# PRML Chapter 1

2021 年 12 月 17 日

## 目次



## 1 Introduction

Pattern recognition とは

- algorithm を用いてデータから自動的に法則を見つけること
- その法則を用いて (データを異なるカテゴリに分類するといったような) 処理を行うこと

**例.** Recognizing Handwritten digits ($0-9$ の手書き数字の分類)

人手による rule や heuristic を用いた分類は、法則や例外が容易に発散してしまいうまくいかない。このように、大量のデータ $\{\boldsymbol{x_1}, \cdots, \boldsymbol{x}_N\}$ (training set) を正解カテゴリ $\boldsymbol{t}$ (target vector) を元にしてモデルの tuning を行うような場合、機械学習的アプローチが有効である。データ $\boldsymbol{x}$ からのモデルの出力 $\boldsymbol{y}(\boldsymbol{x})$ の精度は training (learning) phase にされるが、training set はデータ全体のごく一部でしかないことから、未知のデータを正しく分類できるかというモデルの generalization が常に重要になる。

また、実用上は以下の理由からデータを preprocess することが多い (feature extraction とも呼ばれる)。

- 解くべき問題をより簡単にする (データの variability を減らす)
- 計算をより早くする (問題を解くために必要な情報を取捨選択する)

この際、必要な情報まで落としてしまわないように注意が必要である。

**Figure 1.2** Plot of a training data set of $N = 10$ points, shown as blue circles, each comprising an observation of the input variable $x$ along with the corresponding target variable $t$. The green curve shows the function $\sin(2\pi x)$ used to generate the data. Our goal is to predict the value of $t$ for some new value of $x$, without knowledge of the green curve.

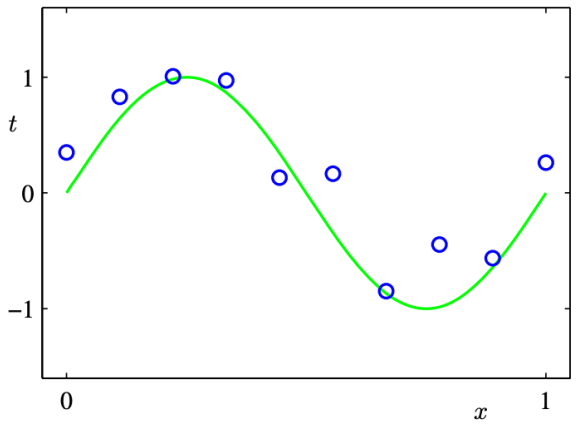


図 1.1　$[0, 1]$ から random に $x$ を 10 箇所抽出し、$\sin 2\pi x$ を計算した後に small Gaussian noise を加えた結果。

以上の話は supervised learning の例であるが、pattern recognition の問題設定には以下のようものがある。

1. Supervised learning
   Input vector を対応する target vector に分類する。出力が有限離散カテゴリなら classification、連続値なら regression と呼ばれる。
2. Unsupervised learning
   Input vector のみから成るデータを用いる。タスクとしては、似たものを grouping する clustering、input space でのデータの分布を見る density estimation、高次元データを低次元に落とし込む visualization などがある。
3. Reinforcement learning
   与えられた状況で reward を最大化する行動を見つけるよう学習する。ただし、reward は一連の行動の結果として得られるため、どの行動がどの程度結果に寄与したのかを正しく割り当てることは難しい。このような問題は credit assignment problem と呼ばれる。一般に reinforcement learning は high reward を求めて新たなことに挑戦する exploration と high reward であることがわかっている行動を取る exploitation のバランスで成り立っている。

本章前半ではこれら手法の根底にある重要な考え方を具体例を交えながら簡単に説明する。また、後半では今後必要な probability theory, decision theory, information theory の基礎を簡潔に導入する。

