ラビットチャレンジ課題　作成者:内山 貴雄

機械学習の各章の要約

●線形回帰

線形回帰モデルとは、ある入力(離散あるいは連続値)から出力(連続値)を予測する問題のうち、説明変数の一次式にて目的変数を示すことができるモデルである。その式の中で、各説明変数に係る係数を回帰係数と呼び、切片と合わせたそれらの推定は最小二乗法にて求められる。

●非線形回帰

非線形モデルとは、基底関数と呼ばれる非線形関数とパラメータベクトルの線型結合により表現されるモデルである。それは、線形回帰モデルと同様に最小2乗法や最尤法により推定される。また基底関数として、多項式関数、ガウス型基底関数、スプライン関数/ Bスプライン関数が用いられる。

モデルの学習の際、学習データとモデル出力値の誤差が小さいが、検証データとモデル出力値の誤差との差が大きい状態を過学習と呼び、その状態を回避ため、主に下記対策を実施する。

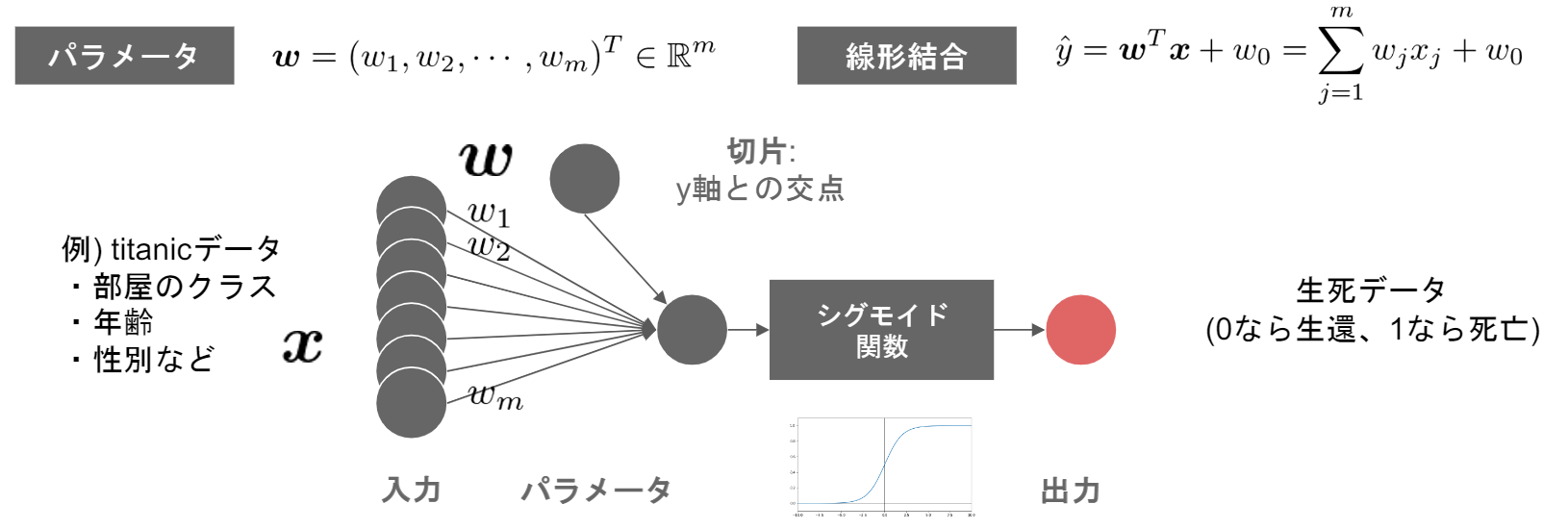
・(対策1)学習データの数を増やす

