パターン認識と機械学習

第1章 序論

1.1 例：多項式曲線フィッティング

1.2 確率論

2012/09/11(Tue)

@daimatz

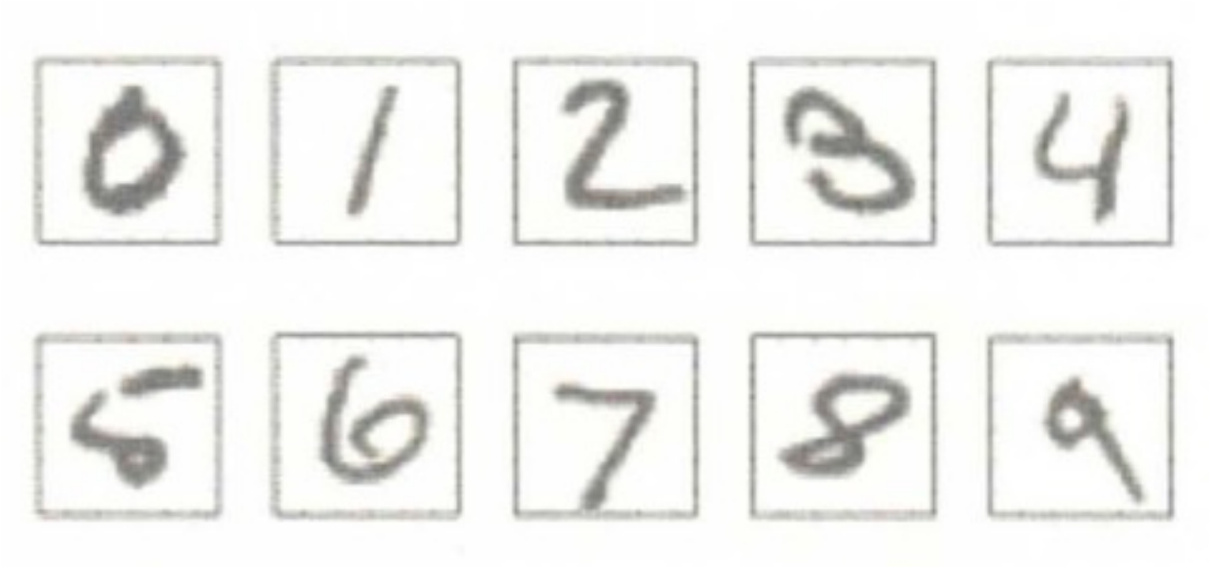
自己紹介

- 松本 大介 (@daimatz)
  - 東京工業大学 大学院情報理工学研究科
    計算工学専攻 修士2年
  - 専門は言語系、型理論とか
  - 最近 Haskell 書こうとしてる
  - 機械学習については全く知りません

第1章 序論

手書き数字認識

- 28x28ピクセルの手書き画像、784次元のベクトル
- そのベクトルを入力として０..９を出力する関数を作りたい
- 筆運びの形を調べて人力によってルールを編み出そうとすると、すぐに発散してしまう。例外も多い。
- 機械学習を採用するとはるかに良い結果が得られる。



機械学習アプローチ

- **訓練集合**
  - N 個の手書き数字の集合 $\{\mathbf{x}_1, \cdots, \mathbf{x}_N\}$
  - これらの手書き数字は人間が判読してラベル付けされているとする
- 1つ1つの数字に対応するカテゴリは**目標ベクトル** $\mathbf{t}$ で表す
  - 画像 $\mathbf{x}$ それぞれに対して1つ $\mathbf{t}$ が対応することに注意

機械学習アプローチ

- 機械学習によって得られるのは関数 $\mathbf{y}(\mathbf{x})$
  - 新しい画像 $\mathbf{x}$ を入力すると目標ベクトルと符号化の仕方が等しい出力ベクトル $\mathbf{y}$ が出力される
  - $\mathbf{y}(\mathbf{x})$ の詳細な形を求める段階を**訓練**または**学習**段階と呼ぶ
  - 一旦モデルが学習されると、**テスト集合**と呼ばれる新たな数字画像に対してカテゴリを決めることができる
  - 訓練で使ったのとは異なる新たな事例を分類する能力を**汎化**と呼ぶ
    - パターン認識における中心的な課題

前処理、特徴抽出

- 実際の応用ではもともとの入力変数は**前処理**によって新しい変数に変換し問題を解きやすくしておく
  - **特徴抽出** (属性抽出)
- 例えば顔検出を行いたい場合
  - 直接膨大な数のピクセルを処理するかわりに、高速に計算でき顔と顔でない物を区別する特徴の情報を保持しておく
- 前処理は一種の次元削減
  - 問題を解くのに重要な情報まで落としてしまわぬよう注意

学習の種類

- **教師あり学習**：訓練データが入力ベクトルとそれに対応する目標ベクトルで構成される場合
  - **クラス分類**：各入力を有限離散カテゴリの1つに対応付ける
  - **回帰**：出力が1つ以上の連続変数の場合
    - 化学プラントにおける生成物の量の予測など
- **教師なし学習**：訓練データが入力ベクトルのみの場合
  - **クラスタリング**：類似したグループを見つける
  - **密度推定**：入力空間のデータの分布を求める

学習の種類

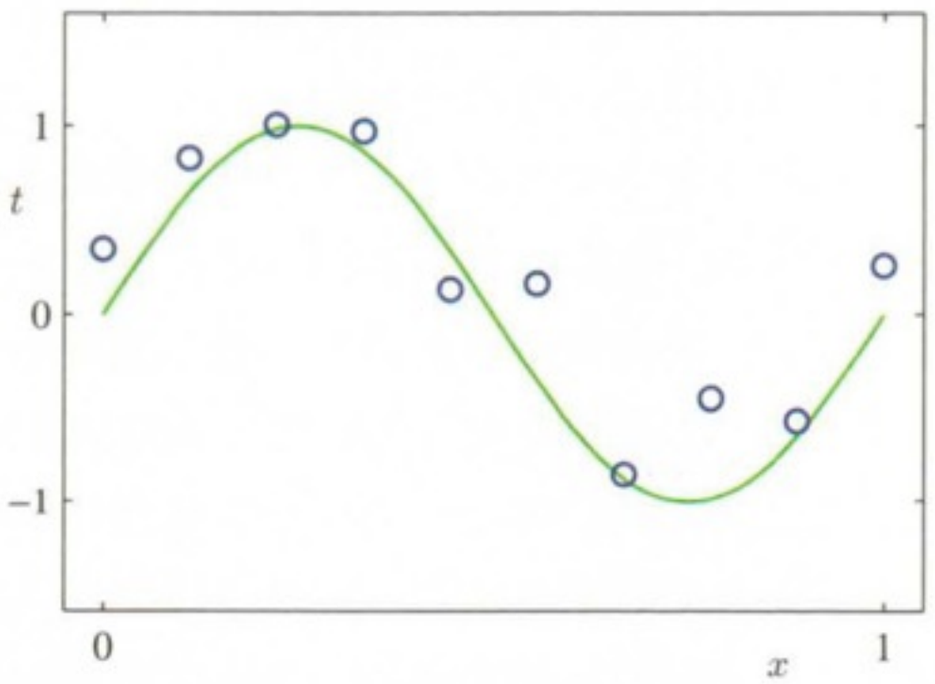
- **強化学習**：与えられた状況かで報酬を最大にするような行動を見つける問題
  - 最適な出力は事例として与えられず、試行錯誤を通じて学習アルゴリズムがそれを発見しなければならない
  - バックギャモンを強化学習で学習した事例
    - 局面ごとに何十もの手があるのに、報酬はゲームの終わりにようやく勝利という形でしか与えられない
    - いくつかの手は良かったかもしれないし、大して良くなかったかもしれない。**信頼度割り当て**問題
  - 新しい手を試す**探査**と、報酬が得られることがわかっている行動をとる**利用**のトレードオフ

