教師あり学習

中田和秀

東京科学大学 工学院 経営工学系

機械学習入門

https://www.nakatalab.iee.e.titech.ac.jp/text/nakata.html

概要

ここでは教師あり学習の一般的な枠組みや基本的な理論について説明をする。深層学習やサポートベクターマシンなどを含む多くの教師あり学習の手法はこの枠組で理解することができる。

目次：

記号の使い方：

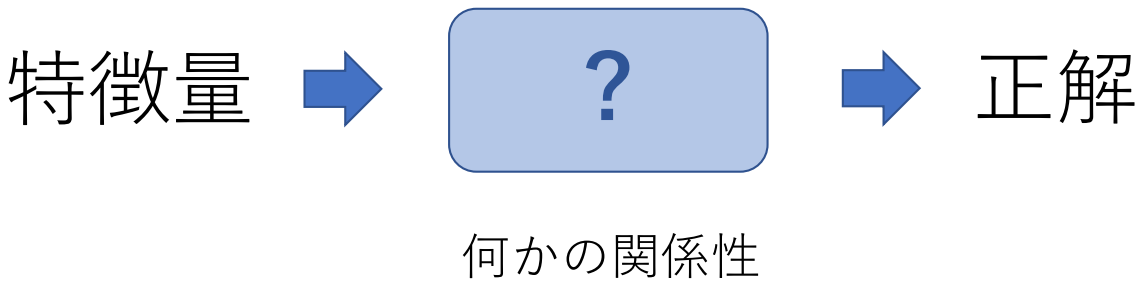
- $A := B$ は、B で A を定義する、B を A に代入することを意味する
- $[n]$ は n までのインデックスの集合を表し $[n] := \{1, 2, \cdots, n\}$

教師あり学習

「正解」がわかっているデータで学習を行い、未知のデータに対して「正解」を予測

手順

1. 多数の { 特徴量, 正解 } の組を使って学習
2. 正解が未知の特徴量に対して正解を予測



２つに大別できる

- 判別・分類：予測するものがカテゴリー
  例：購入の有無判別、故障の予測、スパムメール判断、手書き文字の認識
- 回帰：予測するものが数字
  例：需要の予測、将来の株価の予測、広告効果の推定、入院日数の予測

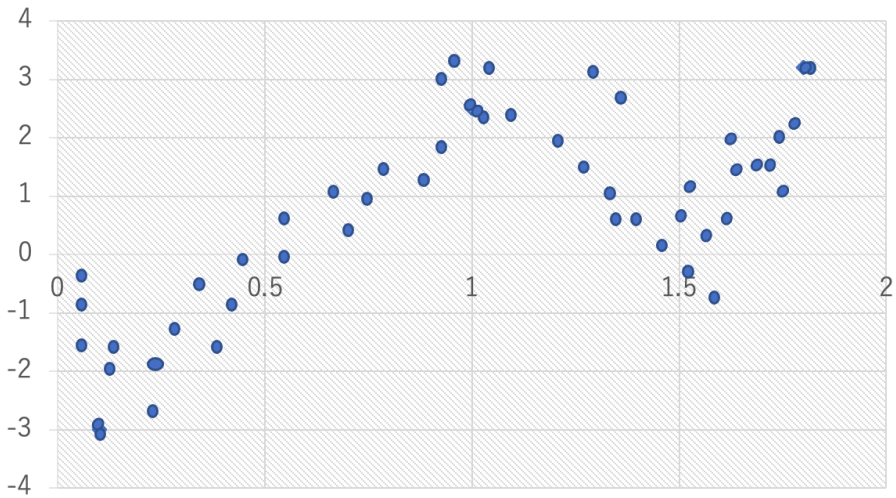
イメージ図

データ：$\{(\boldsymbol{x}_d, y_d)\}_{d\in[D]}$ $[D] := \{1, 2, \cdots, D\}$

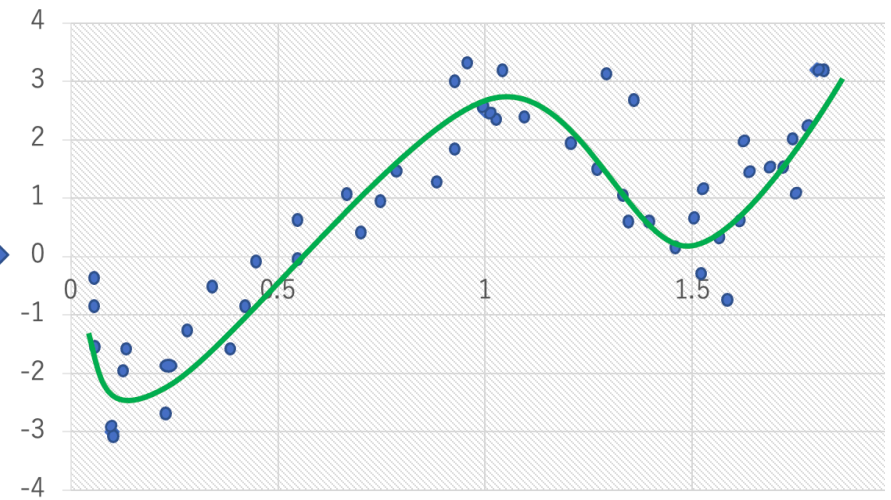
- $\boldsymbol{x}_d \in \mathbb{R}^n$：特徴ベクトル
- y_d：正解。判別のとき有限離散値、回帰のとき連続値。

回帰のイメージ

縦軸：$\boldsymbol{x}_d$，横軸：y_d



緑線（関数）を構築



関数の表現

$$\text{関数} \quad = \quad \text{構造} \quad + \quad \text{パラメータ}$$

- 全ての関数を許すと自由度が大きくなり過ぎる
- 「構造」は決めておき、「パラメータ」分の自由度を持った関数系を考える

関数の例：

パラメタは $\boldsymbol{a}, b, \alpha_d, \boldsymbol{A}_i, \boldsymbol{b}_i$

$$f(\boldsymbol{x}) := \boldsymbol{a}^T \boldsymbol{x} + b$$

$$f(\boldsymbol{x}) := \sum_{d \in [D]} y_d \alpha_d \exp\left\{\|\boldsymbol{x}_d - \boldsymbol{x}\|^2\right\} + y_j - \sum_{d \in [D]} y_d \alpha_d \exp\left\{\|\boldsymbol{x}_d - \boldsymbol{x}_j\|^2\right\}$$

$$f(\boldsymbol{x}) := f_4(\boldsymbol{A}_4 \ f_3(\boldsymbol{A}_3 \ f_2(\boldsymbol{A}_2 \ f_1(\boldsymbol{A}_1 \boldsymbol{x} + \boldsymbol{b}_1) + \boldsymbol{b}_2) + \boldsymbol{b}_3) + \boldsymbol{b}_4)$$

予測器

パラメタ $\boldsymbol{\theta}$ によって決まる関数

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

誤差関数

出来るだけ $y_d = f(\boldsymbol{x}_d; \boldsymbol{\theta})$ $(d \in [D])$ を満たすようなパラメータ $\boldsymbol{\theta}$ を採用したい