## 1.1 Example: Polynomial Curve Fitting

例として、実データ $x$ が与えられたときに real-valued target $t$ を予想するような simple regression problem を考える。ここでは簡単のため $\sin 2\pi x$ を考え、training set を $\boldsymbol{x} \equiv (x_1, \cdots, x_N)^T$、対応する観測値を $\boldsymbol{t} \equiv (t_1, \cdots, t_N)^T$ と表すことにする。また、ここでは観測値 $\boldsymbol{t}$ は $\sin 2\pi x$ に small Gaussian noise を加えたものとする[*1]。

我々のゴールは新たな input $\hat{x}$ の target $\hat{t}$ を予測することである。Uncertainty の定量的な評価などは後にして、まずは直感的な方法を試すことにする。簡単に思いつくのは多項式

$$y(x, \boldsymbol{w}) = \sum_{j=0}^{M} w_j x^j, \tag{1.1}$$

[*1] 現実の観測値は、放射性崩壊のように process そのものが stochastic である場合もあるが、大体の場合はその他の何らかの要因が noise となって本質的な法則の測定を難しくしている。

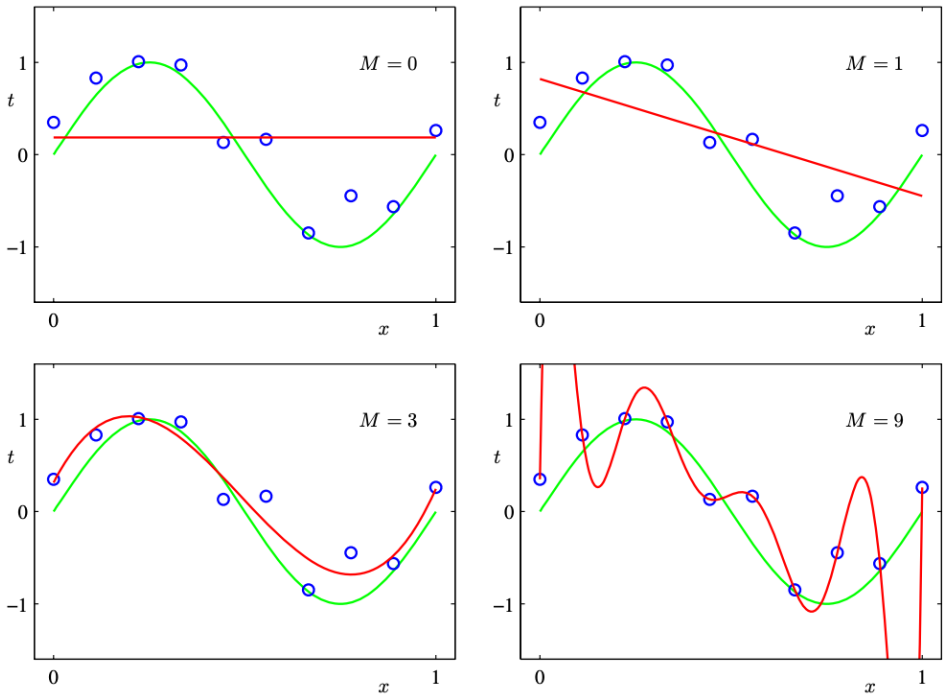


**Figure 1.4** Plots of polynomials having various orders $M$, shown as red curves, fitted to the data set shown in Figure 1.2.

を考え[*2]、二乗誤差

$$E(\boldsymbol{w}) = \frac{1}{2}\sum_{n=1}^{N}\left[y(x_n, \boldsymbol{w}) - t_n\right]^2, \tag{1.2}$$

を最小化する方法である。$E(\boldsymbol{w})$ は $\boldsymbol{w}$ の quadratic function のため、微分すると $\boldsymbol{w}$ の一次関数となり unique な closed form[*3] な解 $\boldsymbol{w} = \boldsymbol{w}^*$ が求まる。