1.1 例：多項式曲線フィッティング

- 単純な回帰問題を考える。実数値の入力変数 x を観測して、実数値の目標変数 t を予測する。
  - $N = 10$ として、データは $\sin(2\pi x)$ にノイズを加えたもの

    $\mathbf{x} \equiv (0.1, 0.2, \cdots, 1)$

    $\mathbf{t} \equiv (t_1, \cdots, t_n)$ は $\sin(2\pi x)$ の値に、ガウス分布に従う小さなランダムノイズを加えて生成



  - 目標はこの訓練集合を使って新たな入力 $\hat{x}$ に対する出力 $\hat{t}$ を予測すること

多項式曲線フィッティング

- 訓練集合から関数 $\sin(2\pi x)$ を見つけるのとほぼ等価
  - 有限個のデータ集合から汎化しなければならない
  - 観測データにはノイズが乗っており、$\hat{x}$ に対する $\hat{t}$ の値には不確実性がある
    - 1.2節の確率論で不確実性を厳密かつ定量的に表現する
- 多項式曲線フィッティング

$$y(x, \mathbf{w}) = w_0 + w_1 x + w_2 x^2 + \cdots + w_M x^M = \sum_{j=0}^{M} w_j x^j$$

  - 係数 $\mathbf{w}$ について線形。**線形モデル** (3章、4章)

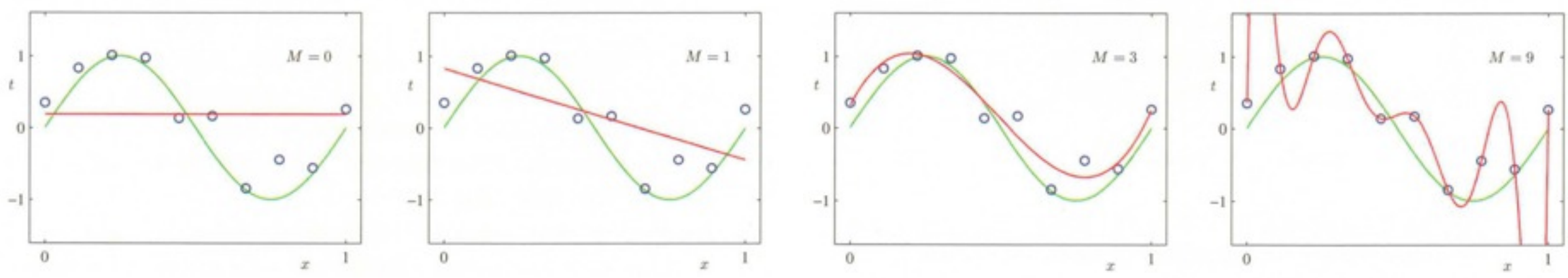
多項式曲線フィッティング

- **誤差関数**の最小化を考える。ここでは単純に二乗和誤差

$$E(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\{y(x_n, \mathbf{w}) - t_n\}^2$$

を最小化する $\mathbf{w}$ を選ぶ。$1/2$ は便宜のための係数。

  - $E(\mathbf{w})$ は $\mathbf{w}$ の2次関数でそれを最小にする唯一の $\mathbf{w}^*$ (演習1.1)
  - 次数 M を選ぶ問題は**モデル比較**あるいは**モデル選択**の一例
    - $M = 0, 1$ は不適切。$M = 3$ は良い感じ。$M = 9$ は**過学習**。



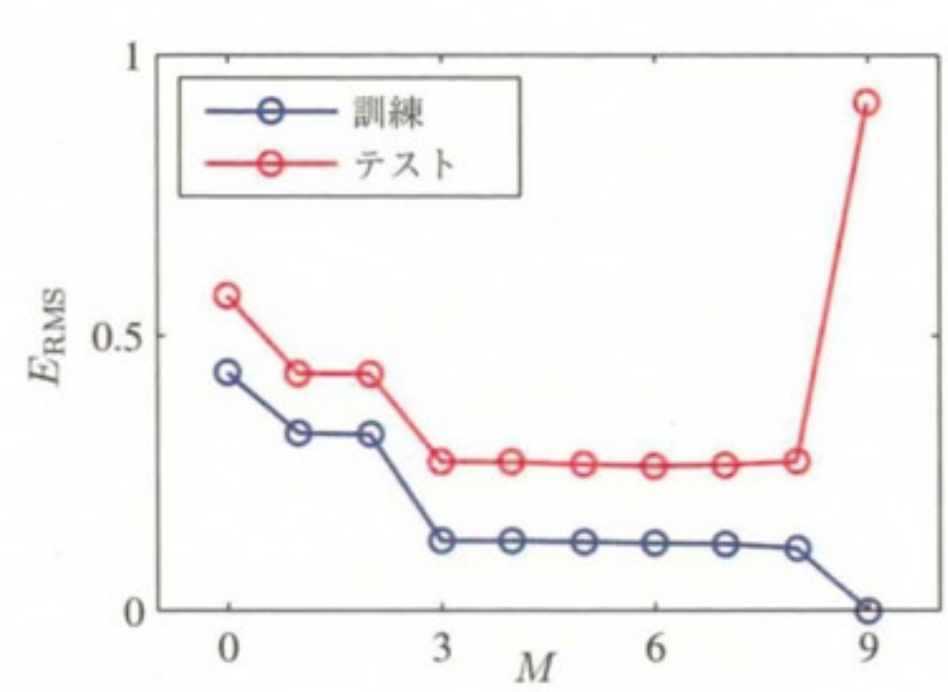
誤差の評価

- 目標は新たなデータに対して正確な予測を行う汎化をすること
- 汎化性能が M にどう依存するかを定量的に評価する
  - 100個のデータ点からなる独立したテスト集合を訓練集合と同様なノイズを加えて生成する
  - 各 M について、テスト集合についての平均二乗平方根誤差 (root-mean-square error) を評価できる