・(対策2)不要な基底関数(変数)を削除して表現力を抑止

・(対策3) 正則化法を利用して表現力を抑止

●ロジスティック回帰モデル

分類問題を解くための教師あり機械学習モデルであり、入力値としてm次元パラメータの線形結合シグモイド関数に与えられ、y=1となる確率の値が出力値となる。

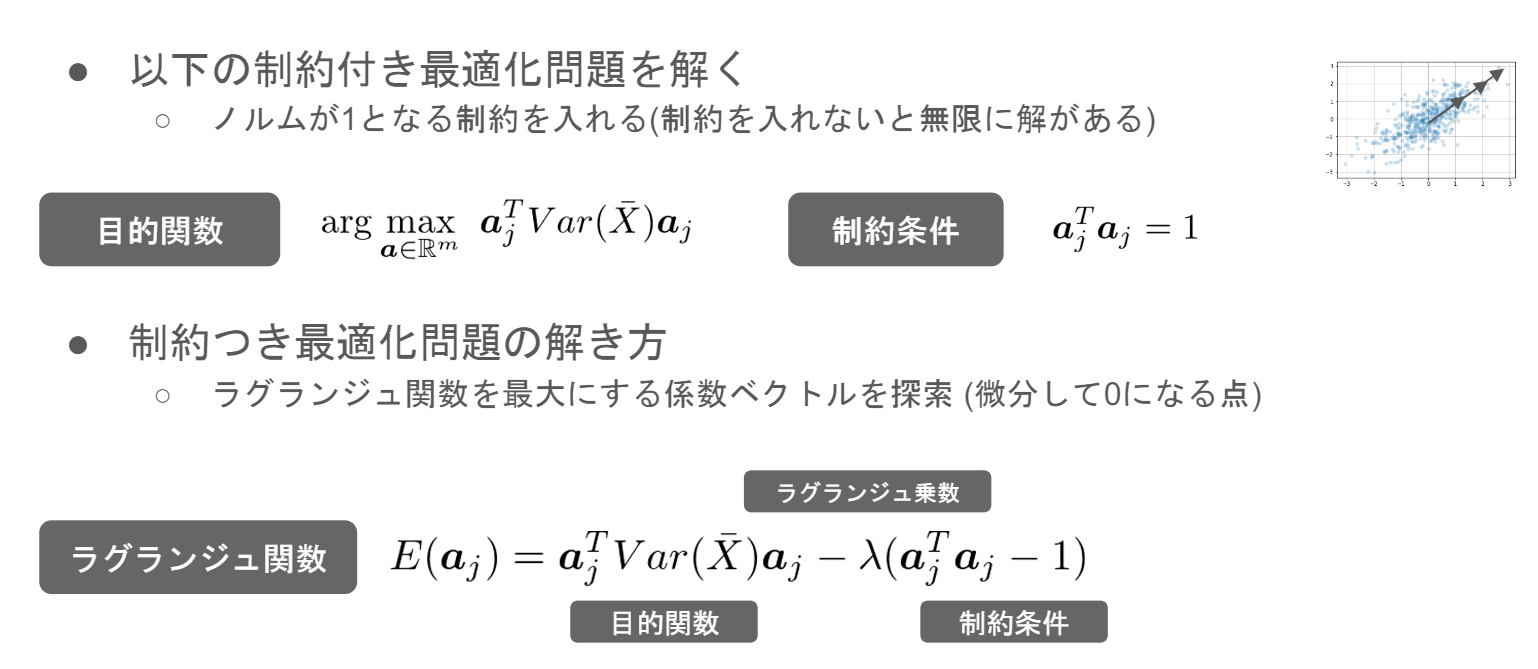


また、本モデルの推定方法（回帰係数の推定）としては、尤度関数の最大化する回帰係数を勾配降下法により求めることが一般的である。

●主成分分析

多変数から構成されるデータを分析する際に、分析精度や結果の可視化の観点から情報の損失はなるべく小さく、少数の変数に圧縮することが可能な本手法が役立つ。

ここで、2次元のデータを1次元のベクトルに射影し圧縮する際には、射影後のデータの分散は大きいほど情報の損失は少ない。これを利用し、その分散値を最大にする射影の際の係数ベクトルを、係数ベクトルのノルムが1となる制約下で求める。その際、前述の最大化する関数をラグランジュ関数と呼ぶ。