一方、この条件からは決めらないモデルの自由度 $M$ を定める問題を model comparison または model selection と呼ぶ。Figure 1.4 より、$M$ が小さすぎるとほとんど予測ができていない一方、大きすぎても $E(\boldsymbol{w}^*) = 0$ ではあるものの $\sin 2\pi x$ とは程遠い形となってしまうことがわかる (over-fitting)。

$M$ と新たなデータに対する予測の精度の関係を確かめるため、training set と同様の手法で新たに 100 個の test set を生成し、$E(\boldsymbol{w}^*)$ の root-mean-square

$$E_{\mathrm{RMS}} = \sqrt{2E(\boldsymbol{w}^*)/N}, \tag{1.3}$$

を計算してみる (Figure 1.5)。すると $3 \leq M \leq 8$ ではどちらも誤差が減少していく一方、$M = 9$ になると自由度が残っていないことから $E(\boldsymbol{w}^*) = 0$ になるが、test set の誤差は大きくなっていることがわかる。

$M$ を大きくしていくと誤差が大きくなっていくのは、高次元のモデルは低次元のモデルも含んでいること、そもそも $\sin 2\pi x$ は全ての次元の項を含んでいることなどを考えると paradoxical にも思える。しかし、$\boldsymbol{w}$ の値を具体的に見てみると $M$ が大きくなるにつれ $\boldsymbol{w}$ は異常に大きくなっていることがわかる (Table 1.1)。つまり、データの微小な違い (noise) に過剰に fitting してしまっているからであると言える。

---

[*2] このモデルは $\boldsymbol{w}$ に関して linear であることから linear model と総称される。

[*3] Closed form に明確な定義はないらしいため「よく知られた形で得られる数式」程度の理解で良さそう。(Wikipedia - Closed-form expression)

**Figure 1.5** Graphs of the root-mean-square error, defined by (1.3), evaluated on the training set and on an independent test set for various values of $M$.

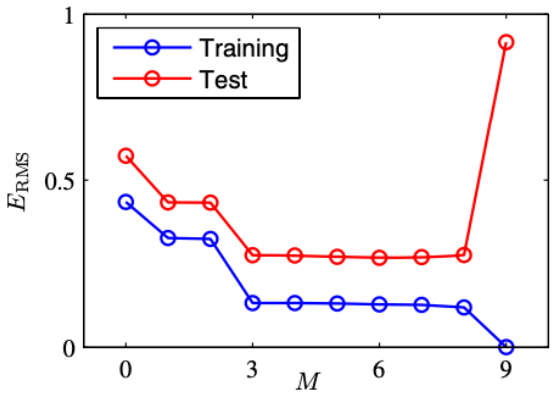


**Table 1.1** Table of the coefficients $\mathbf{w}^\star$ for polynomials of various order. Observe how the typical magnitude of the coefficients increases dramatically as the order of the polynomial increases.

| | $M=0$ | $M=1$ | $M=6$ | $M=9$ |
|---|---|---|---|---|
| $w_0^\star$ | 0.19 | 0.82 | 0.31 | 0.35 |
| $w_1^\star$ | | -1.27 | 7.99 | 232.37 |
| $w_2^\star$ | | | -25.43 | -5321.83 |
| $w_3^\star$ | | | 17.37 | 48568.31 |
| $w_4^\star$ | | | | -231639.30 |
| $w_5^\star$ | | | | 640042.26 |
| $w_6^\star$ | | | | -1061800.52 |
| $w_7^\star$ | | | | 1042400.18 |
| $w_8^\star$ | | | | -557682.99 |
| $w_9^\star$ | | | | 125201.43 |

一方、training set のデータ数を増やした場合、モデルを複雑にしても over-fitting の問題はより起きづらくなる。これはデータが増えて fit できる自由度が増えることによるもので、heurisitic にはモデルの parameter の数はデータ数の数分の一程度に留めておくと良さそうである。しかし、モデルの複雑さは問題設定の複雑さに対応すべきでデータ数に起因すべきではないような気もする。この over-fitting の問題は (最小二乗法を含む) maximum likelihood では一般的な問題であり、このような問題は Bayesian approach では生じないことを後に確認する。