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

平均二乗平方根誤差

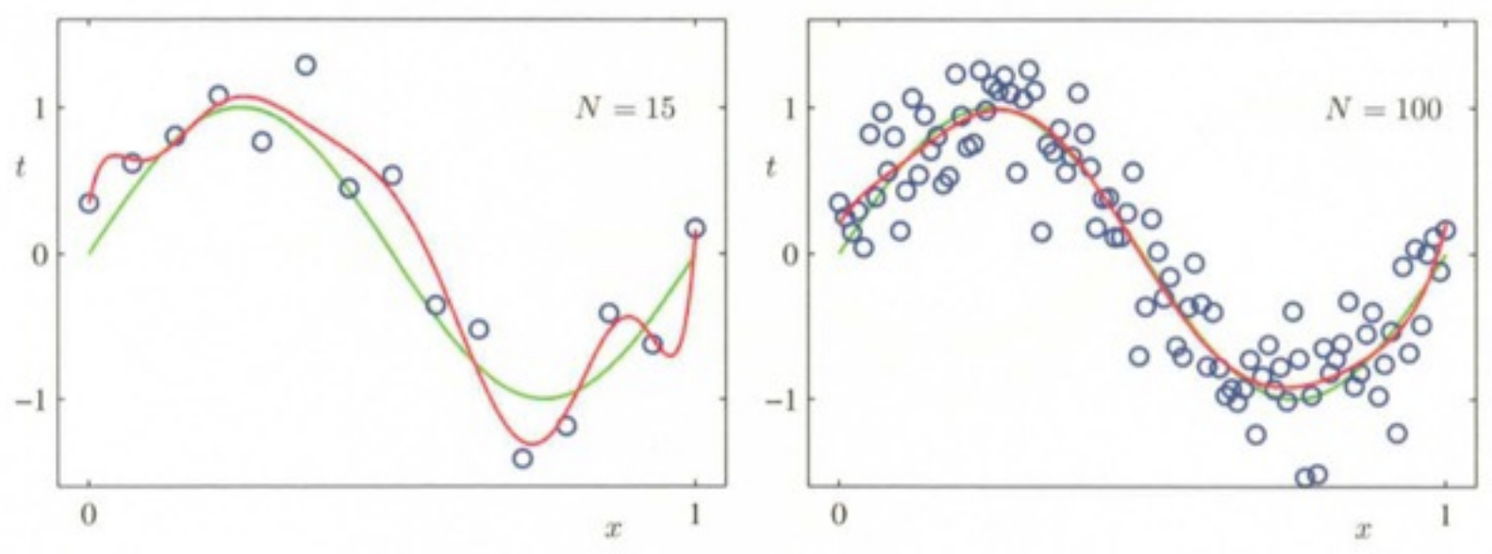
- M の値に対する E_{RMS} の値
  - M が小さいと誤差は大きい
  - $3 \leq M \leq 8$ だと良い感じ
  - $M = 9$ だと訓練集合との誤差は 0 だが、テスト集合に対しての誤差が大きくなる
    - なぜならMが大きくなるにつれて多項式の係数 w_i^* の値が大きくなるから



	$M=0$	$M=1$	$M=3$	$M=9$
$w_0^\star$	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

モデルの固定

- モデルの次数を固定し、データ集合のサイズを変えてみる
  - モデルの複雑さを固定 ($M = 9$) すると、データ集合のサイズが大きくなるにつれ過学習の問題は深刻でなくなる
  - データを多くすれば複雑で柔軟なモデルを当てはめられる
    - 経験則として、データ数はモデル中のパラメータの何倍かよりは小さくなくてはならないと言われている



モデルの固定

- モデルの複雑さは解くべき問題の複雑さに応じて選ぶべき？
  - モデルが単純すぎると過学習の問題を起こす
  - 最小二乗のアプローチは**最尤推定** (1.2.5節) の特別な場合で、過学習の問題は最尤推定の持つ一般的な性質である
  - 過学習の問題を避けるには**ベイズ的アプローチ** (3.4節) をとる。**有効パラメータ数**は自動的にデータサイズに適合する。
- どうやって複雑で柔軟なモデルを限られたサイズのデータ集合に対して使うかを考えてみる

誤差関数の正規化

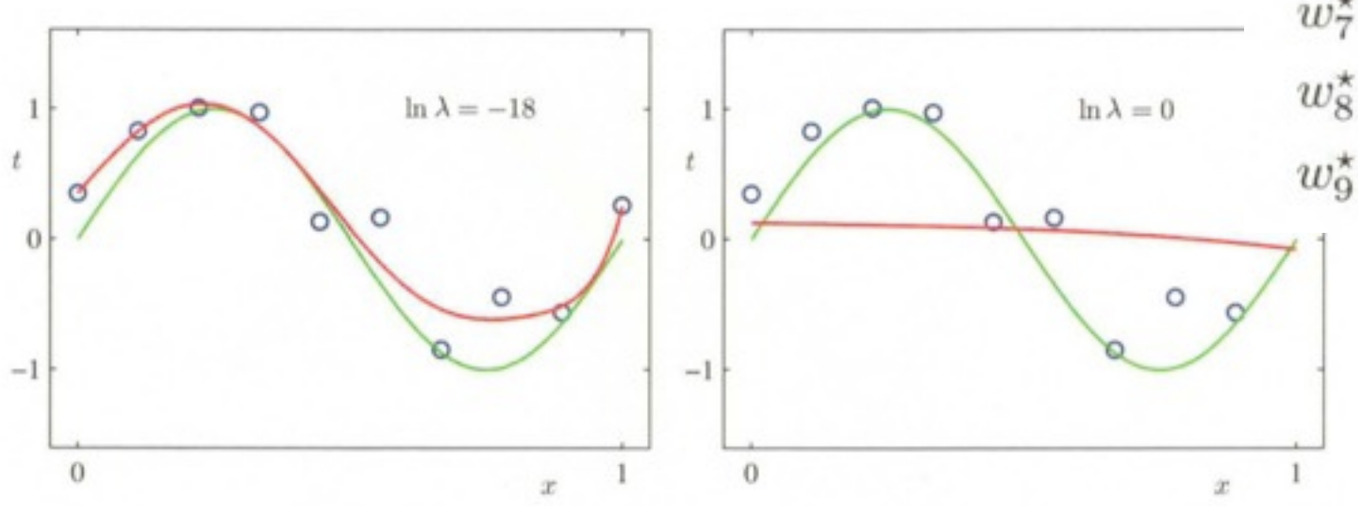
- 過学習の制御のために**正規化**を行う
  - 誤差関数に罰金項を付加して係数が大きくなることを防ぐ
    - 最も単純なのは係数の2乗和。係数 λ は重要度を表す

$$\tilde{E}(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{M}\{y(x_n, \mathbf{w}) - t_n\}^2 + \frac{\lambda}{2}\|\mathbf{w}\|^2$$

