実装演習レポート

科目：機械学習

**【第一章：線形回帰モデル】**

要点

・回帰問題はある入力（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る。直線で予測する場合を線形回帰といい、曲線で予測する場合を非線形回帰という。

・入力の各要素を説明変数または特徴量と呼び、m次元のベクトル（m=1の場合はスカラ）である。出力は目的変数であり、1次元のスカラー値である。

・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、教師あり学習である。入力とm次元のパラメータの線形結合を出力するモデルである。慣例として予測値にはハット(^)を付ける。

・線形結合は入力とパラメータの内積であり、入力ベクトルと未知パラメータの各要素を掛け算して足し合わせたものである。入力ベクトルとの線型結合に加え、切片も足し合わせる。

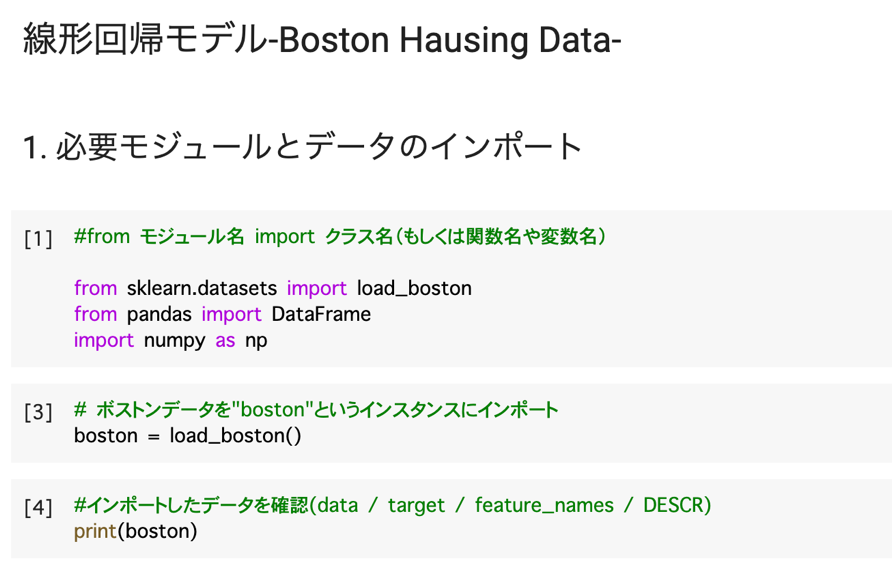
・パラメータは最小二乗法により推定する。最小二乗法では学習データの平均二乗誤差（残差平方和）を最小とするパラメータを探索する。

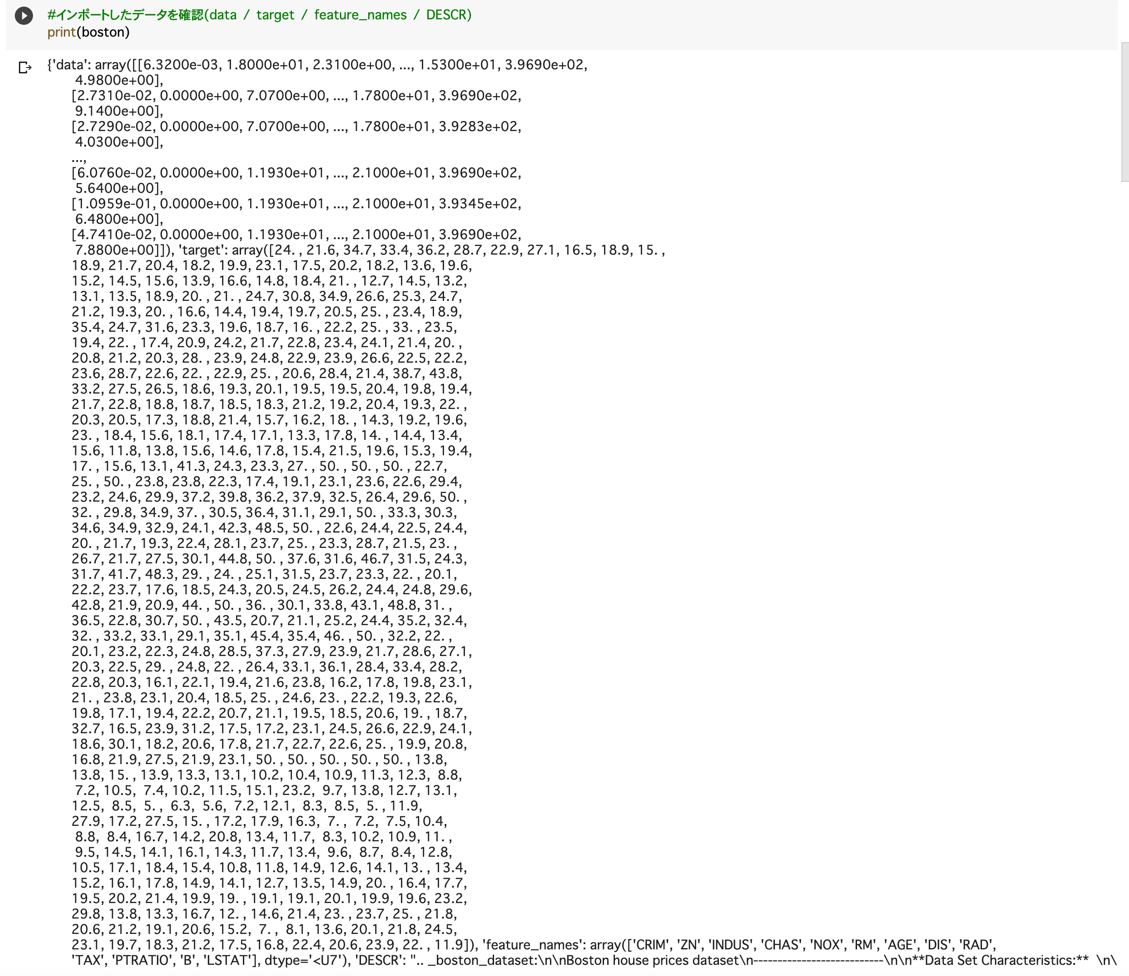
・説明変数が1次元(m=1)の場合は単回帰モデル、多次元(m>1)の場合は線形重回帰モデルと呼ぶ。

・モデルの汎化性能（Generalization）を測定するため、データを学習用データと検証用データに分割する。モデルの未来のデータへの当てはまりの良さ（汎化性能）を測定する。

・誤差を正規分布に従う確率変数を仮定し尤度関数の最大化を利用した推定も可能である。線形回帰の場合には、最尤法による解は最小二乗法の解と一致する。

実装演習結果と考察









ボストンの住宅データセットを線形回帰モデル（2変数の重回帰分析）で分析したところ、部屋数が4で犯罪率が0.3の物件は約4240ドルの価格となることが予測された。

**【第二章：非線形回帰モデル】**

要点

・複雑な非線形構造を内在する現象に対しては非線形回帰モデリングを実施する。

・基底展開法は回帰関数として基底関数と呼ばれる既知の非線形関数とパラメータベクトルの線形結合を使用する。未知パラメータは線形回帰モデルと同様に最小二乗法や最尤法により推定する。

・よく使われる基底関数は多項式関数、ガウス型基底関数、スプライン関数/Bスプライン関数である。

・未学習（underfitting）は学習データに対して、十分小さな誤差が得られないモデルであり、過学習(overfitting)は小さな誤差は得られるが、テスト集合誤差との差が大きいモデルである。