モデルの複雑性と柔軟性を両立する有名な方法は、error function に係数の増大を阻害する penalty 項を導入する regularization と呼ばれる手法である (neural network の文脈では weight decay とも呼ばれる) 。例えば、

$$\tilde{E}(\boldsymbol{w}) = \frac{1}{2}\sum_{n=1}^{N}[y(x_n,\boldsymbol{w}) - t_n]^2 + \frac{\lambda}{2}||\boldsymbol{w}||^2, \tag{1.4}$$

のように重みの二乗を加えておくことで error function の最小化に対する penalty を課すことができ、ridge regression と呼ばれる。ただしこの際、目的変数の「原点」に依存する $\boldsymbol{w}_0$ は penalty に含めないか、もしくはそれに対応した正則化係数をかけることが多い。このように推定にあまり関係ないパラメータの影響を小さくする手法を shrinkage と呼ぶ。

実際に正則化係数 $\lambda$ を適当に設定して再度 $\boldsymbol{w}^*$ を求めると、係数の増大は抑えられ、同時に RMS 誤差も小さくなることがわかる (Table 1.2, Figure 1.8)。このように、単純に誤差を最小化して問題を解く場合、モデルの複雑さは別の枠組みから定めなければならない。そのために、訓練データを事前に validation set として分けておく手法があるが、貴重な訓練データを無駄にしないより洗練されたアプローチを 1.3 でみる。

ここまでで問題解決の発見的アプローチは終わりにして、以降は確率論的な枠組みを用いてより原理的なアプローチを考えていく。

## 1.2 Probability Theory

確率変数 $X, Y$ に対して[*4]、$X$ かつ $Y$ の確率 (同時確率: joint probability) を $p(X, Y)$、$X$ の下での $Y$ の確率 (条件付き確率: conditional probability) を $p(Y|X)$ とすると以下の原則が成り立つ。

The Rules of Probability

$$\text{sum rule} \quad p(X) = \sum_Y p(X, Y), \tag{1.5}$$

$$\text{product rule} \quad p(X, Y) = p(X|Y)p(X). \tag{1.6}$$

ここで、$p(X)$ は $Y$ で和を取っていることから周辺確率 (marginal probability) と呼ばれることもある。

これと $p(X, Y) = p(Y, X)$ から直ちに Bayes' theorem

$$p(Y|X) = \frac{p(X|Y)p(Y)}{p(X)}, \tag{1.7}$$

が成り立つことがわかる[*5]。Bayes の文脈では、単なる事象 $X$ の確率 $p(X)$ などを事前確率 (prior probability)、ある事象 $Y$ が起きた下での $X$ の確率 $p(X|Y)$ などを事後確率 (posterior probability) と呼ぶ。

また、

$$p(X, Y) = p(X)p(Y), \tag{1.8}$$

が成り立つとき、$X, Y$ は互いに独立 (independent) であるという。

### 1.2.1 Probability densities

連続値での確率も考えてみる。実変数 $x$ が $(x, x + \delta x)$ の間の値を取る確率が $\delta x \to 0$ で $p(x)\delta x$ の形で表されるとき、$p(x)$ を $x$ についての確率密度 (probability density) と呼び、以下の性質を満たす。

$$p(x) \geq 0, \tag{1.9}$$

$$\int_{-\infty}^{\infty} p(x)dx = 1. \tag{1.10}$$

従って、定義上、変数変換の際は、微小要素まで含めて一致するよう気をつけなければならない。

$$p_x(x)\delta x \simeq p_y(y)\delta y \tag{1.11}$$

$$\to \quad p_y(y) = p_x(x)\left|\frac{dx}{dy}\right|, \tag{1.12}$$

$$= p_x(g(y))|g'(y)|. \quad (x = g(y)) \tag{1.13}$$

また、以下のようなある値までの確率を累積分布関数 (cumulative distribution function) と呼ぶ。

$$P(z) = \int_{-\infty}^{z} p(x)dx \tag{1.14}$$

複数の変数 $\boldsymbol{x} = (x_1, \cdots, x_D)$ についても同様に定義される。(本文参照)