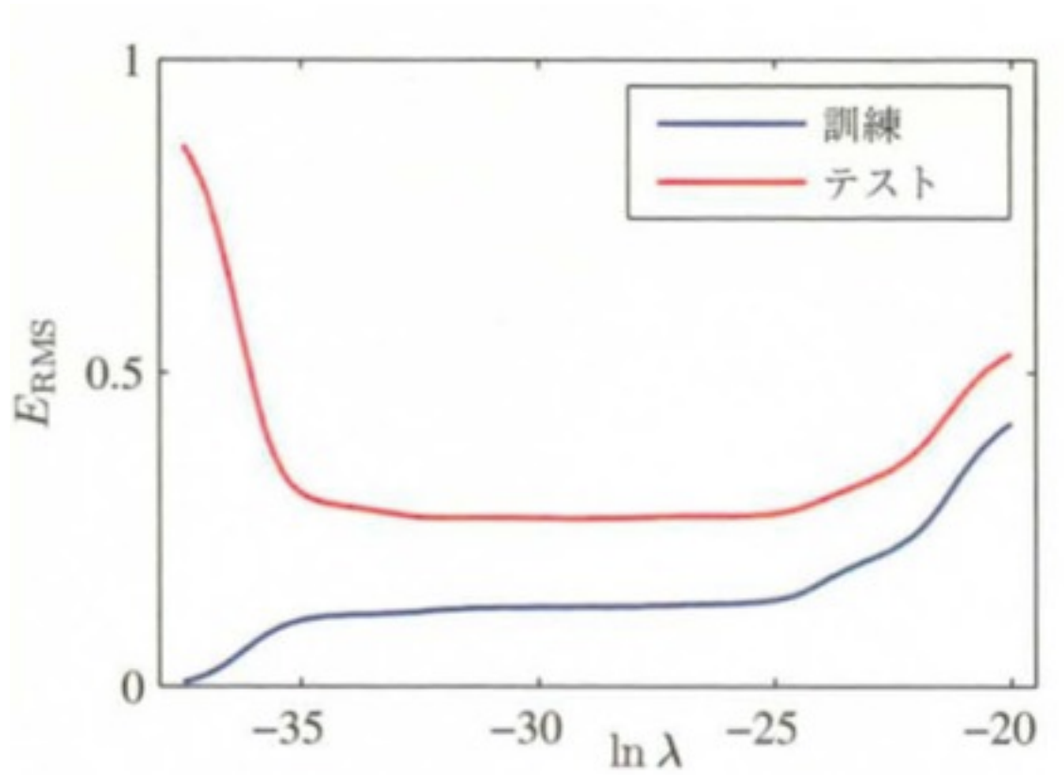
- $\|\mathbf{w}\|^2 = w_0^2 + w_1^2 + \cdots + w_M^2$ だが w_0 は原点の選び方に依存しているため特別扱いすることも多い。いずれにせよ $\tilde{E}(\mathbf{w})$ を最小にする解は完全に閉じた形で求まる
- 係数を小さくする意味で**縮小推定**と呼ばれる。2次の正規化の場合は特に**リッジ回帰**と呼ばれる。

- $M = 9$ 次の多項式を当てはめた結果
  - $\ln \lambda = -18$ にとると
    過学習が抑制されて良い感じ
  - $\ln \lambda = 0$ のように値を大きく
    しすぎると再び悪くなる

	$\ln \lambda = -\infty$	$\ln \lambda = -18$	$\ln \lambda = 0$
$w_0^\star$	0.35	0.35	0.13
$w_1^\star$	232.37	4.74	−0.05
$w_2^\star$	−5321.83	−0.77	−0.06
$w_3^\star$	48568.31	−31.97	−0.05
$w_4^\star$	−231639.30	−3.89	−0.03
$w_5^\star$	640042.26	55.28	−0.02
$w_6^\star$	−1061800.52	41.32	−0.01
$w_7^\star$	1042400.18	−45.95	−0.00
$w_8^\star$	−557682.99	−91.53	0.00
$w_9^\star$	125201.43	72.68	0.01



- 訓練集合とテスト集合の両方に対する RMS 誤差の値を $\ln \lambda$ に対してプロットしてみると、λ の値で過学習の度合いが決定されることがわかる



- 得られたデータを $\mathbf{w}$ を決めるための訓練集合と M や λ を最適化するための**確認用集合 (ホールドアウト集合)** に分ける方法
  - 貴重な訓練データを無駄にすることが多いので、洗練されたアプローチを探す必要がある (1.3節)