・基底関数の数、位置やバンド幅によりモデルの複雑さが変化する。適切な基底関数を用意し、不要な基底関数は削除する。

・モデルの複雑さに伴って、その値が大きくなる正則化項（罰則項）を課した関数を最小化する。L2ノルムを利用する場合をRidge推定量（縮小推定）、L1ノルムを利用する場合をLasso推定量（スパース推定）という。

・適切なモデル（汎化性能が高いモデル）は交差検証法で決定する。

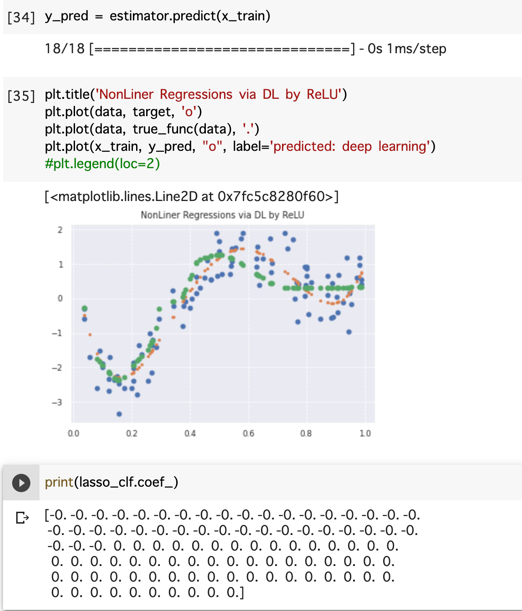
・ホールドアウト法はデータを学習用とテスト用の2つに分割して予測精度や誤り率を推定する。

・クロスバリデーション（交差検証）はデータを学習用と評価用に分割し、検証データで各モデルの精度を測定し、精度の平均（CV値）を汎化性能とする。

・グリッドサーチは全てのチューニングパラメータの組み合わせで評価値を算出し、最も良い評価値を持つチューニングパラメータを持つ組み合わせを、「いいモデルのパラメータ」として採用する。

実装演習結果と考察





非線形回帰モデルはクロスバリデーション（交差検証）で精度を計測し、汎化性能を推定する。

**【第三章：ロジスティック回帰モデル】**

要点

・ロジスティック回帰モデルは分類問題を解くための教師あり機械学習モデルである。

・入力（説明変数又は特徴量）とm次元パラメータの線形結合をシグモイド関数に入力する。

・出力（目的変数）はy = 1になる確率の値になる。

・シグモイド関数の微分はシグモイド関数自身で表現することが可能であり、尤度関数の微分を行う際に計算が容易となる。

・ロジスティック回帰モデルではベルヌーイ分布（確率pで1、確率1-pで0をとる離散確率分布）を利用する。

・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を最尤推定という。

・対数尤度関数にマイナスをかけたものを最小化し、最小二乗法の最小化と合わせる。

・対数尤度関数を係数とバイアスに関して微分する。

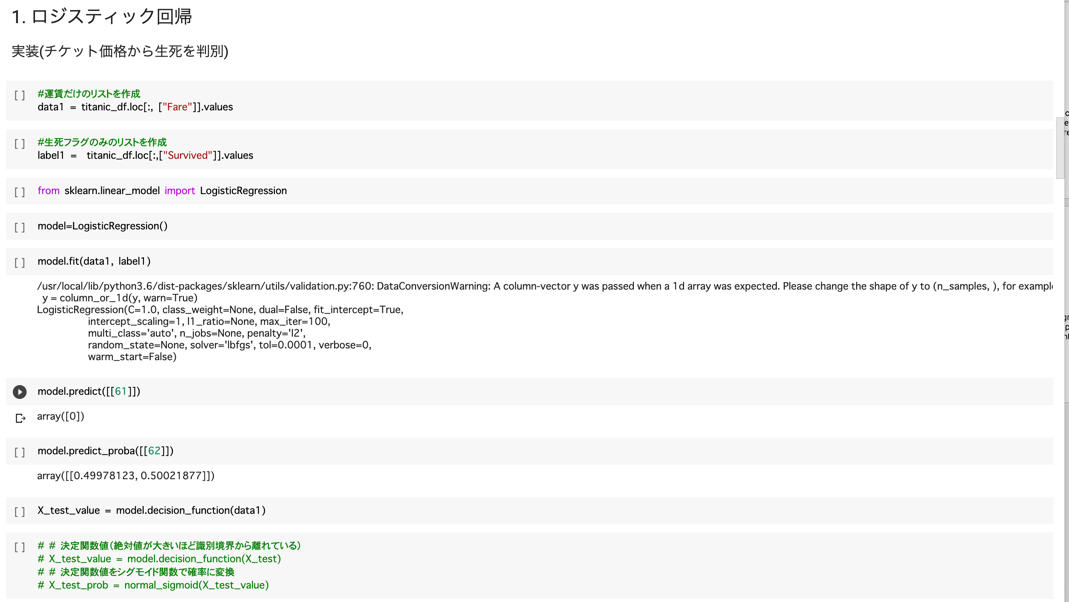
・確率的勾配降下法(SGD)はデータを一つずつランダムに選んでパラメータを更新する。

・再現率(Recall)は本当にpositiveなものの中からpositiveと予測できる割合である。

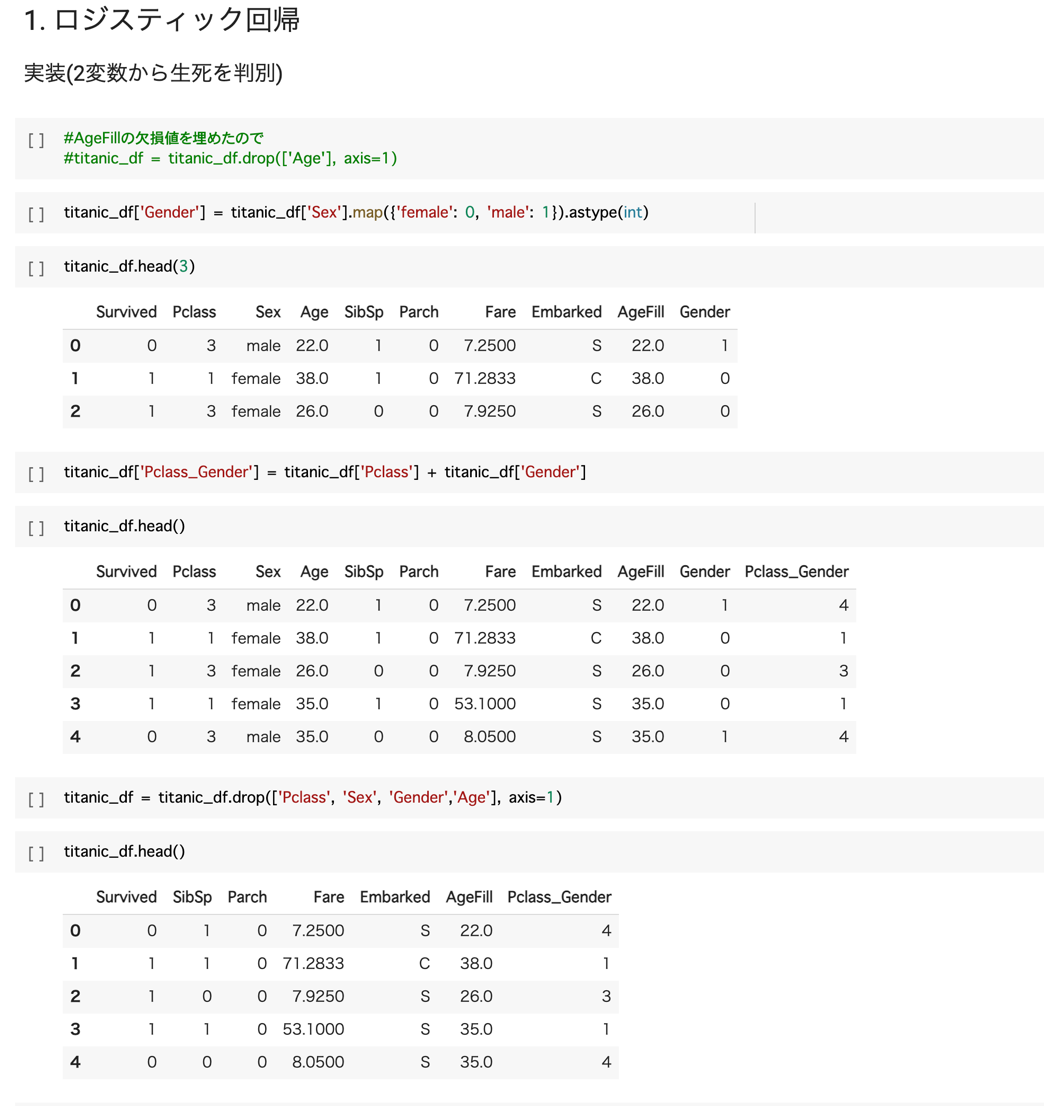
・適合率(Precision)はモデルがpositiveと予測したものの中で本当にpositiveである割合である。