---

[*4] 以降の暗黙の notation として、確率変数は大文字で、その実現時は小文字で表されることに注意。

[*5] 分母は $\sum_Y p(X|Y)p(Y)$ とも書けるので、$Y$ で和をとった場合に $p(Y|X)$ が 1 になるための規格化因子になっていることがわかる。

### 1.2.2 Expectations and covariance

確率分布 $p(x)$ の下でのある関数 $f(x)$ の平均は期待値と呼ばれ、次のように計算される。

Expectation

$$\text{discrete variables} \quad \mathbb{E}[f] = \sum_x p(x)f(x), \tag{1.15}$$

$$\text{continuous variables} \quad \mathbb{E}[f] = \int p(x)f(x)dx. \tag{1.16}$$

これらの期待値はデータから

$$\mathbb{E}[f] \simeq \frac{1}{N}\sum_{n=1}^{N} f(x_n), \tag{1.17}$$

と計算可能で $N \to \infty$ の極限で一致する。

また、以降の notation として、$x$ に関する期待値のみとる場合

$$\mathbb{E}_x\left[f(x,y)\right], \tag{1.18}$$

と表すことにし、条件付き確率の期待値を

$$\mathbb{E}\left[f|y\right] = \sum_x p(x|y)f(x), \tag{1.19}$$

と表すことにする。

Variance and Covariance

$$\text{variance} \quad \text{var}[f] = \mathbb{E}[(f(x) - \mathbb{E}[f(x)])^2] = \mathbb{E}[f(x)^2] - \mathbb{E}[f(x)]^2, \tag{1.20}$$

$$\text{continuous variables} \quad \text{cov}[x,y] = \mathbb{E}_{x,y}[x - \mathbb{E}[x]y - \mathbb{E}[y]] = \mathbb{E}_{x,y}[xy] - \mathbb{E}[x][y]. \tag{1.21}$$

Covariance は vector に対しても同様に

$$\text{cov}[\boldsymbol{x},\boldsymbol{y}] = \mathbb{E}_{\boldsymbol{x},\boldsymbol{y}}\left[\{\boldsymbol{x} - \mathbb{E}[\boldsymbol{x}]\}\left\{\boldsymbol{y}^T - \mathbb{E}[\boldsymbol{y}^T]\right\}\right], \tag{1.22}$$

$$= \mathbb{E}_{\boldsymbol{x},\boldsymbol{y}}\left[\boldsymbol{x}\boldsymbol{y}^T\right] - \mathbb{E}[\boldsymbol{x}]\mathbb{E}[\boldsymbol{y}^T]. \tag{1.23}$$

と求まる。

### 1.2.3 Bayesian probabilities

確率をランダムな繰り返し試行の頻度として捉える古典的 (classical) or 頻度主義的 (frequentist) 確率解釈と呼ぶ一方、確率を不確実性の定量化と捉えることを Bayes 的な見方と呼ぶ[*6]。例えば、南極の氷が今世紀末に消えるかといった問題を考えるとき、極地の氷が溶ける速度などといった量を不確実性まで含めて定量的に評価し、新たな証拠 (例えば極地の詳細な地形など) に応じてこれらを修正していくプロセスは、まさしく Bayes 的な解釈によって実現される[*7]。

[*6] なお、不確実性が満たすべき性質が確率のそれと同じであることは証明さえれているらしい。

[*7] 「確率の記述」という点では頻度主義と Bayes 主義は同義であり、逆にそれが様々な議論を生んでいる。特に Bayes 主義では事前分布を仮定するため、結果がある意味主観に依存してしまう。そのため、事前分布への依存を小さくしたい場合は noninformative prior (無情報事前分布) を使うこともある。

