スライド 1

このチャプターでは、機械学習の基礎と展望を、学習します。

スライド 4

このセクションでは、機械学習とは何か、また機械学習の種類を、整理します。

スライド 5

機械学習とは、データの背後に潜む規則性を見いだし、それを、機械学習 モデルと、呼ばれる数式として、定量化したり将来の予測に用いたりする方法の総称です。いりょう、電力、通信、流通業、製造業、金融業、農業など分野を問わず応用されており、今後もますます広がり続けるといわれています。例えば、インターネットのショッピングサイトで、購入履歴や閲覧履歴のデータを、使用し、機械学習 モデルで、分析して、そのユーザーが興味を持ちそうな商品をお勧めとして提示する。それにより、商品の購買を、促進し、売上を増加させる効果も、機械学習のひとつです。

機械学習には大きくわけて、きょうしありがくしゅう、きょうしなしがくしゅう、強化学習の3つのカテゴリーがあります。

きょうしありがくしゅうは、入力データから正解を予測できるような関数の学習で、使用する数学モデルは回帰問題や分類問題などがあります。

きょうしなしがくしゅうは、ラベルのないデータからパターンや構造を見つけ出す機械学習の手法で、使用する数学モデルはクラスター分析,k-means法、などがあります。

強化学習は、エージェントが、環境と相互作用しながら、報酬を、最大化するように行動を、学習する手法で、ゲームやロボット制御、自動運転などに広く応用されています。使用する数学モデルは、マルコフ決定過程などがあります。次の頁でそれぞれの数学モデルの考え方を見て行きましょう。ここにあげた数学モデル以外にも、いろいろありますので、そのアルゴリズムを調べてみると面白いと思います。

スライド 6

このセクションでは、教師あり学習の線形回帰とロジスティック回帰を勉強して行きましょう。

スライド 7

回帰問題で使われる数学モデルには、目的変数を出力とし、説明変数を入力とする数式で表現するものが多数あります。

ここでは、線形回帰モデルとして、単回帰分析と重回帰分析について説明します。

例えば、単回帰分析の場合、広告費と売上の関係で、売上を目的変数Yとし、広告費を説明変数Xとして、1次関数の形で表します。様々な売上と広告費のデータを学習させ、切片Aと傾きBを求めて数値化します。この行為が学習にあたります。ここで出来上がった式を用いて、広告費をいくらかければ、売上はいくらになるとシミュレーションすることができます。

重回帰分析の場合は、単回帰分析の説明変数が1つであったのに対し、複数の説明変数と売り上げの関係で、式を作ります。先ほどの例で、売上を目的変数Yとし、広告費を説明変数Xいちとすることに加え、価格やブランド認知度も説明変数としてXに、Xさんとおいて関係性を学習させます。この式に売上、広告費、価格、ブランド認知度を学習させ、切片Aと傾きB、傾きC、傾きDを数値化します。この式を用いて、広告費、価格、ブランド認知度を入力値として売り上げをシミュレーションすることができます。

スライド 8

ここで最小二乗法を思い出して下さい。

関数式の値と実測した値がもっとも小さくなるようなモデル関数を、求めるというものでした。

EXCEL、を用いて、手計算してみましょう。

まずは実測値を散布図で表し、その近似式をY イコール A+BXの一次線形として考えて行きます。

スライド 9

実測値の平均値を求め、偏差を実測値マイナス平均値として算出します。

次に、分散の二乗を求めます。分散の二乗は偏差の2乗の平均値でした。

更に、共分散を求めます。共分散は偏差Xと偏差Yの積の平均値でした。

ここから、傾きBと切片Aを計算して行きます。

傾きBは共分散を分散の2乗で割ると求められます。例では1.1と計算されます。

次に切片Aを求めますが、先ほど求めた傾きBを代入した式に、YとXの平均値を入れると、切片が求められます。例ではマイナス5になります。

つまり、この例の場合は、Y イコール マイナス5 + 1.1Xが近似式になります。

EXCELで散布図を描き、近似曲線追加します。ここで線形近似を選んで、近似式を表示のチェックボックスにチェックすると、最小二乗法で計算せずとも、グラフ上に先ほど計算して求めた数式が表示されます。とても便利ですね。

