機械学習特論演習

最終更新日：2024 年 12 月 12 日

授業の概要の方法：

機械学習特論で講述した内容の理解を深めるため、機械学習の代表的手法についての演習を行う。本演習では、予測や分類課題にベイズ的手法を用いて、実際のデータに対して解析を行うことを目的とする。

受講生の到達目標：

本講義を通して学生は以下の知識を取得する。
(1) 代表的な確率密度分布を用いたベイズ推定
(2) (1)に基づく回帰や分類などのデータ解析の実現法
(3) Python を用いた実データ解析のプログラミング

事前に履修しておくことが望まれる科目：

本演習は「機械学習特論」の講義を踏まえた演習であるため、機械学習特論を履修の上、受講すること。「確率・統計」や「多変量解析」の知識を前提とする。Python をプログラミング言語として用いるため、その基本的な文法知識は、関連授業や自習により事前に身につけておくことが望ましい。

授業回数 Lecture	テーマ ／ Theme キーワード ／ Key Word
1	ガイダンス・準備
	演習の概要、Python/Google Colab 操作の復習、確率分布の基礎
2	離散確率分布の学習と予測の演習
	ベルヌーイ分布、二項分布、カテゴリ分布、多項分布、ポアソン分布
3	連続確率分布の学習と予測の演習
	一様分布、ガンマ分布、ベータ分布、ディリクレ分布、ガウス分布
4	応用問題の発表（1）：
	確率分布の学習と予測の演習の発表
5	ベイズ推定を用いた回帰の演習（1）
	多項式回帰、単回帰
6	ベイズ推定を用いた回帰の演習（2）
	重回帰、一般化線形モデル、事後分布と予測分布の計算
7	ベイズ推定を用いた回帰の演習（3）
	回帰の応用問題
8	応用問題の発表（2）：
	実データを用いた回帰分析の発表
9	回帰の階層モデルの演習（1）

	単回帰、重回帰、ロジスティック回帰の階層モデル
１０	回帰の階層モデルの演習（２）
	回帰の階層モデルの応用問題
１１	応用問題の発表（３）：
	実データを用いた階層モデルの発表
１２	ガウス混合モデルによるクラスタリング（１）
	ギブスサンプリング、変分推論、崩壊型ギブスサンプリング
１３	ガウス混合モデルによるクラスタリング（２）
	無限混合ガウスモデル
１４	ガウス混合モデルによるクラスタリング（３）
	混合ガウスモデルなどの応用問題
１５	応用問題の発表（４）：
	実データを用いた総合演習の発表

授業実地形態：

- 対面授業：3, 4, 7, 8, 10, 11, 14, 15（計 8 回）
- Web(ライブ)授業：1, 2, 5, 6, 9, 12, 13（計 7 回）

授業外学習の指示：

前期に実地した講義を元に演習を実地するので、講義内容をいつでも復習できるようにしておくこと。教科書を深く理解しておくこと。Python プログラミングの文法を復習しておくこと。Google Colaboratory の利用するためアカウントを準備しておくこと。

成績評価方法：

定期試験 0, レポート試験 0, 平常点評価 100 ：演習中に出題する課題に対して、レポート提出を 4 回課し、その課題の達成度により評価します。

教科書：

書名：ベイズ推論による機械学習入門 [Amazon]
著者：須山敦志
出版社：講談社
ISBN コード：978-4-06-153832-0
演習時間内にレジュメを配布する。課題となるプログラムの作成の理論的知識として教科書を利用する。

参考書：

書名：Stan と R でベイス統計モデリング
著者：松浦 健太郎
出版社：共立出版
ISBN コード：978-4-320-11242-1
R 版ですが Stan の記述法について参考になります。階層モデルも詳しいです。

書名：イラストで学ぶ 人工知能概論 改訂第 2 版
著者：谷口 忠大
出版社：講談社
ISBN コード：978-4-06-521884-6

授業内外における学生・教員間のコミュニケーションの方法：

- 学生との直接対話（Zoom を含む）
- メール（yukako@fc.ritsumei.ac.jp）

Python 環境についての注意：

演習では Google Colaboratory（https://colab.research.google.com/; 以下、Google Colab）で Python3 を利用します。numpy, matplotlib, pandas, scipy, statsmodels, pystan, arviz ライブラリの利用を予定しています。Anaconda などを使ったローカルの環境は、pystan のインストールで問題が起きる可能性があり、環境設定の問題が生じても時間と TA 数上サポートできません。注意して下さい。

その他：

Zoom は以下の URL を使います。
https://ritsumei-ac-jp.zoom.us/j/97035445036
ミーティング ID: 970 3544 5036

資料は Dropbox に共有し順次更新します。
https://www.dropbox.com/scl/fo/cqtmbafc2iz8z608hykmc/AEvBni0D8ECq-mmzDwSu5eo?rlkey=ty0921kdl7m04mmxwbomoxsrx&st=vnnk77g2&dl=0

本演習をサポートしてもらう TA さんは以下の方です。
- 金村　一輝さん

Google Colab でテキストファイルを利用する手続き：

Google Colab にアップロードしたファイルは最大 12 時間で自動的に消えます。そのため、Google ドライブにファイルを置き、Colab で Google ドライブの領域をフォルダとしてマウントし(ショートカットのようなもの)、Colab からマウントしたフォルダ中にあるファイルを利用する方法を推奨します。他に得意な方法がある人はその方法を使って下さい（より良い方法があれば、情報共有してもらえたらありがたいです）。