- これまでは主に直感に訴えてきたが、これからは確率論を使ってより原理的なアプローチを探っていく

1.2 確率論

- 不確実性に対する定量化と操作に一貫した枠組みを与える
- **加法定理**：

$$p(X) = \sum_{Y} p(X, Y)$$

- **乗法定理**：

$$p(X, Y) = p(Y|X)p(X)$$

  - $p(Y|X)$ は X が起こったもとでの Y が起こる確率。条件付き確率。

ベイズの定理

- **ベイズの定理**： $p(X,Y) = p(Y,X)$ から、乗法定理を使って

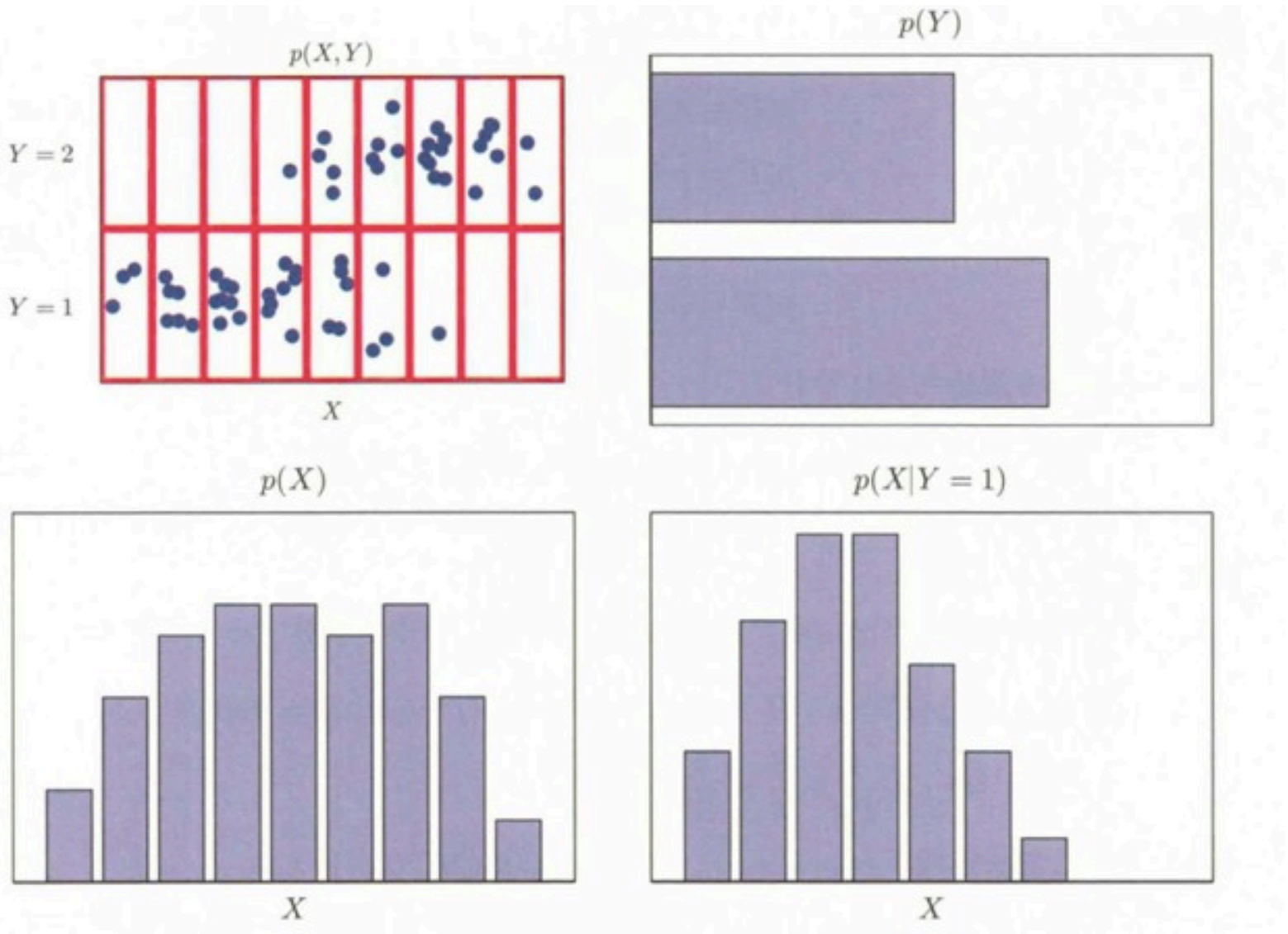
$$p(Y|X) = \frac{p(X|Y)p(Y)}{p(X)}$$

  - 分母は

$$p(X) = \sum_Y p(X|Y)p(Y)$$

  - ベイズの定理の左辺ですべての Y について和をとったときに1になることを保証する正規化定数とみなせる

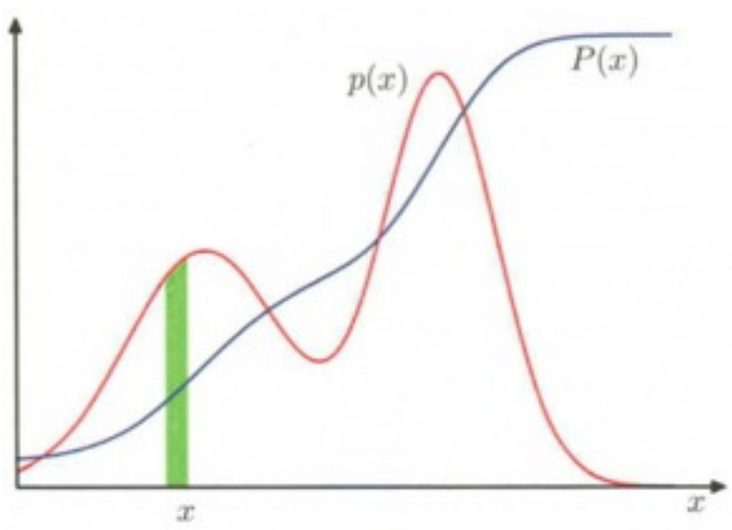
- 9つの値を取りうる X と2つの値を取りうる Y に対する分布
  - 右上の図は、$N \to \infty$ のときその比は確率 $p(Y)$ に一致する



確率密度、累積分布関数

- 実数値をとる変数 x が区間 $(x, x+\delta x)$ に入る確率が $\delta x \to 0$ のとき $p(x)\delta x$ で与えられるとき、$p(x)$ を x の**確率密度**という

$$p(x \in (a,b)) = \int_a^b p(x)dx$$



  - 確率密度 $p(x)$ は次を満たす

$$p(x) \geq 0 \qquad \int_{-\infty}^{\infty} p(x)dx = 1$$

  - **累積分布関数**： x が区間 $(-\infty, z)$ に入る確率。

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

    - $P'(x) = p(x)$ を満たす

同時分布

- いくつかの連続変数 $x_1, \cdots, x_D$ をまとめて $\mathbf{x}$ で表し同時分布 $p(\mathbf{x}) = p(x_1, \cdots, x_D)$ を定義する。$\mathbf{x}$ が $\mathbf{x}$ を含む無限小の体積要素 $\delta\mathbf{x}$ に入る確率は $p(\mathbf{x})\delta\mathbf{x}$ で、確率密度は次を満たす

$$p(\mathbf{x}) \geq 0 \qquad \int p(\mathbf{x}) dx = 1$$

- 確率変数が離散の場合、確率密度は**確率質量関数**とも呼ばれる
  - 取りうる値に「確率の質量」が集中しているとみなす
- 離散変数と連続変数の組み合わせでもこれまでの定理は成立
  - 厳密に示すには測度論が必要

1.2.2 期待値と分散

- **期待値**：確率分布のもとでの平均値。
  - 離散 $E[f] = \sum_x p(x)f(x)$　連続 $E[f] = \int p(x)f(x)dx$
  - どちらも有限個の N 点を用いて近似できる