誤差関数

目標 y と予測 $\hat{y} := f(\boldsymbol{x}; \boldsymbol{\theta})$ とのズレを表す関数

$$L(y, \hat{y}) \geq 0$$

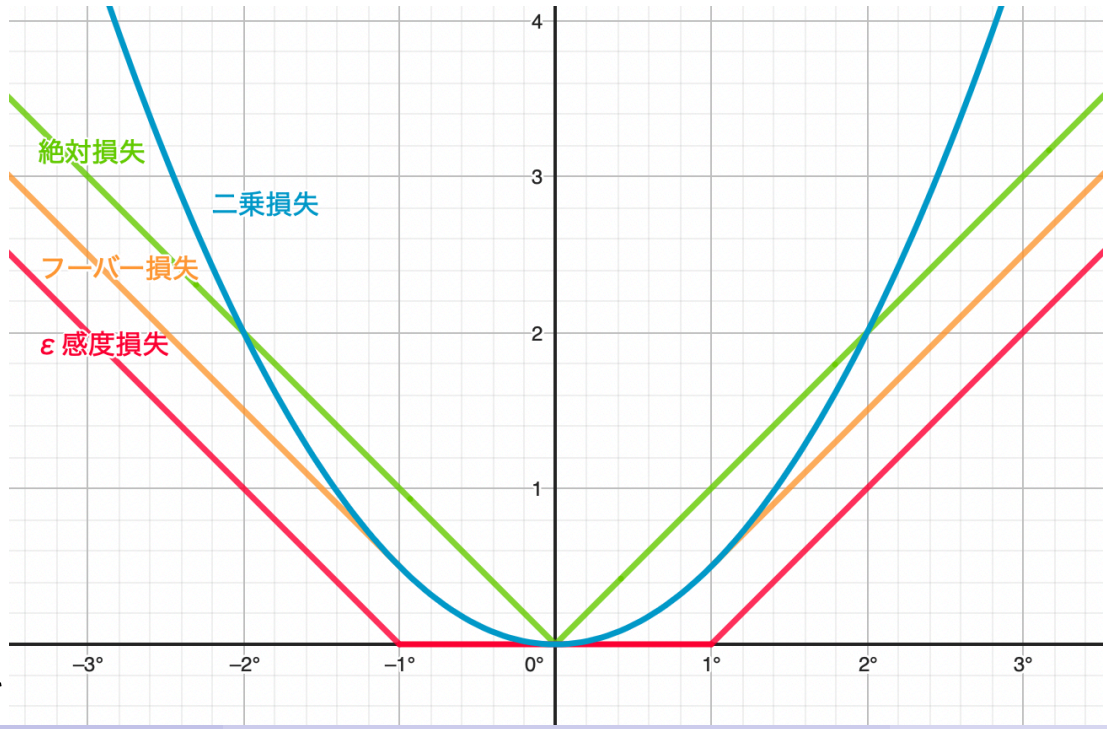
y と $\hat{y}$ が一致している時 0、離れるほど大きな値

誤差関数は損失関数とも呼ぶ

回帰の場合

- ２乗損失： $L(y, \hat{y}) := \frac{1}{2}(y - \hat{y})^2$
- 絶対損失： $L(y, \hat{y}) := |y - \hat{y}|$
- ϵ 感度損失： $L(y, \hat{y}) := \max\{|y - \hat{y}| - \epsilon, 0\}$
- フーバー損失： $L(y, \hat{y}) := \begin{cases} \frac{1}{2}(y - \hat{y})^2 & (|y - \hat{y}| \leq \alpha) \\ \alpha|y - \hat{y}| - \frac{1}{2}\alpha^2 & (|y - \hat{y}| > \alpha) \end{cases}$

$L(0, \hat{y})$ の値



$\epsilon = \alpha = 1$ のとき

判別の予測器

$\hat{y} = f(\boldsymbol{x}_d; \boldsymbol{\theta})$ を y の取る範囲と同じにすると、　　$\{-1, +1\}$, $\{0, 1\}$ など

$f(\boldsymbol{x}_d; \boldsymbol{\theta})$ が不連続になるため学習に都合が悪い（後述）

このため、$\hat{y} = f(\boldsymbol{x}_d; \boldsymbol{\theta})$ と最終的な予測を分ける

- $\hat{y} = f(\boldsymbol{x}_d; \boldsymbol{\theta}) \in \mathbb{R}$　$\cdots$ 実数値

  データを $y \in \{-1, +1\}$ とする

  最終的な予測：$\begin{cases} +1 & (\hat{y} \geq 0) \\ -1 & (\hat{y} \leq 0) \end{cases}$

- $\hat{y} = f(\boldsymbol{x}_d; \boldsymbol{\theta}) \in [0, 1]$　$\cdots$ 確率値

  データを $y \in \{0, 1\}$ とする

  最終的な予測：$\begin{cases} 1 & (\hat{y} \geq \alpha) \\ 0 & (\hat{y} \leq \alpha) \end{cases}$

※ 条件が等号で満たされたときはランダムに選ぶ

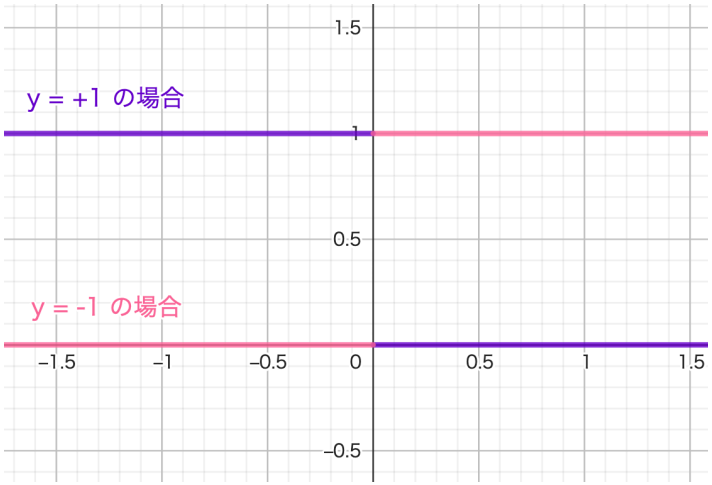
判別の損失関数

例えば、実数値を計算する場合を考える。

データ：$y \in \{-1, +1\}$, 予測器：$\hat{y} = f(\boldsymbol{x}_d; \boldsymbol{\theta}) \in \mathbb{R}$

最終的な予測：$\begin{cases} +1 & (\hat{y} \geq 0) \\ -1 & (\hat{y} \leq 0) \end{cases}$

横軸 $\hat{y}$、 縦軸 $L(y, \hat{y})$



損失関数として、次のものを考えるのが自然。

0-1 損失関数：

$$L(y, \hat{y}) := \begin{cases} 0 & (y\hat{y} > 0) \\ 1 & (y\hat{y} < 0) \end{cases}$$

不連続で非凸となるため学習に都合が悪い（後述）

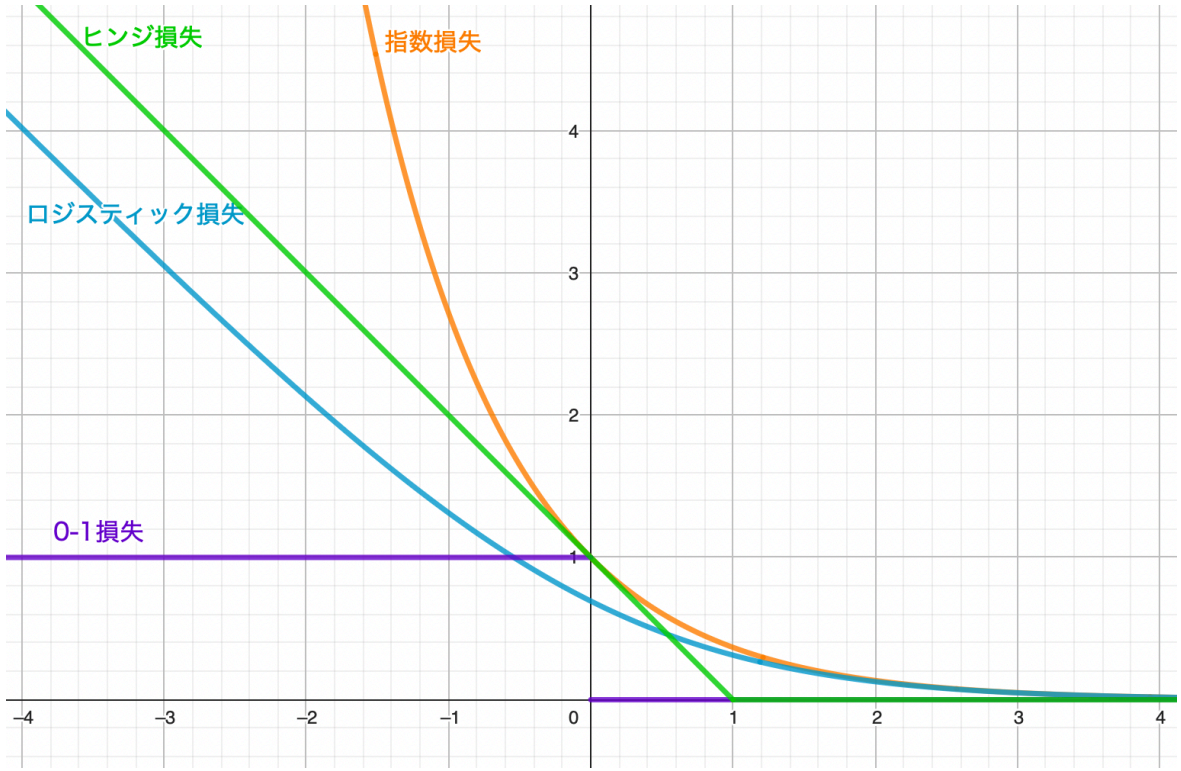
→ 0-1 損失関数の代わりに、連続で凸となる別の関数（サロゲート損失）を使う。

