Deep Learning 輪読会 2017

第5章　機械学習の基礎

2017.11.13

東京大学大学院経済学研究科 D1 森下光之助

理学系研究科附属天文学教育研究センター B4 吉村勇紀

「Deep Learning」本輪読会 2017
5 章：機械学習の基礎

November 13, 2017

東京大学大学院経済学研究科
澤田研究室 D1 森下光之助

Contents

学習アルゴリズム

機械学習アルゴリズムとは

- 機械学習アルゴリズムとは，データから学習ができるアルゴリズムのこと
- Mitchell (1997) による定義：
  「コンピュータプログラムは，性能指標 P で測定されるタスク T における性能が経験 E により改善される場合，そのタスク T のクラスおよび性能指標 P に関して経験 E から学習すると言われている.」

タスク T

- 機械学習のタスクは機械学習システムがどのように事例（example）を実行するべきかという観点で記述される
- 事例とは機械学習システムで実行したい対象や事象から定量的に測定された特徴量（features）の集合である
- 事例はベクトル $\mathbf{x} \in \mathbb{R}^n$ で表す．ここで，ベクトルの各項目 x_i は異なる特徴量である
- 一般的な機械学習のタスクとして，分類（Classification），欠損値のある入力の分類（Classfication with missing inputs），回帰（Regression），転写（Transcription），機械翻訳，構造出力（Structured output），異常検知（Anomaly detection），合成とサンプリング（Synthesis and sampling），欠損値補完（Imputation of missing values），ノイズ除去（Denoising），密度推定（Density estimation）などがある

性能指標 P

- 機械学習アルゴリズムの能力を評価するためには，その性能を測る定量的な尺度を設計しなければならない.
  - たとえば分類タスクではモデルの精度（accuracy）を測定することが多い
- 未知のデータに対して機械学習アルゴリズムがうまく機能するかを知りたいので，学習に使われるデータとは異なるテスト集合（test set）を用いて性能指標を評価する

経験 E

機械学習アルゴリズムは，学習過程においてどのような経験を獲得できるかによって以下の 2 つに大別できる

- 教師あり学習アルゴリズム（Supervised learning algorithms）
  - 特徴量がラベル（label）や目標（target）と関連付けられているデータを用いて学習を行う
  - 特徴量 $\mathbf{x}$ とラベル $\mathbf{y}$ から $p(\mathbf{y}|\mathbf{x})$ を予測
  - 分類や回帰など
- 教師なし学習アルゴリズム（Unsupervised learning algorithms）
  - ラベルが付いていないデータ集合から，そのデータ集合構造の有益な特性を学習する
  - 特徴量 $\mathbf{x}$ とから $p(\mathbf{x})$ や重要な特性を学習
  - 密度推定やクラスタリングなど

データ行列

- 大多数の機械学習アルゴリズムでは，単純にデータ集合を経験する.（例外：強化学習（reinforcement learning））
- 一般的にデータ集合は m 個の要素を含む集合 $\left\{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(m)}\right\}$ として記述できる
- 特に $\mathbf{x}^{(i)}$ と $\mathbf{x}^{(j)}$ のサイズが同じ場合はデータ行列 $\mathbf{X} \in \mathbb{R}^{n \times m}$ として記述できる