・F値はprecisionとrecallの調和平均である。

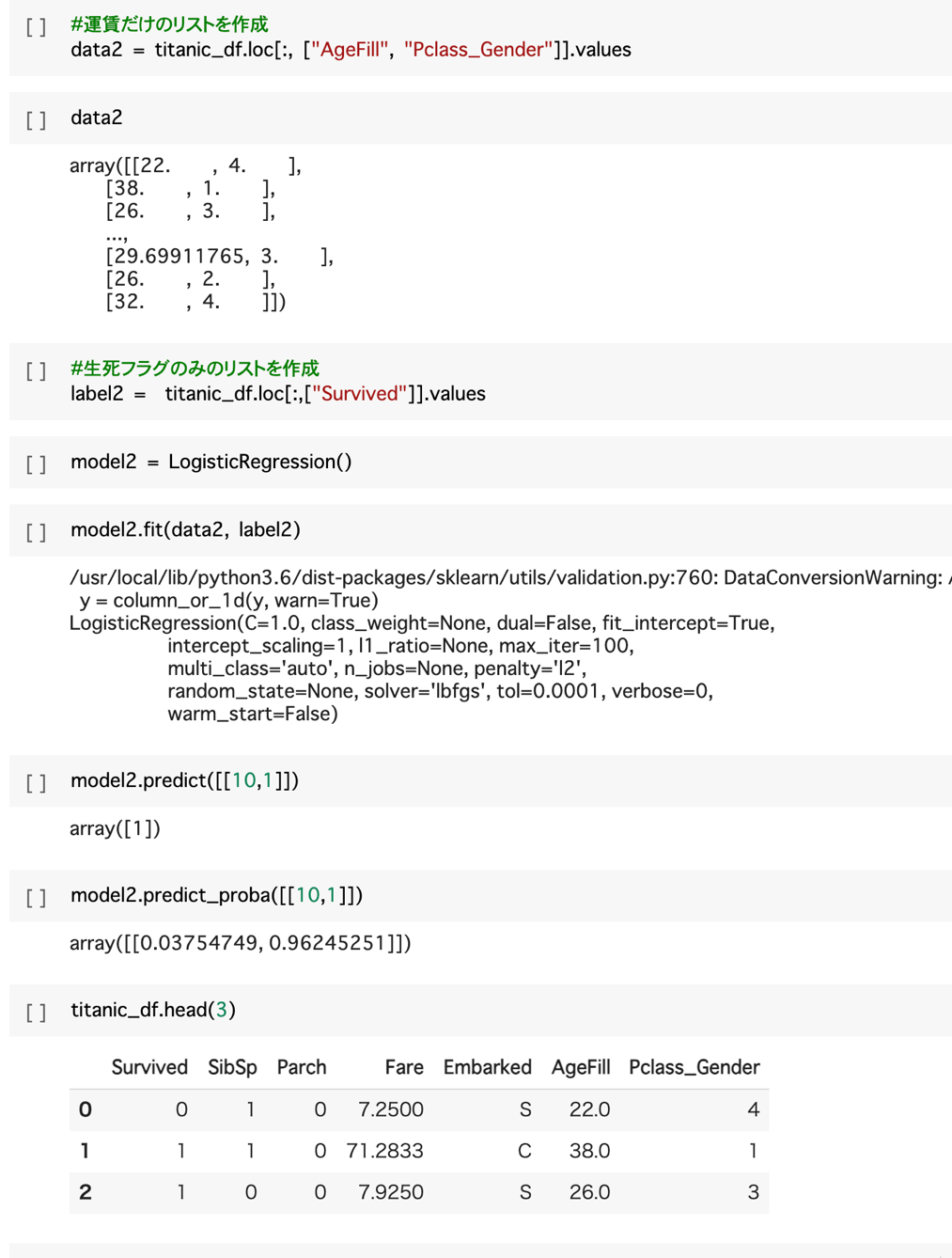
実装演習結果と考察

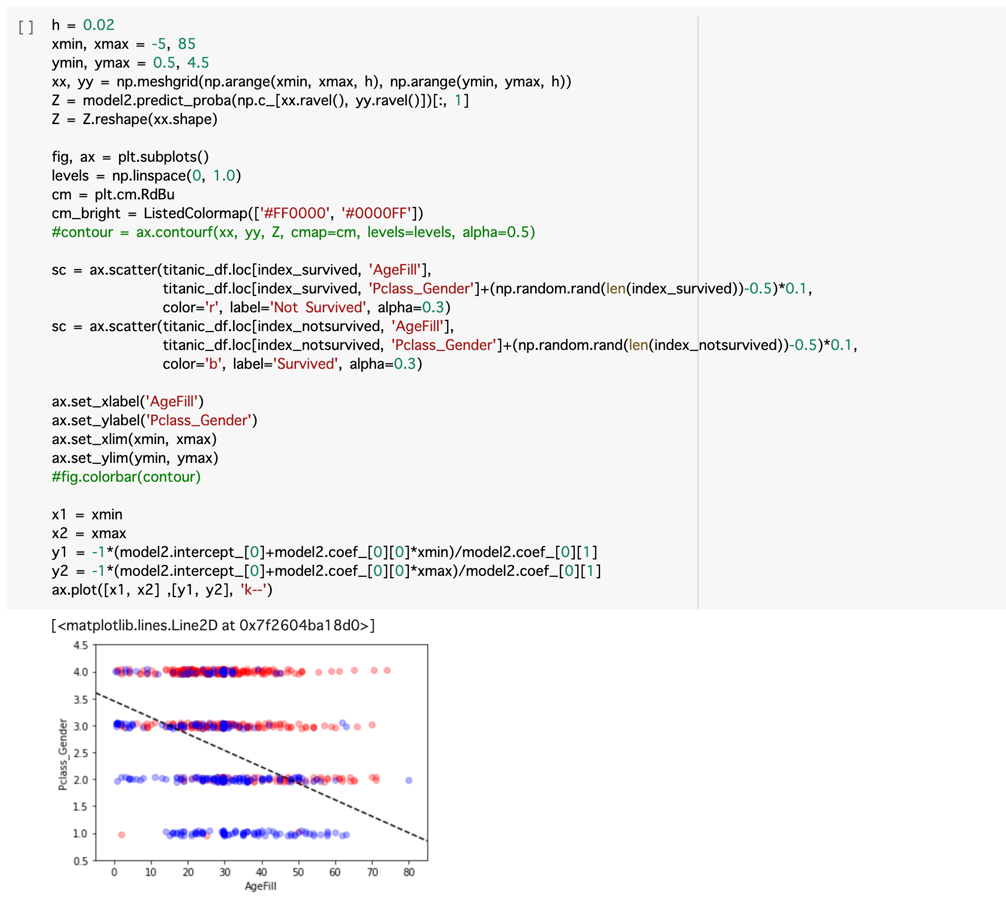




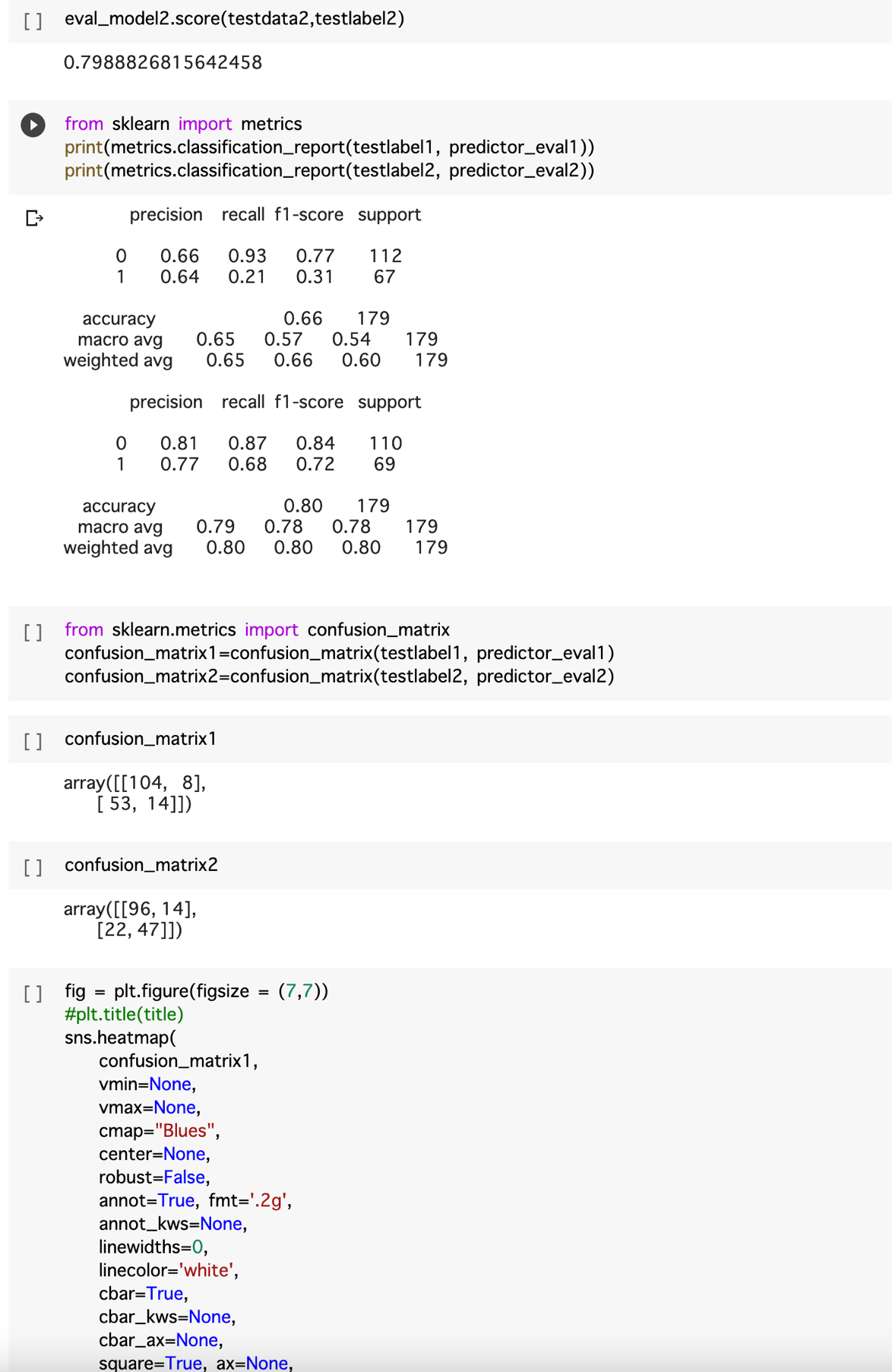


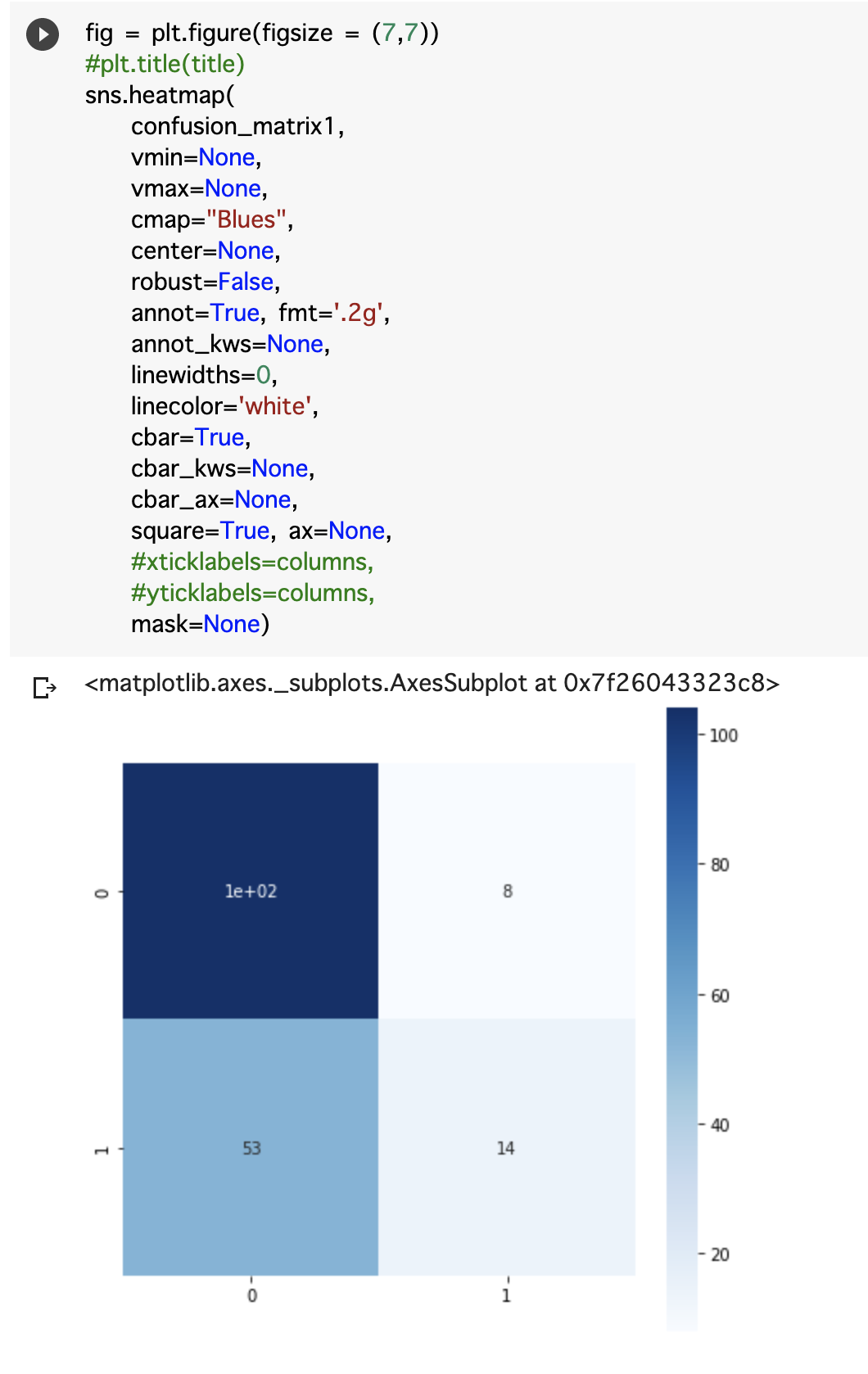


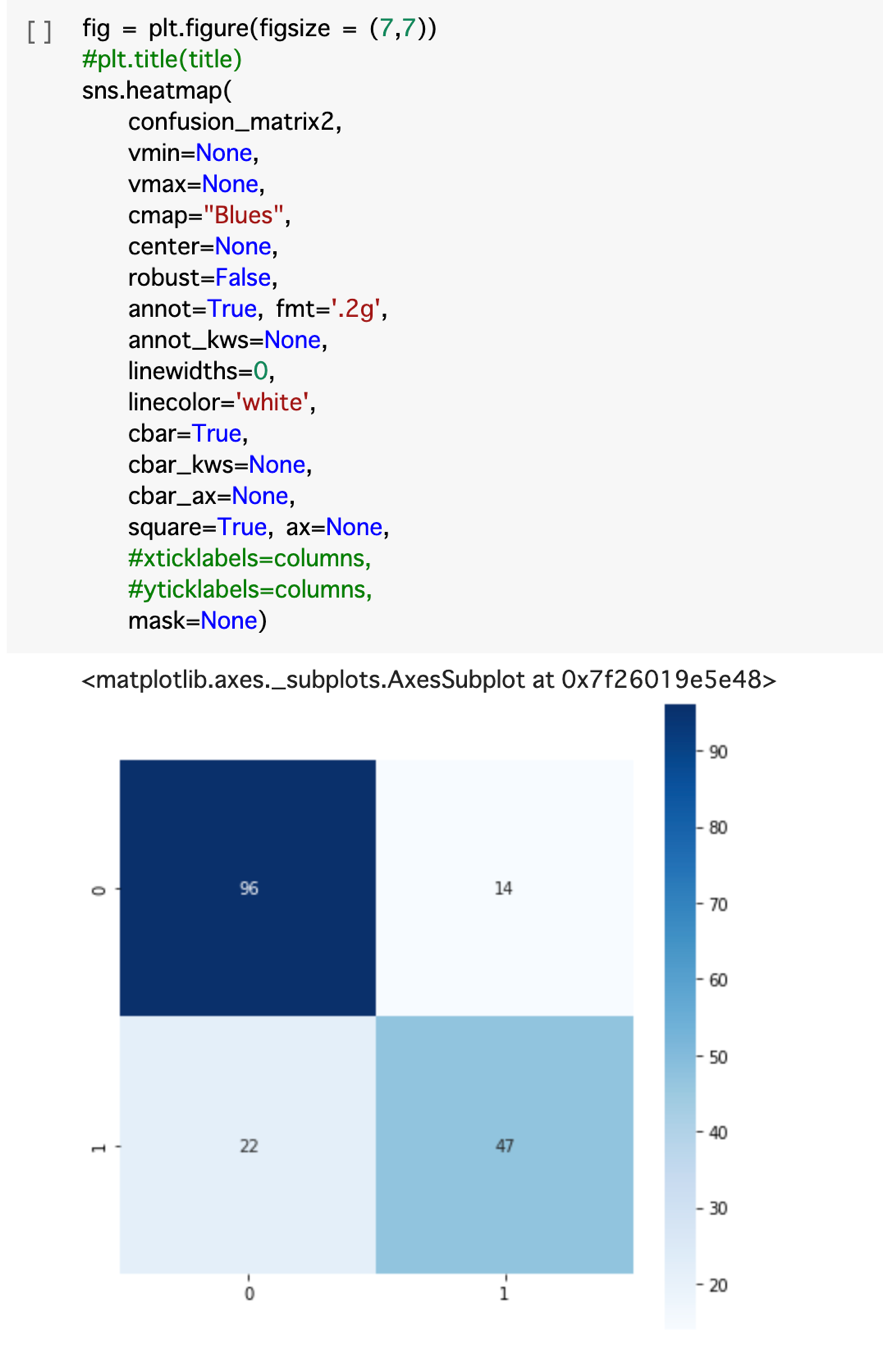
















タイタニックの乗客データからロジスティック回帰モデルを作成したところ、30歳の男性の生き残る確率は0.2（20%）となる。