判別の損失関数（連続値）

$y \in \{-1, +1\},\ \hat{y} \in \mathbb{R}$ とする

- 指数損失：　$L(y, \hat{y}) := \exp(-y\hat{y})$
- ロジスティック損失：　$L(y, \hat{y}) := \log\{1 + \exp(-y\hat{y})\}$
- ヒンジ損失：　$L(y, \hat{y}) := \max\{1 - y\hat{y}, 0\}$

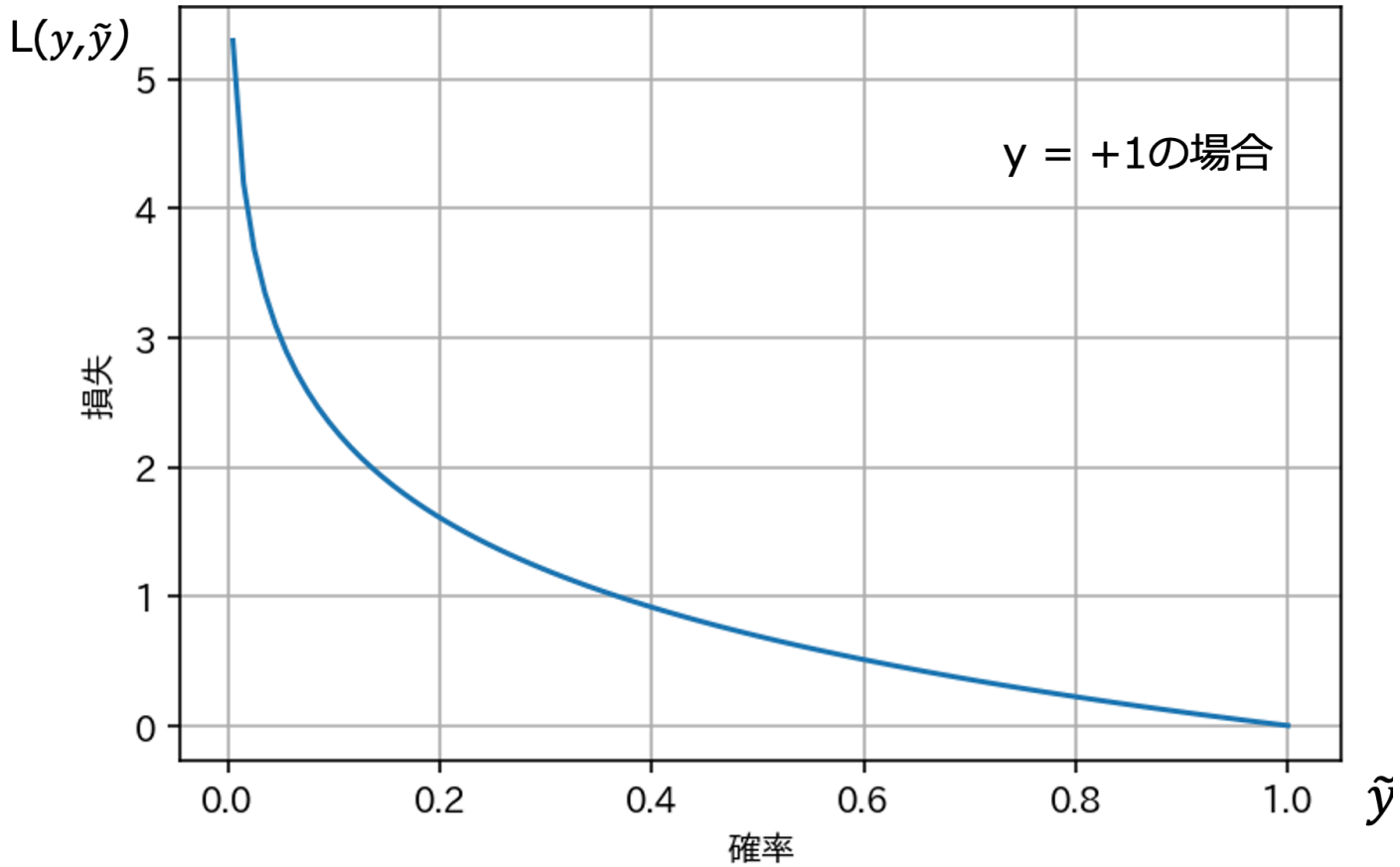
$y = +1$ の場合



判別の損失関数（確率値）

$y \in \{0, 1\},\ \hat{y} \in (0, 1)$ とする

- クロスエントロピー損失：　$L(y, \hat{y}) := -y \log \hat{y} - (1 - y) \log(1 - \hat{y})$



学習

- 誤差 $L(y_d, f(\boldsymbol{x}_d; \boldsymbol{\theta}))$ を小さくするパラメータ $\boldsymbol{\theta}$ を求める
- この計算過程を学習と呼ぶ

学習のフレームワーク

データ $\{(\boldsymbol{x}_d, y_d)\}_{d \in [D]}$ に対する誤差の平均を最小化するパラメータ $\boldsymbol{\theta}$ を計算

$$\min_{\boldsymbol{\theta}} \frac{1}{D} \sum_{d \in [D]} L(y_d, f(\boldsymbol{x}_d; \boldsymbol{\theta}))$$

- 最適解 $\boldsymbol{\theta}^*$ に対して、$f(\boldsymbol{x}; \boldsymbol{\theta}^*)$ が学習で得られた予測器
- 教師あり学習の手法は、予測器の表現 $f(\boldsymbol{x}; \boldsymbol{\theta})$ と誤差関数 $L(y, \hat{y})$ によって決まる

最適化

連続最適化問題

$$\min_{\boldsymbol{\theta}} \frac{1}{D} \sum_{d \in [D]} L(y_d, f(\boldsymbol{x}_d; \boldsymbol{\theta}))$$

- 機械学習における学習は、$\boldsymbol{\theta}$ を変数とする「連続最適化問題」を解くこと
- 最適化問題の性質やサイズによって、最適解 $\boldsymbol{\theta}^*$ を効率よく見つけるためのアルゴリズムは異なる