Pattern recognition においては、Bayes 的な見方は、データではなくモデルパラメータ $\boldsymbol{w}$ (さらにはモデルそのものの選択) 自体の不確実性を定量的に取り扱うことである。

例えば、頻度主義でしばしば使用される maximum likelihood は parameter $\boldsymbol{w}$ が与えられたときにデータ $\mathcal{D}$ が得られる確率 $p(D|\boldsymbol{w})$ (likelihood function) を最大化することで $\boldsymbol{w}$ を求める。ここには根底に **$\boldsymbol{w}$ を固定した何らかの推定すべき量と捉え、その不確実性をデータ $\mathcal{D}$ の分布によって捉えよう**という姿勢が存在している。一方、Bayes 主義的な見方では**データはただ一つであり、パラメータに関する不確実性は $\boldsymbol{w}$ の分布という形で記述される**。従って、likelihood function はあくまで事後分布 $p(\boldsymbol{w}|\mathcal{D})$ を計算するための手段である。

ベイズ的な手法は周辺分布の計算を必要とし計算コストが高いためこれまであまり使われてこなかった。しかし、マルコフ連鎖モンテカルロ法といったサンプリング法の開発や計算機の発展によって実用的に使用することが可能になった。さらに、変分ベイズ法、EP 法 (期待値伝搬法) といった決定論的近似法が開発されたことにより、サンプリング手法を用いることが難しい場合にも応用できる幅が広がっている。

### 1.2.4 The Gaussian distribution

Normal or Gaussian distribution とは、

$$\mathcal{N}(x|\mu,\sigma^2)=\frac{1}{(2\pi\sigma^2)^{1/2}}\exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}. \tag{1.24}$$

ここでそれぞれ、

$$\begin{aligned}\mu&:\ \text{mean},\\ \sigma^2&:\ \text{variance},\\ \sigma&:\ \text{standard deviation},\\ 1/\sigma^2&:\ \text{precision}.\end{aligned}$$

と呼ばれる。

また、$D$-dimensional vector $\boldsymbol{x}$ についても Gaussian distribution は定義でき、

$$\mathcal{N}(x|\boldsymbol{\mu},\boldsymbol{\Sigma})=\frac{1}{(2\pi)^{D/2}}\frac{1}{|\boldsymbol{\Sigma}|^{1/2}}\exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right\}, \tag{1.25}$$

となる。ここで、$\boldsymbol{\Sigma}$ は $D\times D$ の covariance matrix であり、$|\boldsymbol{\Sigma}|$ は determinant を表す。

以下、i.i.d. (independent and identically distributed) に得られたデータ $\boldsymbol{x}=(x_1,\cdots,x_N)^T$ から最尤推定で得られる $\mu$ と $\sigma^2$ の話。最尤推定量が不偏分散ではないことによる bias が over-fitting の問題を生んでいるらしい。(教科書で説明)

### 1.2.5 Curve fitting re-visited

1.1 でみた curve fitting を確率論的見知から再度眺めてみる。

Target 周り不確実性を定量化するため、ここでは Gaussian distribution を考える。つまり、

$$p(t|x,\boldsymbol{w},\beta)=\mathcal{N}\left(t|y(x,\boldsymbol{w}),\beta^{-1}\right). \tag{1.26}$$

これを用いて再度 maximum likelihood を行うと、

$$\ln p(\boldsymbol{t}|\boldsymbol{x},\boldsymbol{w},\beta)=-\frac{\beta}{2}\sum_{n=1}^{N}[y(x_n,\boldsymbol{w})-t_n]^2+\frac{N}{2}\ln\beta-\frac{N}{2}\ln(2\pi), \tag{1.27}$$

となる。従って、$\boldsymbol{w}$ に関する log lilkelihood の最大化は先に定義した error function の最小化と同じであることがわかる。

