3-3　機械学習の基礎と展望

東京大学 数理・情報教育研究センター
2021年5月7日

概要

- 機械学習の基本的枠組みである教師あり学習のデータ分析手法（分類・回帰）と教師なし学習のデータ分析手法（クラスタリング・確率密度推定）を学びます．また機械学習を実行する上で重要な過学習の概念を通して正則化法の重要性を学ます．

本教材の目次

認識と学習

人は物体をみたときに，それが何か認識することができます．

リンゴ

たとえば，周辺環境の変化などがあっても大抵の場合，認識できます．
このような認識機能は経験に基づき学習されます．

例：リンゴの認識

- **経験**　リンゴと呼ばれる物体，それ以外の物体の複数回の観測．
- **学習**　限られた経験からリンゴ間で不偏的な視覚パターンを認知，
  リンゴを認識するルールを獲得．

機械学習

Arthur Samuel

「Field of study that gives computers the ability to learn without being explicitly programmed」（1959）

明示的なプログラムなしにコンピュータに学習能力を与える研究分野.

画像の被写体が車である事をコンピュータに判断させるようなルールを明示的にプログラムすることは非常に困難です.

▶ 代わりに機械学習では蓄積されたデータからルールを発見する学習方法自体をプログラムします.

機械学習は観測データからルールの仮説を出力する過程を実現する技術といえます.

機械学習

- データに潜むルールを自動的に見つけます.
- 人がプログラムするのは認識の仕方ではなく学習方法です.

機械学習でのルール発見に用いられるデータを訓練データと呼びます.
訓練データにない未知データに対しても正しく予測する事（汎化）が目的です.

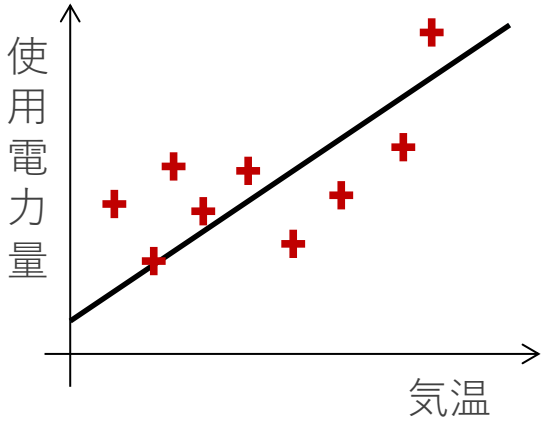
機械学習の学習方式

- **教師あり学習**
  訓練データが入力データと対応する出力値のペアからなります.
  訓練データから真の入出力関係の獲得を目指します.
  - 需要予測： 気温，地域情報から電力使用量を予測.
  - メールのスパム判定：メールの文面からスパムか否かを予測.

- **教師なし学習**
  訓練データが入力データのみからなります.
  訓練データから有益な知識の発見を目指します.
  - クラスタリング：類似したデータのグループを発見.
  - 密度推定：データを生成する確率密度関数を推定.
  - 異常検知：出現頻度の低いデータを検出.

- **強化学習**
  行動に伴い報酬が得られる設定です.
  累積報酬を最大化するような行動規則の獲得を目指します.
  - 将棋・囲碁AI：ある盤面において状況を改善するような打ち方を探索.

実世界で進む機械学習の応用と発展

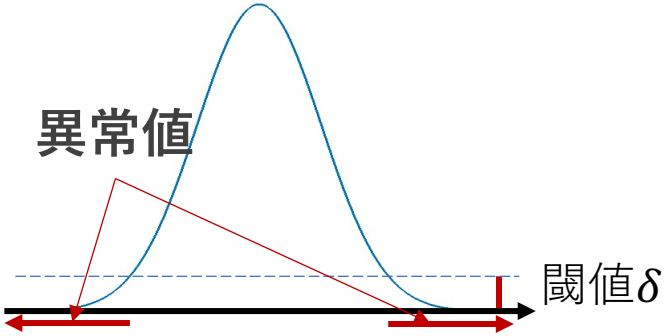


需要予測

日々の気温からその日の電力需要を予測.
教師あり学習の回帰問題で定式化.

異常検知

通常と異なるレアなイベントを検知.
例：カードの不正利用，機器の故障検知.
一般に教師なし学習として定式化.



	映画A	映画B	映画C	…
ユーザ1	4	8	?	…
ユーザ2	2	?	2	…
ユーザ3	2	?	?	…
⋮	⋮	⋮	⋮	…

商品推薦

ユーザ ✖ 映画の嗜好データから推薦.
行列の欠損補完として定式化可能.

機械学習の流れ

1. 行いたいタスクの特定.
   やりたい事は分類，回帰，クラスタリング等のうちどれかを特定します.

2. 分析に必要なデータの確認，対象となるデータの収集.
   収集すべき教師値と予測に効きそうな特徴を検討します.

3. データの分析．機械学習による学習と評価.
   データの前処理・加工と機械学習によるデータ分析を実行します.

4. データ分析結果の共有，課題解決に向けた提案.
   データ分析結果のレポートを作成し，それを材料に次の施策を検討します.

教師あり学習

訓練データは(x, y) = (データの特徴，教師値)というペアの集まりです．

教師値はデータのラベルであり回帰では連続値，分類では離散値です．

教師あり学習による予測：
訓練データから x から y を正しく予測するルールの発見を目指します．

例えば画像認識では (x, y) = (画像データ，被写体名)

$x =$  教師あり学習による予測 → $y =$ 車（正しい予測）

その他の教師あり学習の例：
売上予測 → 店舗の売上高と店舗・地理・気候等のデータで学習する．
罹患予測 → 診断結果と体温等の受診者のデータで学習する．
成約予測 → 成約結果と商品情報・顧客情報のデータで学習する．
離反予測 → 離反情報とユーザーのサービス利用状況データで学習する．

データの収集

予測対象と関連する簡単な説明変数（特徴）の作成をし，データを収集します．

- アヤメの品種予測

教師値：アヤメの品種
　　Setosa, Virginica, Versicolor

特徴：sepal.length（がく片の長さ）
　　sepal.width（がく片の幅）
　　petal.length（花びらの長さ）
　　petal.width（花びらの幅）

特徴ベクトルx　　教師値y

	sepal.length	sepal.width	petal.length	petal.width	variety
0	5.8	4.0	1.2	0.2	Setosa
1	5.0	3.3	1.4	0.2	Setosa
2	7.1	3.0	5.9	2.1	Virginica
3	4.3	3.0	1.1	0.1	Setosa

1 行が 1 データ

教師値・特徴ベクトルは連続値・カテゴリ値（離散値）として収集します．

特徴ベクトルの設計には対象データの専門知識を有効活用しましょう．

データの設計・収集がうまくいけば，あとは汎用の機械学習で分析可能です．

訓練データが多いほど予測精度は向上します，出来るだけ収集しましょう．

データの加工・前処理

機械学習を精度良く正しく実行するにはデータの前処理・加工が必要です.