  性質： 線形性、凸性、微分可能性、疎性など

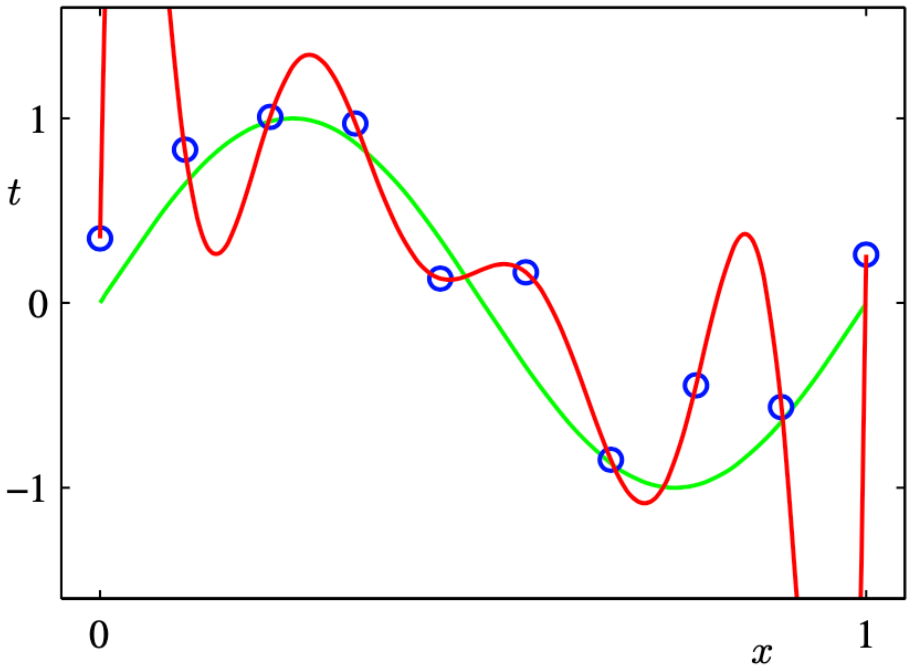
  サイズ： パラメタ数やデータ数など

最適化しやすいように、$f(\boldsymbol{x}; \boldsymbol{\theta})$ や $L(y, \hat{y})$ を設計することが大事

このため、判別のときは予測器 $f(\boldsymbol{x}; \boldsymbol{\theta})$ や損失関数 $L(y, \hat{y})$ をパラメタに対して連続・微分可能・凸などにする。

汎化能力と過学習

- 汎化能力
  未知のデータに対してうまく予測する能力
- 過学習
  学習に使ったデータに過剰に適応してしまい、汎化能力が低くなること



緑線：未知の真値
青丸：データ
赤線：予測値

Pattern Recognition and Machine Learning,
Bishop, 2006, Figure 1.4

過学習の検知法

汎化能力を直接知ることはできないため、次の手段で代替する

1. データを訓練データ $\mathcal{D}_1$ とテストデータ $\mathcal{D}_2$ に分ける。
2. 訓練データ $\mathcal{D}_1$ を使って学習を行い、最適パラメタ $\boldsymbol{\theta}^*$ と訓練データの平均誤差を得る。
$$L_{\text{train}} := \frac{1}{|\mathcal{D}_1|} \sum_{(\boldsymbol{x}_d, y_d) \in \mathcal{D}_1} L(y_d, f(\boldsymbol{x}_d; \boldsymbol{\theta}^*))$$
3. テストデータの平均誤差を計算する。
$$L_{\text{test}} := \frac{1}{|\mathcal{D}_2|} \sum_{(\boldsymbol{x}_d, y_d) \in \mathcal{D}_2} L(y_d, f(\boldsymbol{x}_d; \boldsymbol{\theta}^*))$$
4. - L_{test} を汎化能力の近似値とみなす
   - L_{train} は小さいが L_{test} が大きい場合は過学習

