NICO2AI
パイロット講義 #3
線形回帰

17/07/01

講師：大澤 正彦
教材：八木 拓真

機械学習とは？

$$y = f(x)$$

出力　　　　　学習器　入力

線形回帰

$$y = \boldsymbol{\theta}^T \boldsymbol{x}$$

出力　　　　　学習器　入力

線形回帰

$$y = \begin{pmatrix} \theta_1 & \cdots & \theta_d \end{pmatrix} \begin{pmatrix} x_1 \\ \vdots \\ x_d \end{pmatrix}$$

出力　　　　学習器　　　　入力

目次

- **講義**
  - **確率と統計の復習**
  - **線形回帰モデルの学習**
- 基礎演習
  - Numpyのさらなる活用
  - Matplotlib入門
- 実践演習
  - Numpy活用演習
  - 線形回帰の実装

数式を用いた表記に少しずつ慣れていきましょう

確率と統計の復習

確率と統計をほんのちょっとだけおさらい

確率論、統計学と機械学習の関係

- 統計学と機械学習は、共にばらつきのある、確率的なデータに対する学問である
- ただし、その目的が大きく異なる：
  - 統計学：データの構造を説明する
  - 機械学習：未知の入力に対する予測を行う

- 2者は密接に関係しており、いずれもそのベースに確率論の考え方を含んでいる

→ 今一度、確率とは何だったか思い出してみよう

参考：http://tjo.hatenablog.com/entry/2015/09/17/190000

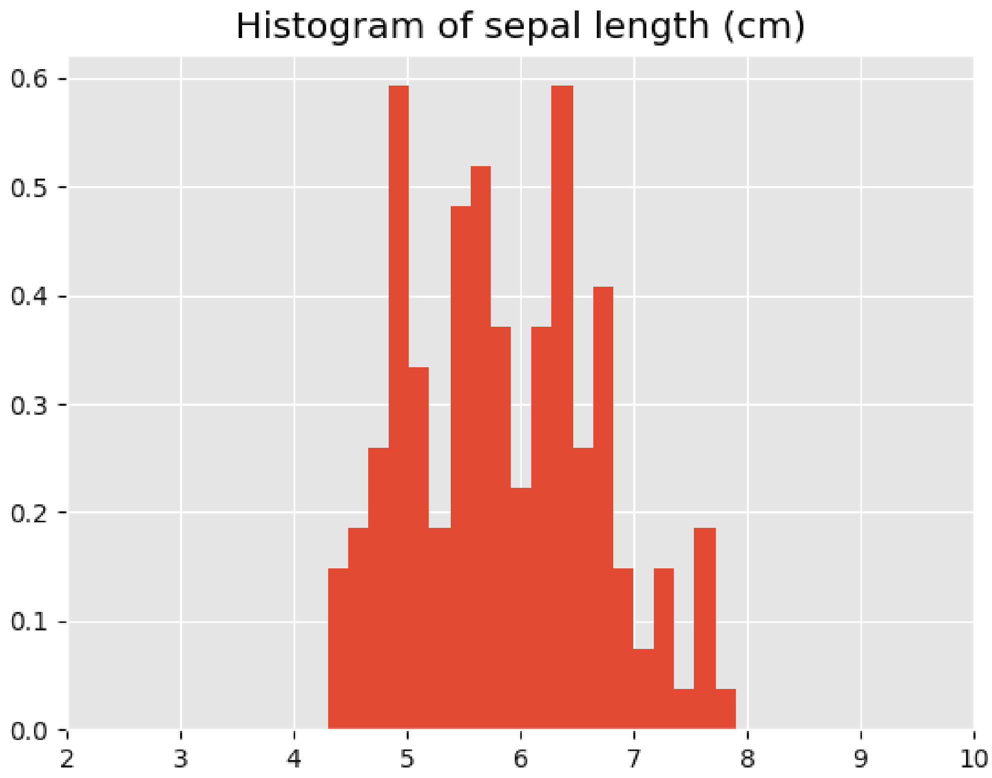
データから何が見える？

例：アヤメのがく片長 (cm)

$\boldsymbol{x}$ = [5.1, 4.9, 4.7, 4.6, 5. , 5.4, 4.6, 5. , 4.4, 4.9, 5.4, 4.8, 4.8, 4.3, 5.8, 5.7, 5.4, 5.1, 5.7, 5.1, 5.4, 5.1, 4.6, 5.1, 4.8, 5. , 5. , 5.2, 5.2, 4.7, 4.8, 5.4, 5.2, 5.5, 4.9, 5. , 5.5, 4.9, 4.4, 5.1, 5. , 4.5, 4.4, 5. , 5.1, 4.8, 5.1, 4.6, 5.3, 5. , 7. , 6.4, 6.9, 5.5, 6.5, 5.7, 6.3, 4.9, 6.6, 5.2, 5. , 5.9, 6. , 6.1, 5.6, 6.7, 5.6, 5.8, 6.2, 5.6, 5.9, 6.1, 6.3, 6.1, 6.4, 6.6, 6.8, 6.7, 6. , 5.7, 5.5, 5.5, 5.8, 6. , 5.4, 6. , 6.7, 6.3, 5.6, 5.5, 5.5, 6.1, 5.8, 5. , 5.6, 5.7, 5.7, 6.2, 5.1, 5.7, 6.3, 5.8, 7.1, 6.3, 6.5, 7.6, 4.9, 7.3, 6.7, 7.2, 6.5, 6.4, 6.8, 5.7, 5.8, 6.4, 6.5, 7.7, 7.7, 6. , 6.9, 5.6, 7.7, 6.3, 6.7, 7.2, 6.2, 6.1, 6.4, 7.2, 7.4, 7.9, 6.4, 6.3, 6.1, 7.7, 6.3, 6.4, 6. , 6.9, 6.7, 6.9, 5.8, 6.8, 6.7, 6.7, 6.3, 6.5, 6.2, 5.9]

→このデータには、どのような法則・構造があるのだろうか？

素朴な方法：ヒストグラム



ばらつきがあって、心なしか山形に見える

→ 見た目ではなく、定量的方法でデータを表せないか？

平均値、分散、中央値

- **平均値** (N：データ数)

$$\mathrm{mean}[\boldsymbol{x}] = \hat{x} = \frac{1}{N}\sum_{i=1}^{N} x_i = 5.84\mathrm{cm}$$

- **中央値**

$$\mathrm{median}[\boldsymbol{x}] = (\text{大きい順に数えて}\frac{N}{2}\text{番目の値}) = 5.80\mathrm{cm}$$