同様に $\mu, \beta$ も計算すると、1.2.4 で求めたものと同様の結果が得られ、以上の結果を用いると、最尤推定量を用いた確率分布

$$p\left(t|x, \boldsymbol{w}_{\mathrm{ML}}, \beta_{\mathrm{ML}}\right) = \mathcal{N}\left(t|y(x, \boldsymbol{w}_{\mathrm{ML}}), \beta_{\mathrm{ML}}^{-1}\right), \tag{1.28}$$

から新たなデータに対する予測できる。

以上の話をより Bayesian 的なアプローチで考えてみる。簡単のため、$\boldsymbol{w}$ の事前分布として Gaussian distribution

$$p(\boldsymbol{w}|\alpha) = \mathcal{N}(\boldsymbol{w}|\boldsymbol{0}, \alpha^{-1}\boldsymbol{I}) = \left(\frac{\alpha}{2\pi}\right)^{(M+1)/2} \exp\left[-\frac{\alpha}{2}\boldsymbol{w}^T\boldsymbol{w}\right], \tag{1.29}$$

を仮定する。Bayes' theorem より事後分布

$$p(\boldsymbol{w}|\boldsymbol{x}, \boldsymbol{t}, \alpha, \beta) \propto p(\boldsymbol{t}|\boldsymbol{x}, \boldsymbol{w}, \beta)p(\boldsymbol{w}|\alpha), \tag{1.30}$$

と表されることを用いてデータから最も起こりえそうな $\boldsymbol{w}$ を選択する方法があり、これを maximum posterior (MAP) と呼ぶ。式 (1.30) の log を取って (1.27) と (1.29) の $\boldsymbol{w}$ に関係する項のみ取ってくると、

$$\frac{\beta}{2}\sum_{n=1}^{N}\left[y(x_n, \boldsymbol{w}) - t_n\right]^2 + \frac{\alpha}{2}\boldsymbol{w}^T\boldsymbol{w}, \tag{1.31}$$

となるため、posterior を最大化する行為は regularization を含めた二乗誤差を最小化する行為と等しいことがわかる。

### 1.2.6 Bayesian curve fitting

以上の話は prior という概念を導入したものの、$\boldsymbol{w}$ を点推定している点で十分な Bayesian approach とは言えない。Fully Baysian approach では $\boldsymbol{w}$ に関する積分を通して事後分布を計算していく。

例えば、新たなデータ $x$ に対する予測は、$\boldsymbol{w}$ についての marginalize

$$p(t|x, \boldsymbol{x}, \boldsymbol{t}) = \int p(t|x, \boldsymbol{w})p(\boldsymbol{w}|\boldsymbol{x}, \boldsymbol{t})d\boldsymbol{w}, \tag{1.32}$$

をすることで求められる。ここで、$p(t|x, \boldsymbol{w})$ は式 (1.26) から、$p(\boldsymbol{w}|\boldsymbol{x}, \boldsymbol{t})$ は式 (1.30) から求められる。

Section 3.3 でみるように今回の場合は analytical に分布を求められて、結果は本文の (1.69) - (1.72) のようになる。特に分散について見てみると、第二項が予め与えた target のゆらぎに新たに加わった Bayes approach による効果だとわかる。

katarash's column

「MAP 推定が Bayesian approach ではない」というのはこの本において重要な主張である。これは、事後分布を用いて $\boldsymbol{w}$ を点推定するだけでは十分でなく、事後分布を用いて新たなデータ $x$ に対する (予測) 分布まで求めて初めて Bayesian approach と呼ぶべきであるという本書のスタンスを如実に表している。

## 1.3 Model Selection

既に見たように maximum likelihood では over-fitting の問題があった。Training set が十分に大きい場合には validation set, test set を確保して学習すれば良いが、多くの場合訓練データは貴重なものである。その際に用いられるのが cross-validation という手法であり、$(S-1)/S$ のデータを学習に使い、性能の評価に全てのデータを用いる方法である。特に $S = N$ のとき、leave-one-out と呼ばれる。