推奨手順：

1. Google ドライブ（https://drive.google.com/）に、例えば input という本演習で使うフォルダを作成し、本演習に必要なファイルをアップロードして下さい。マウスで右クリックすれば「新しいフォルダ」や「ファイルアップロード」が選択できます。
2. Google Colab で、以下のコマンドを実行し Google ドライブをマウントします。
   from google.colab import drive
   drive.mount('/content/drive')
3. 必要に応じて認証作業をします。
   - 認証が必要な場合に表示される URL をクリックします。URL をクリックします。
   - Google のアカウントを選択し「ログイン」ボタンを押してください。または、認証する Google のメールアドレスなどを入力し「次へ」をクリックし、許可項目を確認した上で「許可」ボタンを押します。
   - 「このコードをコピーし、アプリケーションに切り替えて貼り付けてください」というメッセージの後ろにある認証コードをコピーします。
   - Google Colab 側で「Enter your authorization code:」下にある空白に、認証コードを貼り付け実行するとマウントが完了し「Mounted at /content/drive」と表示されます。

```
from google.colab import drive
drive.mount('/content/drive')
```

```
Mounted at /content/drive
```

   - Colab 側で /content/drive/MyDrive/ 以下が Google ドライブの領域になります。ls などで状態を確認して下さい。
     !ls /content/drive/MyDrive

```
!ls /content/drive/MyDrive
```

```
'Colab Notebooks'   input
```

   - 1.でアップロードしたファイルに cat などでアクセスして下さい。
     !cat /content/drive/MyDrive/input/beer-sales-1.csv

4. Python/pandas でファイルを読む方法は、例えば以下の通りです。

   import pandas as pd
   input_dir = "/content/drive/MyDrive/input/"
   data = pd.read_csv(input_dir + 'beer-sales-1.csv')
   data.head()

```
import pandas as pd
input_dir = "/content/drive/MyDrive/input/"
data = pd.read_csv(input_dir + '2-4-1-beer-sales-1.csv')
data.head()
```

	sales
0	87.47
1	103.67
2	83.29
3	131.91
4	106.59

参考：

- Google ドライブの使い方を簡単解説！
  http://one-u.jp/gsuite/2019/08/12/google-drive-howto/
- Google ドライブにマウントファイルへアクセスする方法
  https://blog.kikagaku.co.jp/google-colab-drive-mount

第 1 回「Google Colaboratory で Python プログラミング基礎」課題

Google Colab の操作に慣れることを目的に、Python の基本文法とライブラリの利用法を復習します。

- 基礎課題 1-1：note1-1(Python の基本構文)、note1-2（NumPy と Matplotlib）, note1-3(Pandas によるファイル入出力)を実行し、内容を確認せよ。
- 基礎課題 1-2：NumPy のライブラリの exp 関数を使い、式(1)のロジスティック関数を求める独自関数を作成せよ。入力引数はa, k, x_0, xとすること。

$$F(x) = \frac{k}{1 + \exp\left(-ak(x - x_0)\right)} \tag{1}$$

そして、作成した関数を使い、$a = 1, k = 1$を指定し、$x_0 = 0,2,4$と変化させた場合の図 1 のグラフを作成せよ。

- 基礎課題 1-3：基礎課題 2 の計算結果を Pandas の DataFrame 型の変数に格納し CSV ファイルに保存せよ。
- 基礎課題 1-4：基礎課題 3 で保存した CSV ファイルを読み込み、グラフで表示せよ。
- 基礎課題 1-5： beer-sales-1.csv (共有フォルダに配布)を Google ドライブに格納し、それをプログラムで読み込み、図 2 のグラフを作成せよ。

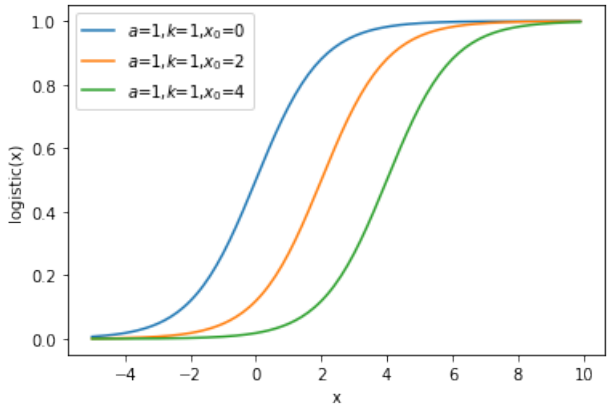


図 1：logistic 関数

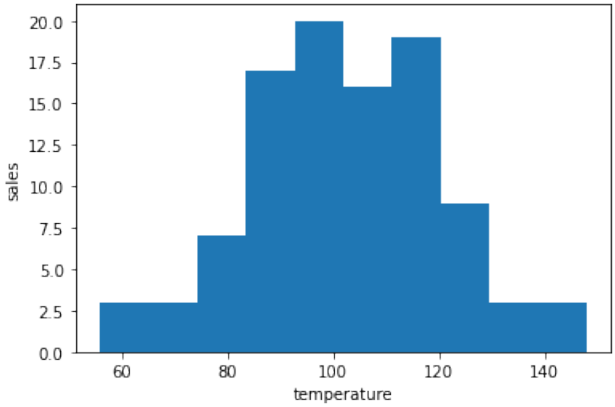


図 2：beear-sales-1 のヒストグラム

Python に慣れている人は、第 2 回の基本課題 2-1 と基本課題 2-2 にも取り組んで下さい。加えて、第 2 回から Python のライブラリで Stan というベイズ統計のライブラリを使います。興味ある人は次回までに調べておいて下さい。確率的プログラミング言語の Stan を Python で利用可能とするインターフェースを提供するライブラリです。R 言語における Stan は Python よりも歴史的に古く、より多くの情報があります。