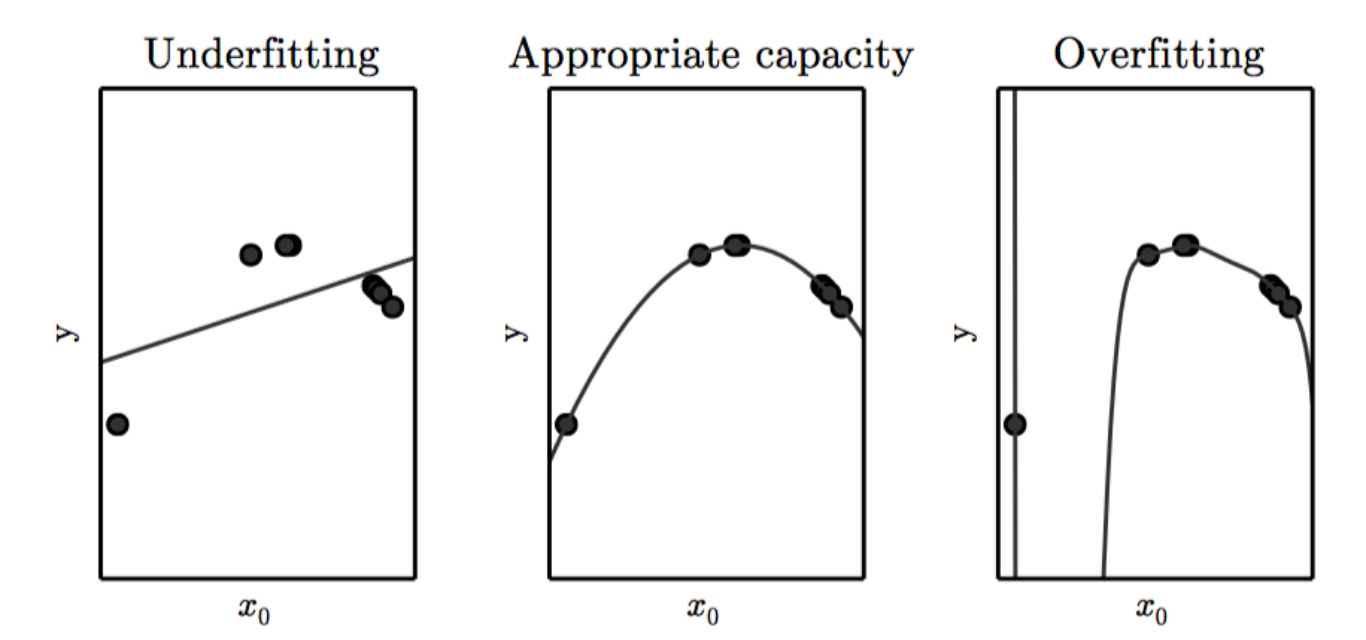
容量・過剰適合・過少適合

汎化

- 機械学習では，モデルの学習に使用した入力だけではなく，これまで見たことのない新たな入力に対しても良い性能を発揮できることが求められる（汎化（generalization））
- 機械学習が最適化と異なるのは，訓練誤差（training error）だけでなく，汎化誤差 （generalization error）（テスト誤差（test error）とも呼ばれる）も小さくしたいという点である
- データ生成過程（data-generating process）に，i.i.d. 仮定（i.i.d. assumptions）とを置くことで訓練誤差とテスト誤差の関係性を数学的に調べることが可能になる

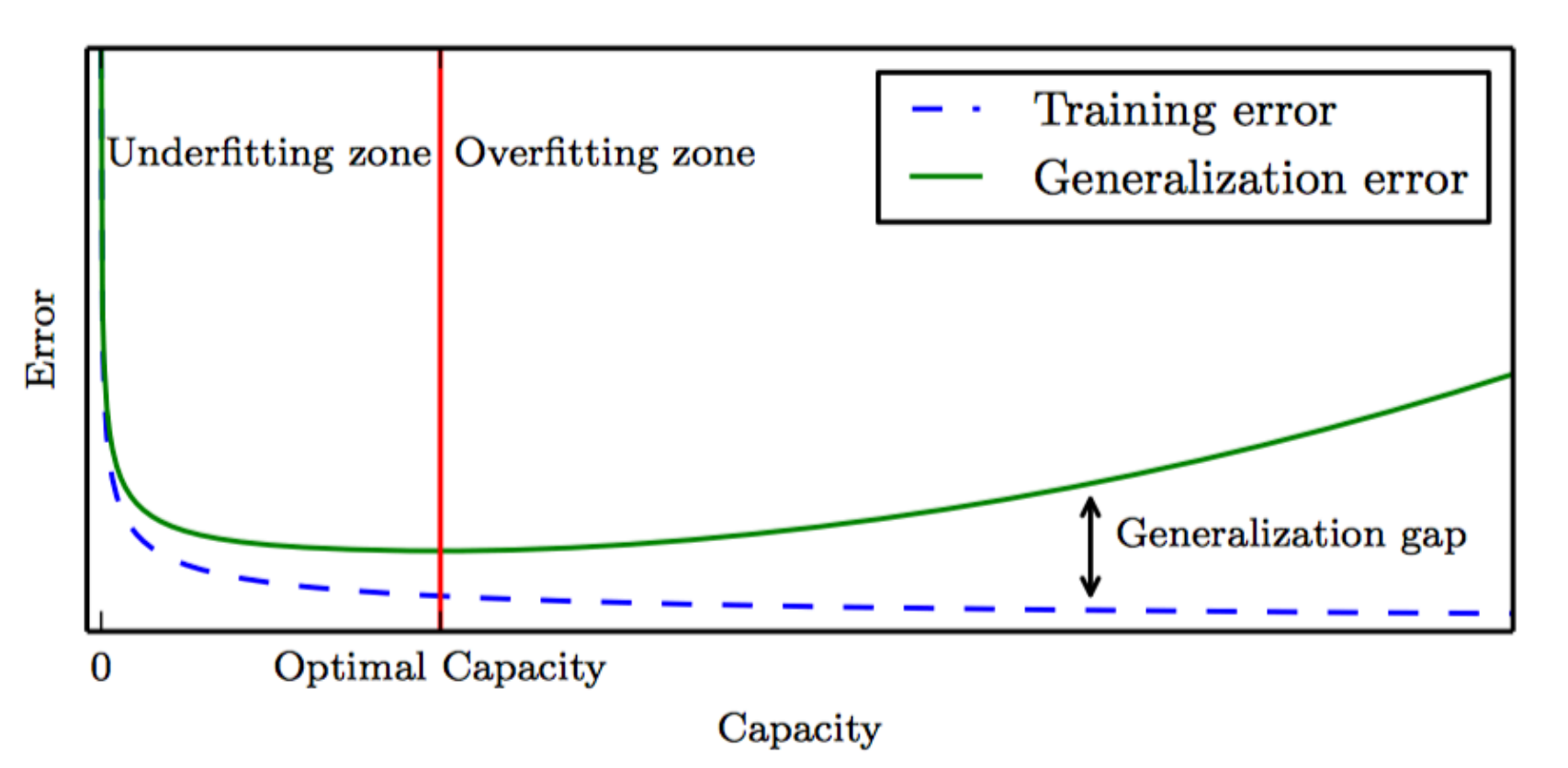
過小適合・過剰適合

- 訓練集合において十分に小さな誤差が得られない場合を過少適合（未学習，underfitting）と呼び，訓練誤差とテスト誤差との差が大きすぎる場合を過剰適合（過学習，overfitting）と呼ぶ
- モデルの容量（capacity）を変更することで過小適合・過剰適合をコントロールする
- 実行する必要のあるタスクの真の複雑さと与えられる訓練データの量に対して適切な容量があるときに，最もよく性能を発揮する



