

ニコニコAIスクール 第3回 線形回帰

18/01/20

講師：電通国際情報サービス (ISID)

小川雄太郎

質問:Slack

AIスクール以外でSlackを使用したことがありますか？



はい



いいえ

クイズ（まだ実行しないでください）

Slackでチームごとにチャンネルを作成してください。

チャンネル名は、チーム1の場合team_01とし、遠隔チームはteam_enkakuとしてください。

上記クイズが解けるようlec_3_4_ogawaを作成します

質問:プログラミング経験

- ・ Java、C#、Pythonなどのオブジェクト指向型プログラムを授業、研究、自学で使ったことがある
- ・ C、Matlab、Octaveなどのプログラミングを授業、研究、自学で使ったことがある
- ・ RとSPSSを除き、プログラミングはニコニコAIスクールが初の体験

質問：チームメンバーの ①研究室・研究内容、②科学的な興味の対象、③なぜAIスクールに参加したのか、AIスクールを通して何ができるようになりたいのか？④呼び名、を知っていますか？

1. 自己紹介
2. 本日の講義概要
3. 講義前半：13:15-15:30（14:30-14:45休憩）
「Numpyの扱い、線形回帰」
4. 講義後半：15:30-16:30
「リスト操作、脳Atlasで結合強度推定」
5. （進行しだいで、オブジェクト指向の解説）

質問はチャネルlec_3_4_ogawaにバンバン貼ってください。
Slackで改行はALT+ENTERです。遠隔の人も同様をお願いします。
※Zoomは履歴が消えてしまうので・・



小川雄太郎（おがわさん、ゆたろさん）
明石高専→東大工学部精密工学科（B）
→東大新領域（M,D）→東大先端研（PD）
→電通国際情報サービス 開発技術部

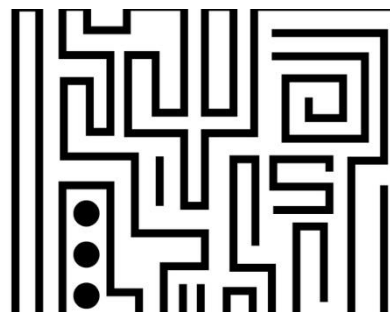
- 神保・小谷研究室で、NIRSと脳波の計測からワーキングメモリ能力の推定、神経細胞集団の数理モデルを縮約して同期現象の解析などを研究。2016年博士号（科学）を取得。
- 現在はディープラーニングをはじめとした機械学習の研究開発・案件支援および社内データ解析に従事
- 元脳科学若手の会代表(2012年)

興味があれば、いつか見てみてください。

- マイナビ出版の技術サイトManateeで強化学習の連載中

<https://book.mynavi.jp/manatee/series/detail/id=87626>

作りながら学ぶ!
強化学習
初歩からPyTorchによる
深層強化学習まで



- Qiita <https://qiita.com/sugulu/>



ユーザーランキング

週間 月間 全て



すぐる
@sugulu

1362

Contributions



株式会社電通国際情報サービス（ISID）

1975年に電通とGEのジョイントベンチャーとして設立されて以来、先進的な情報技術をベースに、アイデアとクリエイティビティを掛け合わせたユニークなIT専門家集団として成長。2000年に東証1部上場。