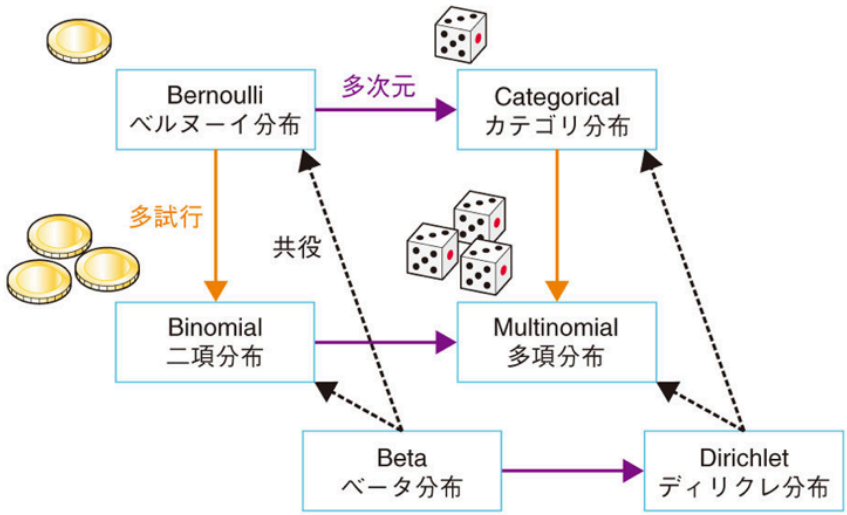
参考：

- Google Colaboratory とは?使い方・メリット・設定などを徹底解説！
  https://blog.kikagaku.co.jp/google-colab-howto
- Stan 超初心者講座@SlideShare
  https://www.slideshare.net/simizu706/stan-62042940

第 2 回「離散確率分布の学習と予測」課題：

教科書の第 2 章「基本的な確率分布」を復習します。さらに、第 3 章「ベイズ推論による学習と予測」に入ります。なお第 3 章では、基本的な確率分布のパラメータを、共役事前分布を使うことで解析的に学習する方法について示しています。しかし、解析的な学習法には限界があります。そこで、事後分布を近似的に解く「マルコフ連鎖モンテカルロ法」を実装したプログラミング言語 Stan の用いる方法も並行して学びます。Stan の補足説明の資料も確認してください。

- 基礎課題 2-1：基本的な離散確率分布として、note2-1(ベルヌーイ分布、二項分布、カテゴリ分布、多項分布)を実行し、内容を確認せよ。



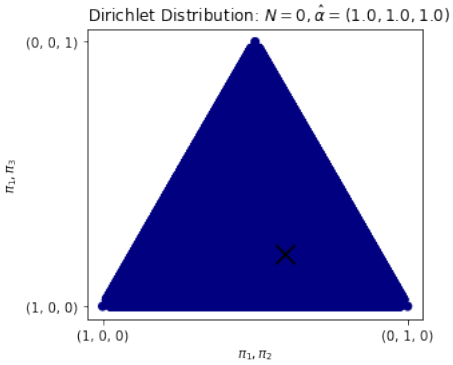
(教科書 2.2.4 節より抜粋)

- 基礎課題 2-2：基本的な連続確率分布として、note2-2(一様分布、ガンマ分布、ベータ分布、ディリクレ分布、正規分布)を実行し、内容を確認せよ。
- 基本問題 2-3：2 次元正規分布のランダムサンプリングと可視化を実現せよ。
- 基礎課題 2-4：note2-4(ベルヌーイ分布の解析的な学習と予測)を実地せよ。
- 基礎課題 2-5：note2-5(ベルヌーイ分布の Stan を用いた学習と予測)を実地せよ。
- 基礎課題 2-6：カテゴリ分布の解析的な学習と予測を実現せよ(図 3)。
- 基礎課題 2-7：カテゴリ分布の Stan を用いた学習と予測を実現せよ(図 4)。
- 基礎課題 2-8：ポアソン分布の解析的な学習と予測を実現せよ。
- 基礎課題 2-9：ポアソン分布の Stan を用いた学習と予測を実現せよ。