過小適合・過剰適合

- 容量が小さい場合は訓練誤差が大きい（過小適合）
- 容量が増大すると訓練誤差は減少するが，訓練誤差と汎化誤差の差は広がる
- 最終的にはその差が訓練誤差の減少量を上回る（過剰適合）



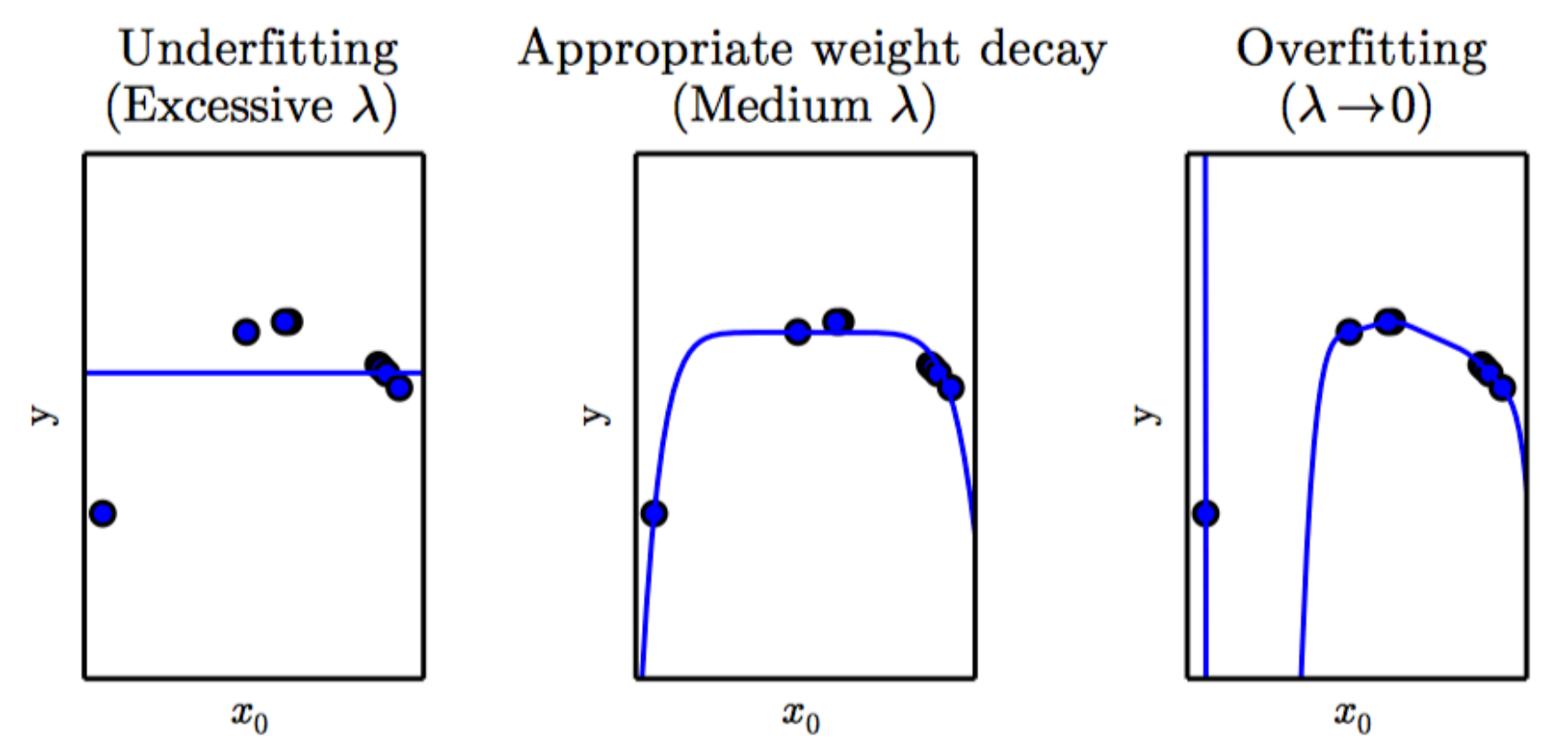
ノーフリーランチ定理

- データを生成する分布すべてを平均すると，どの分類アルゴリズムも，過去に観測されていない点を分類する際の誤差率は同じになる
- つまり，他の機械学習アルゴリズムよりも普遍的に良いと言える機械学習アルゴリズムは存在しない
- よって，普遍的な学習アルゴリズムや絶対的に最良の学習アルゴリズムを探し求めることではなく，対象としたいデータ 生成の分布から抽出されるデータに対して良い性能を発揮する機械学習アルゴリズムがどのようなものであるかを理解することが目標

正則化

- 正則化項（regularizer）と呼ばれるペナルティをコスト関数に追加 することで，関数 $f(x;\boldsymbol{\theta})$ を学習するモデルを正則化できる.
- ex. 重み減衰（weight decay）

$$J(\mathbf{w}) = \mathrm{MES}_{\mathrm{train}} + \lambda \mathbf{w}^\top \mathbf{w}$$



ハイパーパラメータと検証集合

ハイパーパラメータ

- ほとんどの機械学習アルゴリズムには，そのアルゴリズムの挙動を制御するための設定値がある．この設定値はハイパーパラメータ（hyperparameters）と呼ばれる．
- 訓練集合で学習された場合，このようなハイパーパラメータは常に可能な範囲で最大のモデル容量を選択するので，結果的に過剰適合になる
- この問題を解決するためには，訓練アルゴリズムが観察しない検証集合（validation set）が必要になる．