アラヤさんと一緒に、人工知能による画像解析で養殖マグロの個体数を自動カウント



<https://www.isid.co.jp/case/?all>

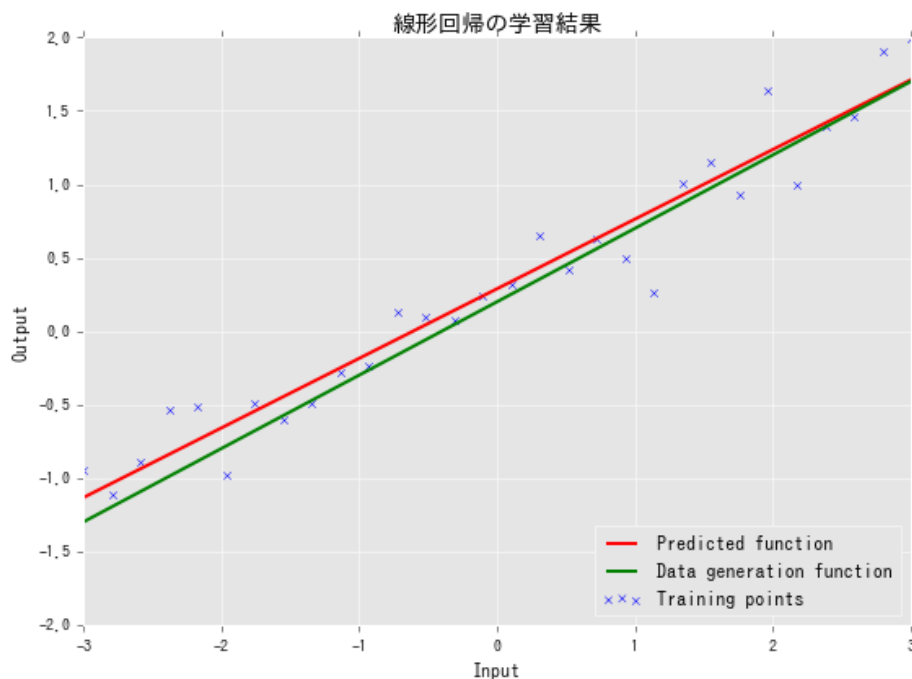
2. 本日の講義概要

8

本日のゴール：以下のことができるようになる

前半

$t = 0.5x + 0.2 + \sigma$ （※ σ はガウス分布に従うノイズ）で表される式から、30個のデータ点を生成し、データ点から生成元の式を、線形回帰を用いて推定する。



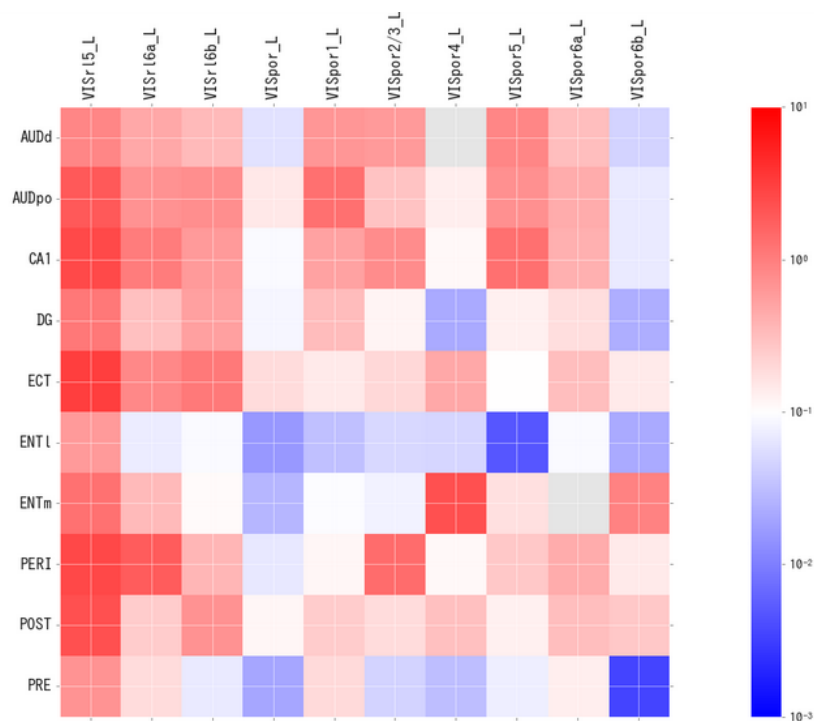
2. 本日の講義概要

9

本日のゴール：以下のことができるようになる

後半

The Allen Mouse Brain Connectivity Atlasの蛍光データから、脳部位間の結合強度を推定する



2. 本日の講義概要

10



※変更：実践演習3.1の途中で休憩を入れます

Numpyの扱い、線形回帰

NumpyとはPythonの数値計算用のライブラリ

Pythonが昨今、人気の理由（個人的感想）

1. PyPI（Python Packaging Index）に登録されたライブラリを個人が自由に利用できる仕組み
2. Travis Oliphant氏が作成したNumPyとSciPyライブラリが強力（※Pythonに行列演算はないが、Numpyが補完。NumpyはC言語ベースで処理が高速）
3. NumPyとSciPyをベースとした、機械学習ライブラリが強力

例）scikit-learn、Chainer、TensorFlowなど

クイズ

-3から3まで等間隔にとった30要素からなるベクトルxと、全要素が1の30次元のバイアスベクトルbを作成し、xとbを結合させて、30行×2列の配列Xを作成してください。

