Rabbit Challenge 機械学習レポート

1. 線形回帰モデル

* 要約

「線形」とは比例関係で、n次元ベクトルの内積で表現できる。回帰問題とは、ある入力(離散あるいは連続値)から出力(連続値)を予測する問題であり、直線で予測する場合は「線形回帰」。線形回帰モデルは、回帰問題を解くための機械学習モデルの一つであり、教師あり学習。入力とm次元パラメータの線結合を出力し、最小二乗法によりパラメータを推定する。データを学習用と検証用に分割し、検証用データで汎化性能を測定する。最後にBostonデータセットを使用し、線形回帰のハンズオンを行った。

* 気づき・追加学習

・気づき

線形回帰は連立方程式であり、行列表現への変換が重要であることを学んだ。

行列の最小二乗法の微分について、行列の式変換が非常に難しく感じた。行列では、転置や微分の公式があることを認識できたことはよい発見だった。今後、行列の微分や変換について、発生ベースで都度調査することとする。

・追加学習

講義で紹介されていた密度比推定について、以下のサイトで追加学習を行った。

URL：「https://speakerdeck.com/oshokawa/mi-du-bi-tui-ding-niyoruyi-chang-jian-zhi?slide=48」

最小二乗法は、平均二乗誤差を利用するため、外れ値に弱い。前処理で対応するか、その他の最適化手法を検討する必要がある。外れ値に頑健な最適化手法として、Huber損失、Tukey損失が紹介されており、Webサイトで確認した。

* ハンズオン

ボストンの住宅データセットを用いて単回帰と重回帰をハンズオン。

単回帰ではRM（一戸あたりの平均部屋数）からPriceを予測。学習結果のモデルに、平均部屋数を1として予測した場合、不正な値が予測された。訓練データにないデータをモデルは予測できないことを確認した。

重回帰では、CRIM（犯罪発生率）を説明変数に追加。単価域と重回帰で回帰係数と切片はほぼ同じであった。また、訓練データと検証データの残差、平均二乗誤差、決定係数の確認を行った。

1. 非線形回帰モデル

* 要約

回帰問題とは、ある入力(離散あるいは連続値)から出力(連続値)を予測する問題であり、曲線で予測する場合は「非線形回帰」。単回帰／重回帰のxをxの基底関数で置き換える。ただし、重みwは線形のままである(linear-in-parameter)。

基底関数として、9次関数、ガウス関数を確認。

次に、未学習と過学習の見分け方(訓練誤差・テスト誤差の状態から判断)と対策を確認した。

・未学習の場合：表現力の高い基底関数を利用する。

・過学習の場合：学習データ増(数の暴力)、不要な基底関数(変数)を削除、正則化法の利用を行う。正則化法は、「MSE＋罰則項」を最小化するよう最適化する

Ridge推定量とLasso推定量を確認。wを軸として、MSEを等高線で表現した図により、Lasso推定量がスパース推定(wが0,1になりやすい)ことを確認。

また、訓練データとテストデータの分割、検証方法として、ホールドアウト法とクロスバリデーションを確認した。ホールドアウト法でデータが少なく、検証データに外れ値や未知データが偏った場合、正しく学習と性能評価ができない。クロスバリデーションは、検証／学習データを設定したイテレーションに分け、精度平均(CV値)で評価する。

* ハンズオン

データセットとして、ランダムに生成した説明変数に、真の関数(4次関数)とノイズを加え目的変数を生成。当該データに対し、単回帰、ガウス型基底、多項式（1～10次関数）、サポートベクトル回帰、Kerasを用いたDLで予測結果を対比した。

1. ロジスティック回帰モデル

* 要約

識別的アプローチ(ロジスティック回帰)と生成的アプローチ(Bayesの定理)があり、ロジスティック回帰やSVNは前者。データとパラメータの線形結合結果をシグモイド関数で変換することにより、分類の確率を算出する。