- **分散**

$$\mathrm{var}[\boldsymbol{x}] = \frac{1}{N}\sum_{i=1}^{N}(x_i - \hat{x})^2 = 0.68\mathrm{cm}$$

平均値や分散は、データの重要な特性を説明するが、
十分ではない→確率分布の導入

(素朴な) 確率分布の定義

- **確率分布：**確率変数 (random variable)の値とそれが出現する確率を対応させたもの
- 確率変数とは、ある確率変数をXとして
  - さいころを振って1が出る ($X = 1$)
  - 神経細胞が1秒間に10回発火した ($X = 10$)

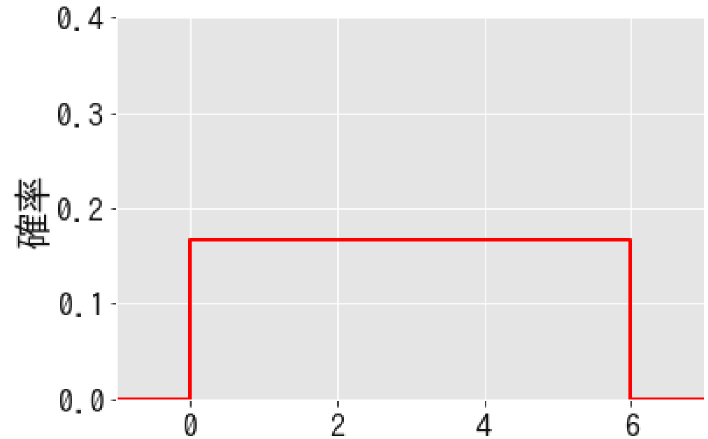
  といったような事象と値の組み合わせを指す

- 確率分布は、(離散/連続かかわらず)次の条件を満たす：
  - 全ての確率変数に対する確率は0以上
  - 全ての確率変数に対する確率の和は1

確率分布の例

- 一様分布

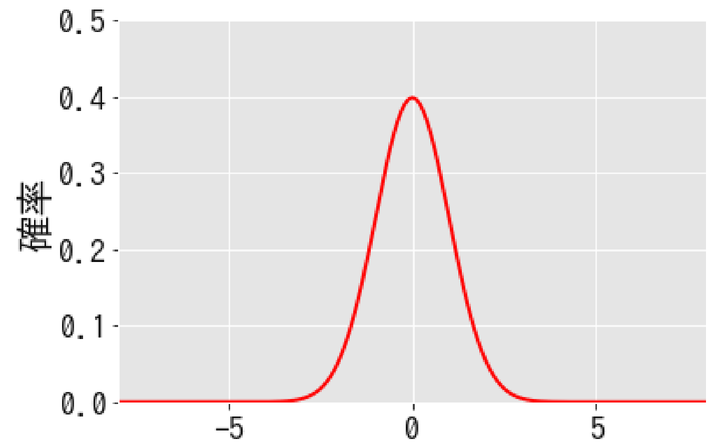
$$p(x;a,b) = \begin{cases} \frac{1}{b-a} & if\ x \in [a,b], \\ 0 & otherwise. \end{cases}$$



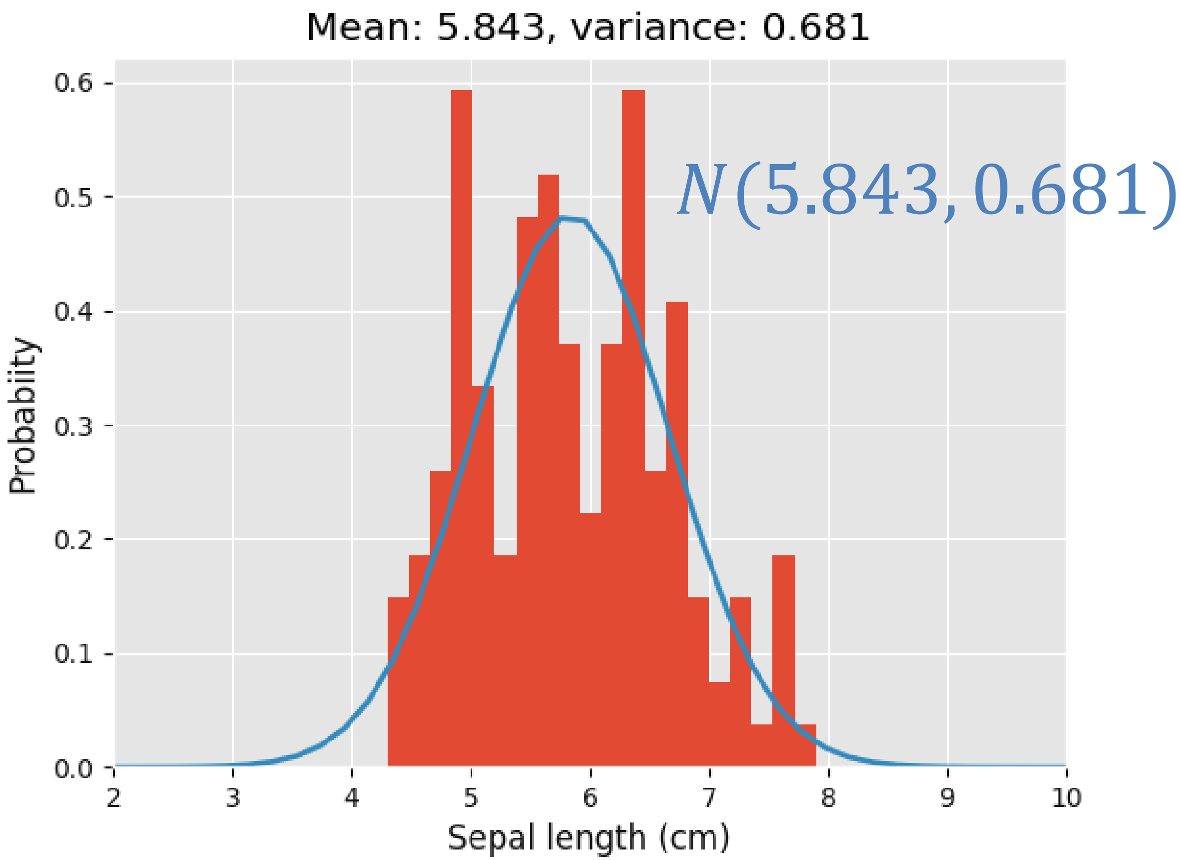
- ガウス分布 (正規分布)

$$p(x;\mu,\sigma^2)$$
$$= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}$$