**【第四章：主成分分析】**

要点

・主成分分析は多変量データの持つ構造をより少数個の指標に圧縮する。

・係数ベクトルが変われば線形変換後の値が変化する。

・ノルムが1となる制約を入れて最適化問題を解く。

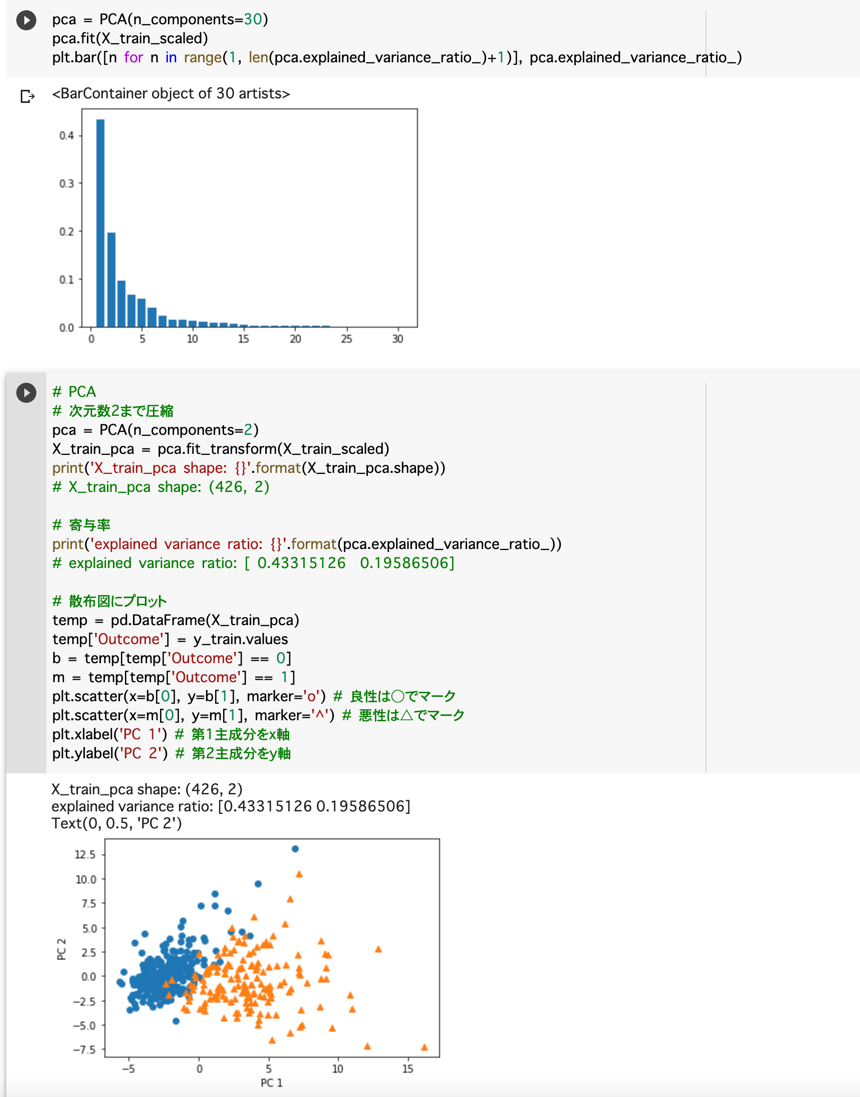
・ラグランジュ関数を微分して最適解を求める。

・分散共分散行列を計算し固有値問題を解くと、（最大）m個の固有値と固有ベクトルのペアが出現する。k番目の固有値（昇順）を並べ対応する固有ベクトルを第k主成分と呼ぶ。

・第k主成分の分散の全分散に対する割合を寄与率という。

実装演習結果と考察