→このクイズが解けるように、NumpyのArray（配列）の結合操作を学習します

→lecture3_1.ipynbを開いてください

Numpyの線形代数パッケージlinalg (Linear algebra) を使用して、線形方程式を解きます

クイズ

$$A = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}, B = \begin{pmatrix} 1 & 2 \\ 2 & 1 \end{pmatrix}, I = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, C = \begin{pmatrix} 9.5 \\ 12.5 \end{pmatrix} \text{において、}$$
$$(A^T B - 3I) \times \theta = C$$

の関係を満たすとき、 θ を求めよ。

→このクイズが解けるように、 Numpyのlinalgの操作を学習します (の前に・・・Slackのコード貼り付けを解説)

困ったら、Slackを使いチームで相談してください

コードの貼り方

- 直に貼るとちょっとでかくて邪魔
- ``````で囲む（SHIFT+@で記号`が出ます）
- code or text snippetを使う

PythonのFigure描画ライブラリであるMatplotlibの使い方を学びます

クイズ

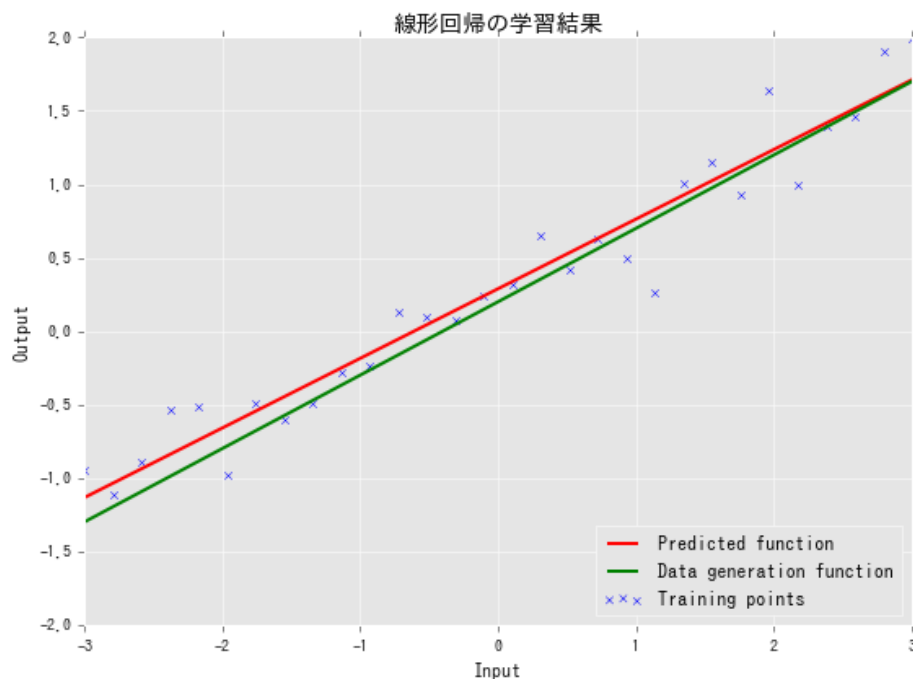
$y = 3x + 4$ を満たす直線と、その直線に平均0、標準偏差2.1のノイズが加わった点15個（ x は-4から4を均等）をプロットせよ。ただし、描画範囲は $-5 < x < 5$ とする。

→このクイズが解けるようにMatplotlibの操作を学習します

クイズ

$t = 0.5x + 0.2 + \sigma$ （※ σ はガウス分布に従うノイズ）で表される式から、30個のデータ点を生成し、データ点から生成元の式を、線形回帰を用いて推定せよ。

ノイズ ϵ は平均0、標準偏差0.2のガウス分布に従うとします。



Slack

 線形回帰、最小二乗法を実装した事がある

 線形回帰、最小二乗法を使った事がある

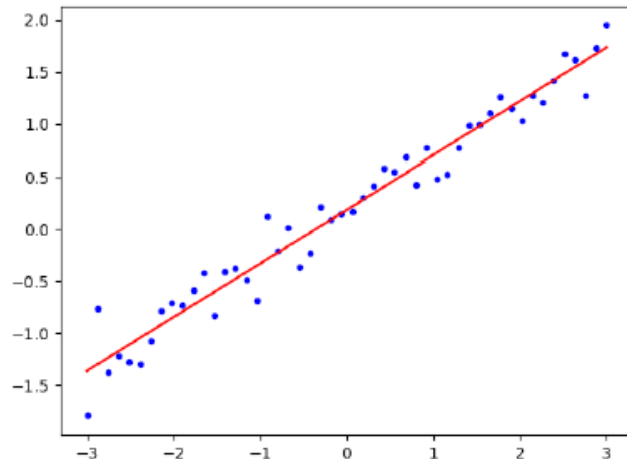
 線形回帰、最小二乗法を聞いたことがある

 線形回帰、最小二乗法は聞いたことがない

線形回帰とは

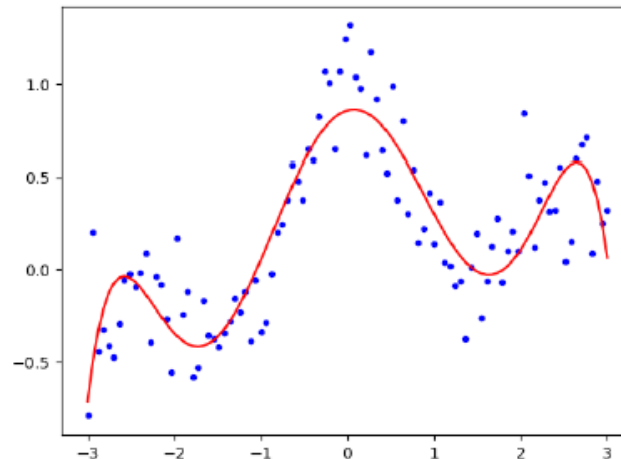