タイタニックの生存結果データ（教師値：Survived）

PassengerId	Survived	Pclass	Name	Sex	Age	SibSp
1	0	3	Braund, Mr. Owen Harris	male	22.0	1
2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1
3	1	3	Heikkinen, Miss. Laina	female	26.0	0
4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1
5	0	3	Allen, Mr. William Henry	male	35.0	0
6	0	3	Moran, Mr. James	male	NaN	0

カテゴリ変数　欠損値

データクレンジング

- 外れ値検出・除去
- 欠損値除去・補完

カテゴリ値の数値への変換

- ダミー変数の追加
  例：Male, Female列を追加し0-1値に変換.

有効な前処理

- 標準化：列を平均0，分散1に変換.
- 正規化：列を$[0,1]$, $[-1,1]$に収める．列or行の平均を0にノルムを1に変換.
- 変数変換
- サンプリング：訓練データの一部を無作為に選び新たに訓練データとします.
  訓練データサイズの調整やアンサンブル学習の際に用います.

回帰と分類

教師あり学習はデータに割り当てられるラベル y が連続値・離散値（識別子，名称，カテゴリ値等）かに依って回帰・分類に分けられます.

回帰の例

- 売上予測：$y =$ 店舗の売上高，$x =$ 店舗・地理・気象・曜日情報
- 価格予測：$y =$ 住宅価格，$x =$ 最寄駅の距離，築年数，間取り，設備，地理情報

分類の例

- 罹患予測 ： $y =$ 診断結果，$x =$ 受診者のレントゲン・血液検査・体温
- 成約予測 ： $y =$ 商品の成約結果，$x =$ 商品，顧客の収入・職種・家族構成
- 離反予測 ： $y =$ 離反状況，$x =$ ユーザーのサービス利用頻度・活動内容

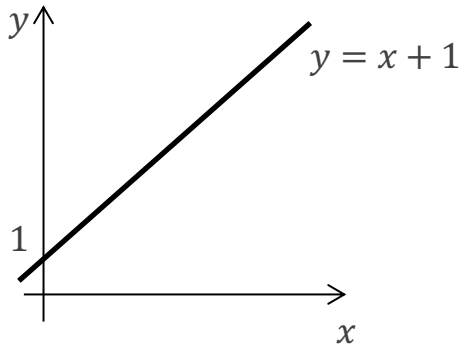
n 個の訓練データ$(x_1, y_1), \ldots, (x_n, y_n)$から写像・関数 $y = f(x)$を学習します.
f には様々なモデルがあり，モデルの選択が精度の担保に非常に重要です.

（参考）写像・関数とは？
入力xを出力yに対応付ける関係fを一般に写像といい$y = f(x)$と呼びます.
特に出力が実数値を取る対応関係は関数と呼びます.

モデル

- 関数・写像は入力xと出力yの関係を記述する：
$$y = f(x).$$

  関数の例：近隣人口xに対し売上高yの増加は線形：
$$y = x + 1.$$



- 一般には正しい係数は不明なので以下の関係を想定する：
$$y = w_1 x + w_0.$$

w_0, w_1をパラメータと呼び，パラメータを持つ関数・写像をモデルと呼びます．

パラメータの値によってこの関数・写像の対応関係が変化しますが機械学習では訓練データ$(x_1, y_1), \ldots, (x_n, y_n)$に適合するようにパラメータを決定します．

（参考）この適合度は損失関数で計算されます．データ分析手法に応じて対応する損失関数が存在します．

モデルで表される関数はデータを記述するルールの候補であることから仮説関数と呼ばれます．

種々のモデル

様々なモデルがあるのでデータに合わせ適切なものを選択します.

深層ニューラルネットも複雑なモデルの一つです.

モデルの例（赤字がパラメータ）

- $f_w(x) = w_0 + w_1 x_1 + w_2 x_2 + \cdots + w_p x_p$
- $f_w(x) = w_0 + w_1 \cos(2\pi x) + w_2 \cos(4\pi x) + w_2 \cos(8\pi x)$
- $f_w(x) = w_0 + \sum_{i=1}^{M} w_i \exp(-(x - \mu_i)^2)$

全パラメータをまとめてwと書きます. 最初の例では$w = (w_0, \ldots, w_p)$です.

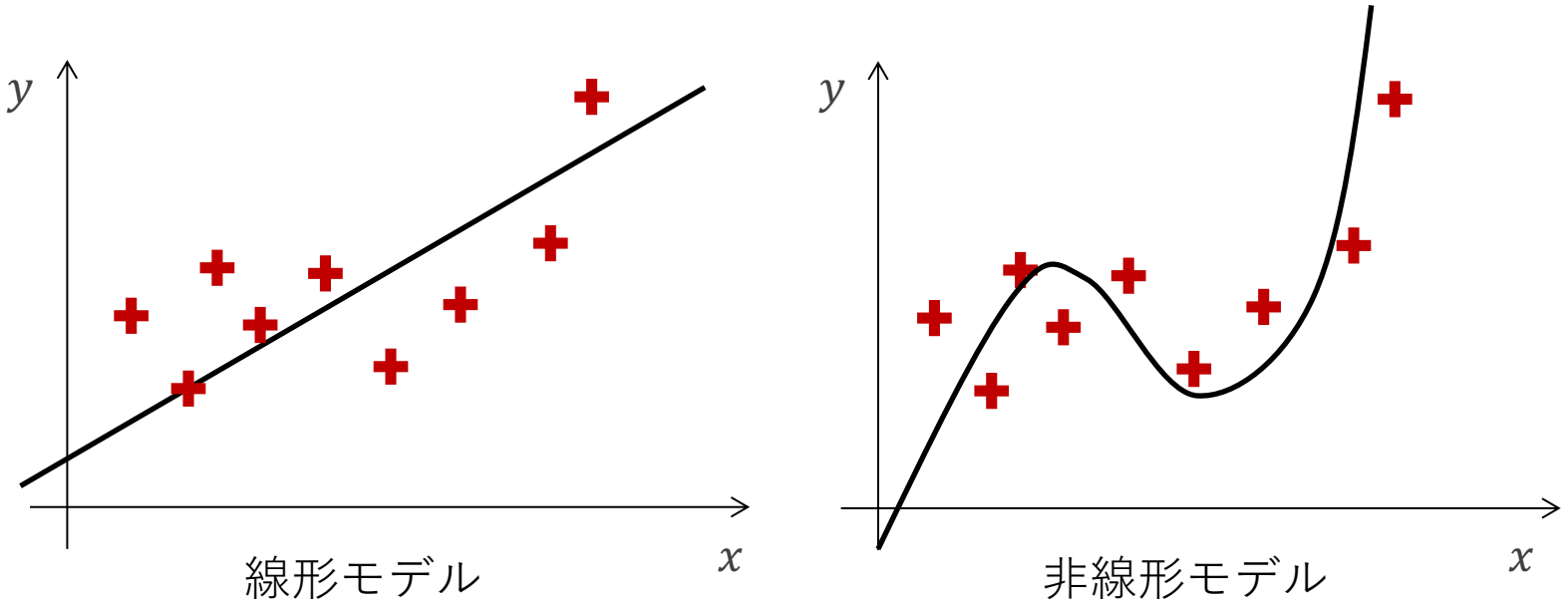
一般にパラメータwを持つモデルをf_wと書くことにします.