乳がん検査データを利用しロジスティック回帰モデルを作成し、主成分を利用して２次元空間上に次元圧縮した際に、良性と悪性がうまく判別可能であると考えられる。

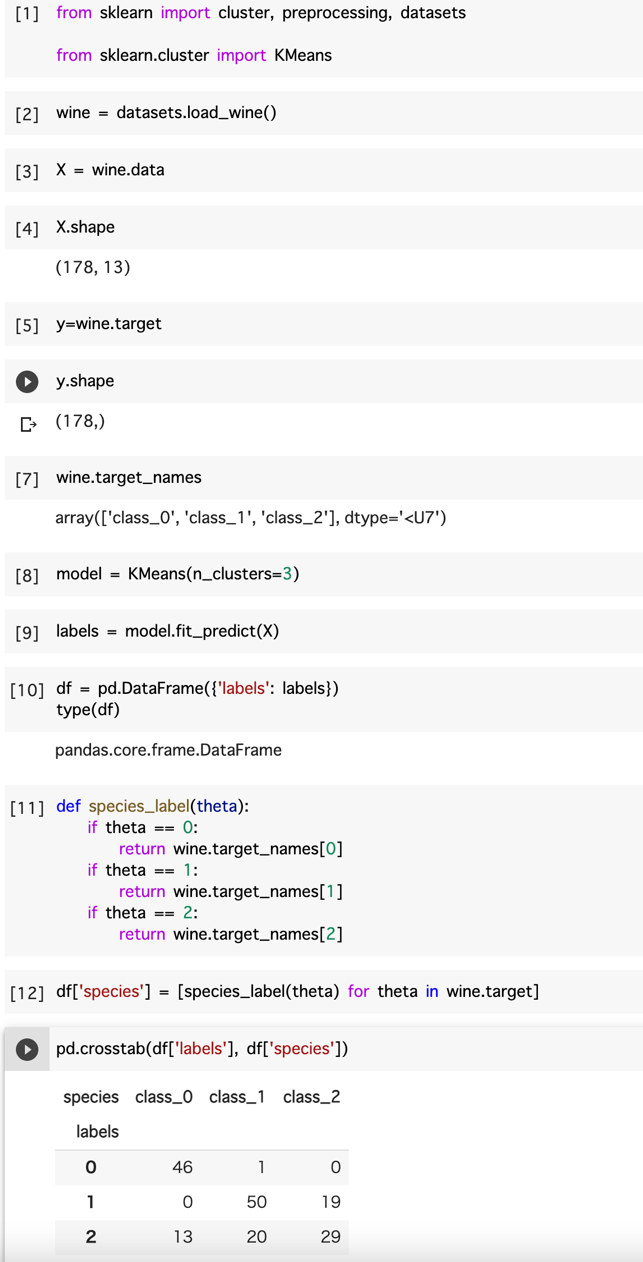
**【第五章：アルゴリズム】**

要点

・k近傍法(kNN)は最近傍のデータをk個とってきて、それらが最も多く所属するクラスに識別する手法である。

・k-平均法(k-means)は各クラスタ中心の初期値を設定し、各データ点に対して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を割り当ててクラスタの平均ベクトル（中心）を計算することを繰り返して収束させる手法である。