データ点（青点）を生み出した生成モデルを、最小二乗法で求める手法

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



(左) $y = ax + b$ とし、 a, b を推定

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



(右) $1, x, x^2, \dots, x^8$ までの線形結合で係数推定

線形回帰って何が線形なの？

生成モデルが基底関数の線形結合（データ生成モデルが線形関数という意味ではない）

例) 基底関数が $(1, x, x^2)$ の場合、 $y = \theta_0 + \theta_1 x + \theta_2 x^2$ となり、係数 θ_k たちを求める

例) こちらは非線形。基底関数の線形結合でない。

$$y = \theta_0 \sin(\theta_1 x)$$

どうやって係数 θ を求めるの？ その 1

推定した生成モデルの出力 $y_i = f(x_i; \theta)$ と、得られたデータ点 (x_i, t_i) の二乗誤差の和が、最小となる θ を求める。この文章を式で書くと次の通り。

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

L_{LS} を誤差関数もしくはは損失関数 (loss function) と呼ぶ。

LSはLeast Squareのこと。

ところで、 $y_i = f(x_i; \theta)$ って何？

どうやって係数 θ を求めるの？ その 2

$\mathbf{x} = (x_1, x_2, \dots, x_N)^T$ とし、基底関数を $f(x_i) = (1, \phi_1(x_i), \phi_2(x_i), \dots, \phi_k(x_i))$ 、係数 θ を $\theta = (\theta_0, \theta_1, \dots, \theta_k)^T$ とすると、組み立てる線形モデルは次のように表される。

$$\mathbf{y} = f(\mathbf{x}; \boldsymbol{\theta}) = \boldsymbol{\phi}(\mathbf{x}) \cdot \boldsymbol{\theta}$$

例えば、基底関数が $(1, x, x^2)$ の場合、 $y = \theta_0 + \theta_1 x + \theta_2 x^2$ となり、係数 θ_k たちを求める。