（参考）Σとは？共通の添字を持つ複数の数式の和を意味する記号です.
三番目の例では $w_i \exp(-(x - \mu_i)^2)$ が添字iを持つ数式で, これを$i = 1$から$i = M$まで足し上げています：
$\sum_{i=1}^{M} w_i \exp(-(x - \mu_i)^2) = w_1 \exp(-(x - \mu_1)^2) + \cdots + w_M \exp(-(x - \mu_M)^2)$.

回帰

特徴ベクトルxから連続値のラベル y を予測することを回帰と呼びます.
予測はパラメータwを持ち実数値に値を取るモデルf_wにより行います.

価格予測の例
$y =$ 住宅価格, $x =$ 最寄駅の距離, 築年数, 間取り, 設備, 地理情報



データの特徴が一変数の時は単回帰分析, 多変数の時は重回帰分析と呼びます.

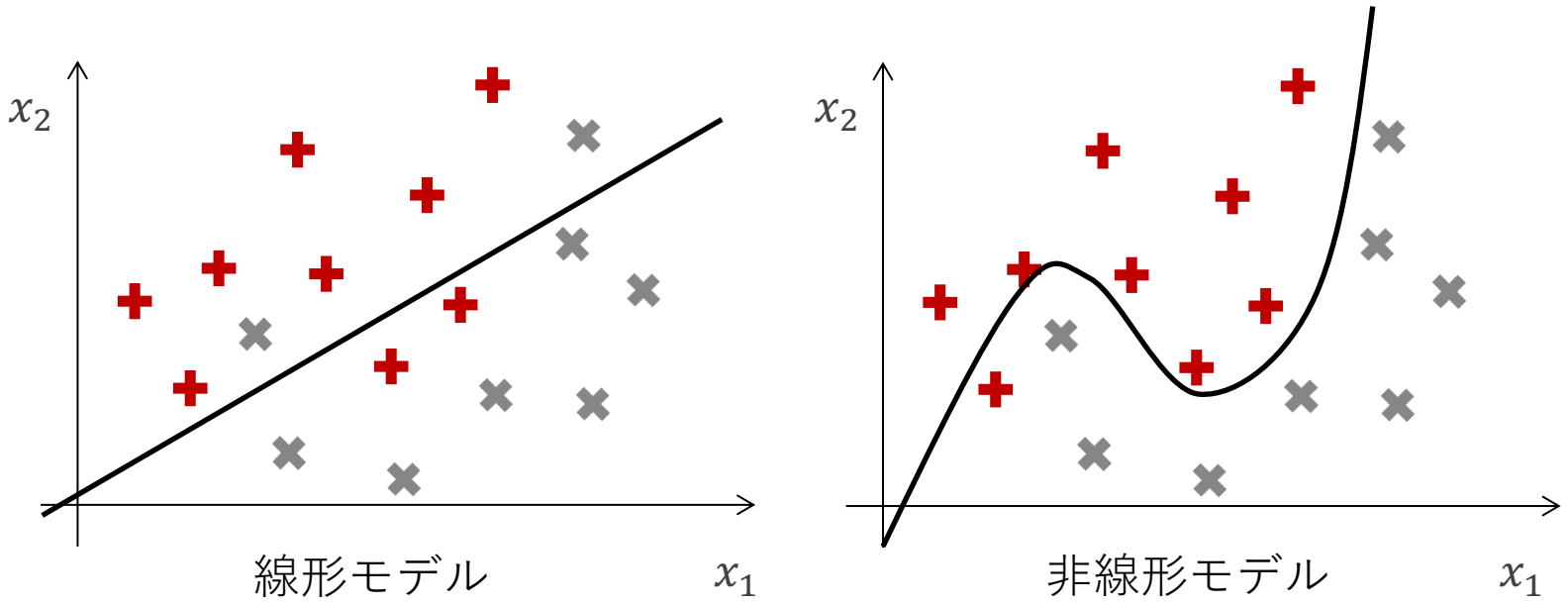
分類

特徴ベクトルxから識別子・カテゴリ等の離散値のラベル y を予測することを分類と呼びます．予測はパラメータwを持ち実数値（又は実数値ベクトル）に値を取るモデルf_wの出力を離散値に丸めて行います．

例えば診断結果等のラベルが1あるいは-1の二値で表されている時，$f_w(x)$が 0 以上であれば 1を割り当て，0 未満 であれば -1を割り当てます．

罹患予測の例

y = 診断結果，x = 受診者のレントゲン・血液・体温



損失関数

パラメータwを持つモデルf_wとデータ(x, y)の適合度合いは損失関数で定めます.
損失関数には様々な種類があり，それぞれ特有の性質を持ちます.

損失関数の例（モデルf_wは実数値に値を取るものとします．）

- 二乗損失関数

$$l(f_w(x), y) = 0.5(y - f_w(x))^2.$$

yと予測値$f_w(x)$が近い時に損失が小さく，主に回帰分析で用います.

- ロジスティック損失

$$l(f_w(x), y) = \log\left(1 + \exp\left(-y f_w(x)\right)\right).$$

yと$f_w(x)$の符号が同じで$y f_w(x)$の値が大きい程，損失は小さくなります.
ロジスティック損失を用いた分類分析をロジスティック回帰分析といいます.

経験損失最小化問題

機械学習では訓練データに平均的によく適合するパラメータを求めます.

これは次の経験損失（訓練誤差）最小化問題を解くことに帰着します:

$$\min_{w\in\Omega}\frac{1}{n}\sum_{i=1}^{n} l(f_w(x_i), y_i)\,.$$

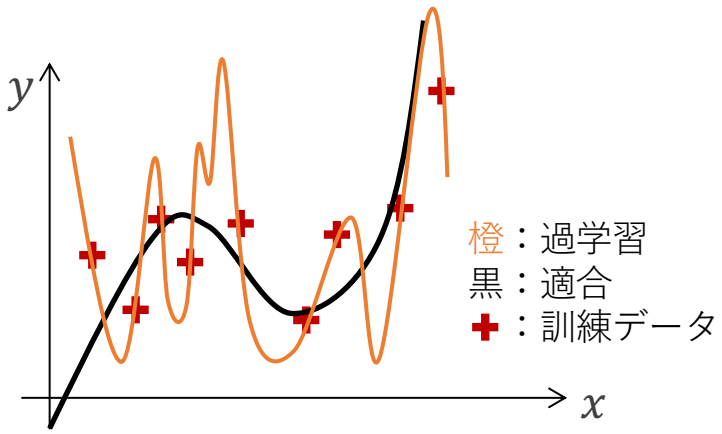
（訓練データの損失関数値の平均）

最小化対象の訓練データ上の平均損失を経験損失（訓練誤差）とよびます.