過学習の抑制

過学習を抑え汎化能力を上げることは、機械学習における重要な課題

過学習を抑制するために次の方法がある

(1) 訓練データの数を増やす

→ 実現可能ならば一番簡単な方法

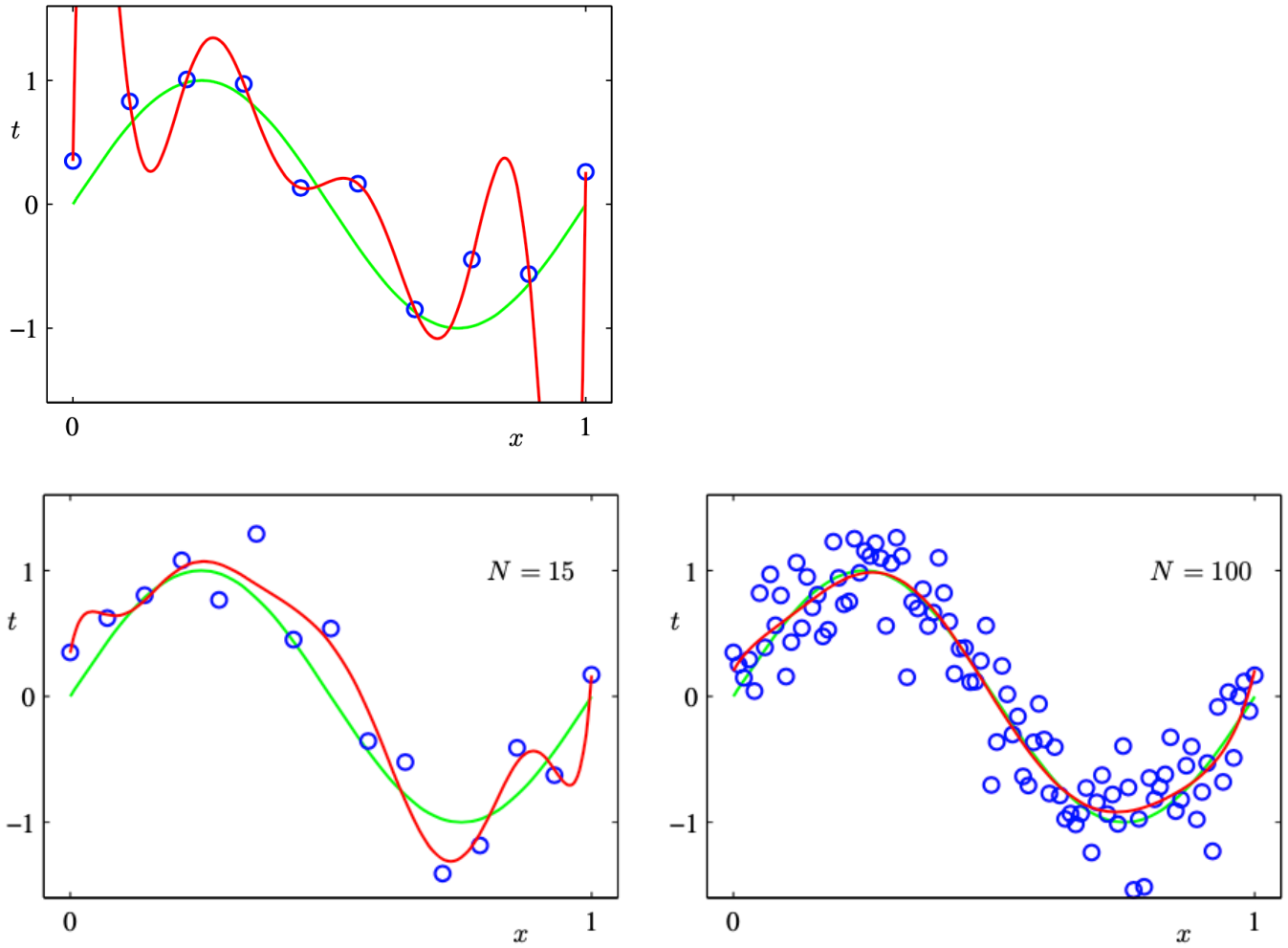
- 表現力（自由度）を抑える

(2) パラメタ数や関数形の制限

(3) 正則化項の導入

(1) 訓練データを増やす

データ数は順に 10,15,100。
データ数が増えるに連れて予測（赤線）は真値（緑線）に近くなる。



Pattern Recognition and Machine Learning, Bishop, 2006, Figure 1.6

(2) パラメタ数の制限

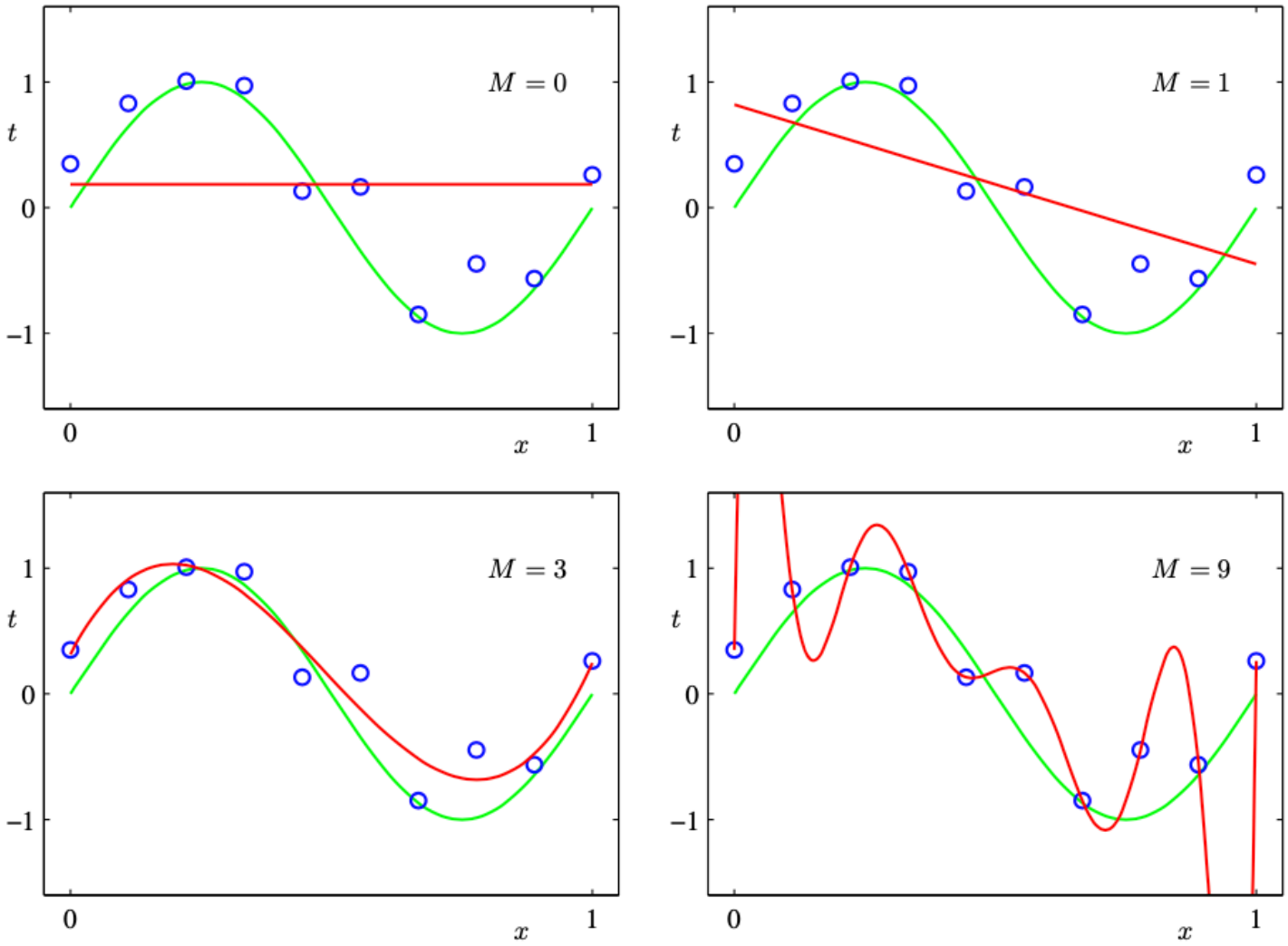
関数系の表現力

低い	$\Longleftarrow$	表現力	$\Longrightarrow$	高い
少ない		パラメタ数		多い
悪い		訓練データへのあてはまり		良い
しにくい		過学習		しやすい
可能		解釈		困難
容易		学習		難しい