スライド 10

重回帰式を数学的に考えてみましょう。

回帰式がY イコール A + BXいち + CXにとします。この回帰式上にある予測値と実測値の差、すなわち残差が最も小さくなるような、切片A,傾きB,傾きCを求めるということになります。

Q,A,B,Cを、実測値 マイナス 予測値の2乗の総和とおいて、これが最も小さくなるABCを計算で求めます。ポイントは、平均値YとそのときのXいち, Xにがこの関数上にあることを利用して式を変形して行きます。

スライド 11

ここで、分散SXYの式を使って、式をシンプルにします。

この時点で、表された式を最小にするための切片Aが算出されます。

スライド 12

最終的には、Q、A、B、Cを、BとCの関数とみて偏微分し、その値がゼロとなるときが最小となることから、BとCを求めます。

かなり計算力が必要になりますね。ただ、実際に活用するときは、EXCELなどを用いて簡単に計算できるので、これを手計算する必要はありません。

実際にあとのセクションでEXCELを用いて重回帰分析を行う例を用意していますので、そちらで考え方を整理できると思います。

スライド 13

教師あり学習でもうひとつ手法を紹介しておきます。

ロジスティック回帰分析です。先ほどの重回帰分析と同じように複数の説明変数があって、それにより求められた目的変数が、「ある」「ない」といった2つの値の結果（目的変数）を予測する統計手法のことです。重回帰分析では目的変数が量的データ（数値）であるのに対し、ロジスティック回帰分析の目的変数は「ある」「ない」、「該当する」「該当しない」のような質的データです。

シグモイド関数の特徴である、入力が小さければ0、大きければ1に近い値を出力することを、利用し、目的変数を２つの分類に振り分けます。

シグモイド関数のUに重回帰式を入れて計算します。大量のデータを仕分けるのにとても便利です。

スライド 14

次にk近傍法を説明します。

例えば、数字の5と6を判断する場合、5の上の線の水平度、5の下の部分の開き具合をその特徴とみなし、ベクトルで表現します。それぞれのベクトル間の距離を計算し、5なのか6なのかを判断します。ここで距離計算はユークリッド距離として計算します。

もう少し具体的にお話しすると、

特徴量は、すいへい度と開き具合で表すことにします。

5と6の正解ラベルを持つ訓練データを、n個、用意します。

近傍として参照するデータの数をk個と決めます。

新しいデータとすべての訓練データとの距離を特徴量から計算します。

計算した距離を昇順に並べ、最も近いk 個のデータを選びます。

選ばれたk個のラベルが5なのか6なのかの多数決で予測結果を出力します。

スライド 15

このセクションでは、重回帰分析の事例として、身の周りの問題を考える上で、重回帰分析を使ってみましょう。

スライド 16

我が国で抱える大きな問題のひとつに少子高齢化があります。この少子高齢化時代において、老後の資金をどうやって確保するのか一緒に考えてみましょう。

日本はお金に関する教育が他国に比べ、弱いです。長野県でみても、投資信託の購入経験が、約２０％程度と小さいです。

投資という考え方を定着させるために、金融機関も様々な施策を検討して行くことが必要になります。

スライド 17

トマピケティーは、２１世紀の資本の中で、資本収益率が経済成長率より大きいという構造があると示しています。

これは格差問題が、経済成長によって解決せず、むしろ経済成長により格差が広がるというショッキングなレポートでした。

この本の中では、累進課税の富裕税で資産の再分配を提案していますが、そんなに簡単に解決する問題ではありません。

つまり、皆さんの老後の資金は皆さん自身で積み上げる必要があることになります。

スライド 18

それでは、テクノロジーの株を考えてみましょう。

ネットでいろいろ調べていたら、マイクロソフトの株が有望であるという資料をみつけました。

2035年に、10兆ドルの企業になるという、推測から、今後11年で、利益成長率が年平均15%と仮定すると、1株当たりの純利益、 EPSが54ドルと計算され、このEPSの何倍まで株が買われているかの指標である、 P E Rを30倍と仮定すると、現在、400ドルである株価が、1620ドルになるというレポートです。

この可能性があるか調べてみましょう。

スライド 19

機械学習は過去のデータを学習させ、シミュレーションすると学びました。目的変数をマイクロソフトの株価とした場合の説明変数を探して行きましょう。

まず、製品の発表と大きく関係があるかみてみます。株価と主要な製品をマッピングしてみると2017年頃からの急激な上昇は、発表された製品が起因しているとは読み取れませんでした。