（参考）最小値記号とは？関数がパラメータを持つとき，パラメータが取り得る範囲での関数の最小値を$\min$記号で記述します．便宜的にこの最小値を求める問題自体を表す事もあります．今回の場合は損失関数値の平均を集合Ωで動くパラメータwについて最小化する問題を表しています.

過学習と正則化

訓練データに対しモデルが複雑な場合,
訓練データには適合しても未知データに
適合しない過学習という現象が起き得ます.



回帰問題での過学習の様子.

得られる関数の複雑さを抑制することで過学習を防ぐ技法を正則化と呼びます.
代表的な正則化手法として正則化付き経験損失最小化があります.
これは経験損失最小化問題に正則化項$\lambda R(w)$を加えた最小化問題です：

$$\min_{w\in\mathbb{R}^p} \frac{1}{n}\sum_{i=1}^{n} l(f_w(x_i), y_i) + \lambda R(w)\,.$$

$w \in \mathbb{R}^p$はパラメータwが要素数pの任意のベクトル値を取れることを意味します.

正則化項の例：

L_2正則化 $R(w) = \|w\|_2^2 = \sum_{j=1}^{p} w_j^2$,　　リッジ回帰 = 二乗損失 + L_2正則化

L_1正則化 $R(w) = \|w\|_1 = \sum_{j=1}^{p} |w_j|$.　　LASSO回帰 = 二乗損失 + L_1正則化

バイアスとバリアンス

機械学習の目標は**汎化誤差**の最小化です.

汎化誤差：未知データに対する平均的な損失で，厳密な計算は一般に不可能.

学習で得られた仮説関数に対する汎化誤差を構成する要素としてノイズ，
バイアス，バリアンスがあります.

とくにリッジ回帰（二乗損失＋L_2正則化）では精緻に分析されています.

- ノイズ
  データにのる観測ノイズです．無規則なためノイズ分の誤差は避けられません.
  リッジ回帰ではラベルの分散に相応する誤差が汎化誤差に含まれます.
- バイアス
  モデルがデータを生成する真の関数を含んでいないことに起因する誤差です.
  リッジ回帰では真の関数と仮説関数の二乗誤差を訓練データの取り方について平均した量に相当します.
- バリアンス
  訓練データが変わるたびに得られる仮説関数が変化することに起因する誤差です.
  リッジ回帰では訓練データを取り替えた際の仮説関数の出力値の分散に相当します.

バイアス・バリアンストレードオフ

モデルに含まれる仮説関数の多様性をモデルの複雑さとよびます．すなわち，
モデル$\mathcal{F}$の関数はすべてモデル$\mathcal{G}$に含まれるようなとき$\mathcal{G}$は$\mathcal{F}$より複雑です．

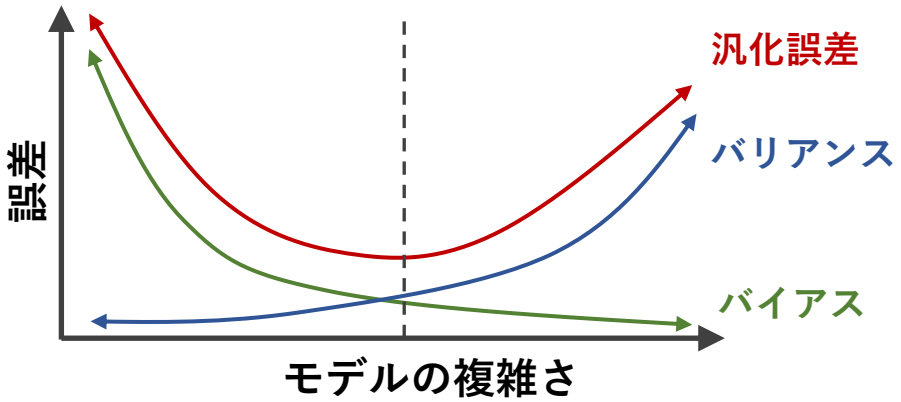
例えば多項式からなるモデルは線形関数からなるモデルより複雑です．

バイアス：モデルが複雑なほど真の関数をよく近似する仮説関数が存在．

→ モデルが複雑なほどバイアスは減少．

バリアンス：モデルが複雑なほど仮説関数の自由度が向上．

→モデルが複雑なほどバリアンスは増大．



正則化付きの学習法では正則化の係数λと得られうる仮説関数のモデルの複雑さに関係性があります．
→ 正則化係数 λ の選択の重要性．

ホールドアウト法

正則化係数を含めモデルの複雑さ（モデルの種類，使用する変数，パラメータ数など）を決定する要素は訓練損失からは定まりません．
（モデルが複雑である程，訓練損失を小さくできてしまうためです．）

ホールドアウト法：検証データで汎化誤差を見積もる手法．

1. データを訓練データ（学習データ）と検証データ（ホールドアウトデータ）に分割
2. 訓練データを用いて学習を実行．
3. 検証データに対する平均損失で汎化誤差を推定．

訓練データ（青）で学習し検証データで精度を評価．

訓練では検証データは用いないので，検証データ上での平均損失は汎化誤差の推定値になるため，この値に従いモデルの複雑さを決定します．

注意：あらゆる複雑さの候補でこの手続きを繰り返すと検証データに対しても過学習してしまいます．そのため最終的な評価は上記手続きとは別に用意したテストデータで行います．テストデータは実際の運用時の未知のデータを想定したものなのでテストスコアに依存したチューニングなどはできません．

交差検証法

正則化係数λの大きさで訓練データへの適合度合いは変わります.

交差検証法は丁度良い適切なλを選ぶ手法です.

k-交差検証法（Cross Validation, CV）

1．データをk個に分割します.

2．分割したデータの一つを検証用とし，残りのデータで学習します.

3．検証用データでの予測誤差を計算します.

4．手順２，３をk個の検証用データの取り方について繰り返します.