交差検証

k-分割交差検証アルゴリズム

Define KFoldXV($\mathbb{D}, A, L, k$):
Require: 与えられたデータ集合 $\mathbb{D}$ とその要素 $\boldsymbol{z}^{(i)}$
Require: 学習アルゴリズム A，データ集合を入力とし，学習済みの関数を出力とする関数.
Require: 損失関数 L，学習済みの関数と事例 $\boldsymbol{z}^{(i)} \in \mathbb{D}$ からスカラー値 $\in \mathbb{R}$ へ変換する関数.
Require: 分割数 k
  $\mathbb{D}$ を k 個の互いに排他的な部分テストセット $\mathbb{D}_i$ に分割する．それらの和集合は $\mathbb{D}$.
  for i from 1 to k **do**
    $f_i = A(\mathbb{D} \backslash \mathbb{D}_i)$
    for $\boldsymbol{z}^{(j)}$in$\mathbb{D}_i$ **do**
      $e_j = L(f_i, \boldsymbol{z}^{(j)})$
    end for
  end for
  Returne

推定量・バイアス・バリアンス

点推定量

- 関心のある量（パラメータなど）について「最良の」予測を 1 つ提示する試みを点推定と呼ぶ
  - パラメータ $\boldsymbol{\theta}$ の点推定を $\widehat{\boldsymbol{\theta}}$ で表す
- $\left\{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(m)}\right\}$ を m 個の独立同一分布（i.i.d.）のデータ点から成る集合とする．点推定量（point estimator）もしくは統計量（statistic）は，データの任意の関数

$$\widehat{\boldsymbol{\theta}}_m = g(\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(m)})$$

  である
- なお，現段階では，真のパラメータの値 $\boldsymbol{\theta}$ は固定であるが未知であり，点推定 $\widehat{\boldsymbol{\theta}}$ は確率変数と仮定する

より望ましい推定量

- 良好な推定量は，訓練データを生成した真の潜在的な $\boldsymbol{\theta}$ に近い出力を持つ関数
- つまり，バイアスや分散が小さい推定量や，一致性を持つ推定量はより望ましい推定量と考えられる

バイアス

- 推定量のバイアスは関数やパラメータの真の値からの期待偏差であり,

$$\text{bias}\left(\widehat{\boldsymbol{\theta}}_m\right) = \mathbb{E}\left(\widehat{\boldsymbol{\theta}}_m\right) - \boldsymbol{\theta}_m$$

  で定義される. バイアスは小さいほうがより良い推定量だと考えられる.

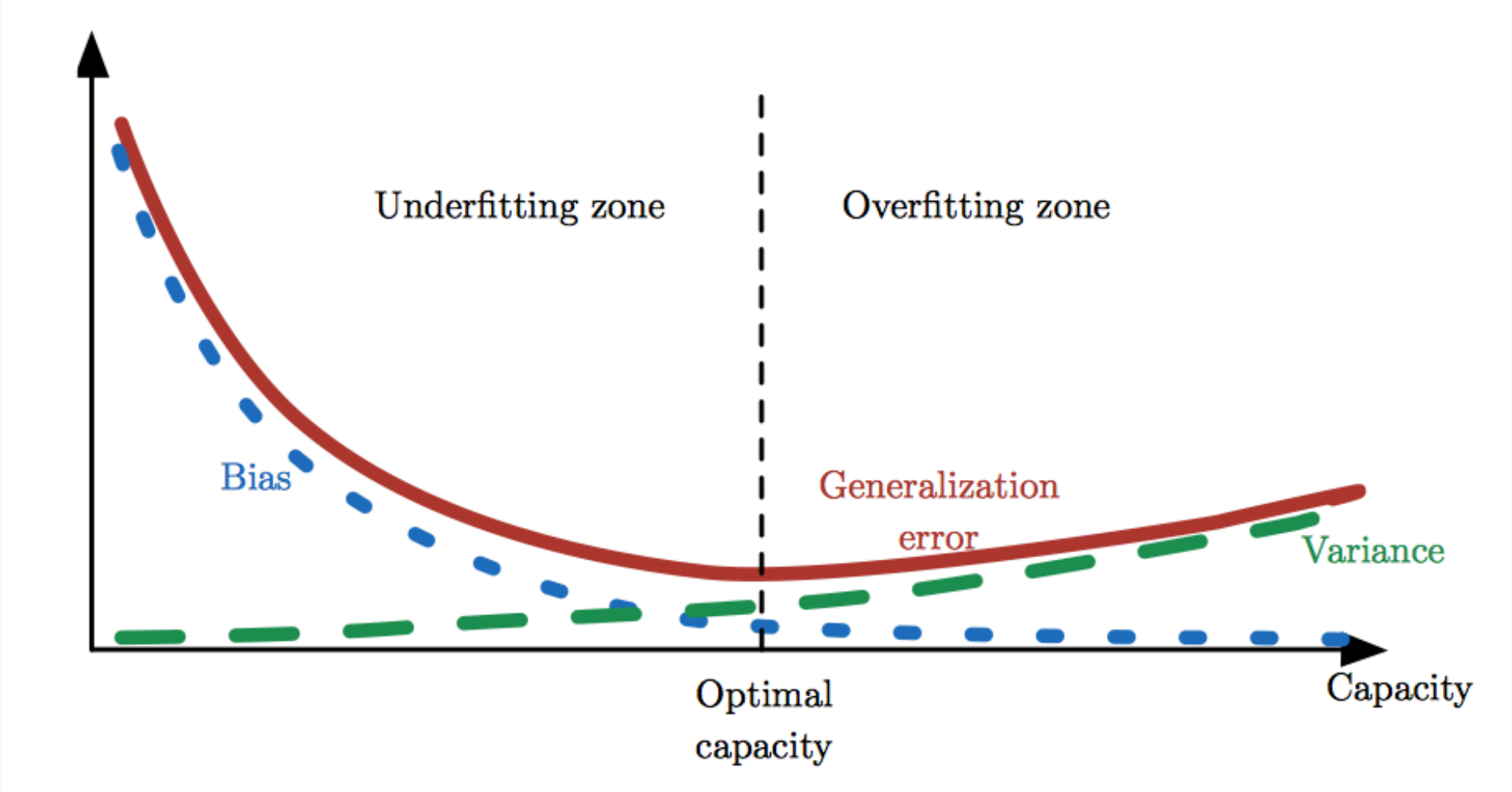
  - $\text{bias}\left(\widehat{\boldsymbol{\theta}}_m\right) = 0$ つまり $\mathbb{E}\left(\widehat{\boldsymbol{\theta}}_m\right) = \widehat{\boldsymbol{\theta}}_m$ のとき推定量は不偏（unbiased）
  - $lim_{m\to\infty}\text{bias}\left(\widehat{\boldsymbol{\theta}}_m\right) = 0$ つまり $lim_{m\to\infty}\mathbb{E}\left(\widehat{\boldsymbol{\theta}}_m\right) = \widehat{\boldsymbol{\theta}}_m$ のとき推定量は漸近不偏（asymptotically unbiased）