経営者がゲーツからバルマー、バルマーからサティアに変わったことにも影響があると思えないですね。

スライド 20

世の中の出来事で、テクノロジーに影響を与えたものを、マッピングしてみました。トレンドとして、インターネット、クラウド、AIというのが見えてきました。

確かに、マイクロソフトは、早いタイミングから、きたるべきAIに対する、準備をしてきたし、OpenAIの主要投資家であり、OpenAIの技術である GPTモデルを、Azure上で独占的に提供しています。これに市場が反応していると考えるのがよさそうです。実は、NVIDIAもAIの波に乗って急激に成長してきた会社です。今後、AIのブームに乗り、市場がどうなっていくのか、これだけだと予測できません。

スライド 21

ここで、他の要素として、米国のGDP、為替、テクノロジーの要素として、回線速度や、プロセス　すなわちゲート長を、調べてみました。

ひと昔前なら、クロック周波数は説明変数の候補として考えられたのですが、デュアルコアなどCPUの構造が変わってきたため、ここでは一旦外すことにします。

スライド 22

ここでプロセスとは、電界効果トランジスタのドレインとソースの間のゲート長をしめします。

現在、2ナノメートルなので、水の分子の大きさが0.38ナノメートルとすると、その約6倍程度です。半導体がいかに小さいか分かると思います。

スライド 23

これも余談ですが、MITのネグロポンテ先生が、1995年にBeing Digitalという本を出しました。今では当たり前ですが、物理的なものからデジタル情報へというテーマで書かれています。その中には、今のSNSの状況を示すデジタルペルソナやプライバシーの侵害など多くのことを予測していました。この先生がかってハーバードビジネスレビューの紙面上で、回線速度が大きく変わると、大きな変革が起きるとお話しされていたことを思い出しましたので、今回、回線速度も重要な要素として説明変数に入れておきます。

スライド 24

EXCELのデータ分析から回帰分析を選んで、データの範囲を指定するだけで計算できます。

とても便利ですね。

スライド 25

得られた結果をみてみると、重相関R,重決定R2および補正R2ともに、0.9以上で、優位 Fも0.05以下のため、この分析結果は有効であると判断します。

二つのGDPは、この式上で同じ意味合いになるため、P値の小さいGDP名目を残し、GDP実質を削除します。

プロセスもP値が大きいため、削除します。

スライド 26

説明変数からGDP（実質）とプロセスを外し、再び計算します。

重相関R，重決定R2（補正R2）とも、0.9以上で、優位Fも0.05以下のため、この分析結果は有効であると判断します。

係数のP値も全て0.05以下なので、これで回帰式を作成することにします。

目的変数を株価、説明変数をGDP名目，為替，回線速度として、切片や傾きがEXCELで計算されました。

スライド 27

さて、得られた式からシミュレーションをしてみましょう。

為替は1ドルあたり140円で固定にしました。回線速度は2030年には5Gが定着し50ギガbps ,2035年には6Gが定着し100ギガbpsと置きました。GDP名目は、別途、単回帰式で2025年、2030年、2035年について算出します。EXCELで散布図を書いて、R2を確認し、SLOPE関数とINTERCEPT関数で傾きと切片を求めました。

それぞれの説明変数を重回帰式に代入し、株価がどうなるのかみてみます。

米国GDPを弱気で読んでいますが、2035年で株価は、1102ドルと計算できました。もともとネットでみつけた、1620ドルに比べ控えめな数字ですが、マイクロソフト株は、長期で持ち続ける価値がある、と考えることができます。

スライド 28

この状況を、日本でも高い有価証券益を出しているけんしんBANKの資金証券部のかたに見せて、アドバイスを頂きました。

GDPや為替レートは、経済時系列データの中でも、単位こん過程が疑われ、そのまま回帰をすると、見せかけの回帰が起こる可能性があるので結果の解決にあたっては注意が必要。つまり、時系列データが時間とともに安定せず、どんどん変化していく性質を持っていることに注意しなさい。