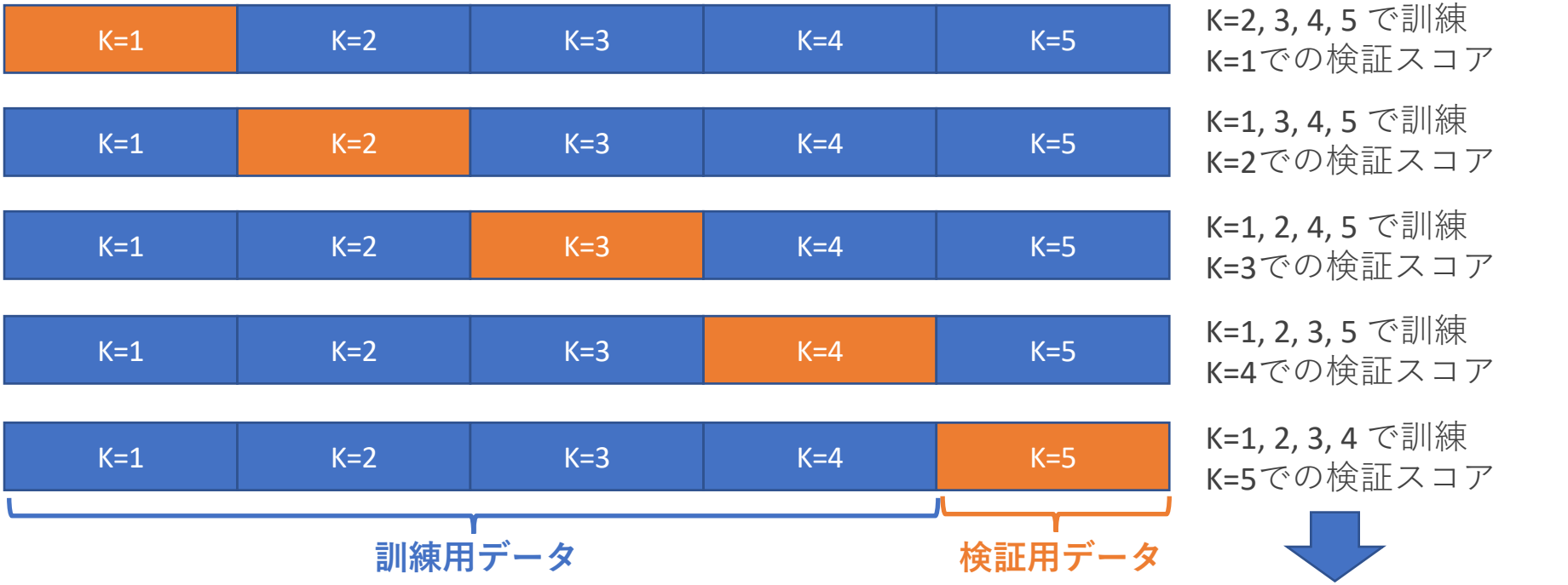
5．k個の予測誤差の平均を計算します.

ステップ5で計算される予測誤差平均をCVスコアと呼びます.

CVスコアは汎化誤差（未知データに対する誤差）の推定値であり，これを最小にするλは過学習・未学習を起こさない丁度良い値であると期待されます.

交差検証

5-交差検証の実行イメージ

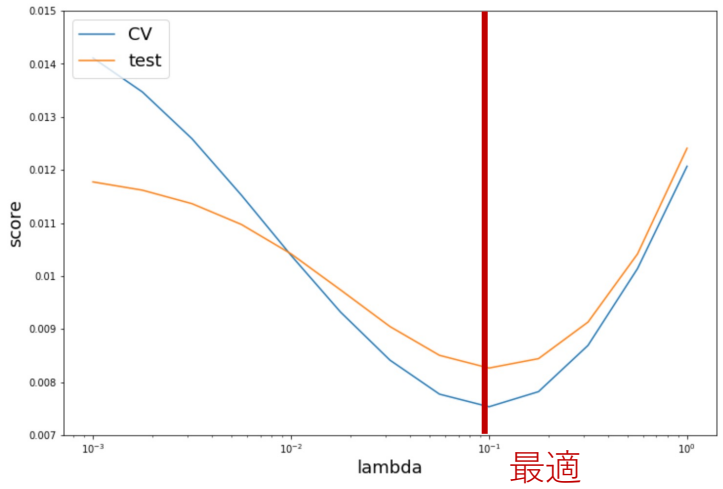


交差検証の実行例

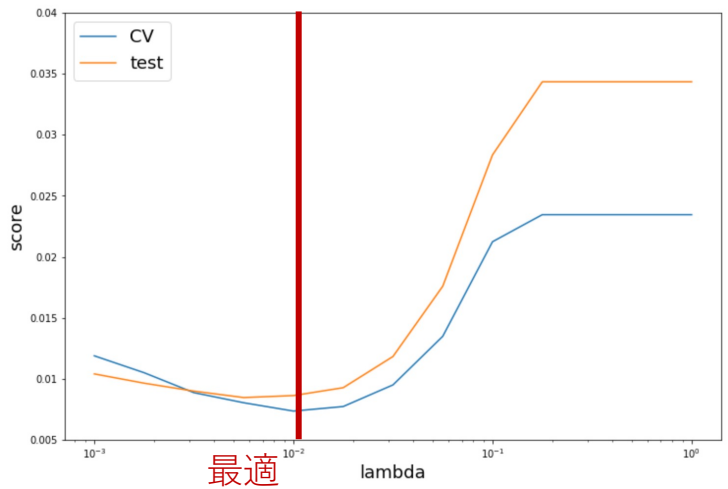
リッジ回帰，LASSO回帰で交差検証の有効性を確認してみます．．

バイアス・バリアンストレードオフから一般に λは小さすぎても，大きすぎても汎化誤差は大きくなることが期待されます.

リッジ回帰



LASSO回帰



図：正則化係数（lambda）に対しCVスコア（青）とテストスコア（橙）をプロット.

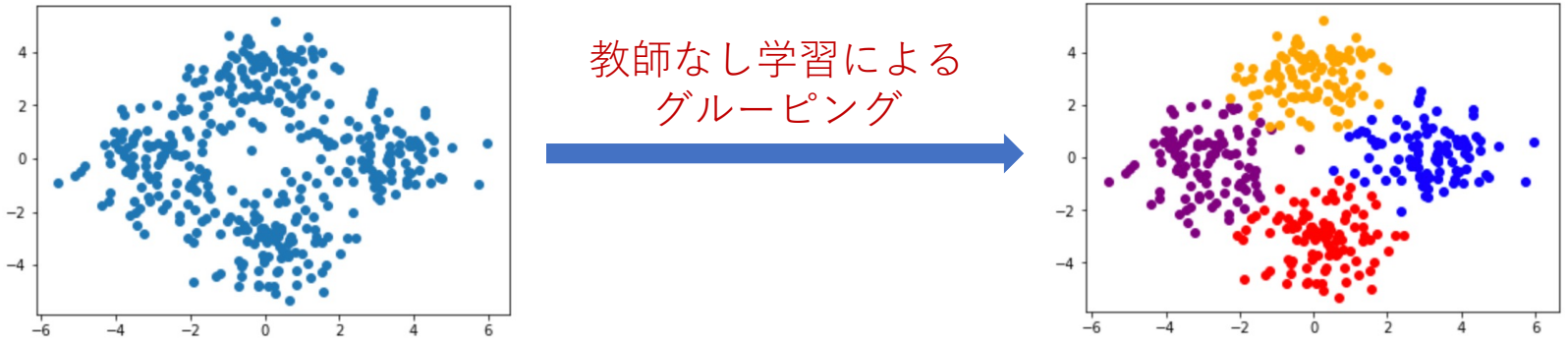
いずれもCVスコアとテストスコアが共通の傾向を示していて，
CVスコアをもとに正則化係数を決定すると良いテストスコアが達成されることが分かります.