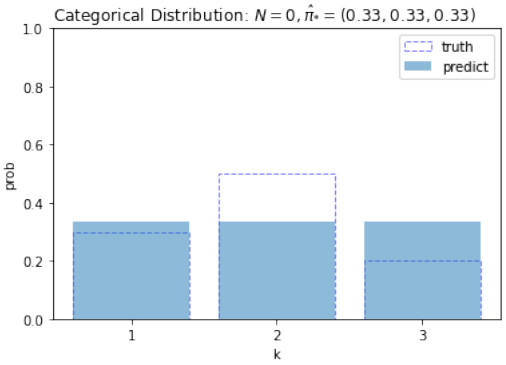
参考：

- Stan モデリング言語：ユーザーガイド・リファレンスマニュアル
  https://stan-ja.github.io/gh-pages-html/

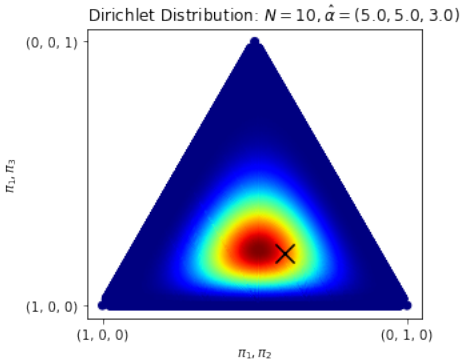
事前分布 N=0：



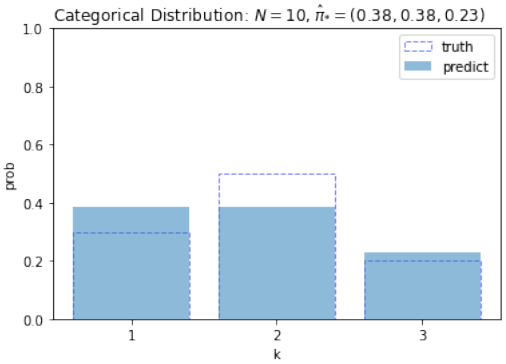
予測分布 N=0：



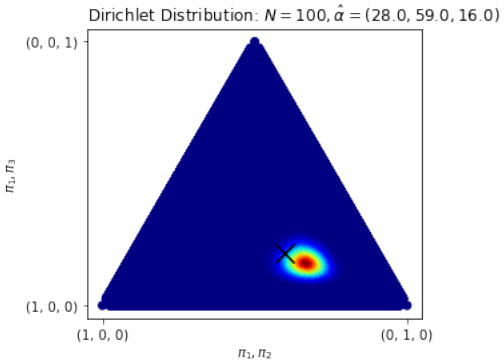
事前分布 N=10：



予測分布 N=10：



事前分布 N=100：



予測分布 N=100：

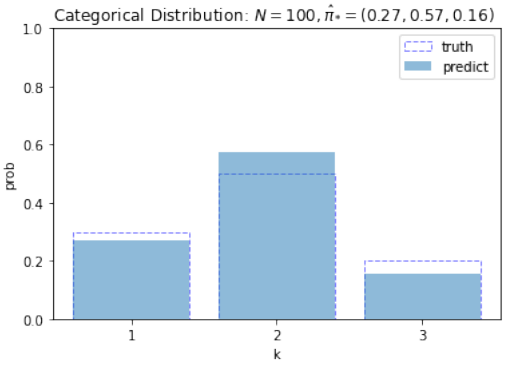


図 3：カテゴリ分布の推定の例

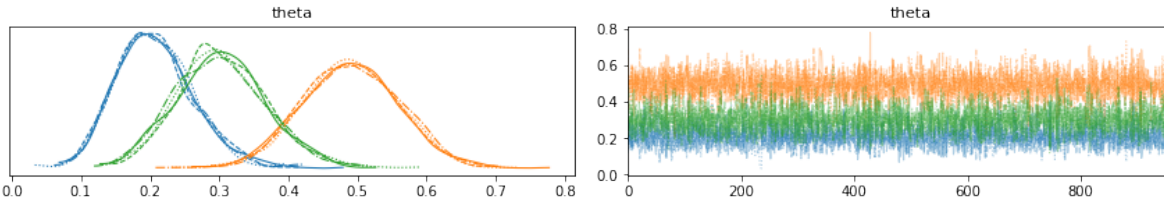


図 4：カテゴリ分布の Stan による推定の例

第３回「連続確率分布の学習と予測」課題：

教科書の第 3 章の 3.3.3 節の内容まで進んだのち、応用課題を提示します。

- 基礎課題 3-1：note3-1(1 次元正規分布の解析的な学習と予測)を確認せよ。
- 基礎課題 3-2：note3-2(1 次元正規分布の Stan を用いた学習と予測)を確認せよ。
- 応用課題 3-1：乳がん診断データ(breast-cancer-wisconsin-data.csv) にベルヌーイ分布を仮定し、ベイズ推定でパラメータを求め、陽性率（検査したとき悪性と判断される人の割合）を解答せよ。なお、CSV ファイルの diagnosis という項目が診断結果に該当し、M = malignant (悪性), B = benign (良性)という意味である。回帰は次回の応用問題の課題となるため分布のみ予測すること。
  - 元データ: https://www.kaggle.com/uciml/breast-cancer-wisconsin-data
- 応用課題 3-2：アヤメの種類のデータ(iris-data.csv)では、CSV ファイルの target という項目がアヤメの種類データに該当し、その種類は、setosa, virginica, versicolor に分類されている。カテゴリ分布を仮定し、パラメータを求めよ。回帰は次回の応用問題の課題となるため分布のみ予測すること。
- 応用課題 3-3：1000g として売っているパンの重量データ（bread-1000g-data.csv）に正規分布を仮定し、パラメータを予想せよ。