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

  - 多変数の場合には添字でどの変数についての平均かを示す $E_x[f(x,y)]$ なら x の分布に関する平均で、y の関数になる

- **条件付き期待値**　$E_x[f|y] = \sum_x p(x|y)f(x)$

期待値と分散

- $f(x)$ の**分散** (演習 1.5)

$$\mathrm{var}[f] = E\left[(f(x) - E[f(x)])^2\right] = E[f(x)^2] - E[f(x)]^2$$

  - 特に確率変数 x 自身の分散は $\mathrm{var}[x] = E[x^2] - E[x]^2$

- 2つの変数 x と y の**共分散**

$$\mathrm{cov}[x, y] = E_{x,y}[\{x - E[x]\}\{y - E[y]\}] = E_{x,y}[xy] - E[x]E[y]$$

  - x と y が同時に変動する度合い。独立なら0になる (演習 1.6)

- 2つの確率変数ベクトル $\mathbf{x}, \mathbf{y}$ に関しては共分散は行列

$$\mathrm{cov}[\mathbf{x}, \mathbf{y}] = E_{\mathbf{x},\mathbf{y}}\left[\{\mathbf{x} - E[\mathbf{x}]\}\{\mathbf{y}^{\mathrm{T}} - E[\mathbf{y}^{\mathrm{T}}]\}\right] = E_{\mathbf{x},\mathbf{y}}[\mathbf{x}\mathbf{y}^{\mathrm{T}}] - E[\mathbf{x}]E[\mathbf{y}^{\mathrm{T}}]$$

  - $\mathbf{x}$ の成分間の共分散は $\mathrm{cov}[\mathbf{x}] \equiv \mathrm{cov}[\mathbf{x}, \mathbf{x}]$

1.2.3 ベイズ確率

- これまで見てきた「確率はランダムな繰り返し試行の頻度である」という見方を**古典的**あるいは**頻度主義的**という
- これから扱うのは**ベイズ的**な解釈
  - 不確かな事象でたくさんの繰り返し観測ができないが、それについて何らかの一般的な知見があるもの
    - 月がかつて太陽を周る軌道上にあったか？
  - 新たな証拠を得た場合、我々の行動に影響を与える
    - 地球観測衛星が集めた新たな形の診断情報など
  - 不確実性を定量的に表して新たな証拠に照らしてそれを修正し、結果として最適な行動や決定を下す

ベイズ確率

- 多項式曲線フィッティングの例で、$\mathbf{w}$のほかモデルそのものの選択に関する不確実性を表すのにベイズ的な観点を採用する
- データを観測する前にあらかじめ $\mathbf{w}$ に関する我々の仮説を事前確率分布 $p(\mathbf{w})$ の形で取り込んでおく。観測データ $\mathcal{D} = \{t_1, \cdots, t_N\}$ の効果は $p(\mathcal{D}|\mathbf{w})$ で表現される
- **ベイズの定理**：

$$p(\mathbf{w}|\mathcal{D}) = \frac{p(\mathcal{D}|\mathbf{w})p(\mathbf{w})}{p(\mathcal{D})}$$

  - $\mathcal{D}$ を観測した事後に $\mathbf{w}$ に関する不確実性を、事後分布 $p(\mathbf{w}|\mathcal{D})$ の形で評価する
  - $p(\mathcal{D}|\mathbf{w})$ は $\mathcal{D}$ に対する評価で$\mathbf{w}$の関数となる。**尤度関数**。$\mathbf{w}$を固定したときに観測データ集合の起こりやすさを表す

ベイズの定理

- ベイズの定理

$$p(\mathbf{w}|\mathcal{D}) = \frac{p(\mathcal{D}|\mathbf{w})p(\mathbf{w})}{p(\mathcal{D})}$$

  - 言葉で書けば　事後確率 $\propto$ 尤度 $\times$ 事前確率
    - この式に現れるすべての値は $\mathbf{w}$ の関数
  - $p(\mathcal{D})$ は積分すると1になることを保証する定数。
    実際ベイズの定理の両辺を $\mathbf{w}$ で積分すると

$$p(\mathcal{D}) = \int p(\mathcal{D}|\mathbf{w})p(\mathbf{w})d\mathbf{w}$$

    ベイズの定理の分母を事前分布と尤度関数で表せる

尤度関数

- 尤度関数 $p(\mathcal{D}|\mathbf{w})$ の扱い方
- 頻度主義
  - $\mathbf{w}$ は固定されたパラメータ。その値は何らかの「推定量」として定められ、誤差範囲は観測データ集合 $\mathcal{D}$ による。
  - **最尤推定**：$\mathbf{w}$ は $p(\mathcal{D}|\mathbf{w})$ を最大にする値。観測データ集合の確率を最大にする $\mathbf{w}$ の値を選ぶ。
- ベイズ主義
  - ただ1つのデータ集合 $\mathcal{D}$ があって、パラメータに関する不確実性は $\mathbf{w}$ の確率分布として表される。

頻度主義とベイズ主義

- ベイズ主義
  - 事前分布が何らかの事前の信念というかは数学的な便宜によって選ばれることが多い
  - 事前分布の選び方によって結果が主観的になってしまう
    - 依存を小さくしたいときは**無情報事前分布**を使う (2.4.3節)
      - しかし異なるモデルの比較が難しい
    - 悪い事前分布を選べば高い確率で悪い結果が得られる
      - 交差確認 (1.3節) などのテクニックがモデル選択に有効

1.2.4 ガウス分布

- **ガウス分布 (正規分布)**

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

  - **平均**：μ、**分散**：σ^2
  - **標準偏差**：σ、**精度パラメータ**：$\beta = 1/\sigma^2$
  - $\mathcal{N}(x|\mu,\sigma^2) > 0$、$\int_{-\infty}^{\infty}\mathcal{N}(x|\mu,\sigma^2)dx = 1$ (演習 1.7)
  - 平均 $E[x] = \mu$
  - 2次のモーメント $E[x^2] = \mu^2 + \sigma^2$ (演習 1.8) より
    分散 $\mathrm{var}[x] = E[x^2] - E[x]^2 = \sigma^2$
  - 分布の最大値を与える x はモード (最頻値) (演習 1.9)