教師なし学習

訓練データは特徴ベクトルのあつまり$x_1, \ldots, x_n$です.

教師値が与えられない点が教師あり学習と異なります.

教師なし学習によるグルーピング：データ同士の類似性に基づくグルーピング（クラスタリング）は教師なし学習の代表例です.



クラスタリングの例

顧客セグメンテーション→ 顧客情報や利用状況によるクラスタリング

店舗クラスタリング→ 店舗や周辺情報からクラスタリング

教師なし学習

クラスタリングの他にも様々な教師なし学習手法があります.

- 確率密度推定
  データの生成分布を推定します．汎用的なタスクでクラスタリング・異常検知・データ生成への応用が可能な手法もあります.
- 異常検知
  センサーデータ等から異常値の検出をします.
- レコメンデーション
  ユーザーの購入履歴や嗜好情報からアイテムの推薦をします.
- データ生成
  データの生成過程を推定しデータの生成・サンプリングをします.

（補足）教師あり学習では特徴ベクトル$\boldsymbol{x}$のラベル$\boldsymbol{y}$の予測を目的としました．一方，教師なし学習では$\boldsymbol{x}$の背後にある構造や生成過程の推定を目的とします.

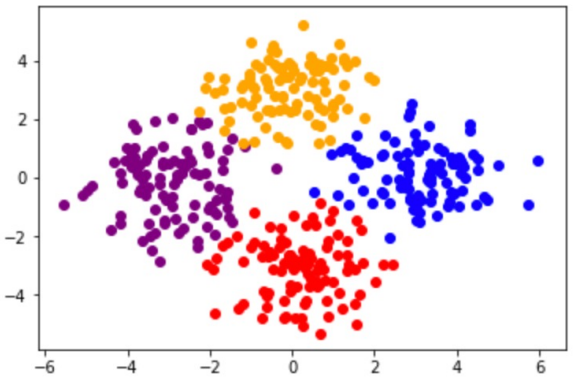
クラスタリング

データには潜在的なカテゴリがあり，それを推定します.

クラスタリングの例

- テキストデータ
  テキストのジャンル（ミステリー・SF・アクション・純文学）
- ECサイトの購入履歴データ
  趣味嗜好やプロフィールに依存した購入パターン

- **データの類似度に基づく方法**
  似たデータは同じクラスタ，似ていないデータは異なるクラスタに属するという志向のもとでクラスタリングを実行します.
  代表例：K-平均法（K-Means法）
- **確率密度推定に基づく方法**
  多峰性を持つ確率密度関数による確率密度推定を行い，各データが属する峰を特定しクラスタリングします.
  代表例：混合正規分布によるクラスタリング



K-平均法によるクラスタリング

K-平均法（K-Means法）

データを以下の方針で指定のK個にクラスタリングします.

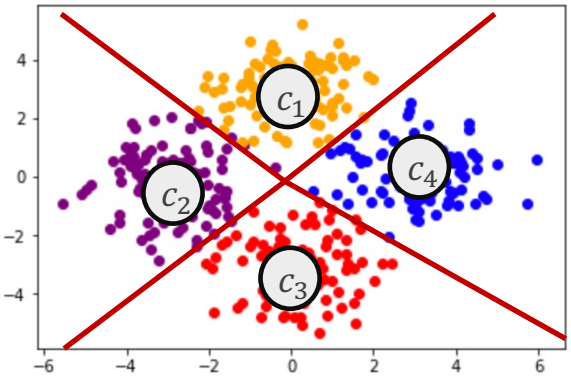
- 各クラスタに$t = 1, \ldots, K$に対応する中心点 c_tを求めます.
- 各データは最も近い c_t が代表するクラスタに割り当てます.

K-平均法の実行にはデータの類似度を示す距離関数を指定する必要があります.

距離関数には様々なものがありますが，代表例はユークリッド距離になります.

（参考）ユークリッド距離は2点 $x_1 = \left(x_1^{(1)}, \ldots, x_1^{(p)}\right), x_2 = \left(x_2^{(1)}, \ldots, x_2^{(p)}\right)$に対し次で定義される距離関数です.

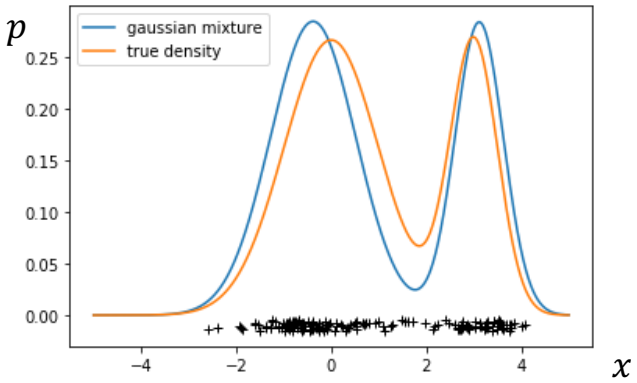
$$d(x_1, x_2) = \left(x_1^{(1)} - x_2^{(1)}\right)^2 + \cdots + \left(x_1^{(p)} - x_2^{(p)}\right)^2.$$



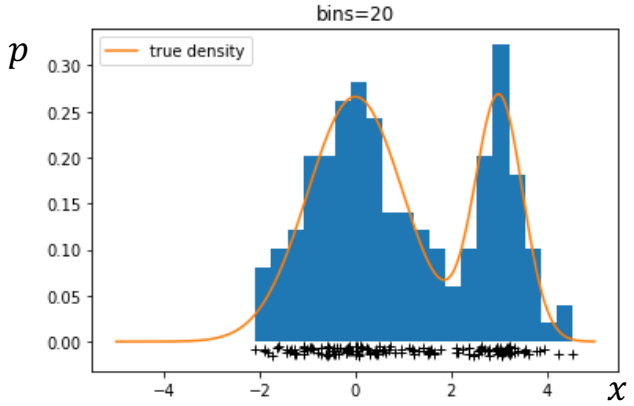
K-平均法によるクラスタリングの様子.
クラスタの中心点とクラスタリング結果が得られます.