実装演習結果と考察



ワインのデータからk-means法を用いて3つのクラスに分類することが可能であった。

**【第七章：サポートベクターマシン】**

要点

・サポートベクターマシン(SVM)は2クラス分類のための機械学習手法であり、線形モデルの正負で2値分類する。

・線形判別関数と最も近いデータ点との距離をマージンという。

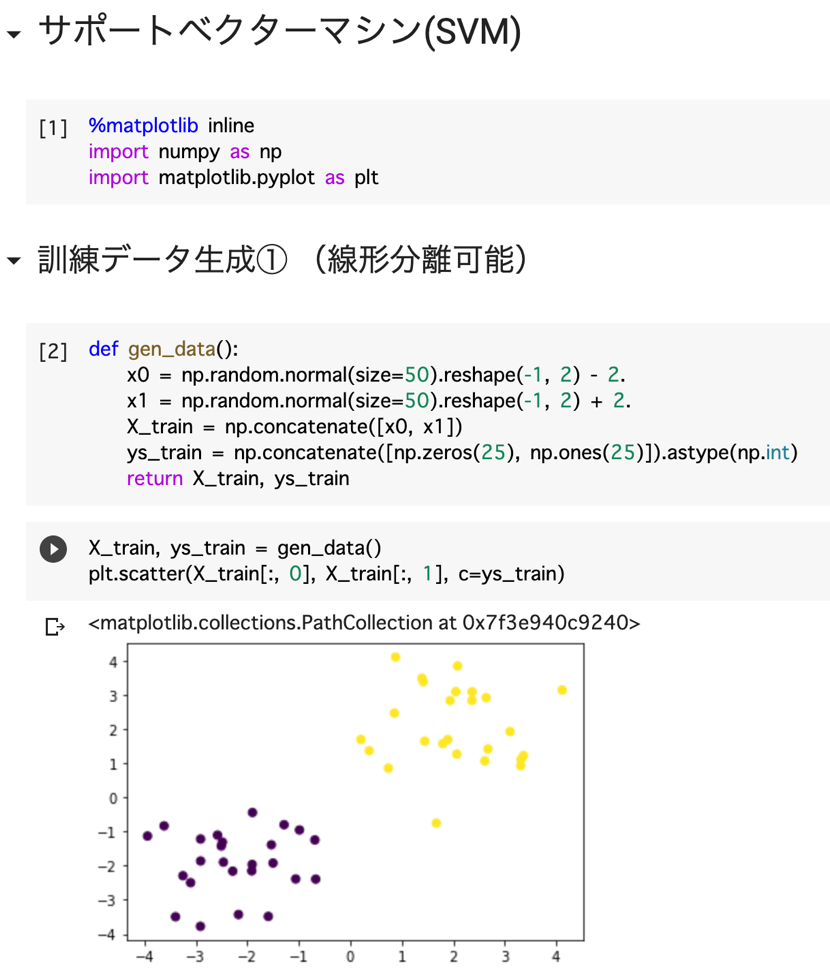