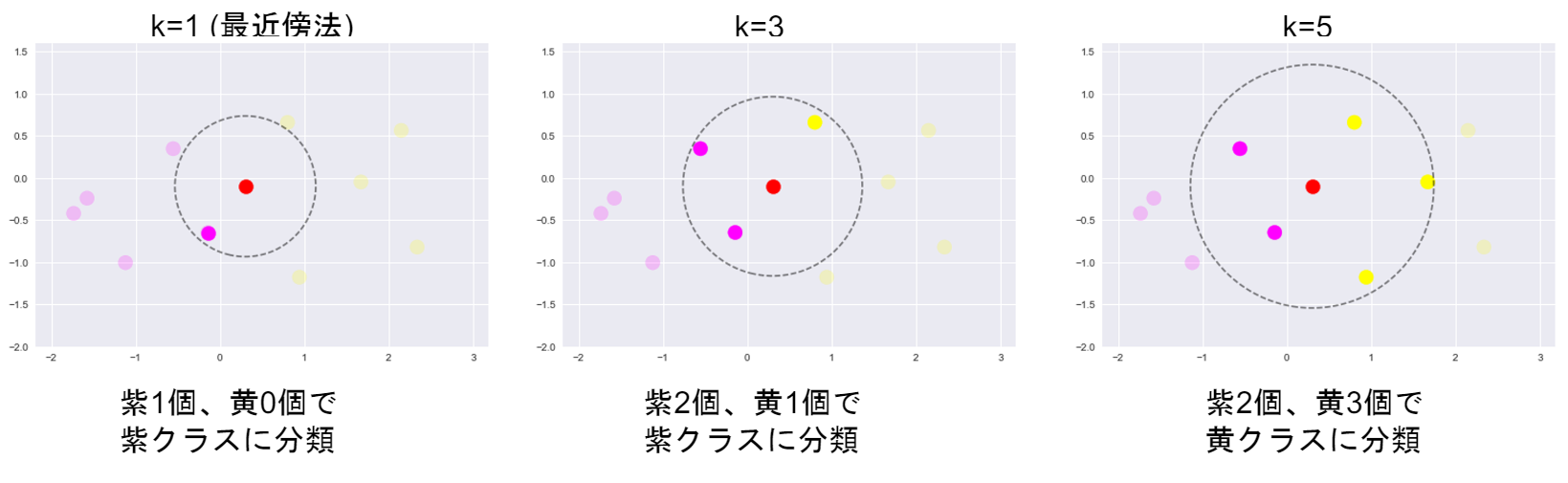


●アルゴリズム

・K近傍法(kNN)

教師なし学習のクラスタリング手法の一つである。

各データにて最近傍データをk個取得し、それらの中で最も多く所属するクラスに識別するアルゴリズムである。また、kを大きくすると決定境界は滑らかとなる習性がある。

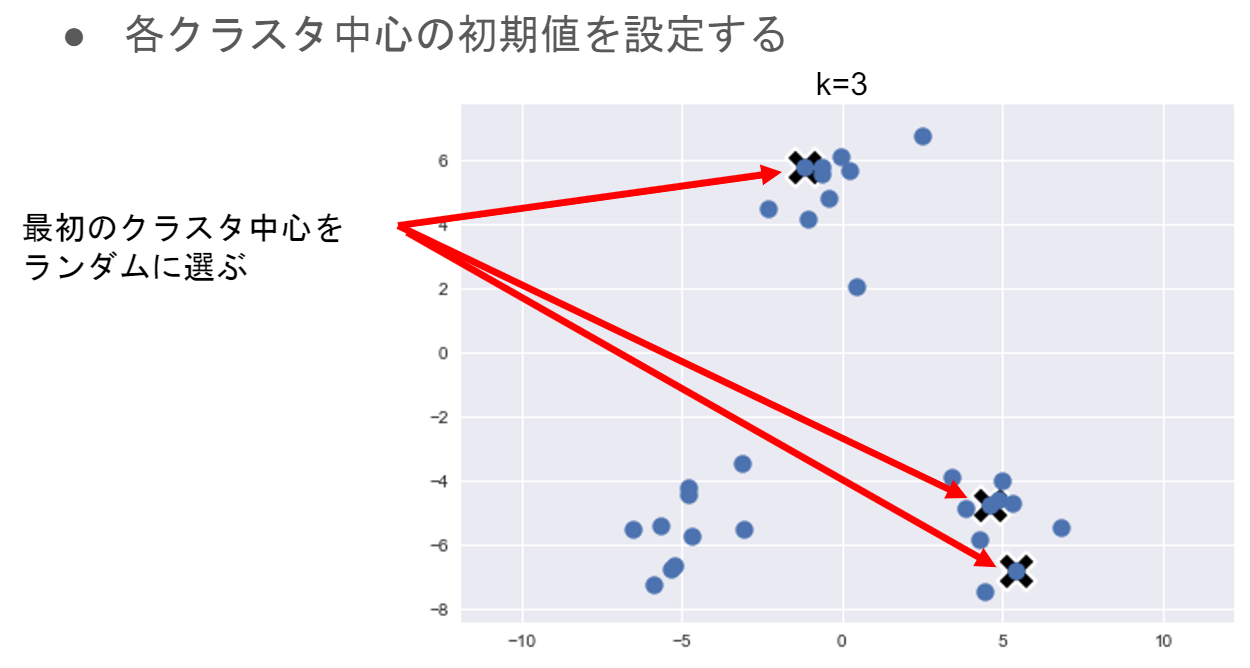


・K平均法(Kmeams)