分散・標準誤差

- 推定量がデータサンプルの関数としてどれだけ変化すると予測されるかを推定量の分散（variance）と呼ぶ．分散の平方根は標準誤差（standard error）と呼ばれる．分散が小さいほうがより良い推定量だと考えられる．
- 推定量の分散を $\mathrm{Var}\left(\widehat{\boldsymbol{\theta}}\right)$，標準誤差を $\mathrm{SE}\left(\widehat{\boldsymbol{\theta}}\right)$ で表す．

バイアスと分散のトレードオフ

一般に，バイアスと分散にはトレードオフの関係がある



二乗誤差

推定量の二乗誤差 (mean squared error) を用いることで，バイアスと分散を同時に考慮することが出来る

$$
\begin{aligned}
\mathrm{MSE}\left(\widehat{\boldsymbol{\theta}}\right) &= \mathbb{E}\left(\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right)^2\right) \\
&= \mathbb{E}\left(\left(\widehat{\boldsymbol{\theta}} - \mathbb{E}\left(\boldsymbol{\theta}\right) + \mathbb{E}\left(\boldsymbol{\theta}\right) - \boldsymbol{\theta}\right)^2\right) \\
&= \mathbb{E}\left(\left(\widehat{\boldsymbol{\theta}} - \mathbb{E}\left(\boldsymbol{\theta}\right)\right)^2\right) + \mathbb{E}\left(\left(\mathbb{E}\left(\boldsymbol{\theta}\right) - \boldsymbol{\theta}\right)^2\right) \\
&\quad + 2\mathbb{E}\left(\left(\widehat{\boldsymbol{\theta}} - \mathbb{E}\left(\boldsymbol{\theta}\right)\right)\left(\mathbb{E}\left(\boldsymbol{\theta}\right) - \boldsymbol{\theta}\right)\right) \\
&= \mathrm{Var}\left(\widehat{\boldsymbol{\theta}}\right) + \mathrm{bias}\left(\widehat{\boldsymbol{\theta}}\right)^2
\end{aligned}
$$

一致性

- ここまで固定サイズの訓練集合における推定量の特性に注目してきたが，訓練データの量が増える場合の推定量の挙動にも注意が必要である.
- 特にデータポイント数 m が増加するにつれて点推定量は真の値に収束することが望ましい．これを数式で表すと

$$\operatorname*{plim}_{m\to\infty} \widehat{\theta}_m \to \theta$$

であり，これを（弱）一致性（consistency）と呼ぶ．ここで plim は確率収束

$$\forall \epsilon > 0, \quad P\left(\left|\widehat{\theta}_m - \theta\right| > \epsilon\right) \to 0 \text{ as } m \to \infty$$

を意味する.
- なお，$P\left(\lim_{m\to\infty} \widehat{\theta}_m = \theta\right) = 1$ つまり概収束（Almost sure convergence）する場合には強一致性をもつという

例：ガウス分布

具体例として，独立同一にガウス分布に従うサンプル集合 $\left\{x^{(1)}, \ldots, x^{(m)}\right\}$ を考える：

$$
\begin{aligned}
p\left(x^{(i)}\right) &= \mathcal{N}\left(x^{(i)}; \mu, \sigma^2\right) \\
&= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{\left(x^{(i)} - \mu\right)^2}{2\sigma^2}\right)
\end{aligned}
$$

ガウス分布のサンプル平均（不偏性）okasii

- このとき，サンプル平均（sample mean）

$$\widehat{\mu}_m = \frac{1}{m}\sum_{i=1}^{m} x^{(i)}$$

はガウス平均パラメータ μ の不偏推定量である

$$\mathbb{E}\left(\widehat{\mu}_m\right) = \mathbb{E}\left(\frac{1}{m}\sum_{i=1}^{m} x^{(i)}\right) = \frac{1}{m}\sum_{i=1}^{m}\mathbb{E}\left(x^{(i)}\right) = \frac{1}{m}\sum_{i=1}^{m}\mu = \mu$$