予測モデルは「適度な」表現力が求められる

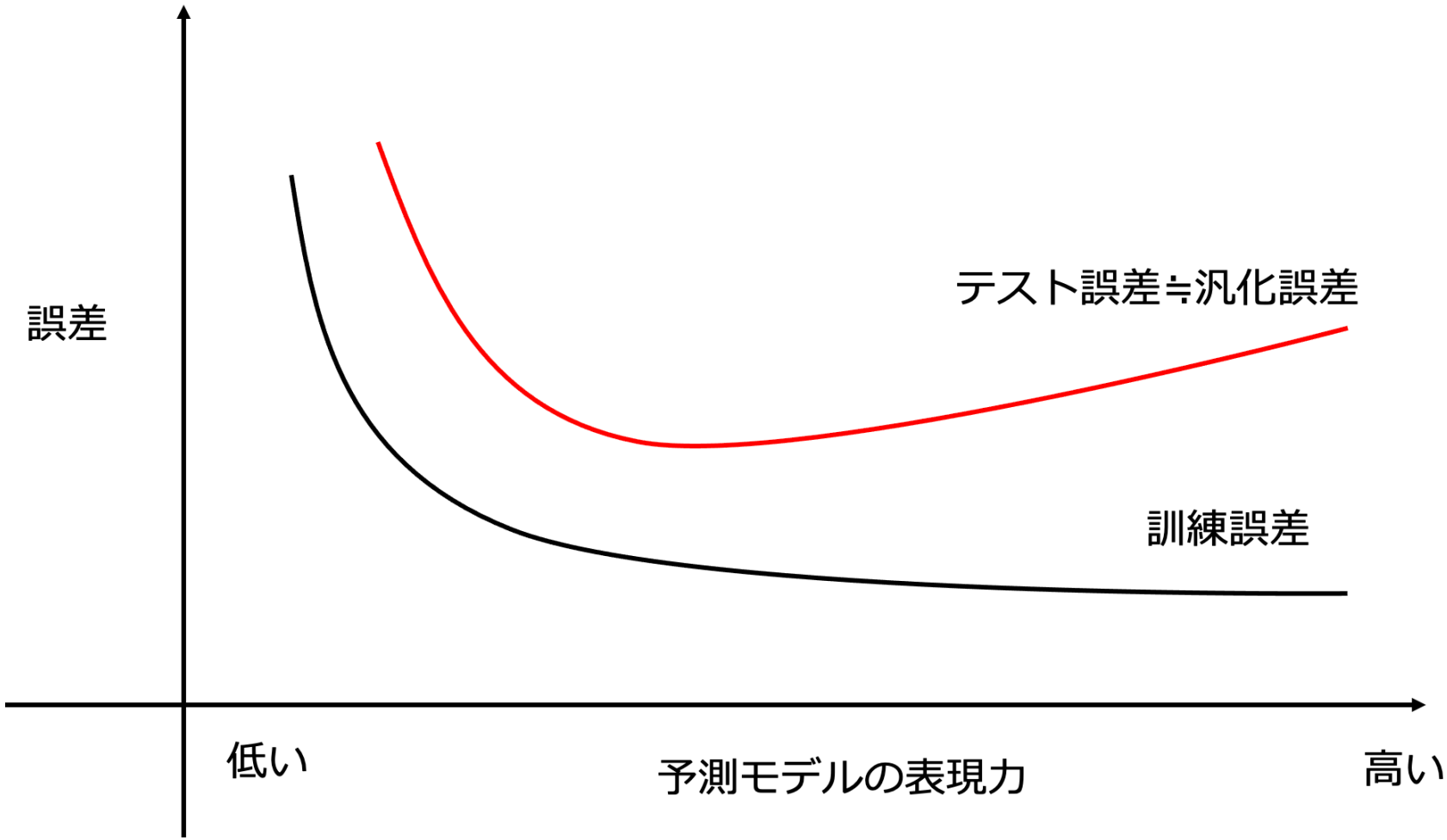
パラメタ数の変化

図中の $M+1$ がパラメタ数。パラメタ数は多くても少なくてもよくない。



Pattern Recognition and Machine Learning, Bishop, 2006, Figure 1.4

モデルの表現力と誤差



(3) 正則化項の導入

正則化項を加えることで、表現力に制限を加える

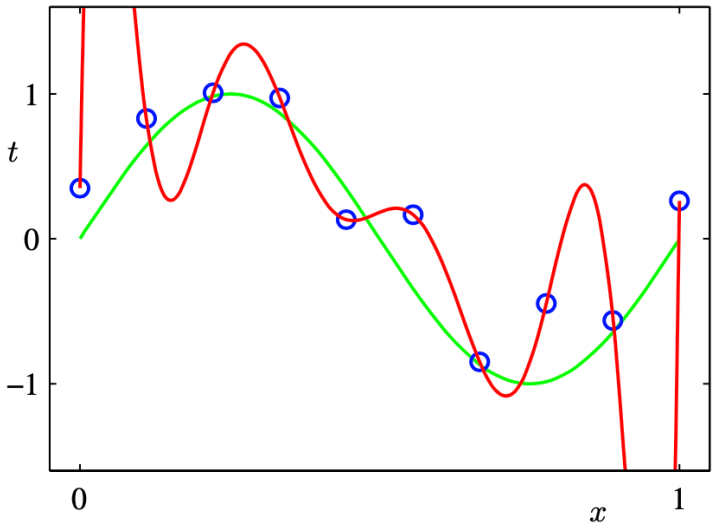
正則化項付き学習

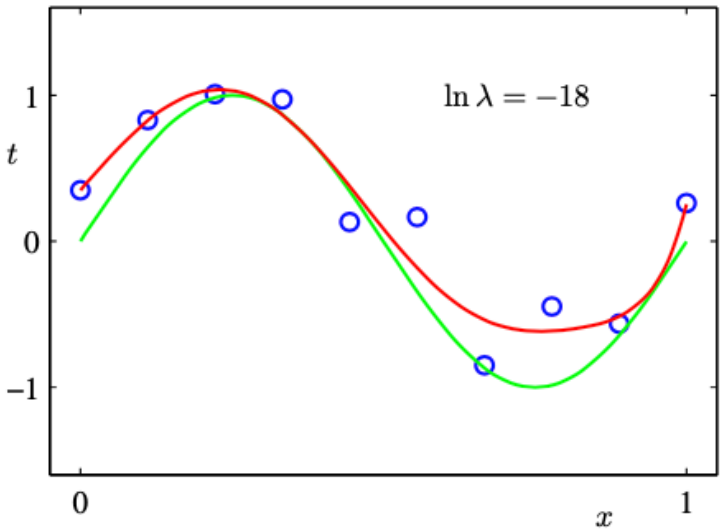
$$\min_{\boldsymbol{\theta}} \frac{1}{D} \sum_{d \in [D]} L(y_d, f(\boldsymbol{x}_d; \boldsymbol{\theta})) + \lambda \|\boldsymbol{\theta}\|^2$$

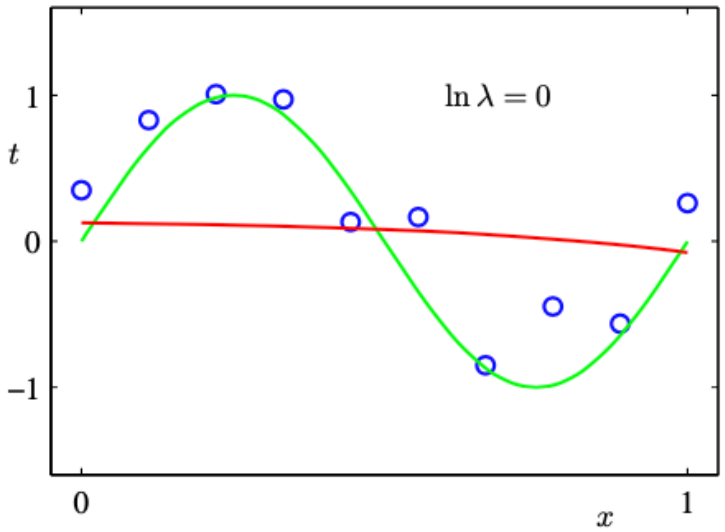
- $\lambda \|\boldsymbol{\theta}\|^2$ が正則化項
- $\lambda \geq 0$ が正則化の強さを決める。
  最適化を行う前に決めておくハイパーパラメータ。
- $\|\boldsymbol{\theta}\|^2$ 以外の正則化項を使うときもある

正則化の強さ

正則化項の重みが順に $\lambda = 0, e^{-18}, 1$. 値は大きくても小さくてもよくない。







Pattern Recognition and Machine Learning, Bishop, 2006, Figure 1.7

統計モデルに基づいた解析

予測器 $f(\boldsymbol{x})$ の性能について考える

汎化誤差

データ $(\boldsymbol{x}, y)$ はある確率分布に従っているとする。

$$\mathbb{E}[L(y, f(\boldsymbol{x}))] = \iint L(y, f(\boldsymbol{x}))p(\boldsymbol{x}, y)d\boldsymbol{x}dy$$

- 汎化誤差を知りたいが、データの生成確率 $p(\boldsymbol{x}, y)$ がわからない
- 代わりに、訓練データやテストデータ $\mathcal{D}$ を経験分布として利用する。

  経験分布：確率 $1/|\mathcal{D}|$ で $(\boldsymbol{x}_d, y_d) \in \mathcal{D}$ が選ばれる
- 経験分布に対する汎化誤差（経験誤差と呼ぶ）は

$$L_{\mathcal{D}} := \frac{1}{|\mathcal{D}|} \sum_{(\boldsymbol{x}_d, y_d) \in \mathcal{D}} L(y_d, f(\boldsymbol{x}_d))$$

- これは、既に述べた訓練データやテストデータの平均誤差である

統計モデルに基づいた解析

汎化誤差： $\mathbb{E}[L(y, f(\boldsymbol{x}))]$

経験誤差： $L_{\mathcal{D}} := \dfrac{1}{|\mathcal{D}|} \displaystyle\sum_{(\boldsymbol{x}_d, y_d) \in \mathcal{D}} L(y_d, f(\boldsymbol{x}_d))$

性質：

- $\mathbb{E}_{\mathcal{D}}[L_{\mathcal{D}}] = \mathbb{E}\left[L(y, f(\boldsymbol{x}))\right]$

  経験誤差は汎化誤差の不偏推定量

- $\mathbb{V}_{\mathcal{D}}[L_{\mathcal{D}}] = \dfrac{1}{|\mathcal{D}|} \mathbb{V}\left[L(y, f(\boldsymbol{x}))\right]$

  - 経験誤差の標準偏差はデータ数の平方根 $\sqrt{|\mathcal{D}|}$ に反比例
  - 一致推定量： $\lim_{D \to \infty} L_{\mathcal{D}} = \mathbb{E}\left[L(y, f(\boldsymbol{x}))\right]$

証明

$$\mathbb{E}_{\mathcal{D}}[L_{\mathcal{D}}] = \mathbb{E}_{\mathcal{D}}\left[\frac{1}{|\mathcal{D}|}\sum_{(\boldsymbol{x}_d, y_d)\in\mathcal{D}} L(y_d, f(\boldsymbol{x}_d))\right]$$

$$= \frac{1}{|\mathcal{D}|}\sum_{(\boldsymbol{x}_d, y_d)\in\mathcal{D}} \mathbb{E}_{\mathcal{D}}\left[L(y_d, f(\boldsymbol{x}_d))\right]$$

$$= \frac{1}{|\mathcal{D}|}\sum_{(\boldsymbol{x}_d, y_d)\in\mathcal{D}} \mathbb{E}\left[L(y, f(\boldsymbol{x}))\right] \quad = \mathbb{E}\left[L(y, f(\boldsymbol{x}))\right]$$

$$\mathbb{V}_{\mathcal{D}}[L_{\mathcal{D}}] = \mathbb{V}_{\mathcal{D}}\left[\frac{1}{|\mathcal{D}|}\sum_{(\boldsymbol{x}_d, y_d)\in\mathcal{D}} L(y_d, f(\boldsymbol{x}_d))\right]$$

$$= \frac{1}{|\mathcal{D}|^2}\sum_{(\boldsymbol{x}_d, y_d)\in\mathcal{D}} \mathbb{V}_{\mathcal{D}}\left[L(y_d, f(\boldsymbol{x}_d))\right] \qquad (\because \text{観測データは独立})$$

$$= \frac{1}{|\mathcal{D}|^2}\sum_{(\boldsymbol{x}_d, y_d)\in\mathcal{D}} \mathbb{V}\left[L(y, f(\boldsymbol{x}))\right] \quad = \frac{1}{|\mathcal{D}|}\mathbb{V}\left[L(y, f(\boldsymbol{x}))\right]$$