どうやって係数 θ を求めるの？ その3

求めたいのは損失関数を最小とする θ で、次のように書く

$$\theta_{LS} = \arg \min_{\theta} L_{LS}(\theta)$$

この θ_{LS} は損失関数を θ で偏微分して、

$$\frac{\partial}{\partial \theta} L_{LS} = 0 \quad \text{となる}\theta\text{です。式で表すと以下となります。}$$

$$\theta_{LS} = \left(\phi(x)^T \phi(x) \right)^{-1} \phi(x)^T T$$

(損失関数の偏微分計算の行列版です。導出は省略するので、時間があるときに調べてみてください)。

演習3.1の前半が終わった方？

Slackで質問

半分弱になったら休憩。

少ししてから、演習3.1追加向けの解説へ

補助課題は、授業最後で

基底関数はどうやって決めるの？

ケースバイケース。1次の直線モデルなら $(1, x)$ 、多項式近似なら $(1, x, x^2 \cdots x^K)$ 。生成メカニズムが分かっている場合もある。ちなみに K をデータ点の数 N と同じにすると全部のデータ点を通るモデルになる。この場合、過学習と呼ばれ、学習データにはきれいにフィッティングするが、未知の点には適合しない（汎化性能が悪い）。そこで、得られたデータを学習データ（test data）と検証データ（validation data）に分けて、学習データでモデルを推定し、検証データによく適合する K を採用する（交差検証法）。

どうして2乗誤差の和を損失関数と考えるの？

推定したモデルと得られた点の間の誤差は、様々なノイズが組み合わさったものだと想定する。

すると、中心極限定理より誤差はガウス分布に従うと考えられ、データ生成モデルもガウス分布に従うと仮定する

(漸近正規性)。ガウス分布に従う生成モデルでデータ点の最尤推定を行うと、対数尤度を求めたときに、ガウス分布のexpの中の2乗項が出てきて、2乗誤差の項と一致する。

機械学習のアルゴリズムは生成モデルの対数尤度を最大化するパラメータを求めるという作戦が基本方針となる。

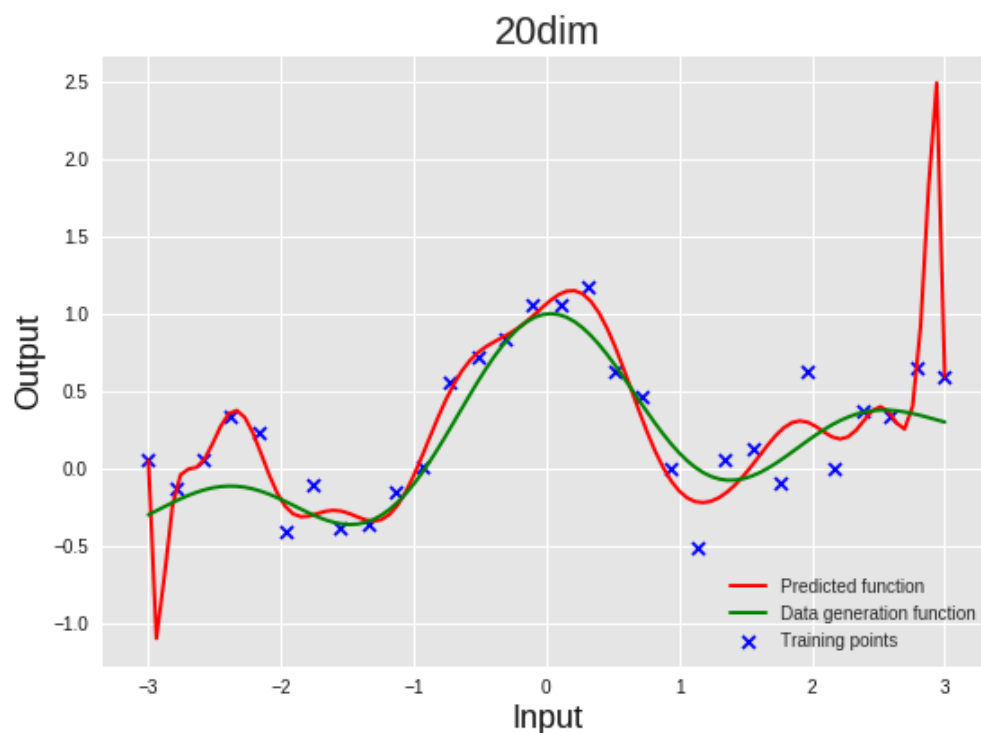
どうして線形回帰は解析的に答えが求まるの？

損失関数が係数 θ の2乗までで表され凸関数であるから。

