Machine Learning tutorial

Takaya KOIZUMI

Mathematical Science, B4

Applied Mathematics and Physics informal seminor

Contents

Contents

機械学習とは

機械学習とは,「関数近似論」である.

世の中で機械学習を使って実現したと言われている技術

1. 翻訳 ({ 全ての日本語 }→{ 全ての英語 } という関数)
2. メール分類 ({ 全てのメールの文章 }→{ 迷惑メール, 非迷惑メール } という関数)
3. 音声認識 ({ 音声 }→{ 文章 } という関数)

もちろん, 間違いを起こすこともある. (大事なメールが, 迷惑メールに入ることも...)

数学的には

前スライドの話を集合論を用いて, もう少し数学的にきちんと書くならば, 以下のようになるだろう.

機械学習?

$\mathcal{X}, \mathcal{Y}$ をそれぞれ $\mathbb{R}^d, \mathbb{R}^m$ の部分集合とする. この時, 良い関数 $f : \mathcal{X} \to \mathcal{Y}$ を見つけることを機械学習という.

しかし, この定義には以下の問題がある.

上の定義の問題点

1. 候補となる関数が多すぎる. (ヒントも何もないのに探せない)
2. 良い関数とは何か, 定義されていない.

Contents

前半の問題解消

では，まず前半の「候補となる関数が多すぎる．」という問題を解決していこう．
この問題の解決方法として，人間がヒント (条件) を与えてあげることで，関数全ての集合ではなく，ある程度絞った集合 $\mathcal{H}$ にするということを考える．この $\mathcal{H}$ のことを仮設空間 (Hyposesis space) と呼ぶ．

Definition (仮設空間)

$\mathcal{X}, \mathcal{Y}$ をそれぞれ $\mathbb{R}^d, \mathbb{R}^m$ の部分集合とする．この時，集合

$$\mathcal{H} := \{ f_w : \mathcal{X} \to \mathcal{Y} \mid f_w \text{ に関する条件} \}$$

のことを仮設空間と呼び，$\mathcal{X}$ を特徴量空間，$\mathcal{Y}$ をラベル空間と呼ぶ．また，$f_w \in \mathcal{H}$ を仮設と呼ぶ．

後半の問題解消

では，後半の「良い関数」というものを定義していこう．機械学習において，良い関数とは，未知のデータ X に対して正しい値 Y を返す関数である．そのために，関数 f に対してその良さを表す指標である汎化誤差を定義する．

Definition (汎化誤差，損失関数)

$\mathcal{H}$ を仮設空間，$(\Omega, \mathcal{F}, \mathbb{P})$ を確率空間，ρ をデータの確率分布とする．この時，汎化誤差 $\ell : \mathcal{H} \to \mathbb{R}$ を，

$$\ell(f_\theta) = \mathbb{E}_{(X,Y)\sim\rho}[l(f_w(X), Y)]$$

と定義する．ここで，$l : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ は損失関数と呼ばれる凸関数である．

損失関数の具体例

損失関数

ここで，　よく使われる損失関数の例をいくつか述べておく．

1. 2 乗損失関数 $l(y_1, y_2) = (y_1 - y_2)^2$
2. 交差エントロピー誤差 $l(y_1, y_2) = -y_2 \log y_1$

これで，「良い関数」を作るためには，汎化誤差 ℓ を最小化させるような仮設空間 $\mathcal{H}$ の元 f を見つけば良いと言うことになったわけだが，汎化誤差には期待値が含まれるため，直接最適化させることが難しい．そのため，持っているデータを利用して別の関数を用意し，その関数を最小化することを考える．

データと経験損失関数

Definition (データ)

$(\Omega, \mathcal{F}, \mathbb{P})$ を確率空間, ρ をデータの確率分布とする.
$\{(X_n, Y_n)\}_{n=1}^N$ を ρ に従う独立な確率変数列とした時,
$\{(X_n, Y_n)\}_{n=1}^N$ の観測値 $\{(X_n(\omega), Y_n(\omega))\}_{n=1}^N$ のことをデータ *(Data)* と呼び, $D = \{(x_n, y_n)\}_{n=1}^N$ と表記する.