- なお，推定量の分散と標準誤差は以下で求まる

$$\mathrm{Var}\left(\widehat{\mu}_m\right) = \mathrm{Var}\left(\frac{1}{m}\sum_{i=1}^{m} x^{(i)}\right) = \frac{1}{m^2}\sum_{i=1}^{m}\mathrm{Var}\left(x^{(i)}\right) = \frac{1}{m^2}\sum_{i=1}^{m}\sigma^2 = \frac{1}{m}\sigma^2$$

$$\mathrm{SE}\left(\widehat{\mu}_m\right) = \sqrt{\mathrm{Var}\left(\widehat{\mu}_m\right)} = \frac{\sigma}{\sqrt{m}}$$

ガウス分布のサンプル平均（一致性）

- サンプル平均は一致推定量でもある
- 証明にはチェビシェフの不等式

$$\forall \epsilon > 0, \quad P\left(|X - \mathbb{E}\left(X\right)| > \epsilon\right) < \frac{\mathrm{Var}\left(X\right)}{\epsilon^2}$$

  を用いる
- $\widehat{\mu}_m$ にチェビシェフの不等式を適用すると

$$\forall \epsilon > 0, \quad P\left(|\widehat{\mu}_m - \mu| > \epsilon\right) < \frac{\sigma^2}{m\epsilon^2} \to 0 \text{ as } m \to \infty$$

  であり，よって $\widehat{\mu}_m$ は一致推定量である

ガウス分布のサンプル分散

- 一方で，サンプル分散（sample variance）

$$\widehat{\sigma}_m^2 = \frac{1}{m}\sum_{i=1}^{m}\left(x^{(i)} - \widehat{\mu}_m\right)^2$$

は不偏推定量ではない：

$$\begin{aligned}
\mathbb{E}\left(\widehat{\sigma}_m^2\right) &= \mathbb{E}\left(\frac{1}{m}\sum_{i=1}^{m}\left(x^{(i)} - \widehat{\mu}_m\right)^2\right) \\
&= \mathbb{E}\left(\frac{1}{m}\sum_{i=1}^{m}\left(x^{(i)} - \mu\right)^2 - \left(\widehat{\mu}_m - \mu\right)^2\right) \\
&= \frac{1}{m}\sum_{i=1}^{m}\mathbb{E}\left(\left(x^{(i)} - \mu\right)^2\right) - \mathbb{E}\left(\left(\widehat{\mu}_m - \mu\right)^2\right) \\
&= \frac{m-1}{m}\sigma^2
\end{aligned}$$

- ただし，漸近不偏ではある：

$$\mathbb{E}\left(\widehat{\sigma}_m^2\right) = \frac{m-1}{m}\sigma^2 \to \sigma^2 \text{ as } m \to \infty$$

ガウス分布のサンプル分散

- サンプル分散は一致性をもつ

$$
\begin{aligned}
\widehat{\sigma}_m^2 &= \frac{1}{m}\sum_{i=1}^{m}\left(x^{(i)} - \widehat{\mu}_m\right)^2 \\
&= \frac{1}{m}\sum_{i=1}^{m}\left(x^{(i)} - \mu\right)^2 - (\widehat{\mu}_m - \mu)^2 \\
&\xrightarrow{p} \sigma^2 \text{ as } m \to \infty
\end{aligned}
$$

ガウス分布のサンプル平均（信頼区間）

- $\left\{x^{(1)}, \ldots, x^{(m)}\right\}$ は独立同一にガウス分布に従うと仮定したので，中心極限定理より

$$\sqrt{m}\left(\widehat{\mu}_m - \mu\right) \xrightarrow{d} \mathcal{N}\left(0, \sigma^2\right) \text{ as } m \to \infty$$

- これと連続写像定理を用いると

$$\begin{aligned}\frac{\widehat{\mu}_m - \mu}{\mathrm{SE}\left(\widehat{\mu}_m\right)} &= \frac{\sqrt{m}\left(\widehat{\mu}_m - \mu\right)}{\widehat{\sigma}_m} \\ &\xrightarrow{d} \frac{1}{\sigma}\mathcal{N}\left(0, \sigma^2\right) \text{ as } m \to \infty \\ &= \mathcal{N}\left(0, 1\right)\end{aligned}$$

  となることがわかる（なお，$\xrightarrow{d}$ は分布収束を表す）

- よって $\widehat{\mu}_m$ の 95%信頼区間は

$$\left(\widehat{\mu}_m - 1.96 \times \mathrm{SE}\left(\widehat{\mu}_m\right),\ \widehat{\mu}_m + 1.96 \times \mathrm{SE}\left(\widehat{\mu}_m\right)\right)$$

  で与えられる

Deep Learning 輪読会 2017

第5章　機械学習の基礎 (5.5-)

2017.11.13

理学系研究科附属
天文学教育研究センター
学部4年 吉村勇紀

構成

5.5 最尤推定

5.6 ベイズ推定

5.7 教師あり学習アルゴリズム

5.8 教師なし学習アルゴリズム

5.9 確率的勾配降下法

5.10 機械学習アルゴリズムの構築

5.11 深層学習の発展を促す課題

5.5 最尤推定

- 最尤推定
  - モデルを固定したとき、事例集合が発現する確率が最大となるパラメータを求める。

$$\boldsymbol{\theta}_{\mathrm{ML}} = \arg\max_{\boldsymbol{\theta}} p_{\mathrm{model}}(\mathbb{X}; \boldsymbol{\theta})$$

$$= \arg\max_{\boldsymbol{\theta}} \prod_{i=1}^{m} p_{\mathrm{model}}(\boldsymbol{x}^{(i)}; \boldsymbol{\theta})$$