非線形回帰の場合や、その他の機械学習手法では解析的に求まらないので、ニュートン法や勾配法で更新し、近似的に θ を求める（勾配法の詳細はニューラルネットワークの講義にて）

過学習を防ぐには?その1

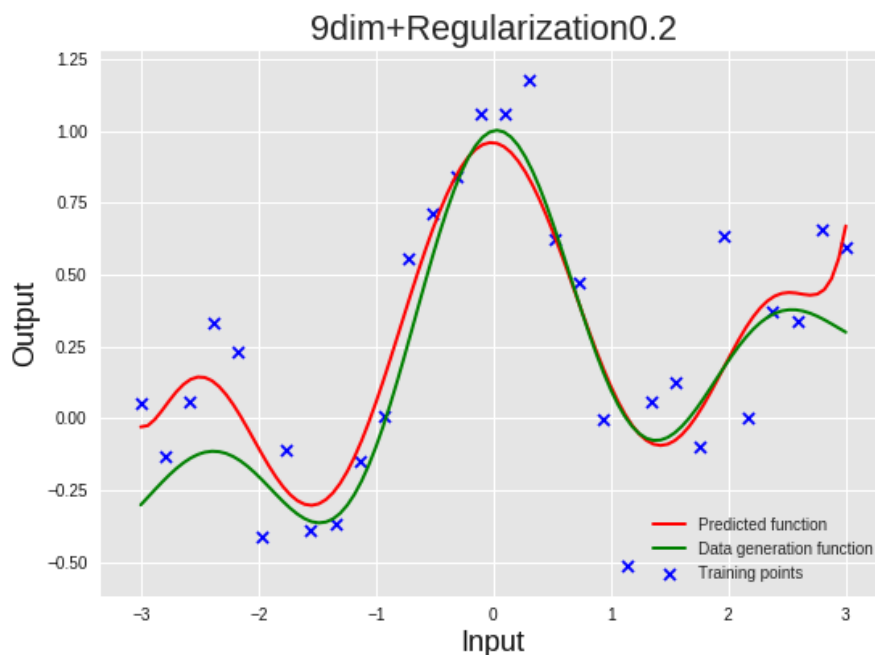
過学習とは、少ないデータ点で無理にモデルをフィットさせるために、 θ が真の値にならず、一部の値がとてもしくなる。



過学習を防ぐには?その2

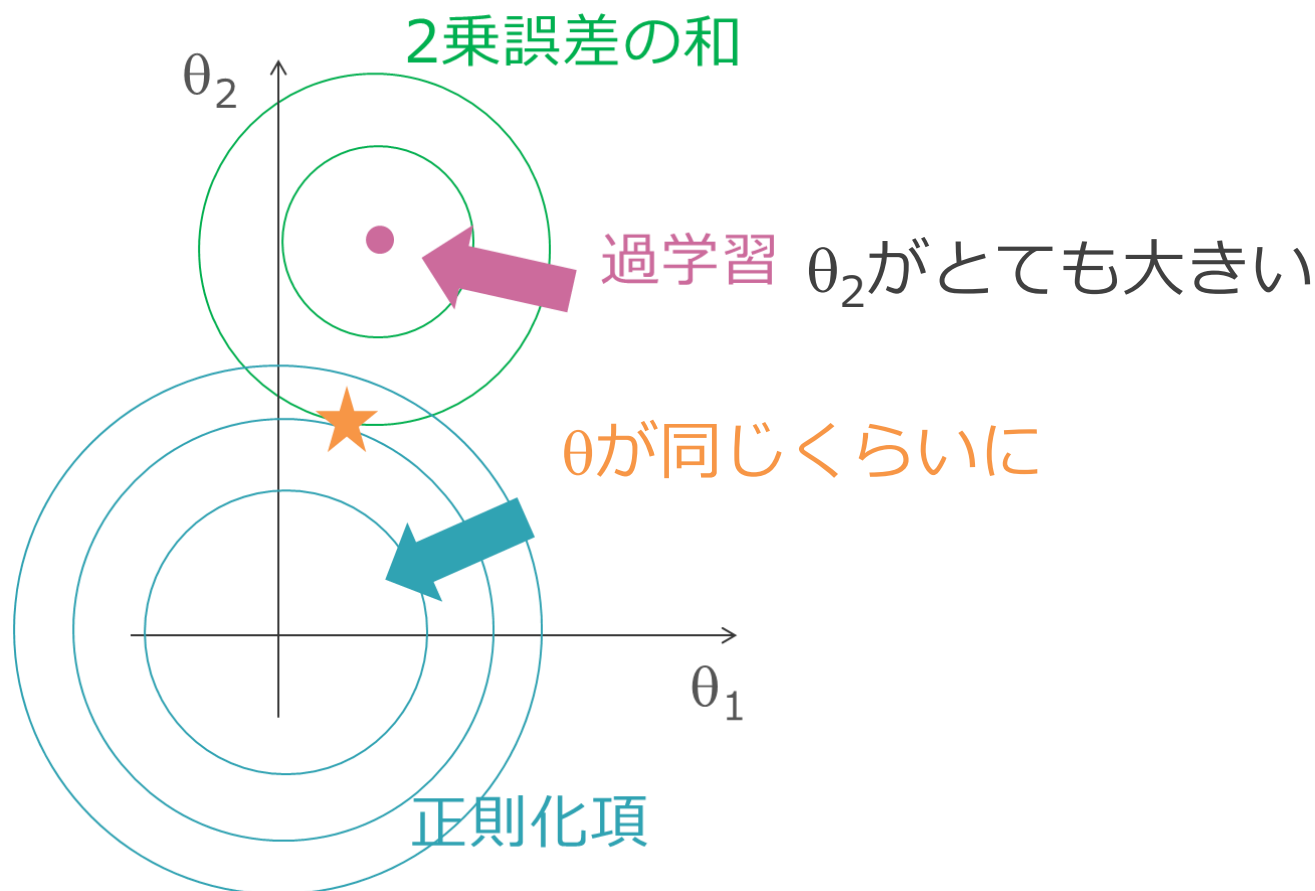
過学習を防ぐために、損失関数に正則化として正則化係数 λ と正則化項（今回はL2ノルム）を追加

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



過学習を防ぐには?その3

L2ノルムのイメージ



クイズ

正則化項の係数である正則化係数 λ は、どうやって決めれば良いですか？

チームごとに相談して、lec_3_4_ogawaにチームの答えを書いてください。（時間5分）

（例）

team_01です

、、、

〇〇をして、正則化係数を決めると良いと考えました。

、、、

map, ラムダ式, reduce, filterを駆使して、効率的にリスト（配列）を加工する方法を学びます

クイズ

$x=[21, 34, 35, 46]$ の各要素を3で割った余りにしたベクトル y を求めよ。

→このクイズが解けるように、map, ラムダ式, reduce, filterの操作を学習します

NumpyのArrayを複数使用し、効率的にArrayを加工する方法「Advanced Indexing」を学びます

クイズ

$x = [10, 20, 40, 60, 50]$, $y = [1, 0, 1, 1, 0]$

において、 y の要素が1である要素番号の x の総和を求めよ。

→このクイズが解けるようにAdvanced Indexingの操作を学習します

3.5 IrisでAdvanced Indexingとヒストグラム 34

Irisデータに対し、Advanced Indexingを使用して、花の種類ごとにがく片と花びらの大きさのヒストグラムを描画

Iris:アヤメの花 (sepetal:がく片、petal:花びら)

Setosa (檜扇菖蒲 : ひおうぎあやめ)

Versicolor (ブルーフラッグ) 、 virginica (Virginia iris)



<https://www.weblio.jp/content/Iris+setosa>