第４回「応用課題」の発表：

応用課題 3-1, 3-2, 3-3 を解き、課題内容、用いた確率分布（モデル）とベイズ推定の過程、結果、考察を説明した資料を PowerPoint で作成して下さい。manaba+R にその説明資料(*.pptx または *.pdf)およびソースコード(*.ipynb または *.py)、（入力データを特殊な形式に変更した場合のみ変更したデータも）提出すること。

プログラム言語は Python を使って下さい。なお、学習や予測の方法は、各自の判断で選んで下さい。解析的に解いても良いですし、PyStan でも、PyMC3 や Edward、TensorFlow Probability など他の得意な確率的プログラミングのライブラリを用いても良いです。

当日は、１課題について 3-4 人に発表していただきます。発表時間は 1 人 5 分程度とします。
提出期限：10 月 19 日(土) 23:59 まで（当日までに発表者を manaba+R で連絡します）

第 5 回「ベイズ推定を用いた回帰（1）」課題：

教科書の第 3 章の 3.5 節に入り、多項式回帰のモデルを構築します。

- 基礎課題 5-1：note5-1(多項式回帰)で、観測モデルを 1 次元正規分布、事前分布を多変量正規分布としたと場合の回帰を確認せよ。
- 基礎課題 5-2：note5-1 を元に、図 5 のように教科書 図 3.8 を再現し、さらに N や M、M_{truth}、を変えた時の予測結果から、データ密度の違いによる予測結果の違いを確認せよ。なお、np.random により生成される乱数は、np.random.seed を使うことで固定することができる。これは、一度発生させた乱数を再現する必要がある場合に便利な機能である。
- 基礎課題 5-3：教科書の式(3.157) の周辺尤度を求め、教科書 図 3.9. を再現し、さらに N や M を変えた予測結果から、周辺尤度の違いを確認せよ。

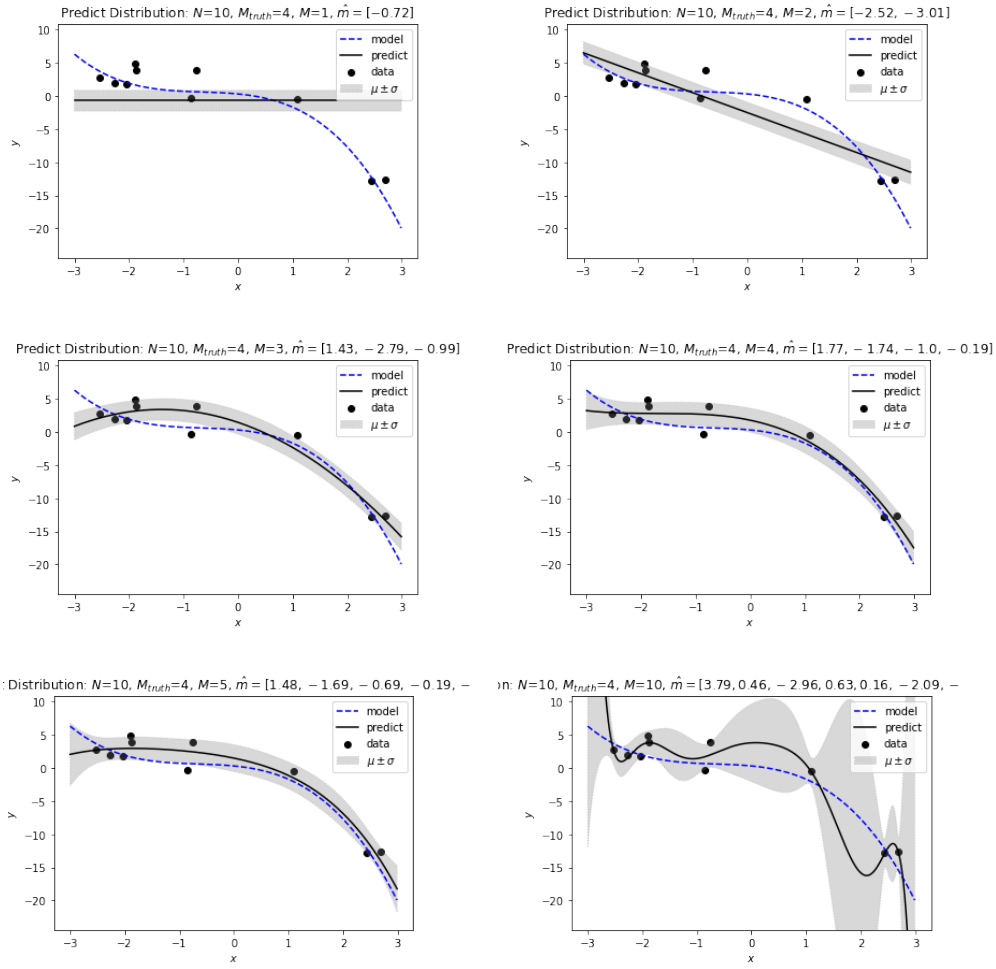


図 5：多項式回帰モデルの予測分布

参考：

- python の乱数シード(seed)を設定する 3 種類の方法
  https://python-ai-learn.com/2021/02/06/seed/