  - アンダーフローを考慮して普通 log をとる

$$\boldsymbol{\theta}_{\mathrm{ML}} = \arg\max_{\boldsymbol{\theta}} \sum_{i=1}^{m} \log p_{\mathrm{model}}(\boldsymbol{x}^{(i)}; \boldsymbol{\theta})$$

5.5 最尤推定

- KLダイバージェンスとの関係
  - 最尤推定は訓練集合で定義される経験分布とモデル分布の差（KLダイバージェンス）を最小化することに相当する

$$D_{\mathrm{KL}}\left(\hat{p}_{\mathrm{data}} \| p_{\mathrm{model}}\right) = \mathbb{E}_{\mathbf{x} \sim \hat{p}_{\mathrm{data}}}\left[\log \hat{p}_{\mathrm{data}}(\boldsymbol{x}) - \log p_{\mathrm{model}}(\boldsymbol{x})\right]$$

  - 上式の次の交差エントロピーの最小化と等しい

$$- \mathbb{E}_{\mathbf{x} \sim \hat{p}_{\mathrm{data}}}\left[\log p_{\mathrm{model}}(\boldsymbol{x})\right]$$

  - 最尤推定はモデル分布を経験分布（真のデータ分布は取り扱えない）に一致させる試みである

5.5.1 条件付き対数尤度と平均二乗誤差

- 条件付き対数尤度
  - 入力xとモデルパラメータに対する出力yの条件確率から最尤推定量を定式化できる

$$\boldsymbol{\theta}_{\mathrm{ML}} = \arg\max_{\boldsymbol{\theta}} P(\boldsymbol{Y} \mid \boldsymbol{X}; \boldsymbol{\theta})$$

  - 確率分布がガウシアンで事例が独立同一分布に従う場合

$$\sum_{i=1}^{m} \log p(y^{(i)} \mid \boldsymbol{x}^{(i)}; \boldsymbol{\theta})$$

$$= -m \log \sigma - \frac{m}{2} \log(2\pi) - \sum_{i=1}^{m} \frac{\left\| \hat{y}^{(i)} - y^{(i)} \right\|^2}{2\sigma^2}$$

5.5.2 最尤法の特性

- 事例数 m → ∞ で最尤推定量は真値に漸近する
  - 真の分布がモデル集合内にあり、かつ
  - 真の分布とモデルパラメータが1対1対応している場合

- 最尤推定量の統計的有効性
  - 同じ真値に対する一致推定量でも有限サンプルに対しては汎化誤差が異なる場合がある
  - mが大きい場合、最尤推定量より小さな平均二乗誤差を持つ一致推定量は存在しない

5.6 ベイズ統計

- パラメータの点推定ではなくパラメータの（事後）確率分布を推定する
  - 事後分布 ∝ 尤度関数 × 事前分布

$$p(\boldsymbol{\theta} \mid x^{(1)}, \ldots, x^{(m)}) = \frac{p(x^{(1)}, \ldots, x^{(m)} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta})}{p(x^{(1)}, \ldots, x^{(m)})}$$

  - パラメータの事後分布を畳み込み積分することで最終的な予測を得る

$$p(x^{(m+1)} \mid x^{(1)}, \ldots, x^{(m)}) = \int p(x^{(m+1)} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid x^{(1)}, \ldots, x^{(m)}) \, d\boldsymbol{\theta}.$$

5.6 ベイズ統計

- 例）ベイズ線形回帰
  - パラメータの事前分布をガウシアンにとる
  - 途中省略（本文参照）
  - 事後分布もガウシアンになる
  - 事後分布の平均値を推定量と思うと、二乗和誤差関数に二次正則化項を加えた評価関数で最尤推定した結果と一致

5.6.1 最大事後確率（MAP）推定

- MAP推定
  - 事後分布の取り扱いは一般に困難である
  - 点推定量の方が扱いが簡単
  - 事後分布が最大となる値を点推定する
  - ガウシアン事前分布+MAP推定 → 二次正則化
  - 混合ガウス分布などの正則化を設計する際にも持ちいる

5.7 教師あり学習アルゴリズム

- 教師あり学習
  - 入力xと出力yの対応を学習する
  - 訓練集合に関してその対応関係は人間の手であらかじめ与えられる（かあるいは自動的に収集される場合もある）

5.7.1 確率的教師あり学習

- 確率的教師あり学習
  - 入力に対して線形なモデルを仮定し、出力はモデル値を平均とするガウシアンに従う

$$p(y \mid \boldsymbol{x}; \boldsymbol{\theta}) = \mathcal{N}(y; \boldsymbol{\theta}^\top \boldsymbol{x}, \boldsymbol{I}).$$

  - 最適な重みは正規方程式を解くことで一意に定まる。
  - クラス分類など2項変数を取り扱う際はロジスティックシグモイド関数などの活性化関数に入れる。

$$p(y = 1 \mid \boldsymbol{x}; \boldsymbol{\theta}) = \sigma(\boldsymbol{\theta}^\top \boldsymbol{x}).$$

