Stock Price Prediction

- Problems in which $t_i$ is continuous are called regression.
- E.g. $t_i$ is stock price, $x_i$ contains company profit, debt, cash flow, gross sales, number of spam emails sent, ...
• Only $x_i$ is defined: unsupervised learning
• E.g. $x_i$ describes image, find groups of similar images
• Suppose we are given training set of $N$ observations $(x_1, \ldots, x_N)$ and $(t_1, \ldots, t_N)$, $x_i, t_i \in \mathbb{R}$

• Regression problem, estimate $y(x)$ from these data
Polynomial Curve Fitting

- What form is $y(x)$?
  - Let’s try polynomials of degree $M$:
    
    $$y(x, w) = w_0 + w_1 x + w_2 x^2 + \ldots + w_M x^M$$

  - This is the hypothesis space.

- How do we measure success?
  - Sum of squared errors:
    
    $$E(w) = \frac{1}{2} \sum_{n=1}^{N} \{y(x_n, w) - t_n\}^2$$

- Among functions in the class, choose that which minimizes this error.
Polynomial Curve Fitting

- What form is \( y(x) \)?
  - Let’s try polynomials of degree \( M \):
    \[
y(x, w) = w_0 + w_1 x + w_2 x^2 + \ldots + w_M x^M
    \]
  - This is the hypothesis space.

- How do we measure success?
  - Sum of squared errors:
    \[
    E(w) = \frac{1}{2} \sum_{n=1}^{N} \{ y(x_n, w) - t_n \}^2
    \]
  - Among functions in the class, choose that which minimizes this error.
Polynomial Curve Fitting

- What form is $y(x)$?
  - Let’s try polynomials of degree $M$:
    $$y(x, w) = w_0 + w_1 x + w_2 x^2 + \ldots + w_M x^M$$
    - This is the hypothesis space.

- How do we measure success?
  - Sum of squared errors:
    $$E(w) = \frac{1}{2} \sum_{n=1}^{N} \left( y(x_n, w) - t_n \right)^2$$
  - Among functions in the class, choose that which minimizes this error.
Polynomial Curve Fitting

- Error function
  \[ E(w) = \frac{1}{2} \sum_{n=1}^{N} \{y(x_n, w) - t_n\}^2 \]

- Best coefficients
  \[ w^* = \arg \min_w E(w) \]

- Found using pseudo-inverse (more later)
Which Degree of Polynomial?

- A model selection problem
- $M = 9 \rightarrow E(w^*) = 0$: This is over-fitting
Generalization

- Generalization is the holy grail of ML
  - Want good performance for new data
- Measure generalization using a separate set
  - Use root-mean-squared (RMS) error: $E_{RMS} = \sqrt{\frac{2E(w^*)}{N}}$