多次元ガウス分布

- D 次元ベクトル連続変数 $\mathbf{x}$ に対して定義されるガウス分布

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

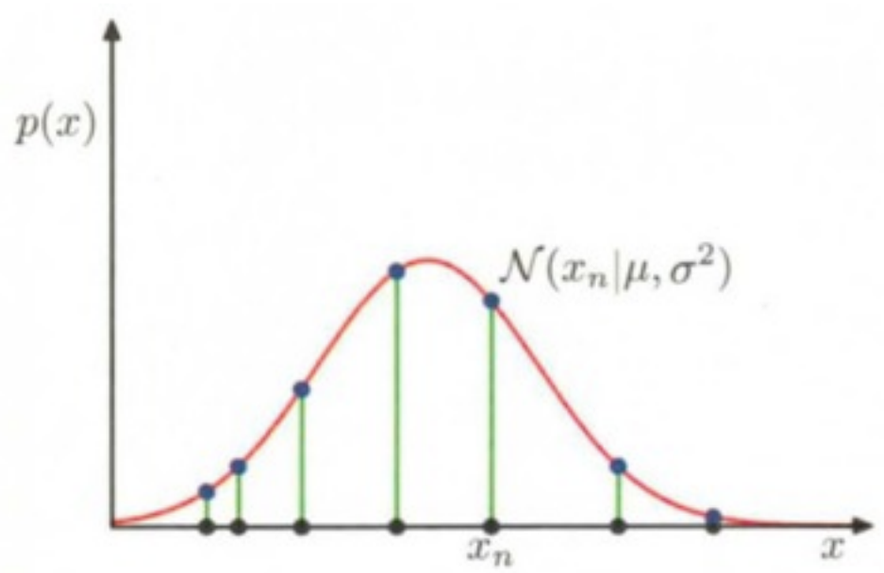
  - 平均：D 次元ベクトル $\boldsymbol{\mu}$
  - 共分散：$D \times D$ 行列 $\boldsymbol{\Sigma}$
  - 詳しくは 2.3 節

ガウス分布に対する尤度関数

- N 個の観測値からなるデータ集合 $\mathbf{x} = (x_1, \cdots, x_N)^{\mathrm{T}}$
  - ここからガウス分布の平均 μ と分散 σ^2 を定める
  - $\mathbf{x}$ は同じ分布から独立に生成された独立同分布 (i.i.d)

$$p(\mathbf{x}|\mu, \sigma^2) = \prod_{n=1}^{N} \mathcal{N}(x_n|\mu, \sigma^2)$$

- μ と σ^2 の関数とみなすと、これはガウス分布に対する尤度関数である



尤度関数の最大化

- 観測されたデータ集合を使ってパラメータを決めるには尤度関数を最大にするようなパラメータの値を求める。
  - パラメータが与えられた下でのデータの確率　でなく、データが与えられた下でのパラメータの確率　を最大化するほうが自然？ →1.2.5 節で議論

- とりあえず尤度関数の最大化を考える。対数をとって

$$\ln p(\mathbf{x}|\mu,\sigma^2) = -\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n-\mu)^2 - \frac{N}{2}\ln\sigma^2 - \frac{N}{2}\ln(2\pi)$$

  - μ に関して最大化　$\mu_{ML} = \frac{1}{N}\sum_{n=1}^{N} x_n$
  - σ^2 に関して最大化 $\sigma^2_{ML} = \frac{1}{N}\sum_{n=1}^{N}(x_n-\mu_{ML})^2$　(演習1.11)

最尤アプローチの限界

- 1変数ガウス分布の最尤パラメータ設定の限界
  - 特に分布の分散が系統的に過小評価されている。**バイアス**
- 最尤解 μ_{ML}、σ^2_{ML} はデータ集合の値 $x_1, \cdots, x_N$ の関数。
  これらの量の、パラメータ μ、σ^2 をもつガウス分布に従うデータ集合に関する期待値は (演習 1.12)

$$E[\mu_{ML}] = \mu$$
$$E[\sigma^2_{ML}] = \left(\frac{N-1}{N}\right)\sigma^2$$

  - 平均は正しい平均だが、分散は $(N-1)/N$ 倍過小評価される
    - N が大きければ重大ではなく、$N \to \infty$ なら真の値に一致
    - しかしモデルが複雑になるとバイアスの問題は難しい

1.2.5 曲線フィッティング再訪

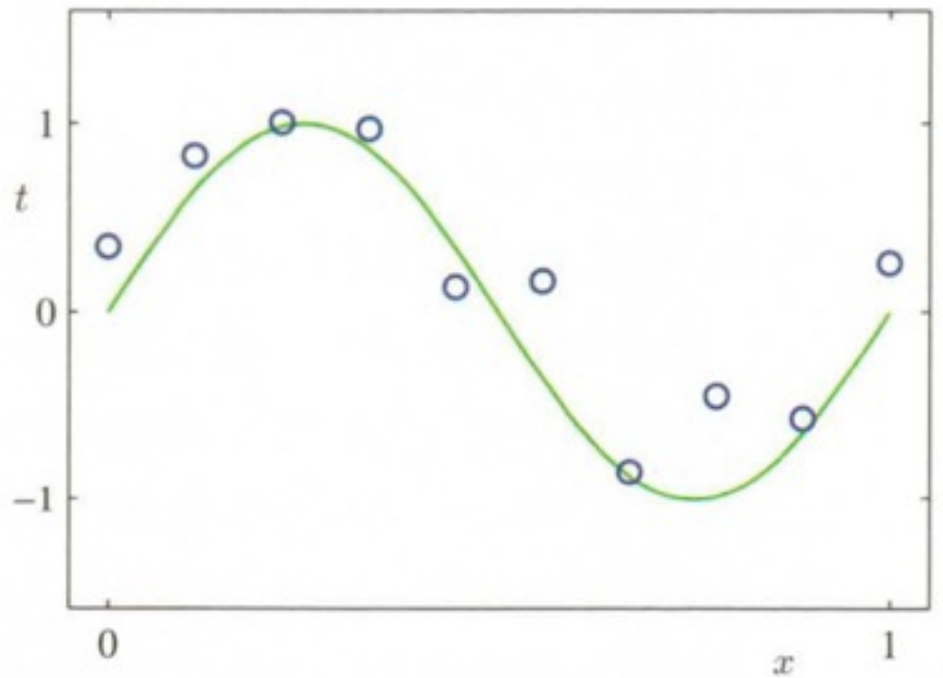
- 曲線フィッティングの例をベイズ的に
- 与えられた x に対し、対応する t は、平均が

