ディープラーニング基礎

Chapter2　機械学習と深層学習

須賀勇貴

茨城大学大学院　理工学研究科　量子線科学専攻　1 年

March 29, 2023

2.1 なぜ深層学習か？

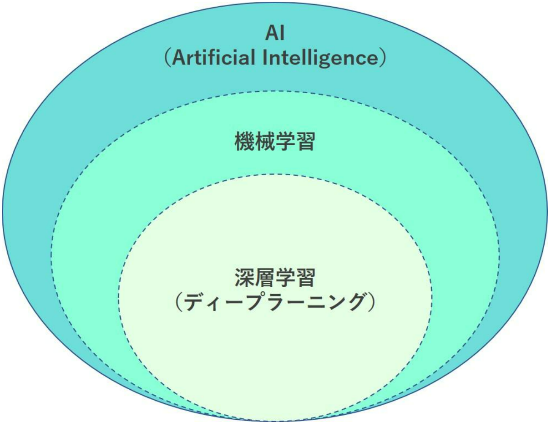

図: AI と機械学習とディープラーニングの関係

https://business.ntt-east.co.jp/content/cloudsolution/column-159.html

2.1 なぜ深層学習か？

深層学習 (Deep Learning) とは

動物の神経回路にヒントを得て提唱された (深層) ニューラルネットワーク計算により，
大量のデータからその背後に潜む知識を自発的に獲得していく手法

- 深層学習の何がすごい？
  1. 汎用的なアルゴリズムを提案してくれる
     → タスクの種類に依存しない
  2. 極めて高い汎化性能がある
     → 手持ちのデータだけの中から，全ての状況に通用する本質的な知識を獲得できる

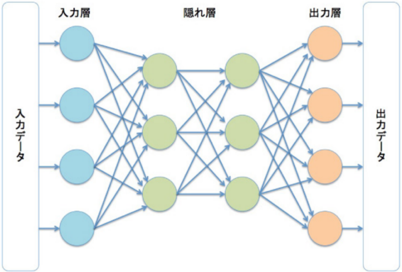

図: 深層ニューラルネットワークの図

2.2 機械学習とは何か

機械学習とは，人間がこなすようなさまざまな学習や知的作業を計算機に実行させるためのアプローチの研究，あるいはその手法そのものを意味する

T. M. ミッシェルによる機械学習の定式化

「コンピュータプログラムが，ある種のタスク **T** と評価尺度 **P** において経験 **E** から学習するとは，タスク **T** におけるその性能を評価尺度 **P** によって評価した際に，経験 **E** によってそれが改善されている場合である」

タスク T $\cdots$ 解きたい問題．
例）回帰，分類，強化学習，パターン認識，クラスタリング分析，最適化分析など
評価尺度 P $\cdots$ モデルの精度．
回帰の場合,「平均二乗誤差」がよく使われる
経験 E $\cdots$ データ集合．

要するに

経験 E の蓄積によってタスク T を解いたときに，評価尺度 P が向上する手法

2.2 機械学習とは何か

2.2.1 代表的なタスク

(1) クラス分類

いくつかのカテゴリ (クラス) に仕分ける作業
与えられた数値データを x，クラス (K 個) を $\mathcal{C}_y$ $(y = 0, 1, \cdots, K)$ で表すことにすると，x をクラス $\mathcal{C}_y$ へ分離するということは，x の所属クラスを表す離散値ラベル y の値を決めることである

$$\boldsymbol{x} \to \boldsymbol{y}(\boldsymbol{x}) \in \{0, 1, \cdots, K\}$$

(2) 回帰

データから，それに対応する実数値 (を並べたベクトル)y を予測する作業. つまり，与えられた x を，対応する y に変換するための関数 $y(x)$ を決定する

$$\boldsymbol{x} \to \boldsymbol{y}(\boldsymbol{x}) \in \mathbb{R}$$

回帰におけるタスクの応用例

機械翻訳，音声認識，異常検知，データ次元削除

2.2 機械学習とは何か

2.2.2 さまざまなデータセット

MNIST

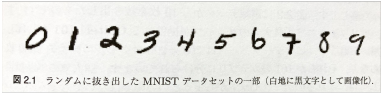

図 2.1 ランダムに抜き出した MNIST データセットの一部（白地に黒文字として画像化）.