https://en.wikipedia.org/wiki/Iris_versicolor

https://en.wikipedia.org/wiki/Iris_virginica

3.5 IrisでAdvanced Indexingとヒストグラム 35

Irisは統計・機械学習で、よく使用されるデータセットです。
その理由は知っていますか？



知っている



知らないな～

3.5 IrisでAdvanced Indexingとヒストグラム 36

Irisのデータは、分散分析や最尤推定を生み出したロナルド・フィッシャーが論文のなかで使用したデータ・セット（作ったのは別の人）であり、以後統計解析の分野ではよく使用されています。



サー・ロナルド・エイルマー・フィッシャー Sir Ronald Aylmer Fisher

（1890年2月17日 – 1962年7月29日）イギリスの統計学者、進化生物学者、遺伝学者で優生学者である。

<https://ja.wikipedia.org/wiki/%E3%83%AD%E3%83%8A%E3%83%AB%E3%83%89%E3%83%BB%E3%83%95%E3%82%A3%E3%83%83%E3%82%B7%E3%83%A3%E3%83%BC>

3.5 IrisでAdvanced Indexingとヒストグラム 37

IrisデータでAdvanced Indexingを使った描画は理解できましたか？



はい



いいえ






ALLEN INSTITUTE FOR BRAIN SCIENCE RESOURCES

<http://brain-map.org/overview/index.html>

※今はアクセスしないでください。

脳・神経科学に関する様々なデータを公開している組織です。脳領域間の結合強度（connectivity）の可視化&定量化データも公開されています。全脳レベルの結合強度の評価（connectivity + -ome = connectome と呼ばれる）

Slack質問：研究分野

-  実験動物を解剖したり、細胞を取り扱うwetな研究
-  サルやマウス等で、行動実験をベースとした研究
-  fMRIや脳波、その他疾患系など、人を対象とした研究
-  数理をベースとした理論神経科学やデータ解析
-  その他（神経倫理学、今は研究はしていない等）

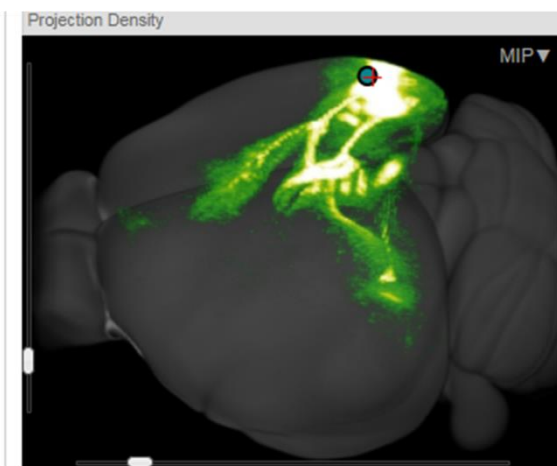
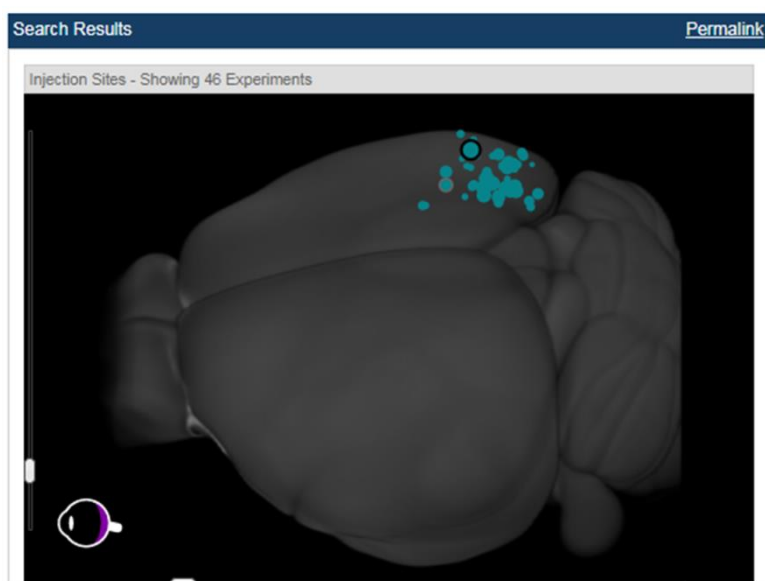