$N(\mu,\sigma^2)$と略記することが多い

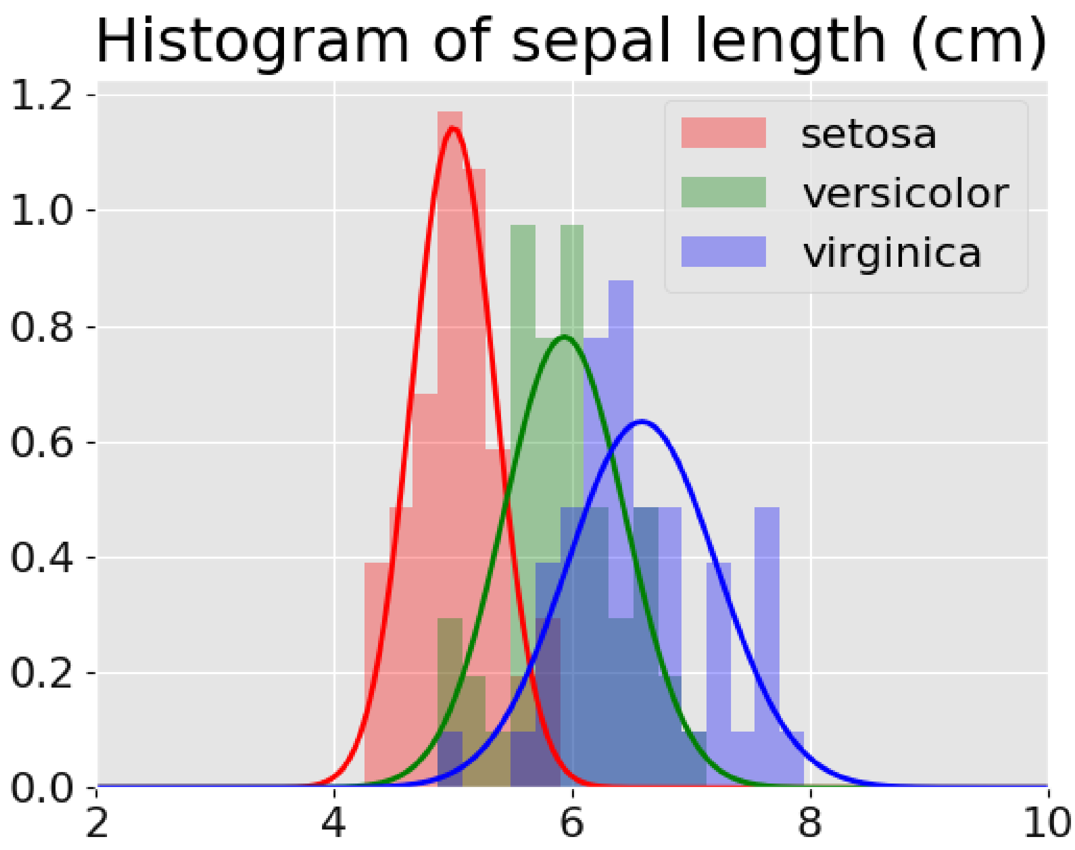


ガウス分布を用いたフィッティング



最尤 (さいゆう) 推定などの手法を用いて、データの
生成分布を推定することができる→予測・分類に使用

ガウス分布を用いたフィッティング



一方、確率分布の仮定が間違っていると、本質を見逃すこともある (前ページは3品種の混合分布だった)

機械学習とガウス分布の関係

- 機械学習においては、多くの場合データの「ばらつき」を表現する場合にガウス分布を利用する
- 例えば、線形回帰 (直線) の場合

$$
\begin{aligned}
y &= ax + b + \epsilon \\
&= ax + b + N(0, \sigma^2) \\
&= N(ax + b, \sigma^2)
\end{aligned}
$$

をデータの生成分布であると仮定するのが一般的

Further reading

- C. M. ビショップ『パターン認識と機械学習 上・下』 (丸善出版、2012)
- 久保拓弥『データ解析のための統計モデリング入門――一般化線形モデル・階層ベイズモデル・MCMC (確率と情報の科学)』 (岩波書店、2012)

線形回帰モデルの学習

数式の表記

- 本講義では、原則として
  - スカラ値を通常の小文字 (e.g. x_i)
  - ベクトルを小文字の太字 (e.g. $\boldsymbol{x}$)
  - 行列を大文字の太字 (e.g. $\boldsymbol{X}$)

という表記ルールに従う

- また、ベクトルは原則縦ベクトルとする

$$\boldsymbol{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{pmatrix} = (x_1, x_2, \cdots, x_N)^T$$

教師あり学習

教師あり学習：

入力と正解のペア (x_i, t_i) $(i = 1, \cdots, N)$ があるとき、

xからtを予測する関数$f(x;\boldsymbol{\theta})$を学習したい

関数fはパラメータ$\boldsymbol{\theta}$を持ち、その最適値を探す

学習 (Learning)

訓練サンプル

(x_1, t_1)

(x_2, t_2) → 学習器 $f(x;\boldsymbol{\theta})$

$\vdots$

(x_N, t_N)

推論 (Inference)

テストサンプル　　予測

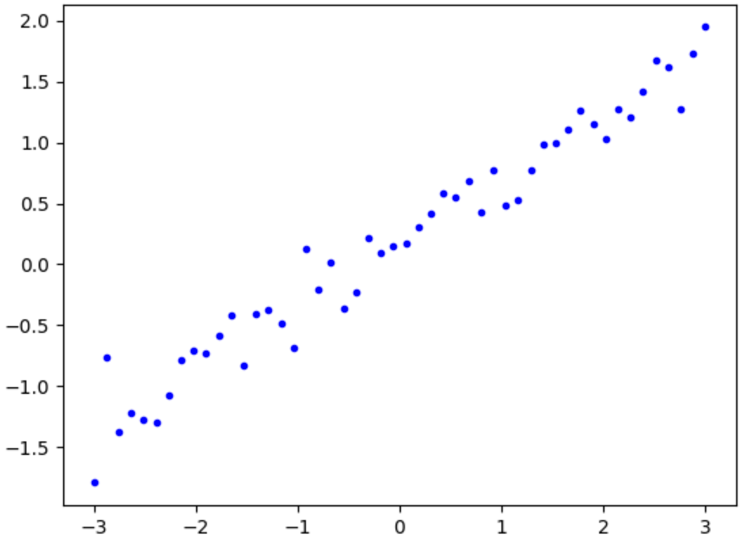
x_{new} → 学習器 $f(x;\boldsymbol{\theta})$ → t_{pred}