- 手書き数字のデータベース
- グレースケールの 28×28 ピクセル画像
- 訓練用に 6 万枚，テスト用に 1 万枚用意されている

Imagenet

- 約 1400 万枚の自然画像からなる巨大データベース
- クラス数は 2 万にも及ぶ

CIFAR-10

図 2.2 各カテゴリからランダムに抜き出した CIFAR-10 データセットの一部.

- 自然画像のデータベース
- 10 のカテゴリに分けられた 32×32 ピクセル画像
- 訓練用に 6 万枚，テスト用に 1 万枚用意されている

2.3 統計入門

2.3.1 標本と推定

どのようにしたらプログラムはデータからタスクをこなすための知識を学び取れるか？

データを科学的に分析 $\to$ 統計学

機械学習の手法も統計を基礎として構築される (統計的機械学習)
まずは用語の確認

データ (集合) やサンプル，標本 $\cdots$ **データ点 (data point) の集まりからなるもの**

手書き文字画像認識の例

データ (集合) $\to$ 画像の集合
データ点 $\to$ 1 枚 1 枚の画像

※これらの用語は乱用されており，データ点を略してデータと言ったり，サンプルをサンプルの要素であるデータ点の意味で用いたりする

$\to$ 文脈から意味を判断

2.3 統計入門

2.3.1 標本と推定

推定を考える

- 統計解析に用いるデータは母集団から無作為に抽出されたものとみなす
- データの分析から母集団についての知識を獲得することが目標
- 母集団の性質はデータ生成確率 $P_{data}(\boldsymbol{x})$ により特徴付けられているものとする (=不確定性のある現象を確率的にモデル化)
- サンプル x はデータ生成確率から抽出されたものであると仮定する

$$\boldsymbol{x} \sim P_{data}(\mathbf{x})$$

$\rightarrow$ 現象を確率的に予測できるようになる

2.3 統計入門

2.3.1 標本と推定

推定を考える

母集団について知る $=$ データ生成分布を知る

データ生成分布を特徴づける量をパラメータと呼ぶ

ガウス分布の例

パラメータは平均値 μ と分散 σ^2

$$P(x) = \mathcal{N}(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

この 2 つの値が決まれば分布が具体的に決定される

実際のデータ生成分布は無数のパラメータを持つ $\rightarrow$ 良く近似できると期待できるモデル分布 $P(\mathrm{X}; \boldsymbol{\theta})$ を仮定し，そのモデルのパラメータ $\boldsymbol{\theta}$ の最適値 $\boldsymbol{\theta}^*$ をデータから推定する
$\rightarrow$ パラメトリックなアプローチ

2.3 統計入門

2.3.2 点推定

点推定とは

手持ちの有限要素のデータ集合 $\mathcal{D} = \{x_1, x_2, \cdots, x_N\}$ から確率分布のパラメータの尤もらしい値を計算すること

点推定のためには，データを決める確率変数 $\{\mathrm{x}_1, \mathrm{x}_2, \cdots, \mathrm{x}_\mathrm{N}\}$ の関数である推定量を作る必要がある

$$\hat{\boldsymbol{\theta}}(\mathrm{x}_1, \mathrm{x}_2, \cdots, \mathrm{x}_\mathrm{N})$$

これに具体的なデータが与えられると，数値としてのパラメータの推定値を得られる

$$\hat{\boldsymbol{\theta}}^*(x_1, x_2, \cdots, x_N) = \hat{\boldsymbol{\theta}}(x_1, x_2, \cdots, x_N)$$

この推定値は考えているパラメータを良く近似するように作る必要がある

2.3 統計入門

2.3.2 点推定

良い推定量の作り方

2.3 統計入門

2.3.2 点推定

良い推定量の作り方

1. バイアスが小さい　バイアス $\cdots$ 推定量の期待値 $E[\hat{\boldsymbol{\theta}}]$ と真の値 θ^* の差

$$b(\hat{\boldsymbol{\theta}}) = E[\hat{\boldsymbol{\theta}}] - \theta^*$$

不偏推定量 $\cdots$ バイアスがゼロのもの (望ましい推定量)
漸近不偏推定量 $\cdots$ データの数が増えるにつれゼロへ漸近するもの
$(\lim_{N\to\infty} b(\hat{\boldsymbol{\theta}}) = 0)$

2.3 統計入門

2.3.2 点推定

良い推定量の作り方

1. バイアスが小さい　バイアス $\cdots$ 推定量の期待値 $E[\hat{\theta}]$ と真の値 θ^* の差

$$b(\hat{\boldsymbol{\theta}}) = E[\hat{\boldsymbol{\theta}}] - \theta^*$$

   不偏推定量 $\cdots$ バイアスがゼロのもの (望ましい推定量)
   漸近不偏推定量 $\cdots$ データの数が増えるにつれゼロへ漸近するもの ($\lim_{N\to\infty} b(\hat{\boldsymbol{\theta}}) = 0$)