Connectome at multiple levels

- ・ Microscale : 個々のニューロン、シナプスレベル (C. elegansなど)
- ・ Macroscale : 拡散テンソル画像 (DTI) による大域的結合 (ヒトなど)
- ・ Mesoscale : GFP発現ウイルスを用いた軸索の可視化など (マウスなど) **今回はこれを使用！！**

The Allen Mouse Brain Connectivity Atlas

注入部(injection)と射影部(projection)の体積を計算し、ある領野からどの領野に注入した蛍光成分が移動したかに基づいて、領野間の結合の強さを測定している。

全脳を295の脳領域に分割し、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。



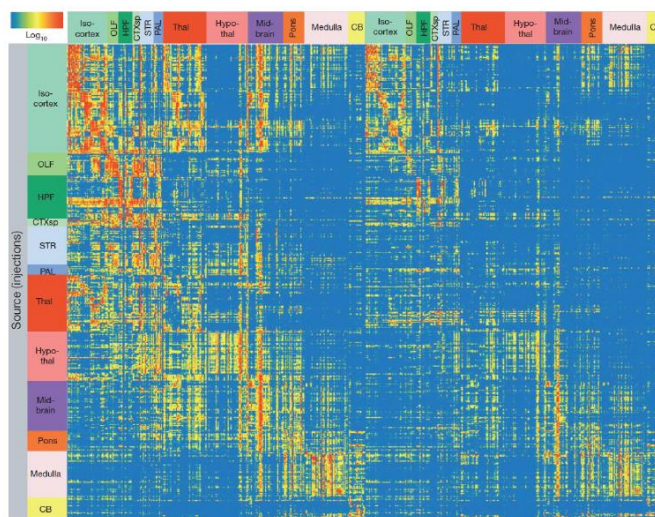
Experiment 116903968 - VISI

The Allen Mouse Brain Connectivity Atlas

Oh, Seung Wook, et al. "A mesoscale connectome of the mouse brain." *Nature* 508.7495 (2014): 207-214.

<https://www.nature.com/articles/nature13186>

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



API:The Allen Mouse Brain Connectivity Atlas

<http://alleninstitute.github.io/AllenSDK/install.html>

<http://alleninstitute.github.io/AllenSDK/connectivity.html>

download_mouse_connectivity_from_ALLEN.ipynb
を見てください。

ファイル情報

injected_data.csv

入力Xに対応する情報のデータ

0列目：実験ID

2列目：注入部位名

3列目：注入部位ID

4列目：注入量

・・・2～4の繰り返し

ファイル情報

target_data.csv

出力Yに対応する情報のデータ

列：部位情報

0行目：index

1行目：部位名

2行目：半球情報

3行目～：各実験での発現量