統計モデルに基づいた解析２

$\boldsymbol{x}$ を固定して考える。

$$\mathbb{E}[L(\boldsymbol{x})] := \int L(y, f(\boldsymbol{x}))p(y|\boldsymbol{x})dy$$

このとき、$\mathbb{E}[L(y, f(\boldsymbol{x}))] = \int \mathbb{E}[L(\boldsymbol{x})]p(\boldsymbol{x})d\boldsymbol{x}$

前提：

- データの確率分布 $p(\boldsymbol{x}, y), p(y|\boldsymbol{x})$ などは分かっているとする
- 誤差関数として２乗誤差を考える

性質：

$$\begin{aligned}\mathbb{E}[L(\boldsymbol{x})] &= \mathbb{E}[\{y - f(\boldsymbol{x})\}^2 |\boldsymbol{x}] \\ &= \{f(\boldsymbol{x}) - \mathbb{E}[y|\boldsymbol{x}]\}^2 + \mathbb{V}[y|\boldsymbol{x}]\end{aligned}$$

データの分布が既知の場合は、$f(\boldsymbol{x}) = \mathbb{E}[y|\boldsymbol{x}]$ という予測器が最良
（２乗誤差を最小化するという意味で）

証明

$$
\begin{aligned}
\mathbb{E}[L(\boldsymbol{x})] =& \mathbb{E}[\{y - f(\boldsymbol{x})\}^2 |\boldsymbol{x}] \\
=& \mathbb{E}[\{f(\boldsymbol{x}) - \mathbb{E}[y|\boldsymbol{x}] + \mathbb{E}[y|\boldsymbol{x}] - y\}^2 |\boldsymbol{x}] \\
=& \mathbb{E}[\{f(\boldsymbol{x}) - \mathbb{E}[y|\boldsymbol{x}]\}^2 |\boldsymbol{x}] \\
&+ \mathbb{E}[\{f(\boldsymbol{x}) - \mathbb{E}[y|\boldsymbol{x}]\} \{\mathbb{E}[y|\boldsymbol{x}] - y\} |\boldsymbol{x}] \\
&+ \mathbb{E}[\{\mathbb{E}[y|\boldsymbol{x}] - y\}^2 |\boldsymbol{x}] \\
=& \{f(\boldsymbol{x}) - \mathbb{E}[y|\boldsymbol{x}]\}^2 \\
&+ \{f(\boldsymbol{x}) - \mathbb{E}[y|\boldsymbol{x}]\} \{\mathbb{E}[y|\boldsymbol{x}] - \mathbb{E}[y|\boldsymbol{x}]\} \\
&+ \mathbb{V}[y|\boldsymbol{x}] \\
=& \{f(\boldsymbol{x}) - \mathbb{E}[y|\boldsymbol{x}]\}^2 + \mathbb{V}[y|\boldsymbol{x}]
\end{aligned}
$$

統計モデルに基づいた解析３

機械学習法自体の汎化性能について考える。

- ある機械学習法において、訓練データ $\mathcal{D}$ が与えられれば、学習によって予測器 $f(\boldsymbol{x})$ が定まる。
- この予測器 $f(\boldsymbol{x})$ は訓練データ $\mathcal{D}$ に依存する。これを $f(\boldsymbol{x};\mathcal{D})$ と書く。
- $f(\boldsymbol{x};\mathcal{D})$ と最良の予測器 $\mathbb{E}[y|\boldsymbol{x}]$ との平均２乗誤差について考える。

$$\begin{aligned}&\mathbb{E}_{\mathcal{D}}[\{f(\boldsymbol{x};\mathcal{D})-\mathbb{E}[y|\boldsymbol{x}]\}^2]\\&=\{\mathbb{E}_{\mathcal{D}}[f(\boldsymbol{x};\mathcal{D})]-\mathbb{E}[y|\boldsymbol{x}]\}^2+\mathbb{V}_{\mathcal{D}}[f(\boldsymbol{x};\mathcal{D})]\end{aligned}$$

バイアス・バリアンス分解

最良の予測器との平均２乗誤差＝バイアスの２乗＋バリアンス

証明

$$
\begin{aligned}
&\mathbb{E}_{\mathcal{D}}[\{f(\boldsymbol{x};\mathcal{D}) - \mathbb{E}[y|\boldsymbol{x}]\}^2] \\
&= \mathbb{E}_{\mathcal{D}}[\{\ (f(\boldsymbol{x};\mathcal{D}) - \mathbb{E}_{\mathcal{D}}[f(\boldsymbol{x};\mathcal{D})])\ +\ (\mathbb{E}_{\mathcal{D}}[f(\boldsymbol{x};\mathcal{D})] - \mathbb{E}[y|\boldsymbol{x}])\ \}^2]
\end{aligned}
$$

外側の $\mathbb{E}_{\mathcal{D}}$ は、$f(\boldsymbol{x};\mathcal{D})$ にのみ影響を与えているため

$$
\begin{aligned}
&= \mathbb{E}_{\mathcal{D}}[\{f(\boldsymbol{x};\mathcal{D}) - \mathbb{E}_{\mathcal{D}}[f(\boldsymbol{x};\mathcal{D})]\}^2] \\
&\quad + 2\mathbb{E}_{\mathcal{D}}[f(\boldsymbol{x};\mathcal{D}) - \mathbb{E}_{\mathcal{D}}[f(\boldsymbol{x};\mathcal{D})]] \times (\mathbb{E}_{\mathcal{D}}[f(\boldsymbol{x};\mathcal{D})] - \mathbb{E}[y|\boldsymbol{x}]) \\
&\quad + \{\mathbb{E}_{\mathcal{D}}[f(\boldsymbol{x};\mathcal{D})] - \mathbb{E}[y|\boldsymbol{x}]\}^2
\end{aligned}
$$

第一項　$\mathbb{E}_{\mathcal{D}}[\{f(\boldsymbol{x};\mathcal{D}) - \mathbb{E}_{\mathcal{D}}[f(\boldsymbol{x};\mathcal{D})]\}^2] = \mathbb{V}_{\mathcal{D}}[f(\boldsymbol{x};\mathcal{D})]$

第二項

$$
\begin{aligned}
&2\mathbb{E}_{\mathcal{D}}[f(\boldsymbol{x};\mathcal{D}) - \mathbb{E}_{\mathcal{D}}[f(\boldsymbol{x};\mathcal{D})]] \times (\mathbb{E}_{\mathcal{D}}[f(\boldsymbol{x};\mathcal{D})] - \mathbb{E}[y|\boldsymbol{x}]) \\
&= 2\left\{\mathbb{E}_{\mathcal{D}}[f(\boldsymbol{x};\mathcal{D})] - \mathbb{E}_{\mathcal{D}}[f(\boldsymbol{x};\mathcal{D})]\right\}\ \ (\mathbb{E}_{\mathcal{D}}[f(\boldsymbol{x};\mathcal{D})] - \mathbb{E}[y|\boldsymbol{x}]) \\
&= 0
\end{aligned}
$$