2. 分散が小さい (つまり，真の値に対して推定値のばらつきが小さい)

$$Var(\hat{\boldsymbol{\theta}}) = E\left[\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\right)^2\right]$$

2.3 統計入門

2.3.2 点推定

良い推定量の作り方

1. バイアスが小さい　バイアス $\cdots$ 推定量の期待値 $E[\hat{\boldsymbol{\theta}}]$ と真の値 θ^* の差

$$b(\hat{\boldsymbol{\theta}}) = E[\hat{\boldsymbol{\theta}}] - \theta^*$$

   不偏推定量 $\cdots$ バイアスがゼロのもの (望ましい推定量)
   漸近不偏推定量 $\cdots$ データの数が増えるにつれゼロへ漸近するもの
   $(\lim_{N\to\infty} b(\hat{\boldsymbol{\theta}}) = 0)$

2. 分散が小さい (つまり，真の値に対して推定値のばらつきが小さい)

$$Var(\hat{\boldsymbol{\theta}}) = E\left[\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\right)^2\right]$$

3. 一致性がある
   データ点の数が増えるにつれて統計量が真のパラメータに近づいていくという性質
   一致推定量 $\cdots$ $N \to \infty$ に従い $\hat{\boldsymbol{\theta}} \to \boldsymbol{\theta}^*$ となる推定量

2.3 統計入門

2.3.2 点推定

(1) ガウス分布

$$P(x) = \mathcal{N}(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

ガウス分布から無作為に取り出した N 個のデータからパラメータを推定するにはどうしたらよいか？任意のデータ点 x_n の期待値

$$E_{\mathcal{N}}[\mathrm{x}_n] = \int_{-\infty}^{\infty} x_n P(x_n) dx_n = \mu$$

$\rightarrow$ μ の推定値 $\hat{\mu}$ をサンプル平均としてみる

$$\hat{\mu} = \frac{1}{N} \sum_{n=1}^{N} \mathrm{x}_n$$

これは見事に不偏推定量となっている

$$E_{\mathcal{N}}[\hat{\mu}] = \frac{1}{N} \sum_{n=1}^{N} E_{\mathcal{N}}[\mathrm{x}_n] = \mu$$

2.3 統計入門

2.3.2 点推定

(1) ガウス分布

$$P(x) = \mathcal{N}(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

ガウス分布から無作為に取り出した N 個のデータからパラメータを推定するにはどうしたらよいか？$(\mathrm{x}_n - \mu)^2$ の期待値

$$E_{\mathcal{N}}[(\mathrm{x}_n - \mu)^2] = \int_{-\infty}^{\infty} (x_n - \mu)^2 P(x_n) dx_n = \sigma^2$$

$\rightarrow$ 平均値のときと同様に σ^2 の推定値 $\hat{\sigma}^2$ をサンプル平均による近似で表す

$$\hat{\sigma}^2 = \frac{1}{N} \sum_{n=1}^{N} (\mathrm{x}_n - \mu)^2$$

これは不偏推定量ではなく，漸近的不偏推定量になる

$$E_{\mathcal{N}}[\hat{\sigma}^2] = \left(\frac{N}{N-1} \right) \sigma^2$$

2.3 統計入門

2.3.2 点推定

(2) ベルヌーイ分布

$$P(x) = p^x(1-p)^{1-x}$$

パラメータは p のみ．期待値と分散は以下

$$E_P[\mathrm{x}] = \sum_{x=0,1} xP(x) = P(1) = p$$

$$E_P[(\mathrm{x}-p)^2] = \sum_{x=0,1} (x - 2px + p^2)P(x) = p(1-p)$$

$\rightarrow$ 推定量はサンプル平均とするのが良さそう

$$\hat{p} = \frac{1}{N}\sum_{n=1}^{N} \mathrm{x}_n$$

これは不偏推定量になっている

$$E_P[\hat{p}] = \frac{1}{N}\sum_{n=1}^{N} E_P[\mathrm{x}_n] = p$$

2.3 統計入門

2.3.2 点推定

(2) ベルヌーイ分布

$$P(x) = p^x(1-p)^{1-x}$$

分散の大きさはどうか

$$E_P[(\hat{p}-p)^2] = \frac{1}{N^2}\sum_{n=1}^{N} E_P[(\mathrm{x}_n - p)^2] = \frac{1}{N}p(1-p)$$