確率密度推定

訓練データ $x_1, \ldots, x_n$ からデータを生成する確率密度関数 $p(x|w)$ を推定します．
確率密度推定法として最尤推定法やヒストグラムによる推定があります．



混合正規分布による最尤推定



ヒストグラムによる確率密度推定

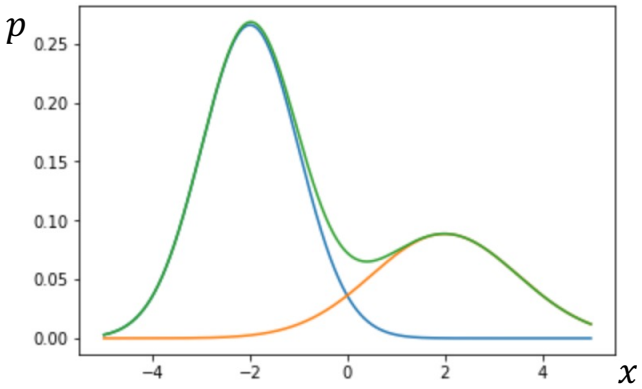
横軸はデータxの座標，縦軸は確率密度 $p(x|w)$ です．橙のグラフは正解の確率密度関数，左図の青のグラフと右図の青のヒストグラムはそれぞれの推定結果です．

（参考）確率モデルの最尤推定とは？
確率モデル $p(x|w)$はパラメータwを持つデータxについての確率密度関数で，wに依存して確率密度は変化します．最尤推定は訓練データ $x_1, \ldots, x_n$ を最も高い確率で発生させるようなパラメータwを求める手法です．最尤推定によりデータの背後に潜む確率密度関数が推定されることになります．

混合正規分布による確率密度推定

混合正規分布とは複数の峰を持つ確率分布（確率密度関数）で，パラメータにより峰の位置と裾の広さが調整されます．

データが複数のまとまったクラスタを持つ場合の確率密度推定に適した確率分布といえます．



緑：2つの峰を持つ混合正規分布
橙，青：各々の峰を描いたグラフ．

横軸はデータ x の座標，縦軸は確率密度を示しています．

確率密度推定を行う際に混合数はユーザーが指定する必要があります．

混合正規分布による確率密度推定ではデータが属す峰も推定する事ができ，結果としてクラスタリングも実行されます．

（参考）峰が一つの混合正規分布は単に正規分布と呼びます．

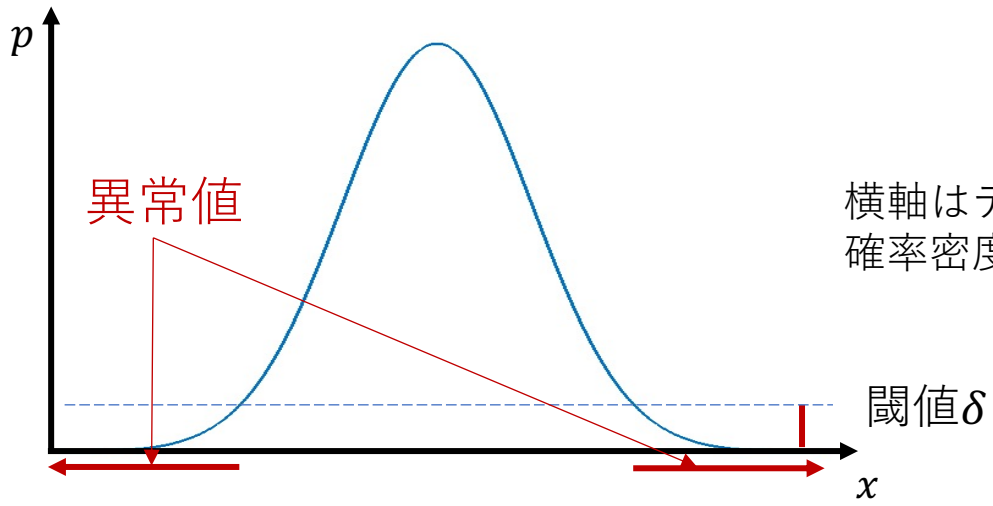
確率密度推定の異常検知への応用

異常検知は非常に多くの応用例があります：

工場のセンサーデータ，カードの使用履歴，監視カメラ，為替取引

正常データ・異常データが共に十分あれば教師あり学習が適用可能ですが一般に異常データは非常に少ないため教師なし学習を適用します．

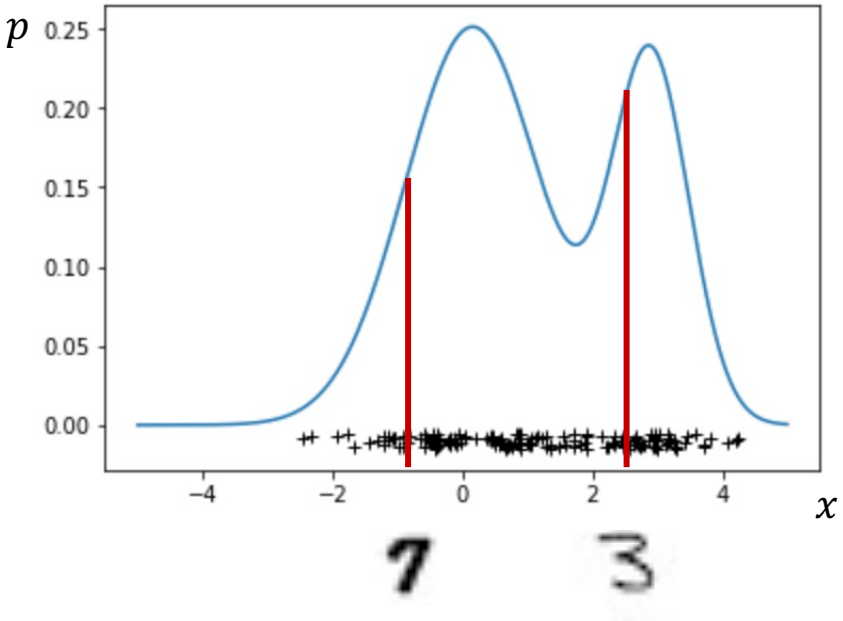
典型的なアプローチ：確率密度推定の結果 $p(x|\hat{w})$ と閾値 δ を設定し $p(x|\hat{w}) \leq \delta$ を満たすデータ x を異常とみなします．

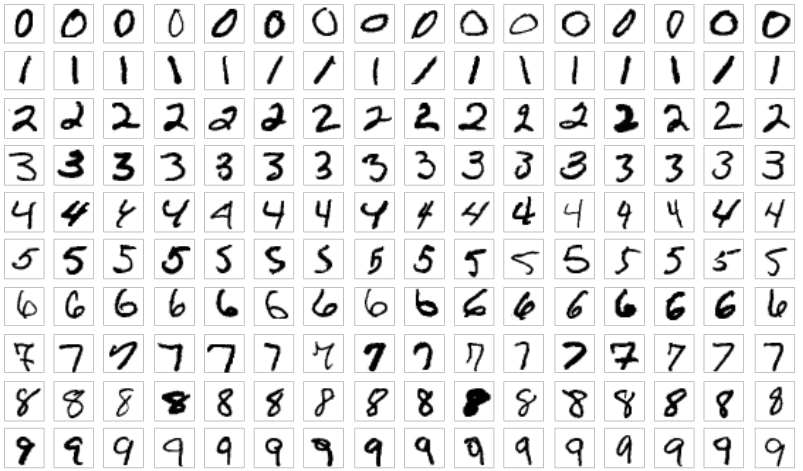


横軸はデータ x の座標，縦軸は推定された確率密度関数です．

確率密度推定のデータ生成への応用

確率密度推定によりデータの生起確率が分かるため，推定された確率密度関数からサンプリングする事で尤もらしいデータの生成も可能です．





MNIST: 手書き文字データセット

MNISTで学習した確率モデルから確率の高いデータをサンプリングするとそれらしい手書き文字が得られます．

レコメンデーション

ユーザーによるアイテムの評価値情報が部分的にある時，欠損値を推定することで評価値を推定しアイテムの推薦をします．

	アイテムA	アイテムB	アイテムC	…	アイテムZ
ユーザー1	4	8	*	…	2
ユーザー2	2	*	2	…	*
ユーザー3	4	7	*	…	*
⋮	⋮	⋮	⋮	…	⋮