教師あり学習アルゴリズムの分類

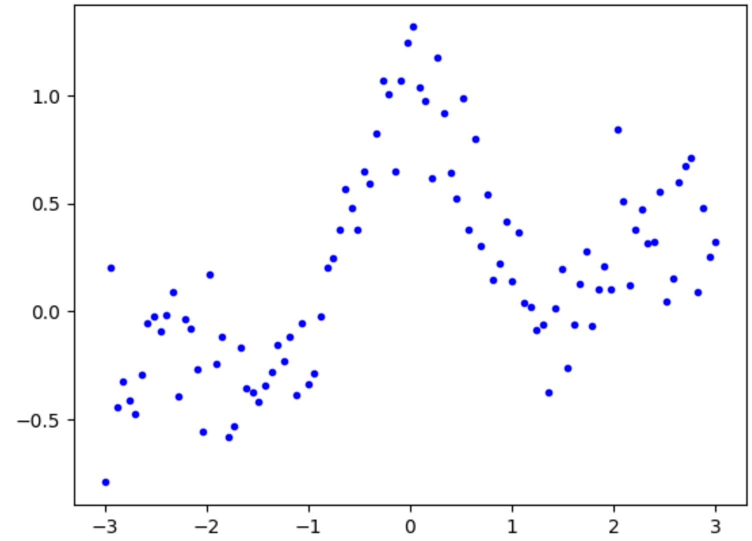
- **回帰問題**
  - 線形回帰←今日のテーマ
  - 非線形回帰
  - スパース回帰
- **分類問題**
  - 決定木
  - サポートベクトルマシン
  - ランダムフォレスト
  - (多くの) ニューラルネット
  - etc.

回帰問題

$$y = 0.5x + 0.1 + \epsilon$$

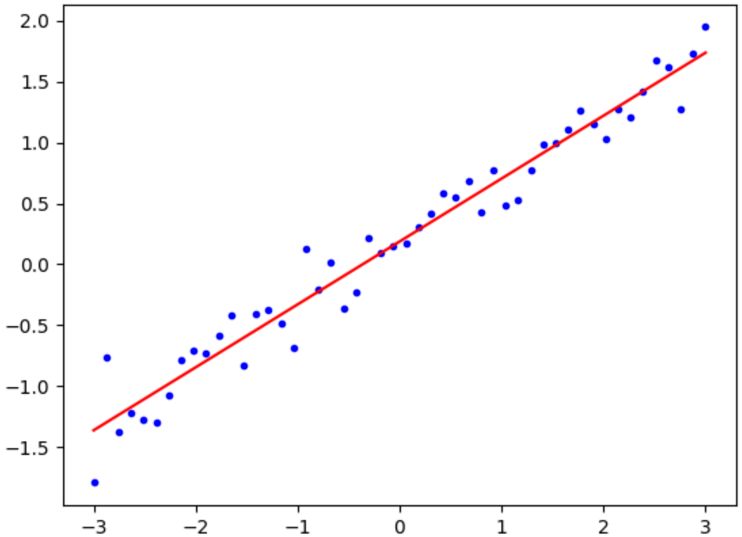


$$y = \frac{\sin \pi x}{\pi x} + 0.1x + \epsilon$$

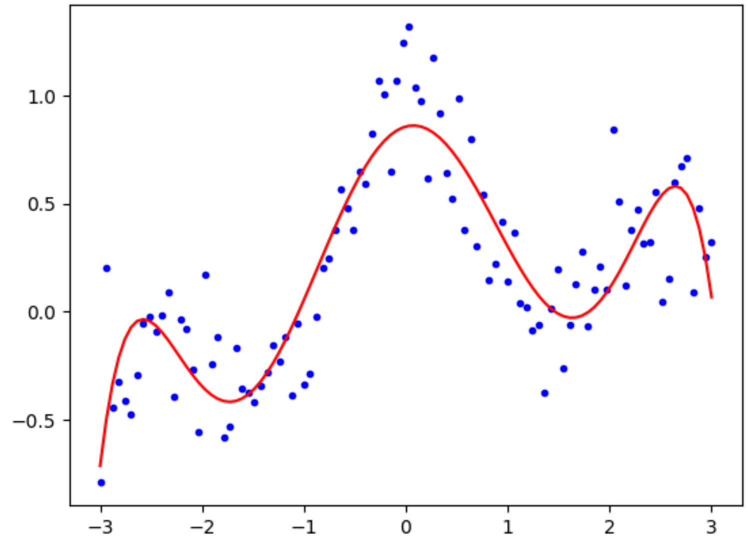


回帰問題

$$y = 0.5x + 0.1 + \epsilon$$



$$y = \frac{\sin \pi x}{\pi x} + 0.1x + \epsilon$$



回帰問題の目標：データ点に適合する関数の発見
→連続量の予測に利用 (e.g. 農作物の収量 (t) の予測)

線形/非線形回帰問題

- 学習器がパラメータθに対して線形 (1次) であるとき、線形回帰問題と呼ぶ (入力データに対しては非線形でも良い)

$$y = \theta_0 + \theta_1 x + \theta_2 x^2$$

- パラメータに対して非線形な関数の例：
  多層パーセプトロンのシグモイド関数

$$\sigma(\theta^T \boldsymbol{x}) = \frac{1}{1+\exp(-\theta^T \boldsymbol{x} - \gamma)}$$

非線形関数中にパラメータが含まれる

(θをa増やすと、fがb増えるという線形の関係が破綻する)

機械学習問題の設計

- 学習モデル (Model)
  - 加法モデル、多項式回帰、カーネル回帰、etc…
- 損失関数 (Loss function)
  - 平均二乗誤差、クロスエントロピー誤差、ヒンジ損失、KLダイバージェンス損失、etc…
- 最適化手法 (Optimization method)
  - 解析解の導出、勾配法、EMアルゴリズム、焼きなまし法、遺伝的アルゴリズム、etc…

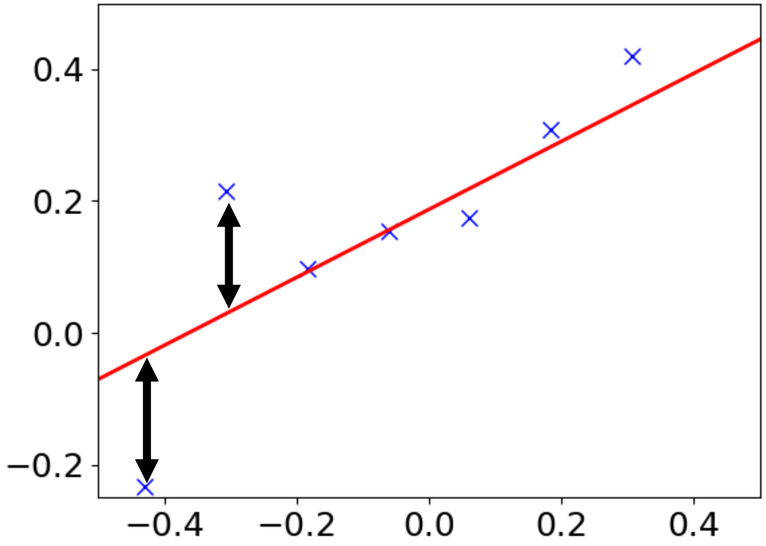
今後扱う機械学習問題においても、データに合わせて上の3つを如何にして組み合わせるかが重要になる