第 6 回「ベイズ推定を用いた回帰（２）」課題：

Stan で回帰分析をします。Stan によるベイズ推定は拡張性が高いところが利点です。そこで、正規分布を仮定した正規線形モデルとしての単回帰と重回帰(多項式回帰)、正規分布以外の分布を仮定した一般化線形モデルとして、ロジスティック回帰モデルを学びます。

- 基礎課題 6-1：note6-1(単回帰、ベイズ信用区間、ベイズ予測区間)で、気温とビールの売り上げデータ(beer-sales-data.csv)を用いて stan による回帰の基礎を確認せよ。
- 基礎課題 6-2：note6-2(単回帰/重回帰、デザイン行列の利用)を確認せよ。なお、n個のデータに対し、目的変数をyとし、m個の説明変数$x_1, x_2, \ldots, x_m$を線形結合であらわした重回帰$y_i = b_0 + b_1 x_{1i} + \cdots + b_m x_{mi} + \varepsilon_i$の式$(i = 1, \ldots, n)$は、行列で以下のように$\boldsymbol{Y} = \boldsymbol{X}b + \varepsilon$で表すとことができる。ここで行列$\boldsymbol{X}$をデザイン行列と呼ぶ。

$$\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{m1} \\ 1 & \vdots & \ddots & \vdots \\ 1 & x_{1n} & \cdots & x_{mn} \end{bmatrix} \begin{bmatrix} b_0 \\ \vdots \\ b_m \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

- 基礎課題 6-3：note6-3(JSON ファイルを読み込み、train_test_split 関数の利用)として、Stan によるデザイン行列を利用した多項式回帰(重回帰)を実現せよ(図 6)。note 5-1 で保存したパラメータと生成したノイズ付きのデータを利用し、学習データと予測データに適切に分けて用いること。
- 基礎課題 6-4：note6-4(二項ロジスティック回帰モデル）を確認し、タイタニックのデータ(titanic_converted.csv)を学習データと予測データに適切に分け、ベイズ信用区間と予測区間を追加し完成せよ。なお、二項ロジスティック回帰は、目的変数のデータが 2 値の以下の式で定義される。

$$p_i = \text{logit}^{-1}(b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_m x_m)$$

$$y_i \sim Bernoulli(p)$$

ここで logit^{-1} はロジット関数 logit=1/(1+exp(−x))の逆関数を意味し、ロジスティック関数またはシグモイド関数とも呼ばれ、Stan において、inv_logit 関数が利用できる。説明変数に含まれる離散データはダミーコード化、例え、1(あり)か 0(なし)のいずれかに値、にすれば利用できる（詳細は、学部授業の多変量解析を見直すこと）。

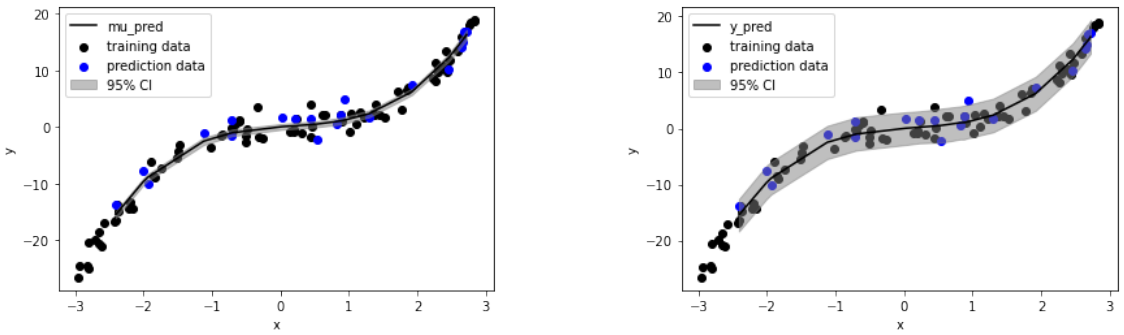


図 6：多項式回帰モデルの予測分布。95%ベイズ信用区間(左)と 95%ベイズ予測区間(右)

第７回「ベイズ推定を用いた回帰（３）」課題：

- 応用課題 7-1：ポアソン分布は、一定期間に平均してλ回「でたらめに」起こる事象が、ある区間にちょうど k 回起きる確率を意味する。ビールの売り上げや生き物の個体数など、0 以上の整数を取る離散型のデータを対象とする場合に、用いることができる。データにポアソン分布を仮定したポアソン回帰モデルを作成し、魚の釣獲尾数と天気および気温のデータ（fish-data.csv）に対する、回帰の 95%ベイズ信用区間と予測区間を示せ。CSV ファイルの fish_num という項目の釣獲尾数を目的変数として、説明変数に気温(temperature)と天気(weather, その種類は cloudy と sunny に分けられている)を用いること。ポアソン回帰モデルは、以下の式で定義される。

$$\lambda_i = \exp(a + b_1 x_1 + b_2 x_2 + \cdots + b_m x_m)$$
$$y_i \sim P\text{oisson}(\lambda_i)$$