オーバーフィッティング（過学習）にも注意が必要。学習データを間引いて、一部を検証データとして扱うことで、過学習にならないように注意しなさい。

スライド 29

マイクロソフト株をベースに投資シミュレーションを行ってもらいました。在学中のよ年間は、月２万円、社会人になってから月５万円、これを３０年間持ち続けると、老後に、2326万円程度の資産が出来ていることになります。いずれにしても株価は急上昇していくので早めに投資を開始し、投資金額も大きくしていくことが重要とのアドバイスを頂きました。また、このシミュレーションには配当金が入っていないので、実際には配当金を株増しに使う運用で更に効果を高めることができるとも。

スライド 30

最後に一般的な投資のアドバイスも頂きました。

このように身近な課題において考える手段としても、重回帰分析は簡単に活用できることがお判りいただけたと思います。

また、課題を解決する上で、様々な知識が役立ちます。そのために、今の時点から、専門外の書籍を読む習慣をつけることをお勧めします。知識の幅が広いほど、しっかりとした判断ができるようになります。

スライド 31

このセクションでは教師なし学習を学びます。

スライド 32

教師なし学習の代表的な例は、クラスター分析になります。クラスター分析は、データを類似度や距離に基づいてグループ分けする手法です。これにより、似た特性を持つデータをまとめて分析しやすくします。

ここに示されている分析の流れは、階層型クラスターの例で、顧客の購入頻度と平均購入額があった場合、どういうグループに分かれるかを分析したものです。

顧客を購入頻度と平均購入額でベクトル化し、顧客同士の距離を計算します。その際、扱うデータは最大値の相対的なパーセントに変換し、極端に大きい数字に影響されないようにします。

顧客間の距離計算ができたら、小さいものからカップリングして行きます。この例だとGとHが６％でもっとも小さいのでカップリングされます。HとIは18％となっていますが、既にHがGとカップリングされているため、そのカップリングされたGとHの階層に対してカップリングされます。このように作業を進めると、階層構造が出来上がります。これをデンドログラムと呼びます。最後にいくつのグループに分けるかを基準値として上から階層を降りてグルーピングします。この例では基準値を４としているため、G　H　I　J　がひとつのグループにわけられます。

スライド 33

このセクションでは強化学習を学びます。

スライド 34

強化学習は、試行錯誤を繰り返して、最適な行動や操作を決定していく、AIの、学習アルゴリズムです。

例えば、ロボットに、基本動作だけ教えておき、転んだとき、どの動作をすれば、立ち上がれるかを、ロボット自身で、試していきながら、学習して、最終的には、転んでも、自力で立ち上がれるようになるなど。

この原理を説明する際に、エージェントと環境ということばを使います。エージェントは指示を出すプログラム、環境はその結果を反映するプログラムとします。

まず、エージェントは、この状態のときに、この行動をする、というポリシーを、持っています。図ではこれを、方策と呼んでいます。方策π tのとき、状態がS tで、その行動をA tと定義しておくと、それに従い、行動A tをとります。

行動A tをとった結果、状態がS t+1に変化します。そして、それに対して報酬R t+1がもらえます。この報酬R t+1の結果から、その方策がよかったのか悪かったのか、を評価します。

これは、マルコフ決定過程といって、現在の状態と行動によって次の状態が決まるというマルコフ性に基づいています。

もう少し具体的にみてみましょう。

スライド 35

Q学習と呼ばれるアルゴリズムを考えてみましょう。

最適行動価値Qは、ある状態で特定の行動を取った場合に得られる「期待される価値」を表し、このQを高めることでより良い選択ができるようになります。

この最適行動価値Q は、スライド上にある式であらわされます。式の中の学習率と割引率はパラメータです。学習率は、今の価値に対して次の報酬や未来の価値を、どのくらい適応させるかというもので、学習が始まる段階では高く設定し、時間の経過とともに少しずつ減少させていきます。割引率は、将来の報酬をどれくらい重視するかというもので、1に近いほど将来の報酬を当てにするように設定します。

実装イメージとしては、Q テーブルと呼ばれる表を作り、状態と行動とQの組み合わせを管理させます。報酬が出た時点でテーブルを更新して行きます。

例えば、5×5 のボード上で、ロボットがゴールまでたどり着く最適な経路を、学習させたいとします。ボードには障害物が設置してあり、ロボットはそこを通過できないものとします。

最初に東に1マス進めたとき、Qを2と設定します。障害物がなかったので、報酬を、10もらいます。次の状態のQは5と設定し、学習率、割引率をそれぞれ0.1、0.9と設定して、Qを計算すると、3.25 になります。現在地から東へ移動するQは、2 から 3.25 に更新されます。これを繰り返すことで、最適な経路をロボットは学んで行くことになります。