最小二乗回帰

モデルの予測と正解の二乗和誤差 (Sum Squared Error) が最小になるようなパラメータ$\boldsymbol{\theta}$を探す

$$\boldsymbol{\theta}_{LS} = \mathrm{argmin}_{\boldsymbol{\theta}} L_{LS}(\boldsymbol{\theta})$$

$$L_{LS}(\theta) = \frac{1}{2}\sum_{i=1}^{N}\left(f(x_i;\boldsymbol{\theta}) - t_i\right)^2$$



二乗誤差である理由：

・予測から遠くなればなるほど大きなペナルティを与える

・ノイズがガウス分布に従う場合最尤推定と等価

(つまり、自然な定式化)

最小二乗回帰の解析解

$$L_{LS} = \frac{1}{2} \parallel \boldsymbol{X\theta} - \boldsymbol{t} \parallel^2 = \frac{1}{2} (\boldsymbol{X\theta} - \boldsymbol{t})^T (\boldsymbol{X\theta} - \boldsymbol{t})$$ (行列表記)

L_{LS} は θ に関して2次関数の形 (凸関数) をしているので、
偏微分の値が0になる点が最小点であり

$$\frac{\partial L_{LS}}{\partial \theta} = \boldsymbol{X}^T (\boldsymbol{X\theta} - \boldsymbol{t}) = 0 \ \left(\because \frac{\partial \boldsymbol{a}^T \boldsymbol{b}}{\partial \boldsymbol{a}} = \frac{\partial \boldsymbol{b}^T \boldsymbol{a}}{\partial \boldsymbol{a}} = \boldsymbol{b}\right)$$

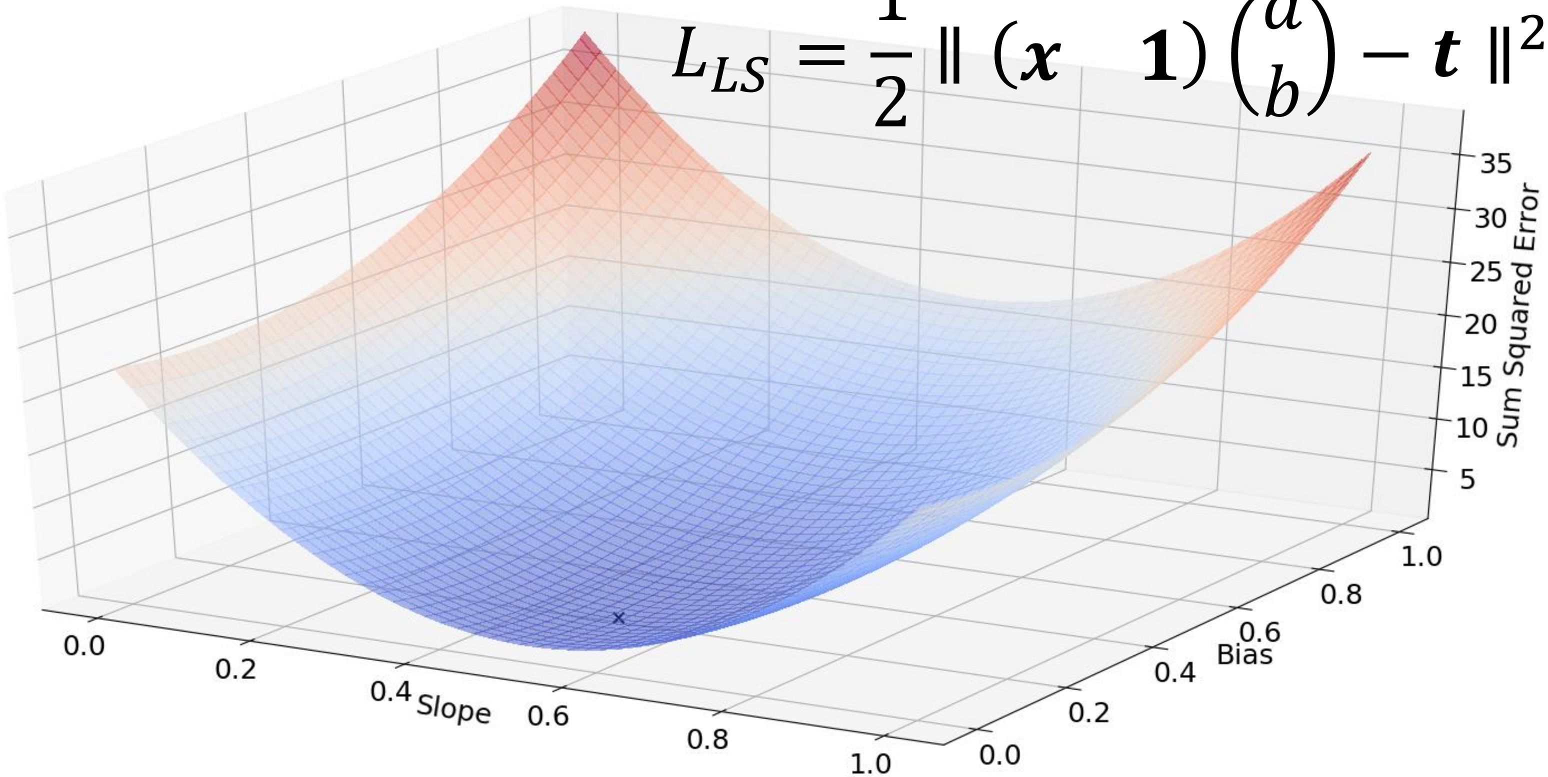
$$\boldsymbol{X}^T \boldsymbol{X\theta} = \boldsymbol{X}^T \boldsymbol{t}$$

両辺に$\left(\boldsymbol{X}^T \boldsymbol{X}\right)^{-1}$をかけて解析解$\boldsymbol{\theta}_{LS}$が求まる

$$\boldsymbol{\theta}_{LS} = \left(\boldsymbol{X}^T \boldsymbol{X}\right)^{-1} \boldsymbol{X}^T \boldsymbol{t}$$