・マージンが最大となる線形判別関数（wTx + b = 0）を求める。

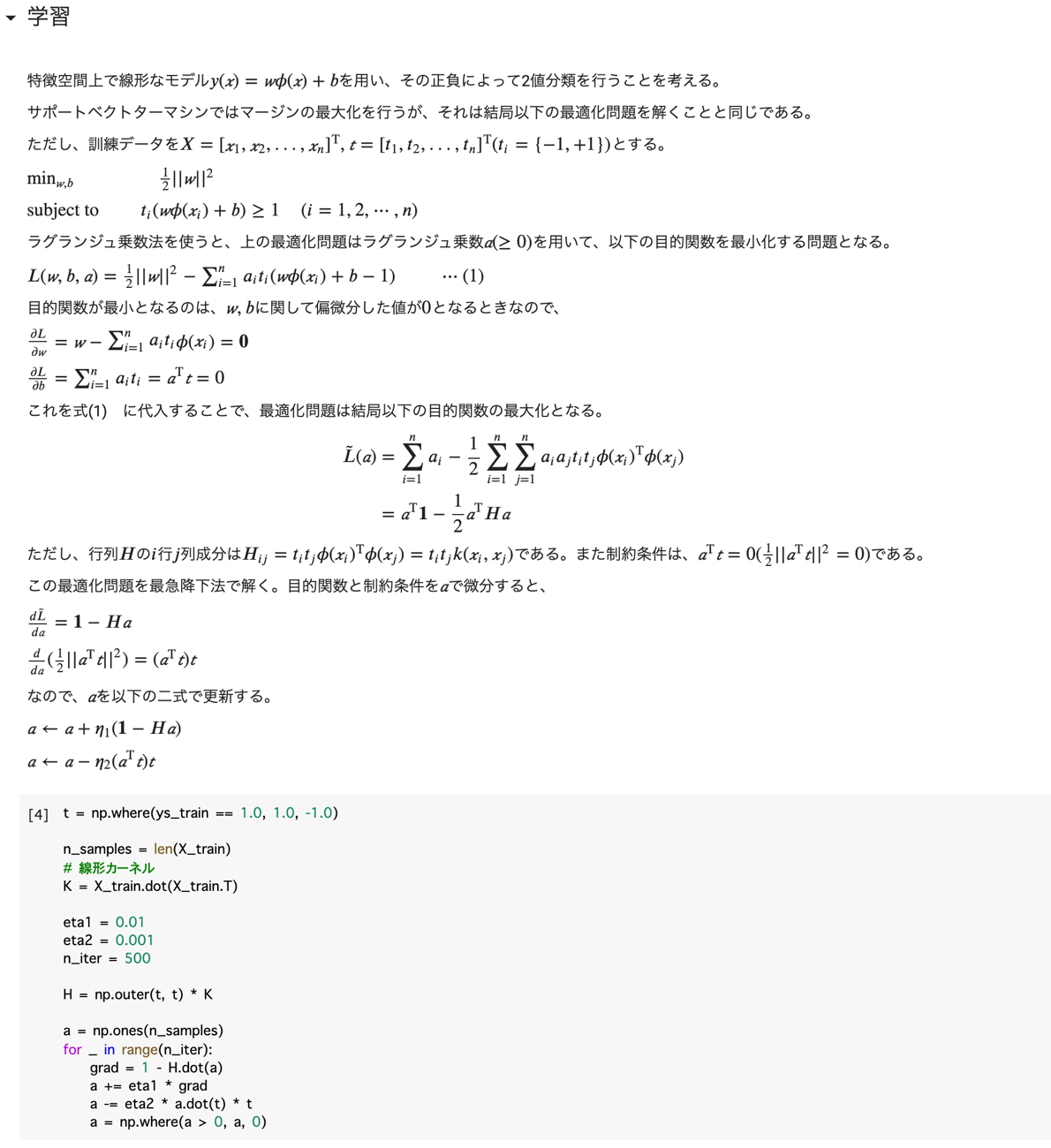
・SVMの主問題の目的関数は、制約条件はであり、この最適化問題をラグランジュ未定乗数法で解く。

・ラグランジュの未定乗数法

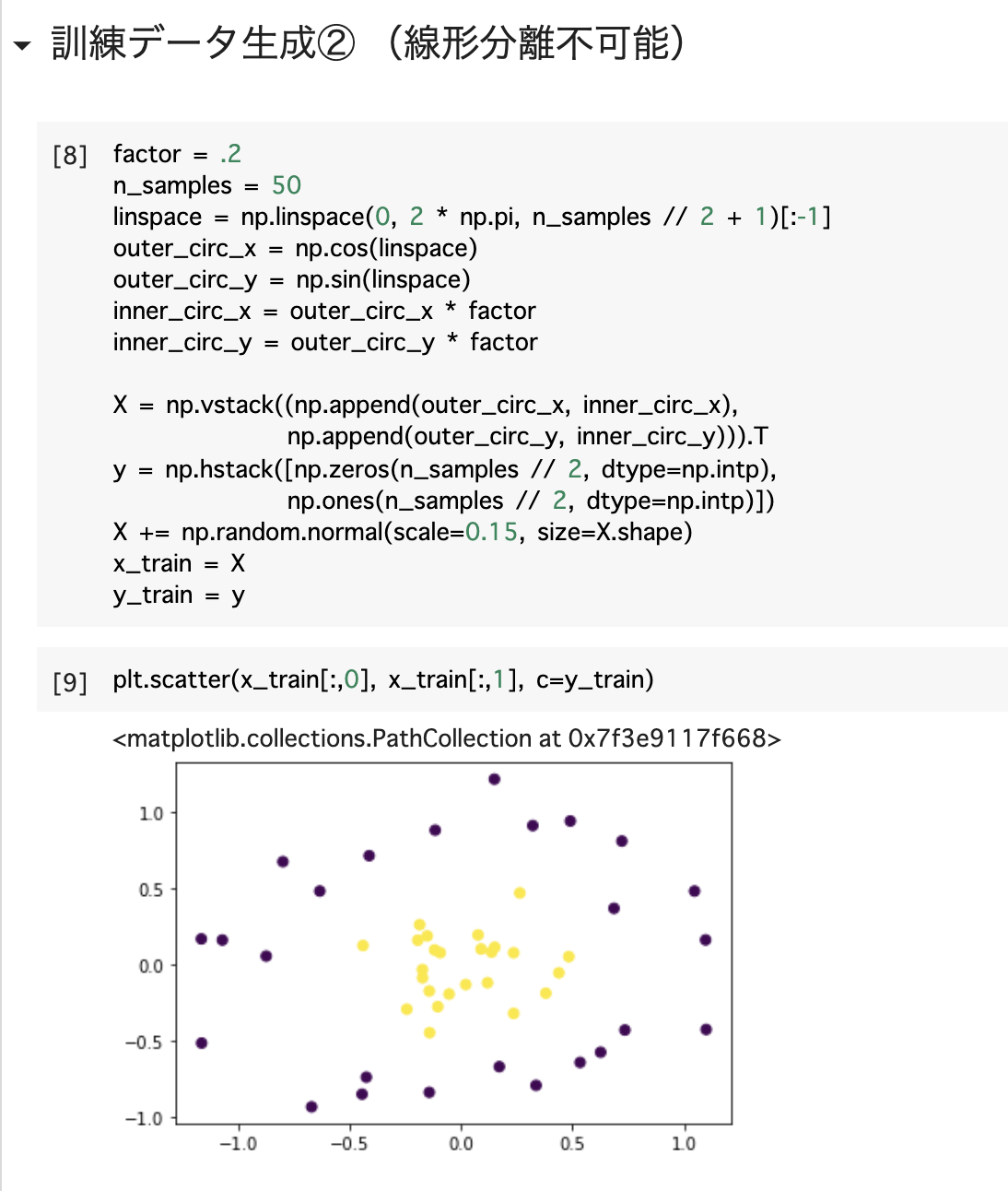
制約のもとで、f(x)が最小となるxは、変数 を用いて新たに定義した **ラグランジュ関数** において、 を満たす。