なお、シグモイド関数の微分はシグモイド関数で表現できる。

最尤推定法は尤度関数を最大化するようなパラメータを選ぶ推定方法。

尤度関数の対数を取る最大の理由は、確率の積による桁落ちを防ぐためである。

シグモイド関数で確率に変換しているため、線形回帰モデルのように解析的に求めることができないため、勾配降下法によりパラメータを最適化していく。

勾配降下法は全データを用いて更新するため、データ量が多い場合に計算できない。その場合は、確率的勾配降下法(SGD)を利用する。

動画では紹介されていなかったが、分類モデルの性能を測る指標として混同行列と評価指標を確認。

・正解率：分類したいクラスに偏りがある場合、意味をなさない場合が多い。

・再現率(Recall)：病気の検診で「陽性であるものを陰性と誤診(False Negative)してしまうのを避けたい」場合に有用。

・適合率(Precision)：「重要なメールをスパムメールと誤判別」されるより「スパムと予測したものが確実にスパム」である方が便利。この場合に有用。

・F値：再現率と適合率のバランスをとった指標。

* 気づき・演習・追加の学習

最適化アルゴリズムとして、モメンタムの追加学習を行った。

URL：「https://qiita.com/omiita/items/1735c1d048fe5f611f80」

* ハンズオン

タイタニックデータに対し前処理（不要データ削除、欠損値補完）を実施し、ロジスティック回帰（1説明変数、2説明変数）を実施。

また、混同行列とクロスバリデーションの確認を行った。

1. 主成分分析

* 要約

可視化した場合に人が判断できるのは、3次元が限界である。多変量データの持つ構造をより少数個の指標に圧縮。（変数を減らすことに伴う情報損失をなるべく小さくする。）情報の量を分散の大きさと捉え、線形変換後の変数の分散が最大となる射影軸を探索。

射影後ベクトルの分散最大化を、ノルム1の制約条件の下で考え（ラグランジュの未定乗数法）、結果、分散共分散行列の固有値問題となる。分散共分散行列は対象行列であり、固有値は必ず0以上、固有ベクトルが直行する。また、分散共分散行列の分散が、固有値の和に相当するため、寄与率、累積寄与率の算出が容易であり、次元削減の判断を行いやすい。

* 気づき・演習・追加の学習

射影後ベクトルの分散最大化が、ラグランジュ関数の停留点を求めることになる理由が理解できなかった。ラグランジュの未定乗数法について、理解を深めるため以下のサイトを参考にした。

https://k-san.link/pca-01/#Lagrange-01

* ハンズオン

cancerデータセットを用いて、PCAのハンズオンを実施。第二主成分で7割、第三主成分で8割以上の寄与率があることを確認。その後、第二主成分までを用いて散布図を描画。

1. k近傍法（kNN）

* 要約

分類問題のための機械学習手法。最近傍のデータを個取ってきて、それらがもっとも多く所属するクラスに識別。パラメータであるkを変化させることで結果が変わり、kの値を大きくすると決定境界は滑らかになる。

1. k-平均法（k-means）

* 要約

教師なし学習のクラスタリング手法。与えられたデータをk個のクラスタに分類する。アルゴリズムは以下の通り。

1. 各クラスタ中心の初期値を設定
2. 各データ点に対し、各クラスタ中心との距離を計算し、最も距離が近いクラスタに割り当てる
3. 各クラスタの平均ベクトル(中心)を計算する
4. 収束するまで2,3を繰り返す

注意事項は、1の初期値が離れている必要があること、適切なkを設定することである。

* ハンズオン

ワインデータ(13説明変数、178レコード)を用いて、k-means(クラスタ数3を設定)によるクラスタリングをハンズオン。1クラスタは目的変数のclassが集中し、うまく分類できている。他の2クラスはclassが混在し、うまく分類できていない状況であった。初期値のパラメータとして、random\_stateをチューニングしたが、結果に大きな変化が見られず、特徴量選択が必要であると考えた。