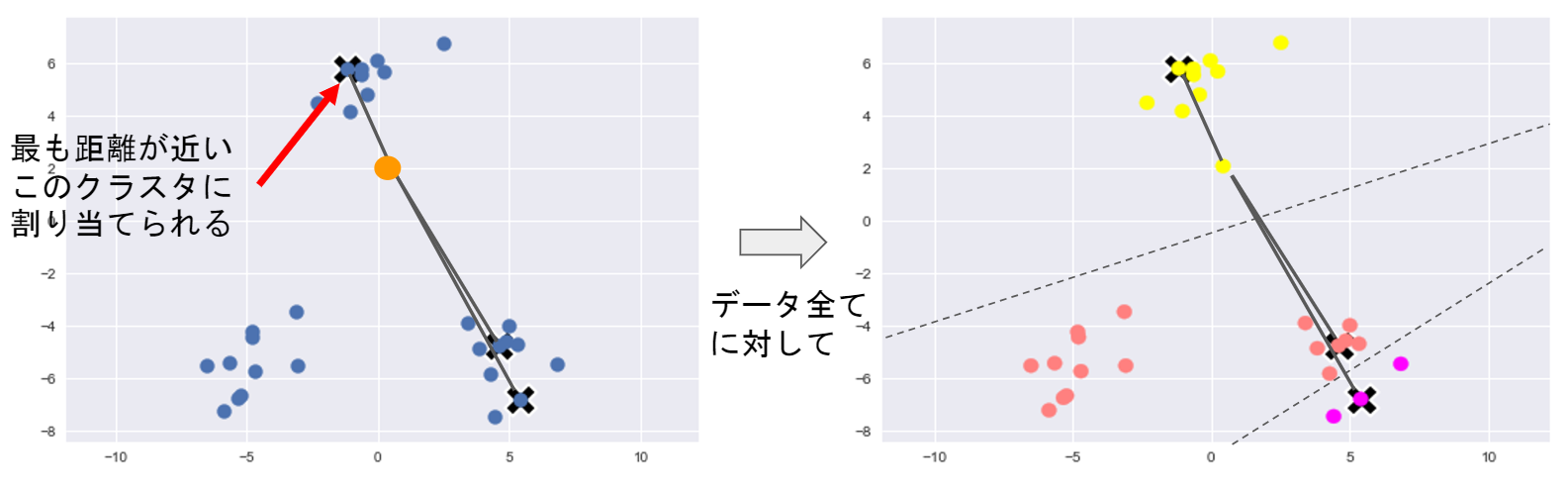
教師なし学習のクラスタリング手法の一つである。

**k平均法(k-means)のアルゴリズム**

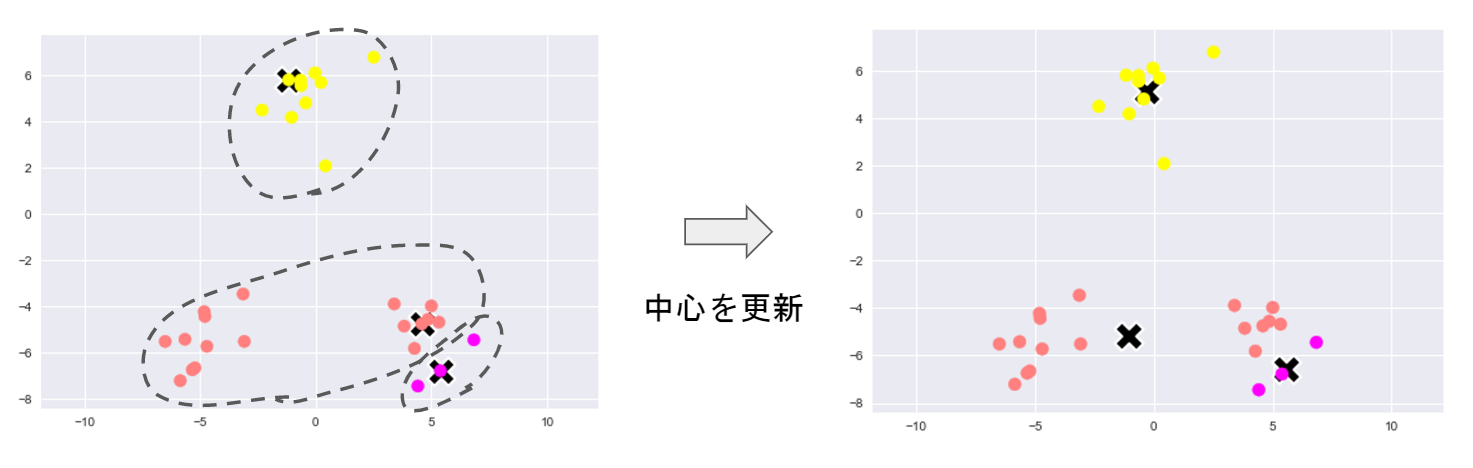
1. 各クラスタ中心の初期値を設定する。



1. 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る。
2. 各クラスタの平均ベクトル（中心）を計算する。



1. 収束するまで2, 3の処理を繰り返す。



●サポートベクターマシン

2クラス分類のための機械学習手法である。

データとのマージンが最大となる識別面を学習によって獲得する。このとき、学習において必要なデータをサポートベクトルと呼ぶ。

学習データが線形分離不可能な場合は、不等式制約にスラック変数を導入する。この様なSVMをソフトマージンSVMと呼ぶ。