最小二乗回帰の視覚的解釈

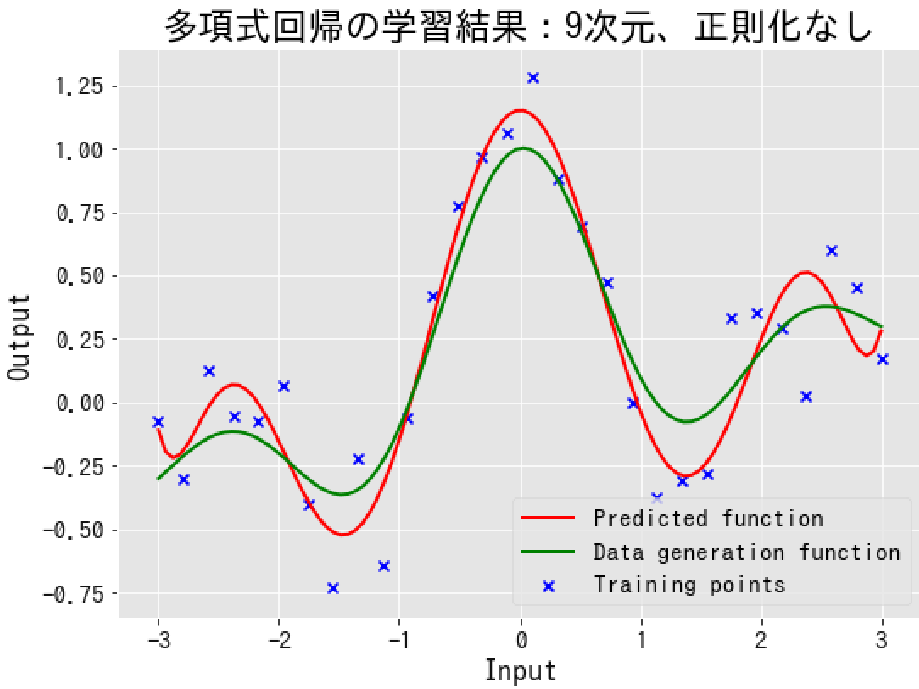
$$L_{LS} = \frac{1}{2} \| (\boldsymbol{x} \quad \boldsymbol{1}) \begin{pmatrix} a \\ b \end{pmatrix} - \boldsymbol{t} \|^2$$

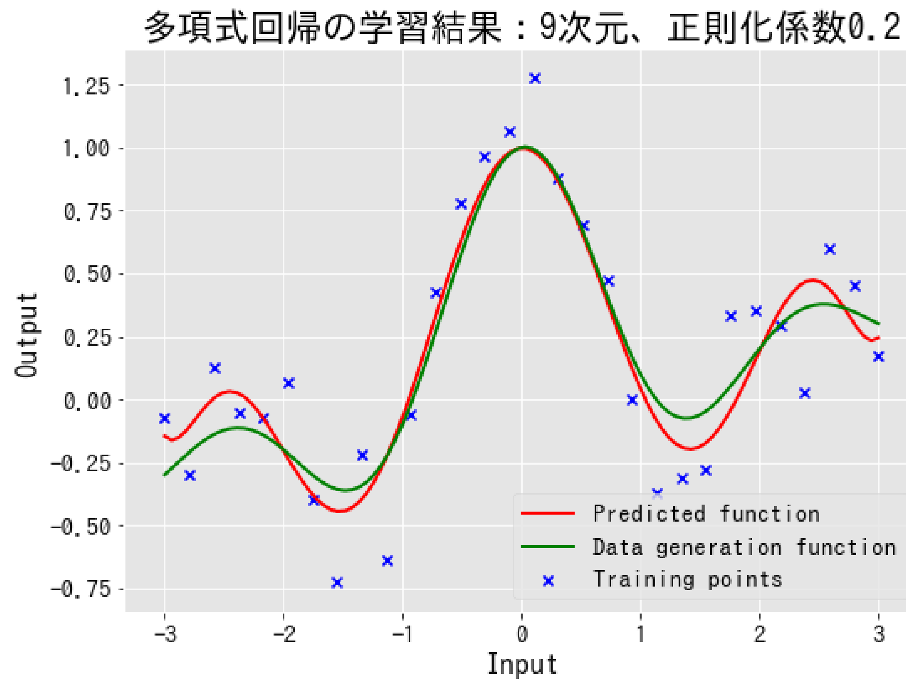


2変数の線形回帰($y = ax + b$)について、傾きaとバイアスbを変化させたときの二乗誤差L_{LS}をプロット

→確かに凸形であり、偏微分が0=平らな点が最小点

L2正則化 (過学習の回避)





9次元多項式回帰 $\theta_0+\theta_1 x + \cdots + \theta_8 x^8$

モデルの表現力が高いため、本来の曲線から外れてノイズの載ったデータにフィットしてしまう：過学習

→L2正則化 (regularization)で緩和できる

L2正則化最小二乗回帰

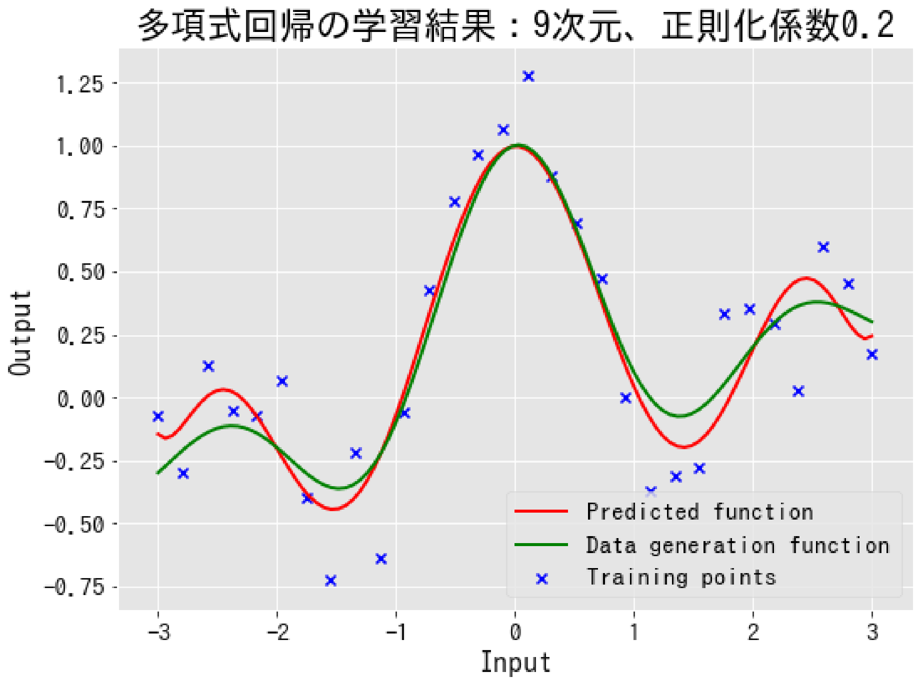
誤差関数を次のように変更：

正則化項
(L2ノルム)

$$L_{LS} = \frac{1}{2} \parallel \boldsymbol{X\theta} - \boldsymbol{t} \parallel^2 + \lambda \parallel \boldsymbol{\theta} \parallel^2$$