Definition (経験損失関数)

$\mathcal{H}$ を仮設空間, $D = \{(x_n, y_n)\}_{n=1}^N$ をデータ, $l : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ を損失関数とする. この時, 経験損失関数 $\mathcal{L}_D : \mathcal{H} \to \mathbb{R}$ を,

$$\mathcal{L}_D(f_w) = \sum_{n=1}^{N} l(f_\theta(x_n), y_n)$$

と定義する.

機械学習と学習アルゴリズム

さて，ここで改めて機械学習を定義しよう．

Definition (機械学習，学習アルゴリズム)

$\mathcal{H}$ を仮設空間，D をデータ，$\mathcal{L}_D$ を経験損失関数とする．この時，アルゴリズム $\mathcal{A}$ を用いて，$\mathcal{L}_D$ を最小化・最大化させる過程のことを機械学習 (あるいは単に学習) と呼び，その時のアルゴリズム $\mathcal{A}$ のことを学習アルゴリズムと呼ぶ．また，最適解 $f^* \in \mathcal{H}$ を最適仮設と呼び，その時のパラメータ w^* を最適パラメータと呼ぶ．

これ以降，5 つ組 $(\mathcal{X}, \mathcal{Y}, D, \mathcal{H}, \mathcal{L}_D)$ を ML 空間と呼ぶことにする．

Contents

最も基礎的なモデル

まず, 最も基礎的な機械学習モデルである単回帰分析を紹介する.

Example (単回帰分析)

$(\mathcal{X}, \mathcal{Y}, D, \mathcal{H}, \mathcal{L}_D)$ を以下で定義する. $\mathcal{X} = \mathbb{R}$, $\mathcal{Y} = \mathbb{R}$,

$$\mathcal{H} = \{f : \mathcal{X} \to \mathcal{Y} \mid f(x) = wx, w \in \mathbb{R}\},$$

$$\mathcal{L}_D(f) = \sum_{i=1}^{N} (f(x_i) - y_i)^2.$$

この ML 空間 $(\mathcal{X}, \mathcal{Y}, D, \mathcal{H}, \mathcal{L}_D)$ 上で,

$$\mathop{\arg\min}_{f \in \mathcal{H}} \mathcal{L}_D(f)$$

を求める問題を単回帰分析という.

SRA の学習

SRA は勾配降下法などを用いて解くこともできるが，今回は解析的に最適パラメータを求める．SRA の経験損失関数は w に関して 2 次関数となっているので，平方完成を用いると，

$$f^*(x) = w^* x, \qquad w^* = \frac{\sum_{i=1}^{N} x_i y_i}{\sum_{i=1}^{N} x_i^2}$$

と定義される $f^* \in \mathcal{H}$ が最適であることがわかる．

Contents

重回帰分析

Example (重回帰分析)

ML 空間 $(\mathcal{X}, \mathcal{Y}, D, \mathcal{H}, \mathcal{L}_D)$ を以下のように定義する.
$\mathcal{X} = \mathbb{R}^d (N \geq d)$, $\mathcal{Y} = \mathbb{R}$,

$$\mathcal{H} = \{f : \mathcal{X} \to \mathcal{Y} \mid f(x) = W^\top x, W \in \mathbb{R}^d, b \in \mathbb{R}\},$$

$$\mathcal{L}_D(f) = \sum_{i=1}^{N} (f(x_i) - y_i)^2.$$

この ML 空間 $(\mathcal{X}, \mathcal{Y}, D, \mathcal{H}, \mathcal{L}_D)$ 上で

$$\arg\min_{f \in \mathcal{H}} \mathcal{L}_D(f)$$

を求める問題を重回帰分析という.

MRA の学習

SRA の時と同様に解析的に最適パラメータを求める．$X \in \mathbb{R}^{N \times d}$, $\mathrm{y} \in \mathbb{R}^N$ を $X = [x_1 x_2 \dots x_N]^T$, $\mathrm{y} = (y_1, y_2, \cdots, y_N)^\top$ と定義する．
もし X が可逆であるとすると，

$$f^*(x) = W_*^\top x, \qquad W_* = (X^\top X)^{-1} X^\top \mathrm{y}$$

と定義される $f^* \in \mathcal{H}$ が最適であることがわかる．

Contents

非線形データへの対応