実装演習結果と考察

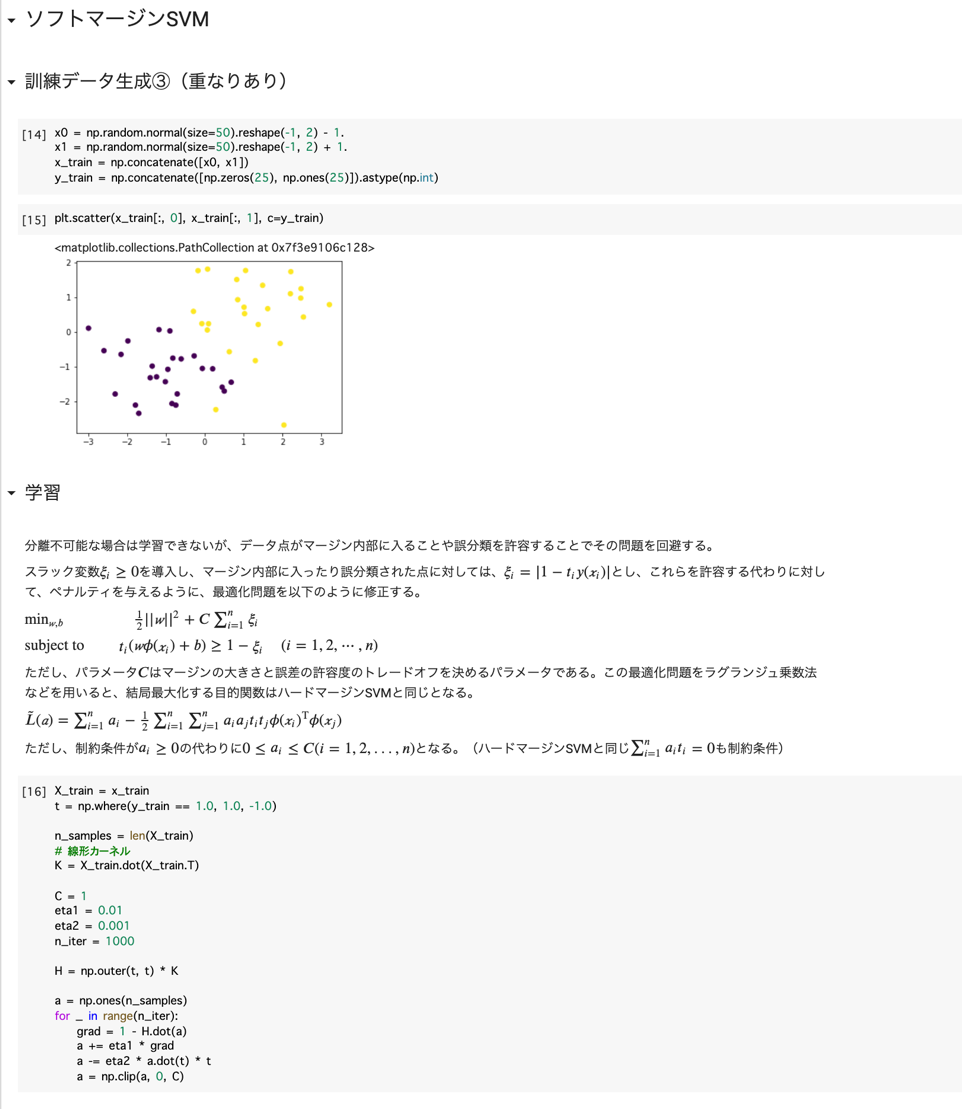


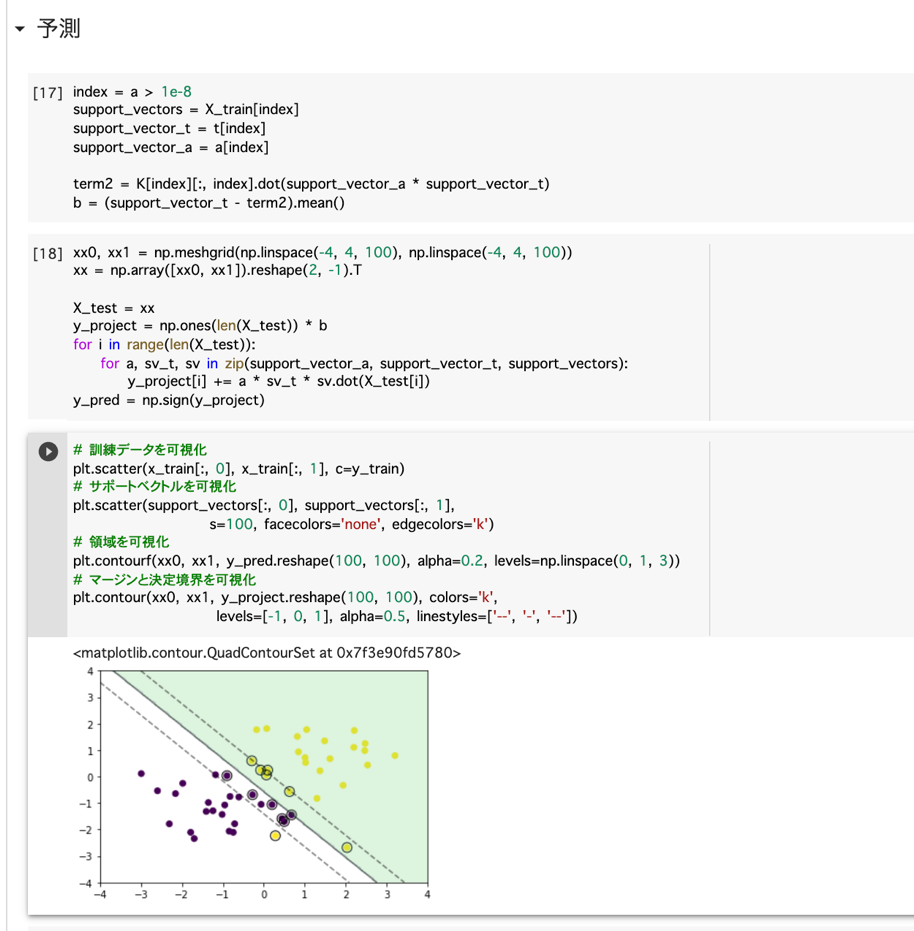












SVMではマージンの最大化を、ラグランジュ未定乗数法を用いて最適化問題を解くことにより行う。元のデータ空間で線形分離不可能な場合は、カーネルを利用して特徴空間上で線形分離すること試みる。分離不可能な場合は、データ点がマージン内部に入ることや誤分類を許容することで問題を回避する。これをソフトマージンSVMという。