- 直感的には、
  過学習＝(フィットするために$\boldsymbol{\theta}$が大きくなりすぎる)
  という状態なので、大きい$\boldsymbol{\theta}$にペナルティを与える
- 通常最適なλは交差検証で決定し、解析解は次のように書き直せる：

$$\boldsymbol{\theta}_{LS} = \left(\boldsymbol{X}^T\boldsymbol{X} + \lambda \boldsymbol{I}\right)^{-1}\boldsymbol{X}^T\boldsymbol{t}$$



Further reading

- Pythonチュートリアル (公式)

https://docs.python.jp/3/tutorial/

- Numpy 100 exercises (英語)

http://www.labri.fr/perso/nrougier/teaching/numpy.100/

Numpyの基礎文法を速習できる100個の小演習。

- Dive Into Python 3 日本語版 (経験者向け)

http://diveintopython3-ja.rdy.jp/index.html

- 統計的機械学習入門

https://www.nii.ac.jp/userdata/karuizawa/h23/111104_3rdlecueda.pdf

- Numpy Talk at SIAM

https://www.slideshare.net/enthought/numpy-talk-at-siam

Allen Brain Atlas:
Mouse Connectivity

水口 智仁

Allen Brain Atlas

- Allen Brain Instituteによる様々な脳画像等のデータベース
  - Allen Mouse Brain Atlas
    各脳領域の遺伝子発現パターンを可視化
    - Allen Developing Mouse Brain Atlas
  - Allen Human Brain Atlas
  - Allen Brain Observatory
    視覚刺激呈示中の神経活動（Ca imaging）

etc. (See http://www.brain-map.org/overview/index.html)

Allen Mouse Brain Connectivity Atlas

- 脳領域間の結合強度 (connectivity) の可視化&定量化
- 全脳レベルの結合強度の評価
  connectivity + -ome = connectome
- "A mesoscale connectome of the mouse brain" (*Nature*, 2014)

Connectome at multiple levels

- **Microscale:** 個々のニューロン・シナプスレベル
  e.g.) C. elegans

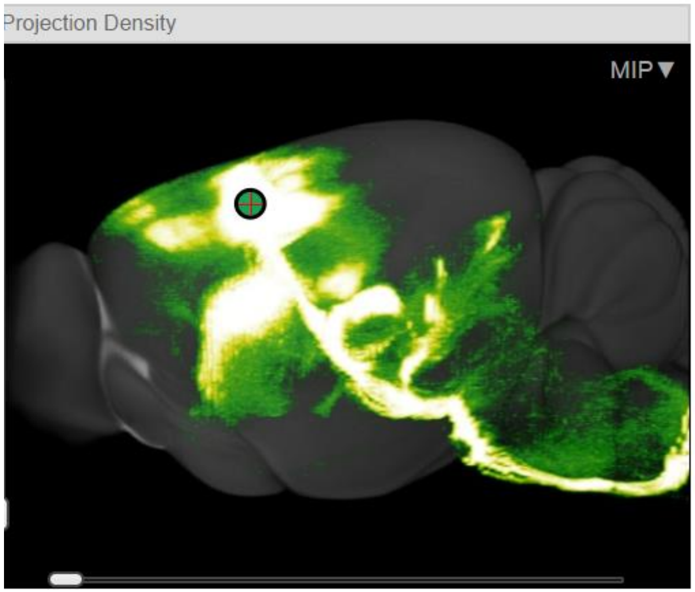
- **Macroscale:** 大域的な領域間の結合強度
  e.g.) 拡散テンソル画像（DTI）によるヒト脳

- **Mesoscale:** Macroscaleより細かく脳領域を分割して結合強度を評価する
  → Allen Mouse Brain Connectivity Atlas

Experimental design (in *Nature* 2014)

基本原理：
GFP発現ウイルスを用いて神経細胞の軸索を可視化し、特定の領域から脳全体への投射パターンを調べる。

- 全脳を295の脳領域に分割
- 個々の領域にGFP発現ウイルスを注入
- 切片化して顕微鏡撮影



Experiment 127084296 - MOp

Fig.3 Adult mouse brain connectivity matrix (in *Nature* 2014)