スライド 36

このセクションでは予測評価について学びます。

スライド 37

これまで機械学習のモデルを作る上で、学習データを用いて学習させるお話をしてきました。実際には出来上がったモデルが、ちゃんと機能するかを評価する必要があります。

このときに使用するデータをテストデータと呼びます。この評価の際に、学習データには適合しているのにテストデータを流すと適合しなくなる現象が、みられます。これは学習データに過度に当てはまるモデルのため、それ以外のデータが来た時に予測精度が悪くなるという状態です。これを過適合とか過学習といいます。先ほどの重回帰分析の事例で、けんしんBANKがコメントしてくれたものです。

じゃあ、どうすれば回避できるのでしょうか？

まずは学習データを平準化または正則化します。このスライドのグラフも学習データにぴったりと当てはまるような曲線が描かれていますが、テストデータのことを考えると実は一次線形の方が正しい気もします。

データを間引いたりすることで、モデルの精度が上がることがあります。また、テストデータの差を損失関数を使って是正し、勾配や切片を微修正して行き精度を上げるという方法もあります。更にはモデルが多項式の場合、その次数が誤っていることも考えられ、その次数をハイパーパラメータとして触ることも視野に入れます。

スライド 38

次にテストデータ（検証データ）はどうするのかをお話しします。

もっともシンプルなのは、入手したデータの７～６割を学習データとして、３～４割を検証データとして進めるホールドアウト法があります。

他には交差検証法といって、学習データと検証データを入れ替えながら検証誤差を計算する方法もあります。

スライド 39

ここで、予測の評価は、どうするのかをお話ししたいと思います。

回帰問題の場合は、平均二乗誤差（MSE）という方法があります。正解のデータから予測値を引いて２乗したものの平均値をとって評価するというやりかたです。

これは、よくみると最小二乗法の式に似ていませんか？そうです、最小二乗法で作った関数式は、この平均二乗誤差（MSE）が最も小さくなるということがわかると思います。

スライド 40

次に分類問題の場合、どのように評価するか見て行きましょう。

この例では、ある試薬による検査結果を混同行列として表に示しています。簡単に言えば、陽性の患者を正しく陽性と予測できたかを判断する際の考え方です。

このマトリクスに示すように全てのパターンは、TP+FN+FP+TNになります。この場合、正解率は全パターン中で、正しく陽性と判断できた人と正しく陰性と判断できた人、つまり、TP+TNの割合ということになります。

次に適合率は、試薬で陽性と判断された人の中で正しく陽性と判断された人の割合になり、再現率は、実際の陽性の人のうち、正しく陽性と判断された人の割合になります。

また、特異度は実際の陰性の人のうち、正しく陰性と判断された人の割合になります。

ここで、F値は、ゼロから1の数字で1に近い場合、性能がよいと評価する指標です。これは、適合率×再現率を2倍し、 適合率+再現率で割った数値です。

スライド 41

このマトリクスを、イメージ図で書いてみました。

本当に陰性の人を正常群、本当に陽性の人を、疾患群として分布をイメージしました。

この図から分かるように、正常群の中にも、ギ陰性として疾患をもった人が紛れ込むし、疾患群の中にも、実際は正常な、ギ陽性の人が紛れ込みます。

ここに境界線、しきいち シータの軸を立てます。この しきいちを右に動かせば、ギ陽性が増え、左に動かせばギ陰性が増えることが理解できると思います。

ここで、しきいちをゼロから1まで動かしたときの再現率と、1マイナス特異度でグラフを書きます。

つまり、実際の疾患群のうち、正しく陽性と判断された人と、正常群のうち正しく陰性と判断されなかった人の数を、シータを変化させながら、プロットしたものです。

この曲線を、ROC曲線と言います。1マイナス特異度がゼロから1の範囲で、この曲線によって作られる面積を、AUCと呼びます。このAUCが、1に近いほど機械学習モデルが2つのカテゴリを分離できていることを示します。

スライド４２

付録として１つのチュートリアルを用意しています。

長野市内に飲食店を作るという計画の中で、EXCELを用いて、簡単な重回帰分析を行います。

皆さんに興味を持ってもらえそうなシナリオを準備しましたので、是非、やってみて下さい。