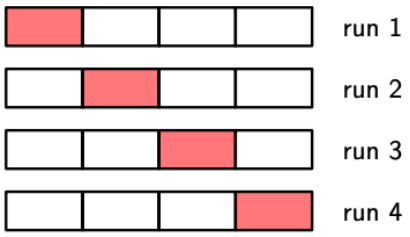


**Figure 1.18** The technique of $S$-fold cross-validation, illustrated here for the case of $S = 4$, involves taking the available data and partitioning it into $S$ groups (in the simplest case these are of equal size). Then $S - 1$ of the groups are used to train a set of models that are then evaluated on the remaining group. This procedure is then repeated for all $S$ possible choices for the held-out group, indicated here by the red blocks, and the performance scores from the $S$ runs are then averaged.

しかし、cross validation はデータ数が大きくなるほど training に必要な回数が多くなる。さらに調整すべき parameter が大量にある場合は指数関数的に学習量が増えていくことを考えると、training data のみで完結して一度に複数の parameter を最適化できるような model selection の基準があることが望ましい。

このような観点から歴史的に様々な information criteria が提案されている[*8]。有名なものは Akaike information criterion (AIC)

$$-\,\mathrm{AIC} = \ln p(\mathcal{D}|\boldsymbol{w}_{\mathrm{ML}}) - M, \tag{1.33}$$

である。ここで $p(\mathcal{D}|\boldsymbol{w}_{\mathrm{ML}})$ は best-fit likelihood であり $M$ は model の parameter 数を表す。このようにモデルのパラメータ数に応じた penalty を課すことによって、モデルのパラメータ数も同時に最適化することができる。

この発展形として Bayesian information criterion (BIC) というものもあり、これは 4.4.1 で議論する。

## 1.4 The Curse of Dimensionality

これまでの curve fitting の例では iput は 1 次元だったが、一般に input はより高次元である。この high dimensionality が pattern recognition では問題を引き起こしうることをみる。

ある input の空間に存在するデータを分類する単純なやり方は、領域を適当な mesh に区切ってその中に最も存在するデータを予測値とすることである。一方でこのやり方は input の次元が大きくなるにつれ分類に必要なデータ数が $D$ の指数乗で大きくなることで容易に破綻する[*9]。

このように高次元空間で生じる問題は次元の呪い (curse of dimensionality) と呼ばれる。しかし、実際には

- 実際のデータは、effective にはより低い次元で表される空間に confine されていることが多い[*10]
- 実際のデータは (少なくとも local には) 滑らかに繋がっていることが多く、input のわずかな変更は interpolation で予測することができる

といった理由から、必ずしも high dimension では効果的な pattern recognition を行えないというわけではない。

---

[*8] ここで述べられているのは maximum-likelihood が含む bias (モデルの parameter 推定に用いたデータを再利用して分布を計算していること) をうまく取り除こうとすると様々な項が出てくるよというものである。

[*9] 直感と反する高次元の面白い性質がある。半径 $r$ の $K$ 次元球の体積を $V_D(r) = K_D r^D$ と定義すると、半径 1 の円の表皮 $1-\epsilon$ に存在する相対的な体積は、

$$\frac{V_D(1) - V_D(1-\epsilon)}{V_D(1)} = 1 - (1-\epsilon)^D, \tag{1.34}$$

となることから、次元が大きければ大きいほど体積のほとんどは球の表皮に存在することがわかる。

[*10] 本文では、ベルトコンベア上の物体の方向を画像から認識する例があげられている。この場合、input は pixel 数で与えられる高次元空間だが、本質的には $(x, y, z)$ といった方向を決める 3 つの自由度のみが重要である。つまり、高次元空間に埋め込まれた低次元の多様体 (manifold) の性質さえわかれば良いということである。