訓練集合として，N 個の観測地 x を並べた $\boldsymbol{x} := (x_1, \cdots, x_n)^T$ とそれぞれ対応する観測値 t を並べた $\boldsymbol{t} := (t_1, \cdots, t_N)^N$ が与えられたとする．そして目標データ集合 $\boldsymbol{t}$ は，まず $\sin(2\pi x)$ の関数値を作成したのち，ガウス分布に従う小さなランダムノイズを加えて対応する t_n を作った．すなわち

$$t_n \sim sin(2\pi x_n) + \varepsilon \quad \varepsilon \sim N(0, \sigma^2)$$

となるデータを作成している．このようにして生成されたデータは多くの現実データ集合の持つ性質をよく表している．すなわち，データはこれから学習しようとする規則性を保持してはいるが，それぞれの観測値はランダムノイズによって不正確なものになっている．このノイズは，放射性崩壊のように，本質的に確率的なランダムプロセスによる場合もあるが，多くはそれ自身は観測されない信号源の変動によるものである．

我々の木費用は，この訓練集合を利用して，新たな入力変数の値 $\hat{x}$ に対して $\hat{t}$ の値を予測することである．後で見るように，これは背後にある関数 $\sin(2\pi x)$ を暗に見つけようとすることとほぼ等価であるが，有限個のデータ集合から汎化しなければならない点で，本質的に難しい問題である．さらに，観測データはノイズが乗っており，与えられた $\hat{x}$ に対する $\hat{t}$ の値には不確実性がある．1.2 説では，そのような不確実性を厳密かつ定量的に評価する枠組みを与える．また 1.5 説で議論する決定理論は，確率論的な枠組みを利用して，適切な基準の下での最適な予測をすることを可能にする．

ただしここでは話を先に進めるために，曲線フィッティングに基づく単純なアプローチを，あまり形式ばらない形で考えよう．ここでは特に，以下のような多項式を使ってデータへのフィッティングを行うことにする．

$$y(x, \boldsymbol{w}) = w_0 + w_1 x + w_2 x^2 + \cdots + w_M x^M = \sum_{j=0}^{M} w_j x^j \tag{1}$$

ただし，M は多項式の次数で，x^j は x の j 乗を表す．多項式 $y(x, \boldsymbol{w})$ は x の非線形関数であるものの，係数 $\boldsymbol{w}$ の線形関数であることに注意する．すなわち，相異なる x_1, x_2 に対して

$$y(x_1, \boldsymbol{w}) + y(x_2, \boldsymbol{w}) = y(x_1 + x_2, \boldsymbol{w})$$

は常には成り立たないものの，相異なる $\boldsymbol{w}_1, \boldsymbol{w}_2$ に対しては

$$y(x, \boldsymbol{w}_1) + y(x, \boldsymbol{w}_2) = y(x, \boldsymbol{w}_1 + \boldsymbol{w}_2) \tag{2}$$

が成立する．多項式のように，未知のパラメータに関して線形であるような関数は非常に重要な性質を持つ．それらは線形モデルと呼ばれ，のちの章で詳細に議論する．
訓練データに多項式を当てはめることで係数の値を求めてみよう．これは $\boldsymbol{w}$ を任意に固定した時の関数 $y(x, \boldsymbol{w})$ の値と訓練集合のデータ点との間のずれを測る誤差関数の最小化で達

成できる．誤差関数の選び方として，単純で広く用いられているのは，各データ点 x_n における予測値 $y(x_n, \boldsymbol{w})$ と対応するもく票値 t_n との二乗和誤差

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

となり，これを最小化することになる．のちにこの関数を選ぶ理由について議論するが，利点の 1 つは，これが微分可能な凸関数であることにある．また，上の関数が 0 となるのは $y(x, \boldsymbol{w}_n)$ が全訓練データを通るとき，かつその時に限ることに注意する．