統計モデルに基づいた解析４

統計モデルに基づいた解析２と３をまとめると

$$\begin{aligned}\mathbb{E}[\mathbb{E}_{\mathcal{D}}[\{f(\boldsymbol{x};\mathcal{D})-y\}^2]] = &\int \{\mathbb{E}_{\mathcal{D}}[f(\boldsymbol{x};\mathcal{D})]-\mathbb{E}[y|\boldsymbol{x}]\}^2 p(\boldsymbol{x})d\boldsymbol{x}\\ &+\int \mathbb{V}_{\mathcal{D}}[f(\boldsymbol{x};\mathcal{D})]p(\boldsymbol{x})d\boldsymbol{x}\\ &+\int \mathbb{V}[y|\boldsymbol{x}]p(\boldsymbol{x})d\boldsymbol{x}\end{aligned}$$

機械学習法の期待損失＝バイアスの２乗＋バリアンス＋ノイズ

人が理解しやすい評価指標

平均誤差（損失）は学習するのに便利な指標

- 「よさ」「悪さ」を一つの実数値の大小関係で表現
- パラメータに対して微分可能

予測性能を人が解釈するには、別の指標を用いる。

- 判別・分類
  混同行列、F 値、AUC など
- 回帰
  平均 2 乗平方根誤差 (RMSE)、平均絶対誤差 (MAE) など

判別の評価指標

混同行列

	予測 偽	予測 真
正解 偽	TN(True Negative) 真陰性	FP(False Positive) 偽陽性
正解 真	FN(False Negative) 偽陰性	TP(Ture Positive) 真陽性

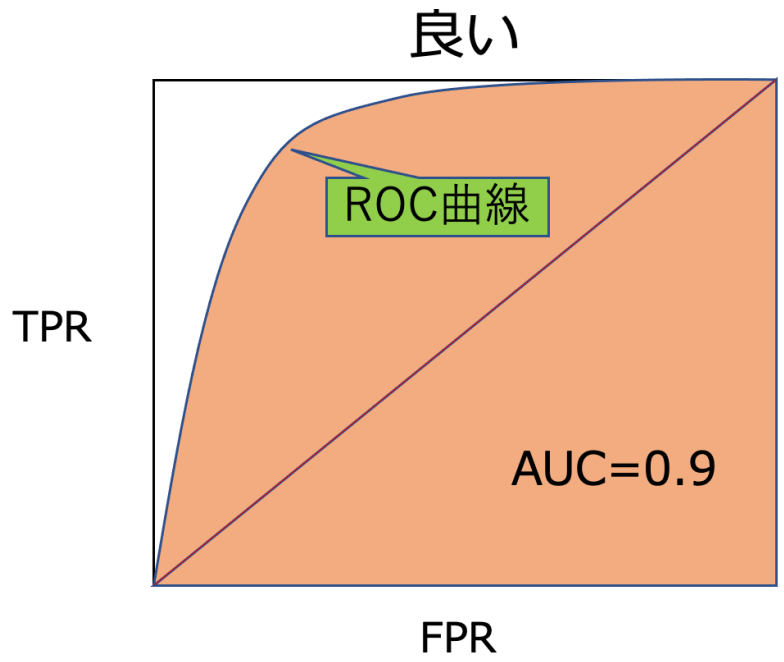
評価指標

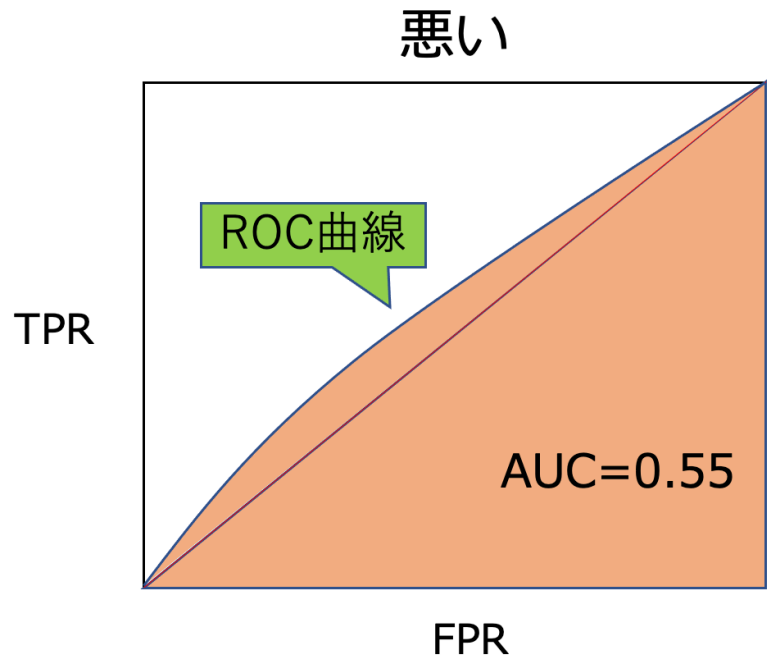
- 正解率 (Accuracy)：$\dfrac{TP+TN}{TP+TN+FP+FN}$
- 適合率 (Precision)：$\dfrac{TP}{TP+FP}$
- 再現率 (Recall)：$\dfrac{TP}{TP+FN}$
- F 値 (F-score)：適合率と再現率の調和平均　　$2 \times \dfrac{\text{適合率} \times \text{再現率}}{\text{適合率} + \text{再現率}}$

※ 適合率と再現率はトレードオフの関係にあるため、その下に偏った平均である調和平均を考える
※ 不均衡データでは、正解率が指標にならないため、F 値などを使う

AUC

- 閾値を変えることで、TPR=$\frac{TP}{TP+FN}$、FPR=$\frac{FP}{TN+FP}$ が変化する
- ROC 曲線 [1]：縦軸に TPR、横軸に FPR を取った曲線
- AUC：ROC 曲線の右下の面積





AUC は 0 以上 1 以下で大きな方がよい。ランダムに予測すると AUC は 0.5。

[1] 適合率と再現率を軸に取る PR 曲線を使う場合もある

AUC の解釈 1

正例と負例をランダムに選択したとき、AUC は次の値と一致する。

「正例の予測値 $>$ 負例の予測値」となる確率
$+ 0.5\times$「正例の予測値 $=$ 負例の予測値」となる確率

予測値が全て異なる値を取るときを説明：

- 正例の予測値：$a_1, a_2, \cdots, a_S$
- 負例の予測値：$b_1, b_2, \cdots, b_T$

※ 昇順に並んでいるとする。

$$\mathrm{TPR} = \frac{TP}{TP + FN} \in \left\{0, \frac{1}{S}, \frac{2}{S}, \ldots, \frac{S-1}{S}, 1\right\}$$
$$\mathrm{FPR} = \frac{FP}{TN + FP} \in \left\{0, \frac{1}{T}, \frac{2}{T}, \ldots, \frac{T-1}{T}, 1\right\}$$

ROC 曲線は横軸を T 等分、縦軸を S 等分した格子上を動く。

AUC の解釈 2

横軸が $\frac{T-t}{T}$ と $\frac{T-t+1}{T}$ の間を通る ROC 曲線

- 閾値は $b_t + \epsilon$ から $b_t - \epsilon$
- ROC 曲線の高さは $\frac{1}{S}\sum_{s\in[S]} I(a_s > b_t)$

よって

$$\begin{aligned}
\text{AUC} :=& \sum_{t\in[T]} \frac{1}{T} \left\{ \frac{1}{S} \sum_{s\in[S]} I(a_s > b_t) \right\} \\
=& \frac{1}{ST} \sum_{(s,t)\in[S]\times[T]} I(a_s > b_t) \\
=& P[a > b]
\end{aligned}$$

回帰の評価指標

$\mathcal{D}$：テストデータ

- 平均二乗平方根誤差 (Root Mean Squared Error: RMSE)

$$\sqrt{\frac{1}{|\mathcal{D}|}\sum_{(\boldsymbol{x}_d,y_d)\in\mathcal{D}}(y_d - f(\boldsymbol{x}_d;\boldsymbol{\theta}^*))^2}$$

- 平均絶対誤差 (Mean Absolute Error: MAE)

$$\frac{1}{|\mathcal{D}|}\sum_{(\boldsymbol{x}_d,y_d)\in\mathcal{D}}|y_d - f(\boldsymbol{x}_d;\boldsymbol{\theta}^*)|$$