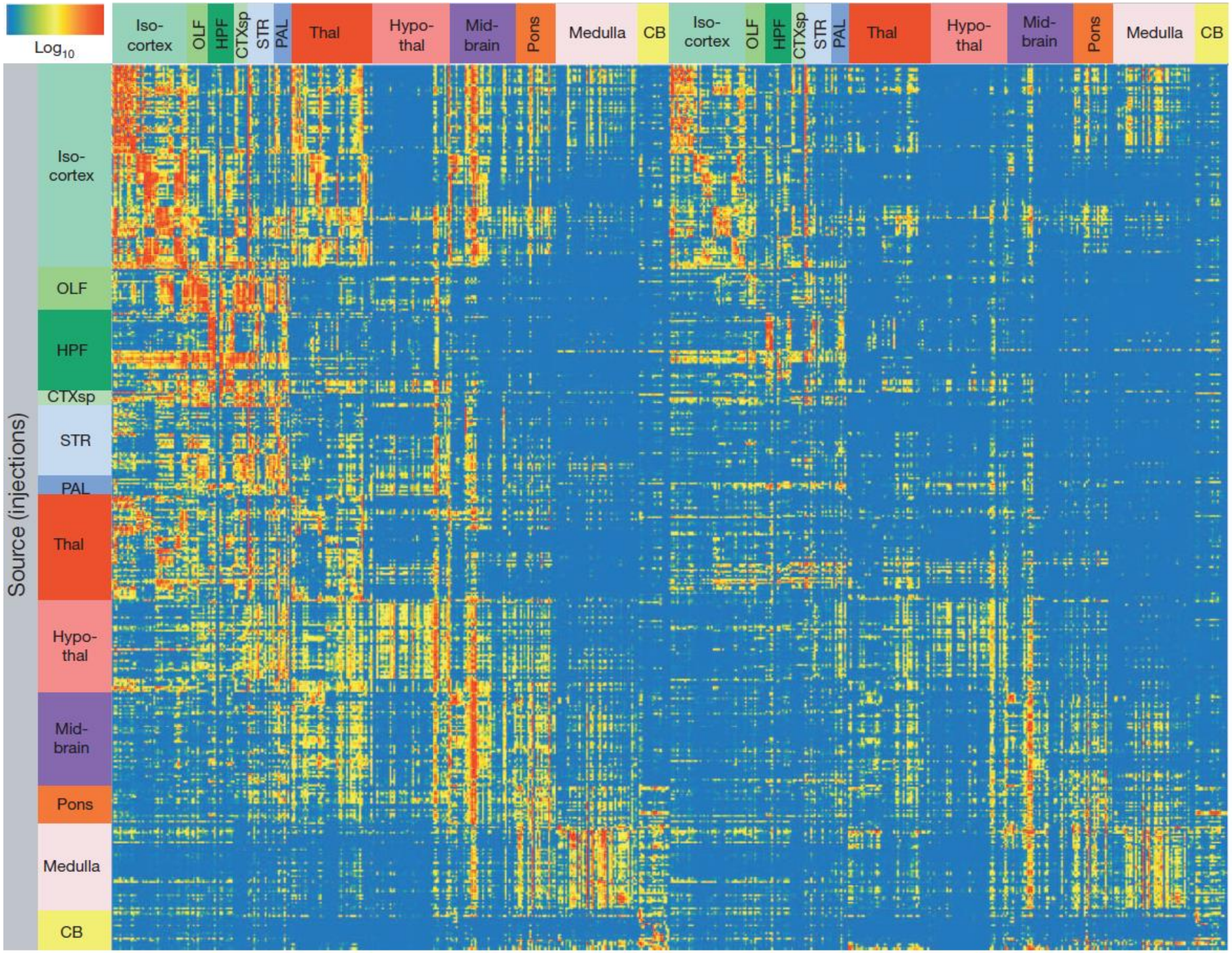
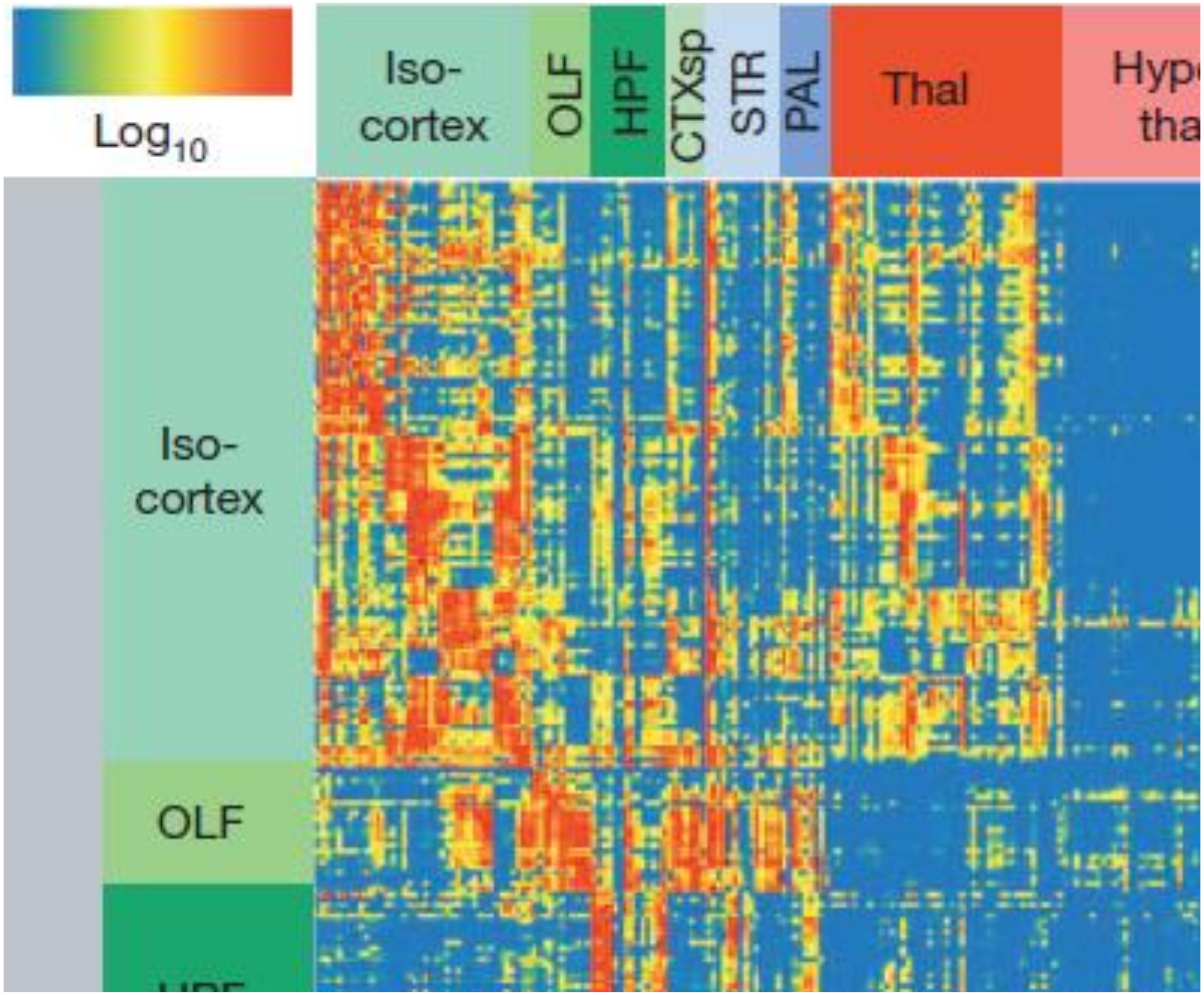


Fig.3 Adult mouse brain connectivity matrix



Supp. Table 2 (Fig. 3の元データ)

各列の説明
primary-injection-structure: ウイルス注入部位と最も重なる脳領域
secondary-injection-structure: 次に重なりの多い脳領域
injection volume: ウイルス注入部位の体積

					FRP	MOp
画像ID image-series-id	標本ID specimen-name	primary-injection-structure	secondary-injection-structures	injection volume (mm^3)	R	R
100141273.00	378-671	MOp	SSp-ll	0.18	0.00	2.62
100141563.00	378-697	MOp	SSp-ll	0.14	0.01	3.09
100141780.00	378-795	MOp	SSp	0.15	0.16	10.06

例：
378-795の標本では、MOp（一次運動野, primary motor area）とSSp（一次体性感覚野）に重なる位置にウイルスを注入したところ、FRP（大脳皮質前頭極）の蛍光強度は0.16だった。