが，多大な協力をしていただいた研究室の同輩諸氏に感謝致します.

2022 年 2 月

滝沢 光介

参考文献

[1] 松田岳士, 渡辺雄貴, “教学 IR，ラーニング・アナリティクス，教育工学”, 日本教育工学会論文誌, Vol. 41, No. 3, pp. 199–208, 2017.

[2] 近藤伸彦, 畠中利治, “学士課程における大規模データに基づく学修状態のモデル化”, 教育システム情報学会誌, Vol. 33, No. 2, pp. 94–103, 2016.

[3] Marceron, A., Blikstein, P. and Siemens, G., “Learning Analytics: From Big Data to Meaningful Data”, *Journal of Learning Analytics*, Vol. 2, No. 3, pp. 4–8, 2015.

[4] 浅野茂, “データベースの構築と IR の課題”, 高等教育研究, 第 19 集, 2016.

[5] Daniel, B., “Big Data and Analytics in Higher Education: Opportunities and Challenges”, *British Journal of Educational Technology*, Vol. 46, No. 5, pp. 904–920, Sept 2015.

[6] Ifenthaler, D. and Yau, J. Y.-K., “Utilising learning analytics to support study success in higher education”, *Educational Technology Research and Development*, Vol. 68, pp. 1961–1989, 2020.

[7] Long, P. and Siemens, G., “Penetrating the Fog: Analytics in Learning and Education”, *EDUCAUSE Review*, Sept/Oct 2011.
https://er.educause.edu/articles/2011/9/penetrating-the-fog-analytics-in-learning-and-education, 閲覧日 2025.12.19.

[8] 鶴田美保子, “大学生の就職活動を成功させる要因”, 金城学院大学論集 人文科学編, 第 15 巻第 1 号, pp. 109–119, Sept 2018.

[9] 畔津憲司, “九州市立大学経済学部 2012 年度卒業生の卒業後進路及び就職活動実態等に関する調査報告”, 北九州市立大学『商経論集』, 第 49 巻第 1・2 号, pp. 75–120, Dec 2013.

[10] Ricci, F., Rokach, L. and Shapira, B. (eds.), *Recommender Systems Handbook (2nd ed.)*, Springer, 2015.

[11] Resnick, P. and Varian, H. R., “Recommender systems”, *Communications of the ACM*, Vol. 40, No. 3, pp. 56–58, 1997.

[12] 神嶌敏弘, “推薦システムのアルゴリズム”,
`https://www.kamishima.net/archive/recsysdoc.pdf`, 閲覧日 2025.12.19.

[13] Drachsler, H. and Greller, W., “Privacy and analytics: it’s a DELICATE issue. A Checklist for Trusted Learning Analytics”, *Proc. 6th International Learning Analytics & Knowledge Conference (LAK 2016)*, pp. 89–98, 2016.

[14] Hosmer, D. W., Lemeshow, S. and Sturdivant, R. X., *Applied Logistic Regression (3rd ed.)*, Wiley, 2013.

[15] Tibshirani, R., "Regression Shrinkage and Selection via the Lasso", *Journal of the Royal Statistical Society: Series B (Methodological)*, Vol. 58, No. 1, pp. 267–288, 1996.

[16] Baker, R. S. and Inventado, P. S., "Educational Data Mining and Learning Analytics", in *Learning Analytics*, pp. 61–75, Springer, 2014.