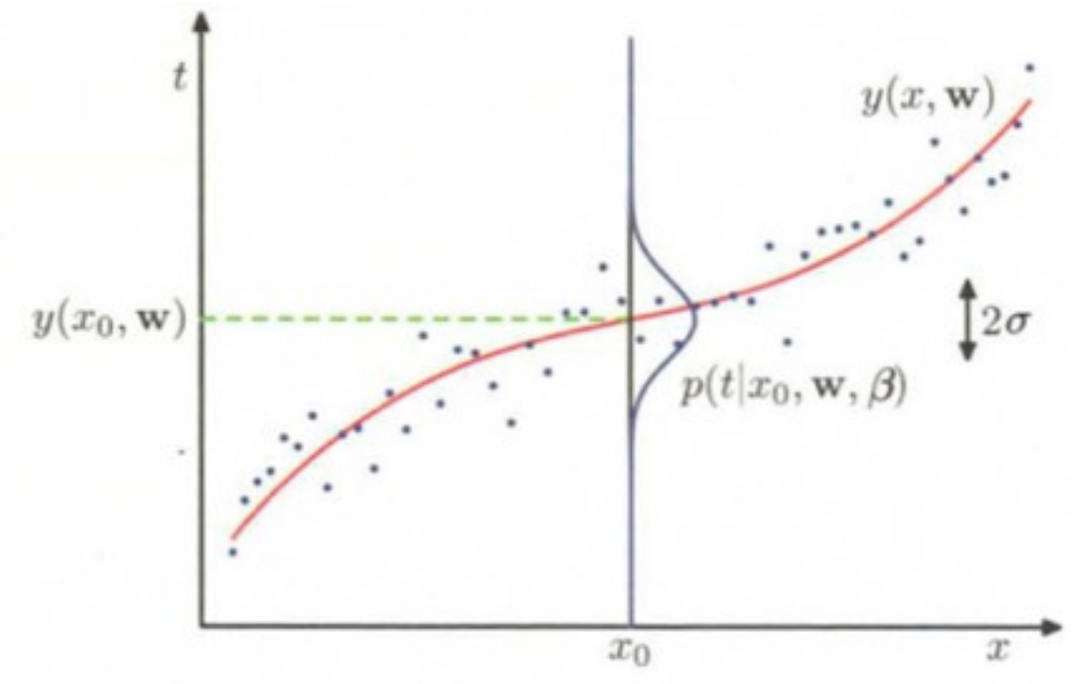
$$y(x, \mathbf{w}) = \sum_{j=0}^{M} w_j x^j$$

に等しいガウス分布に従うとする。
つまり

$$p(t|x, \mathbf{w}, \beta) = \mathcal{N}(t|y(x, \mathbf{w}), \beta^{-1})$$

- β は精度パラメータ。
  分散の逆数





最尤推定

- 訓練データ $\{\mathbf{x}, \mathbf{t}\}$ を使ってパラメータ $\mathbf{w}, \beta$ を最尤推定で求める
- 尤度関数　$p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta) = \prod_{n=1}^{N} \mathcal{N}(t_n|y(x_n, \mathbf{w}), \beta^{-1})$　を最大化
- 対数をとって

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

  - $\mathbf{w}$ についての最大化は実は二乗和誤差の最小化と同じ
    - 二乗和誤差関数はノイズがガウス分布に従うという仮定のもとで尤度の最大化の結果とみなせる
  - β について最大化すると　$\frac{1}{\beta_{ML}} = \frac{1}{N} \sum_{n=1}^{N} \{y(x_n, \mathbf{w}_{ML}) - t_n\}^2$

分布を予測

- $\mathbf{w}, \beta$ が決まれば新しい x に対して予測できる
  - 確率モデルで定式化したので**予測分布**で与えられる

$$p(t|x, \mathbf{w}_{ML}, \beta_{ML}) = \mathcal{N}(t|y(x, \mathbf{w}_{ML}), \beta_{ML}^{-1})$$

- よりベイズ的なアプローチ。簡単に事前分布としてガウス分布

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

を考える。α は分布の精度パラメータ、$M+1$ は $\mathbf{w}$ の要素数

  - α のようにモデルパラメータの分布を制御するパラメータを**超パラメータ**と呼ぶ

- ベイズの定理から $\mathbf{w}$ の事後分布$\propto$事前分布$\times$尤度関数

$$p(\mathbf{w}|\mathbf{x}, \mathbf{t}, \alpha, \beta) \propto p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta)p(\mathbf{w}|\alpha)$$

  - これで与えられたデータに基いて最も確からしい $\mathbf{w}$ を見つける、つまり事後分布を最大化する $\mathbf{w}$ を決めることができる
    - **最大事後確率推定 (MAP推定)** という
  - 実際この事後確率の最大値は

$$\frac{\beta}{2}\sum_{n=1}^{N}\{y(x_n, \mathbf{w}) - t_n\}^2 + \frac{\alpha}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w}$$

    の最小値として与えられる
    - スライド17枚目、正則化最小二乗和誤差の最小化と同じ

1.2.6 ベイズ曲線フィッティング

- 今のところ $\mathbf{w}$ の点推定を行なっているだけでベイズ的でない
- 完全なベイズアプローチでは確率の加法・乗法定理を矛盾なく適用して $\mathbf{w}$ のすべての値に関して積分する必要がある
- 予測分布 $p(t|x, \mathbf{x}, \mathbf{t}) = \int p(t|x, \mathbf{w})p(\mathbf{w}|\mathbf{x}, \mathbf{t})d\mathbf{w}$
  - ただし $p(t|x, \mathbf{w}) = \mathcal{N}(t|y(x, \mathbf{w}), \beta^{-1})$
  - $p(\mathbf{w}|\mathbf{x}, \mathbf{t})$ は事後分布 $p(\mathbf{w}|\mathbf{x}, \mathbf{t}, \alpha, \beta) \propto p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta)p(\mathbf{w}|\alpha)$

ベイズ曲線フィッティング

- 先の積分は解析的に計算できて、結局予測分布は

$$p(t|x, \mathbf{x}, \mathbf{t}) = \mathcal{N}(t|m(x), s^2(x))$$

  - $m(x) = \beta \boldsymbol{\phi}(x)^{\mathrm{T}} \mathbf{S} \sum_{n=1}^{N} \boldsymbol{\phi}(x_n) t_n$
  - $s^2(x) = \beta^{-1} + \boldsymbol{\phi}(x)^{\mathrm{T}} \mathbf{S} \boldsymbol{\phi}(x)$
  - $\mathrm{S}^{-1} = \alpha \mathrm{I} + \beta \sum_{n=1}^{N} \boldsymbol{\phi}(x_n) \boldsymbol{\phi}(x_n)^{\mathrm{T}}$
  - ベクトル $\boldsymbol{\phi}(x)$ は $\phi_i(x) = x^i \ (i = 0, \cdots, M)$ を満たす

- 分散や平均は x に依存
- 分散の第1項は t のノイズによる不確実性
- 分散の第2項はベイズ的な扱いで出てきたもの。$\mathbf{w}$ の不確実性