$\rightarrow$ 大きなデータに対して分散が小さくなるような推定量

2.3 統計入門

2.3.3 最尤推定

データ生成分布のパラメトリックモデル $P_{model}(\mathrm{x};\boldsymbol{\theta})$ が与えられているとして，
サンプル $\mathcal{D}=\{x_1,x_2,\cdots,x_N\}$ はこの分布から無作為に抽出されているとする．
このとき，このデータ集合が得られる同時確率密度は

$$P(x_1,x_2,\cdots,x_N;\boldsymbol{\theta})=\prod_{n=1}^{N}P_{model}(x_n;\boldsymbol{\theta})$$

これを変数 θ に対する量 $L(\boldsymbol{\theta})$ とみなして，尤度関数と呼ぶ．

$$L(\boldsymbol{\theta})=P(x_1,x_2,\cdots,x_N;\boldsymbol{\theta})$$

データ $\{x_1,x_2,\cdots,x_N\}$ は $L(\boldsymbol{\theta})$ を最大にするように実現されると解釈
$\rightarrow$ パラメータの値は $L(\boldsymbol{\theta})$ を最大化したもの

2.3 統計入門

2.3.3 最尤推定

最尤推定法

尤もらしいパラメータの値 $\boldsymbol{\theta}_{ML}$ は，尤度を最大化するものである

$$\boldsymbol{\theta}_{ML} = \underset{\theta}{\mathrm{argmax}} L(\boldsymbol{\theta})$$

実際は対数尤度を最大化，または負の対数尤度を最小化することが多い (結果は変わらない)

$$\boldsymbol{\theta}_{ML} = \underset{\theta}{\mathrm{argmax}} \log L(\boldsymbol{\theta})$$

$$\boldsymbol{\theta}_{ML} = \underset{\theta}{\mathrm{argmin}} (-\log L(\boldsymbol{\theta}))$$

2.3 統計入門

2.3.3 最尤推定

(1) ガウス分布の例
N 個のデータに対する尤度関数

$$L(\boldsymbol{\theta}) = \prod_{n=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_n-\mu)^2}{2\sigma^2}} \quad (\boldsymbol{\theta} = (\mu, \sigma^2))$$

対数尤度関数は

$$\log L(\boldsymbol{\theta}) = -\frac{N}{2}\log\sigma^2 - \sum_{n=1}^{N}\frac{(x_n-\mu)^2}{2\sigma^2} + \text{const.}$$

これの最大値は，パラメータに対する微分係数がゼロの場所を求めればよい

$$0 = \left.\frac{\partial \log L(\boldsymbol{\theta})}{\partial \mu}\right|_{\boldsymbol{\theta}_{ML}} = \frac{1}{\sigma_{ML}^2}\sum_{n=1}^{N}(x_n - \mu_{ML})$$

$$0 = \left.\frac{\partial \log L(\boldsymbol{\theta})}{\partial \sigma^2}\right|_{\boldsymbol{\theta}_{ML}} = \frac{N}{2}\frac{1}{\sigma_{ML}^2} + \frac{1}{2(\sigma_{ML}^2)^2}\sum_{n=1}^{N}(x_n - \mu_{ML})^2$$

$\rightarrow$ これを解けば最尤推定量 $(\mu_{ML}, \sigma_{ML}^2)$ は点推定の時と一致することがわかる

2.3 統計入門

2.3.3 最尤推定

(2) ベルヌーイ分布の例
N 個のデータに対する尤度関数

$$L(p) = \prod_{n=1}^{N} p^{x_n}(1-p)^{1-x_n}$$

対数尤度関数は

$$L(p) = \sum_{n=1}^{N} (x_n \log p + (1-x_n)\log(1-p))$$

これの最大値は，パラメータに対する微分係数がゼロの場所を求めればよい

$$0 = \left.\frac{\partial \log L(p)}{\partial p}\right|_{p_{ML}} = \frac{\sum_n x_n}{p_{ML}} - \frac{\sum_n (1-x_n)}{1-p_{ML}} = \frac{\sum_n x_n - Np_{ML}}{p_{ML}(1-p_{ML})}$$

$\rightarrow$ これを解けば最尤推定量 p_{ML} は点推定の時と一致することがわかる

2.4 機械学習の基礎

2.5 表現学習と深層学習の進展