  なお、Stan ではポアソン回帰モデルの作成において、exp と poisson を同時に使うと不安定になることが知られ、これら 2 つを組み合わせた関数として poisson_log 関数の利用が推奨されている。poisson_log 関数を利用する場合は、次のような変数変換となる。

$$\lambda_i = a + b_1 x_1 + b_2 x_2 + \cdots + b_m x_m$$
$$y_i \sim P\text{oissonLog}(\lambda_i)$$

- 応用課題 7-2：応用問題 3-1 で用いた乳がん診断データ(breast-cancer-wisconsin-data.csv）に、ベイズ推定を用いた回帰を適用し、パラメータを求め、その結果（予測精度）を述べ、考察せよ。癌の組織の種類 M = malignant (悪性), B = benign (良性）と特徴間にどのような回帰を設定するかは、各個人の判断にまかせます。発表資料では、どのような考え(判断)で回帰を行なったかを述べて下さい。
  - データの項目は URL(http://taustation.com/breast-cancer-dataset/)が参考になる。
- 応用課題 7-3：応用問題 3-2 で用いたアヤメの種類のデータ(iris-data.csv)に、ベイズ推定を用いた回帰を適用し、パラメータを求め、その結果（予測精度）を述べ、考察せよ。アヤメの種類と特徴量間にどのような回帰を設定するかは、各個人の判断にまかせます。発表資料ではその判断、どのように考えて回帰を行なったかを述べて下さい。
  - データの項目は URL(https://archive.ics.uci.edu/ml/datasets/iris)が参考になり、sepal length in cm (がくの長さ), sepal width in cm (がくの幅), petal length in cm (花弁の長さ), petal width in cm （花弁の幅）, target (アヤメの種類：setosa, virginica, versicolor)を意味している。

注意：重回帰分析などで変数を複数扱う場合に、値のスケール(単位)の違いの影響をなくし、特徴量が持つ値の重みを平等にし、学習のコストを削減するために、データの前処理を行う必要がある。その代表的な手法に、**(1)** 標準化（standardization）：平均($\bar{x_i}$)や分散(s_i)（標準偏差）を使い、$x_i' = (x_i - \bar{x_i})/s_i$データの平均を 0 に分散を 1 に補正する、や、**(2)** Min-Max 正規化(min-max normalization)：特徴量内の最小値($x_{\min}$)と最大値($x_{\max}$)を使い$x_i' = (x_i - x_{\min})/(x_{\max} - x_{\min})$とデータを 0～1 に補正する、がある。標準化は最小値や最大値が決まっていない場合や外れ値が存在する場合に利用する場合が多く、正規化は最小値や最大値が決まっている画像処理などで用いられる場合が多い。

第８回「応用課題」の発表：

応用課題 7-1, 7-2, 7-3 を解き、課題内容、用いたデータ(学習データ数，予測データ数，データの選択法、前処理法)と、回帰モデル、ベイズ推定の過程、結果、考察を説明した資料を PowerPoint で作成して下さい。manaba+R にその説明資料(*.pptx または *.pdf)およびソースコード(*.ipynb または *.py)、（入力データを特殊な形式に変更した場合のみ変更したデータも）提出すること。

プログラム言語は Python を使って下さい。なお、学習や予測の方法は、各自の判断で選んで下さい。解析的に解いても良いですし、PyStan でも、PyMC3 や Edward、TensorFlow Probability など他の得意な確率的プログラミングのライブラリを用いても良いです。

当日は、１課題について数名に発表していただきます。発表時間は１人５分程度とします。

提出期限：11 月 16 日(土) 23:59 まで（当日までに発表者を manaba+R で連絡します）

第９回「回帰の階層モデル（１）」課題：

少し教科書からはなれ、より高度なモデル構築ができるようになるため、Stan で「階層モデル」を行います。階層モデルは、各サンプルに説明変数だけ独立して推定するだけでは説明がつかない、グループに由来する差（グループ差）や個人に由来する（個人差）を扱うための手法一つです。

- 基礎課題 9-1：note9-1(単回帰の階層モデル)で、note 内の説明などをよく読み、stan による回帰の階層モデルの基礎を確認せよ。
- 基礎課題 9-2：note9-2(ロジスティック回帰の階層モデル，多変量回帰，複数の階層)で、stan による回帰の階層モデルの拡張法を確認せよ。

第 10 回「回帰の階層モデル（２）」課題：

Stan で回帰の階層モデルを構築し、階層がない場合と比較し、分析してみましょう。

- 応用課題 10-1：応用問題 7-2 で用いた乳がん診断データ(breast-cancer-wisconsin-data.csv）に、ベイズ推定を用いた回帰の階層モデルを適用し、パラメータを求め、階層を用いない場合と比較し考慮した階層の良し悪しを考察せよ。癌の組織の種類 M = malignant (悪性), B = benign (良性）と特徴間にどのような回帰・階層を設定するかは、各個人の判断にまかせます。発表資料では、どのように考えてモデルを作成したかを述べて下さい。
- 応用課題 10-2：応用問題 7-3 で用いたアヤメの種類のデータ(iris-data.csv)のうち virginica という種類のアヤメが先頭の 10 サンプルしかない場合を想定し作成した抜粋データ(iris-data2.csv）がある。この抜粋データ(iris-data2.csv)にベイズ推定を用いた回帰の階層モデルを適用し、パラメータを求め、一部のカテゴリのデータの数が少ない場合の階層を用いない場合と比較し考慮した階層の良し悪しを考察せよ。アヤメの種類と特徴量間にどのような回帰・階層を設定するかは、各個人の判断にまかせます。発表資料ではどのように考えてモデルを作成したか述べて下さい。