  - この場合、最適な重みは閉形式の解は存在しない

5.7.2 サポートベクトルマシン

- サポートベクトルマシン
  - 線形関数でモデル化する
  - クラス識別情報のみ出力する

- カーネルトリック
  - モデルは事例との内積で表現できる
  - $$\boldsymbol{w}^\top \boldsymbol{x} + b = b + \sum_{i=1}^{m} \alpha_i \boldsymbol{x}^\top \boldsymbol{x}^{(i)}$$

$$f(\boldsymbol{x}) = b + \sum_i \alpha_i k(\boldsymbol{x}, \boldsymbol{x}^{(i)}).$$

5.7.2 サポートベクトルマシン

- カーネルの利点
  - カーネル関数を計算する方が内積を取るより、計算が簡単
  - カーネルトリックによる改良を施したものを「カーネルマシン」あるいは「カーネル法」と呼ぶ

- カーネルマシンの欠点
  - 決定関数の評価コストが事例数に比例する
  - スパースなαを学習することで評価コストを抑える → サポートベクトル
  - データ集合が大きいと訓練の計算コストが高くなる

5.7.3 その他の教師あり学習アルゴリズム

- k近傍法
  - データ集合のうち入力xに近いk個の平均をとる
  - 訓練集合が小さいと汎化性能は悪い
  - 訓練集合が大きくなるほど精度が上がる一方、同時に計算コストは上がる
  - 特徴量間の優劣がつけられない

- 決定木
  - 決定木によって入力空間を（軸に沿って）分割
  - 決定木の訓練は本書の範囲を超える
  - 軸に沿わない決定境界は困難

5.8 教師なし学習アルゴリズム

- 教師なし学習
  - 訓練集合の特徴量を抽出するが教師信号は学習しない
  - データの「最良」の表現を見つける
  - 単純な表現 → 低次元、スパース、独立

5.8.1 主成分分析

- 主成分分析
  - 低次元で相関の小さい表現を学習する
  - 共分散行列が対角行列になるようなデータセットへの線形変換を学習する

$$\boldsymbol{z} = \boldsymbol{x}^\top \boldsymbol{W}$$

  - 特異値分解（または対角化）によって変換行列を得る

$$\boldsymbol{X}^\top \boldsymbol{X} = \left(\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{W}^\top\right)^\top \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{W}^\top = \boldsymbol{W}\boldsymbol{\Sigma}^2\boldsymbol{W}^\top$$

5.8.2 k平均クラスタリング

- クラスタリング
  - 入力x をone-hotコードベクトルh に対応させる
  - i番目のクラスに属する場合 h_i=1、それ以外の成分は0

- k平均クラスタリング
  - k個あるセントロイドの最適化とクラス割り当ての最適化を交互に行い収束させる

- クラスタリングの問題点
  - クラスタリングと現実世界との対応が明らかではない
  - 1つの特徴に対応する複数のクラスタリングが有りうる
  - 分散表現が好まれることもある

5.9 確率的勾配降下法

- 確率的勾配降下法(SGD)
  - ほとんどの深層学習はSGDで動作している
  - 勾配を用いた最適化

$$\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) = \frac{1}{m} \sum_{i=1}^{m} \nabla_{\boldsymbol{\theta}} L(\boldsymbol{x}^{(i)}, y^{(i)}, \boldsymbol{\theta})$$

  - 計算コストは通常O(m)
  - 勾配は期待値なので少ないサンプルだけで評価しても良い
  - 最終的に勾配方向にパラメータを進める

$$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \epsilon \boldsymbol{g}$$

5.10 機械学習アルゴリズムの構築

- 機械学習を構成するもの
  - データ集合の仕様
  - 誤差関数
    - 負の対数尤度（＋正則化項）
  - モデル
    - 線形 or 非線形
  - 最適化手順

5.11深層学習の発展を促す課題

- 高次のデータに対する困難
  - 本章で扱う機械学習アルゴリズムでは音声認識や物体認識などの問題を解決できない
  - 高次のデータを扱う際は汎化が指数関数的に困難になる
  - 従来のアルゴリズムは高次データの汎化に適さない

5.11.1 次元の呪い

- 次元の呪い
  - 次元が高いデータには困難が伴う
  - コンポーネントの数 > 事例数
  - 例えば単純なm次多項式フィッティングだと係数がD^m個

5.11.2 局所一様と平滑化

- 平滑化事前分布
  - 「学習する関数は小さな領域であまり変化してはならない」という仮定を反映した事前分布

$$f^*(\boldsymbol{x}) \approx f^*(\boldsymbol{x} + \epsilon)$$

  - k平均法や決定木も多かれ少なかれこの仮定を置いている

- 複雑な関数
  - 高次元な関数や領域毎に挙動が違う複雑な関数ではこの仮定は適切でない
  - タスク毎に固有の仮定を導入して解決する
  - 複数の階層を考えることで仮定を軽くする → ニューラルネット

5.11.3 多様体学習

- 多様体仮説
  - 現実的なデータは高次元空間中の低次な領域に押し込まれている
  - ノイズデータ中に実際のデータを見出す確率は極めて小さい
  - 多様体を描く変換は、実際に想像可能なことが多い
    - 平行移動、回転、ぼかし、変色、…

参考文献

- Deep Learning
  - Ian Goodfellow, Yoshua Bengio, Aaron Courville
  - 日本語版
    https://www.amazon.co.jp/%E6%B7%B1%E5%B1%A4%E5%AD%A6%E7%BF%92-Ian-Goodfellow/dp/4048930621