嗜好の近いユーザー

嗜好の近いユーザー同士は欠損値でも近い評価をするだろうというモデリングに従い欠損補完が可能です．

（参考）この様な欠損値補完手法として行列分解等があります．

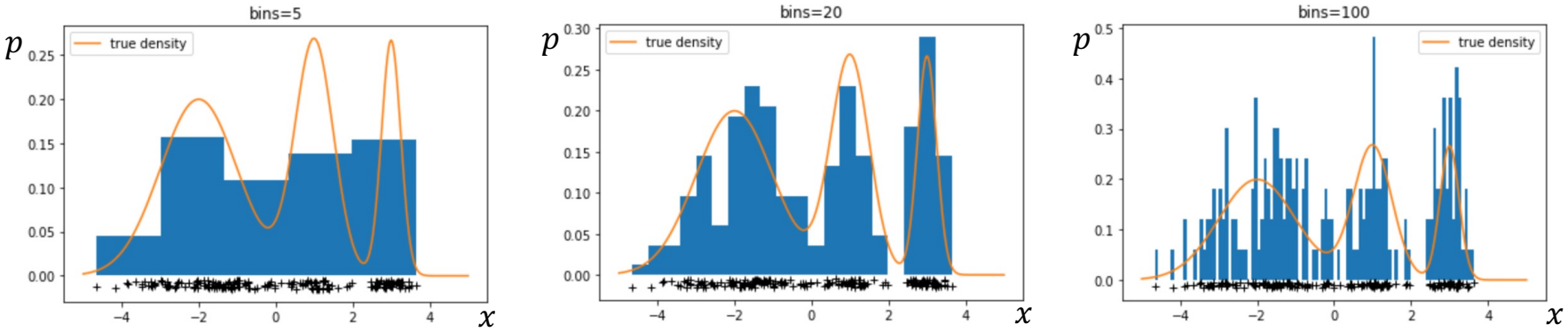
応用例：商品販売サイトでの商品推薦，動画サイトでの動画推薦．

過学習とモデル選択

教師なし学習でも有限個の訓練データにモデルを適合するため,
訓練データに対しモデルが複雑過ぎる場合に過学習が生じます.
過学習を防ぐためにモデルの表現力をコントロールする必要があります.

モデルの表現力をコントロールするハイパーパラメータの例

- K-平均法の中心点の数
- 混合正規分布の混合数
- ヒストグラムのビン幅（ヒストグラムの各帯をビン，その幅をビン幅と呼びます）



ヒストグラムのビン幅（左から5, 20, 100）で未学習・過学習が生じる様子.
横軸はデータの座標，縦軸は確率密度，橙のグラフは正解の確率密度関数です.

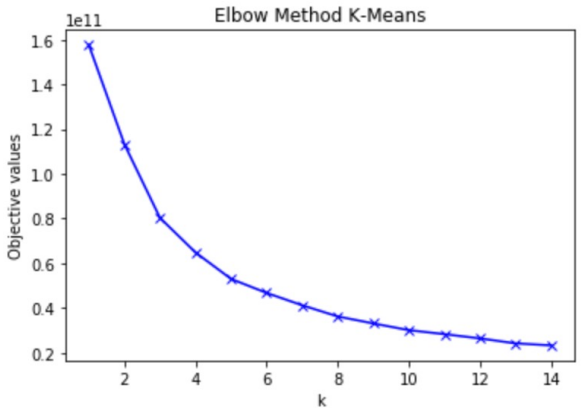
モデル選択の方法

教師なし学習でも適当な評価指標を用いた交差検証で適用すべきモデルの選択やハイパーパラメータの決定は可能です.

しかし教師あり学習におけるラベルの推定精度のようなモデル非依存の評価指標がない問題（クラスタリングや確率密度推定）では，手法毎にモデル選択法が提案されています.

例：K-平均法のエルボー法

Kの値毎にK-平均法を実行し目的関数値（データへの適合度）をプロットします. 基本的にKについて単調に減少しますが減少幅が十分に小さくなってきたKを適切なものとし選択します.

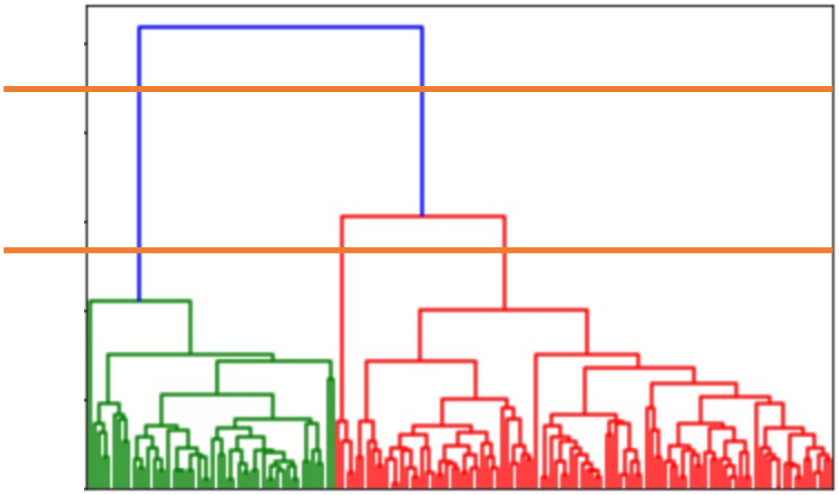


エルボー法の実行例.
縦軸はK-平均法の目的関数値です.
K=5あたりで減少が落ち着いてきます.

階層クラスタリング

これまで紹介したクラスタリング手法は非階層クラスタリングです.
一方，階層的にクラスタを形成する階層クラスタリングという手法もあり，
データ間の関係性に階層構造がある場合のクラスタリングに有用です.
例：生物や遺伝子配列の分類.

階層クラスタリング手法では各データが一つのクラスタを形成している状態から近いクラスタ同士を順次併合していき，一つのクラスタになるまで繰り返します.
結果，右図のような系統樹が得られます.



階層クラスタリングの利点：
系統樹の高さを指定する事でクラスタ数
を学習後に調整する事が出来ます.