第 11 回「応用課題」の発表：

応用課題 10-1 と 10-2 を解き、用いたデータ(学習データ数，予測データ数，データの選択法)と、用いた回帰の階層モデル、ベイズ推定の過程、結果、考察を説明した資料を PowerPoint で作成して下さい。manaba+R にその説明資料(*.pptx または *.pdf)およびソースコード(*.ipynb または *.py)、（入力データを特殊な形式に変更した場合のみ、変更したデータも）提出すること。

プログラム言語は Python を使って下さい。なお、ベイズ推定の方法は、各自の判断で選んで下さい。解析的に解いても良いですし、PyStan でも、PyMC3 や Edward、TensorFlow Probability など他の得意な確率的プログラミングのライブラリを用いても良いです。

当日は、１課題について数名に発表していただきます。発表時間は１人 5～10 分程度とします。
提出期限：12 月 6 日(土) 23:59 まで（当日までに発表者を manaba+R/Dropbox/メールで連絡します）

第 12 回「混合モデルによるクラスタリング(1)」課題：

教科書の第 4 章の 4.1 節に戻り、4.4 節のガウス混合モデルにおける推論を実践します。

- 基礎課題 12-1：note12-1(ガウス混合モデル、ギブスサンプリング)で、2 次元ガウス混合モデルのパラメータのギブスサンプリングを確認せよ。
- 基礎課題 12-2：note12-2(ガウス混合モデル、変分推論)で、2 次元ガウス混合モデルのパラメータの変分推論を確認せよ。
- 基礎課題 12-3：note12-3(ガウス混合モデル、崩壊型ギブスサンプリング)で、2 次元ガウス混合モデルのパラメータの変分推論を確認せよ。

第 13 回「混合モデルによるクラスタリング(2)」課題：

Stan でガウス混合モデルとその拡張を構築します。

- 基礎課題 13-1：note13-1(一次元ガウス混合モデル、弱情報事前分布)で、stan によるガウス混合モデルの基礎と、MCMC の安定化の方法を確認せよ。
- 基礎課題 13-2：note13-2(D 次元ガウス混合モデル)で、多次元ガウス混合モデルの PyStan を用いた学習と予測を実現法と、note12-1 で保存したデータ(note12-1_output.json)を用いた、2 次元混合モデルの動作を確認し、GMM の結果の集計方法を工夫せよ。
- 基礎課題 13-3：note13-3(無限混合ガウスモデル，一次元)で、PyStan による棒折り過程(SBP、図 7)を用いた一次元ガウス混合モデルとその利点を確認し、GMM の結果の集計方法を工夫せよ。

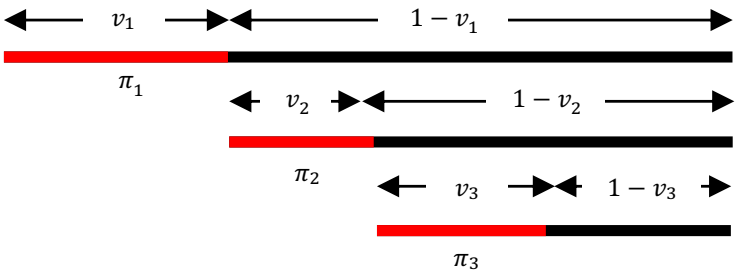


図 7：棒折り過程の概念図。説明は note13-3 を参照のこと。

第 14 回「混合モデルによるクラスタリング(3)」課題：

最後の応用問題です。

- 応用課題 14-1：応用問題 10-2 でも用いたアヤメの種類のデータ(iris-data.csv)に、アヤメの種類は明らかではない状態を仮定し、第 12 回や第 13 回で学んだガウス混合モデルを適用し、得たクラスタリング結果を可視化し精度を求めよ。どのように特徴量を選択するかは、各個人の判断にまかせます。発表資料ではどのように考えて特徴選択したかを述べて下さい。
- 応用課題 14-2：教科書の第 5 章にあるモデルのうちいずれか 1 つ、もしくは、教科書にはない、より発展的なモデルのベイズ推定を実装し、なんらかのデータに適用し、その結果を解説して下さい。本演習で実地済みのモデルは実装の対象外です。

第 15 回「応用課題」の発表：

応用課題 14-1 と 14-2 を解き、用いたデータと、モデル、ベイズ推定の過程、結果、考察を説明した資料を PowerPoint で作成して下さい。manaba+R にその説明資料(*.pptx または *.pdf)およびソースコード(*.ipynb または *.py)、独自の入力データを用いた場合はそのデータも提出すること。

プログラム言語は Python を使って下さい。なお、ベイズ推定の方法は、各自の判断で選んで下さい。解析的に解いても良いですし、PyStan でも、PyMC3 や Edward、TensorFlow Probability など他の得意な確率的プログラミングのライブラリを用いても良いです。

当日は、１課題について数名に発表していただきます。発表時間は１人 5〜10 分程度とします。
提出期限：1 月 18 日(土) 23:59 まで（当日までに発表者を manaba+R/Dropbox/メールで連絡します）