Parametric Methods

CptS 570 – Machine Learning
School of EECS
Washington State University
Assume a model of the underlying distribution $p(x|\theta)$

Estimate the parameters $\theta$ of the model based on the training set $X$

Regression

Bias/variance dilemma

Model selection
Maximum Likelihood Estimation

- **Likelihood** of \( \theta \) given the sample \( X \)
  \[
  l(\theta|X) = p(X|\theta) = \prod_t p(x^t|\theta)
  \]

- **Log likelihood**
  \[
  L(\theta|X) = \log l(\theta|X) = \sum_t \log p(x^t|\theta)
  \]

- **Maximum likelihood estimator (MLE)**
  \[
  \theta^* = \arg\max_{\theta} L(\theta|X)
  \]
\( p(x) = \mathcal{N}(\mu, \sigma^2) \)

\[
p(x) = \frac{1}{\sqrt{2\pi\sigma}} \exp\left[ -\frac{(x-\mu)^2}{2\sigma^2} \right]
\]

- \( \mu \) and \( \sigma^2 \) MLE:
  \[
  m = \frac{\sum x^t}{N}
  \]
  \[
  s^2 = \frac{\sum (x^t - m)^2}{N}
  \]
Bayes’ Estimator

- Treat $\theta$ as a random variable with prior $p(\theta)$
- Bayes’ rule: $p(\theta | X) = \frac{p(X | \theta) p(\theta)}{p(X)}$
- Maximum a Posteriori (MAP):
  - $\theta_{\text{MAP}} = \arg\max_{\theta} p(\theta | X)$
- Maximum Likelihood (ML):
  - $\theta_{\text{ML}} = \arg\max_{\theta} p(X | \theta)$
Regression

\[ r = f(x) + \varepsilon \]

estimator: \[ g(x | \theta) \]

\[ \varepsilon \sim \mathcal{N}(0, \sigma^2) \]

\[ p(r | x) \sim \mathcal{N}(g(x | \theta), \sigma^2) \]

\[ \mathcal{L}(\theta | X) = \log \prod_{t=1}^{N} p(x^t, r^t) \]

\[ = \log \prod_{t=1}^{N} p(r^t | x^t) + \log \prod_{t=1}^{N} p(x^t) \]
Regression: From LogL to Error

\[ \mathcal{L}(\theta | X) = \log \prod_{t=1}^{N} \frac{1}{\sqrt{2\pi \sigma}} \exp \left[ -\frac{(r^t - g(x^t | \theta))^2}{2\sigma^2} \right] \]

\[ = -N \log \sqrt{2\pi \sigma} - \frac{1}{2\sigma^2} \sum_{t=1}^{N} [r^t - g(x^t | \theta)]^2 \]

\[ E(\theta | X) = \frac{1}{2} \sum_{t=1}^{N} [r^t - g(x^t | \theta)]^2 \]
Linear Regression

\[ g(x^t | w_1, w_0) = w_1 x^t + w_0 \]

\[ \sum_t r^t = N w_0 + w_1 \sum_t x^t \]

\[ \sum_t r^t x^t = w_0 \sum_t x^t + w_1 \sum_t (x^t)^2 \]

\[ A = \begin{bmatrix} N & \sum_t x^t \\ \sum_t x^t & \sum_t (x^t)^2 \end{bmatrix}, \quad w = \begin{bmatrix} w_0 \\ w_1 \end{bmatrix}, \quad y = \begin{bmatrix} \sum_t r^t \\ \sum_t r^t x^t \end{bmatrix} \]

\[ w = A^{-1} y \]
Polynomial Regression

\[ g(x^t | w_k, \ldots, w_2, w_1, w_0) = w_k (x^t)^k + \cdots + w_2 (x^t)^2 + w_1 x^t + w_0 \]

\[ D = \begin{bmatrix} 1 & x^1 & (x^1)^2 & \cdots & (x^1)^k \\ 1 & x^2 & (x^2)^2 & \cdots & (x^2)^k \\ \vdots \\ 1 & x^N & (x^N)^2 & \cdots & (x^N)^k \end{bmatrix} \]

\[ r = \begin{bmatrix} r^1 \\ r^2 \\ \vdots \\ r^N \end{bmatrix} \]

\[ w = (D^T D)^{-1} D^T r \]
Estimating Bias and Variance

- $M$ samples $X_i = \{x^t_i, \ r^t_i\}, \ i=1,...,M$
  are used to fit $g_i(x), \ i=1,...,M$

\[\text{Bias}^2(g) = \frac{1}{N} \sum_t \left[ \bar{g}(x^t) - f(x^t) \right]^2\]

\[\text{Variance}(g) = \frac{1}{NM} \sum_t \sum_i \left[ g_i(x^t) - \bar{g}(x^t) \right]^2\]

\[\bar{g}(x) = \frac{1}{M} \sum_t g_i(x)\]
Bias/Variance Dilemma

- Example
  - $g_f(x) = 2$ has no variance and high bias
  - $g_f(x) = \sum_t r_t / N$ has lower bias with variance

- As we increase complexity,
  - Bias decreases (a better fit to data) and
  - Variance increases (fit varies more with data)

- Bias/Variance dilemma: (Geman et al., 1992)
Example

- \( f(x) = 2 \sin (1.5x) \)
- Noise \( N(0,1) \)
- Five samples taken (one below)
- Fit order 1, 3 and 5 polynomials
(a) Function and data

(b) Order 1

(c) Order 3

(d) Order 5

\[ f \]

\[ g_i \]

\[ \bar{g} \]

bias

variance
Polynomial Regression

Best fit “min error”

- Bias
- Error
- Variance
Model Selection

- **Cross-validation**
  - Measure generalization accuracy by testing on data unused during training (validation set)

- **Regularization**
  - Penalize complex models
  - $E' = \text{error on data} + \lambda \times \text{model complexity}$

- **Minimum description length (MDL)**
  - Best model minimizes description of model plus description of data given model
(a) Data and fitted polynomials

$$f(x) = 2 \times \sin(1.5x)$$

Order 1–8 polys

(b) Error vs polynomial order

*Best fit, “elbow”*
Coefficients increase in magnitude as order increases:
1: \([-0.0769, 0.0016]\]
2: \([0.1682, -0.6657, 0.0080]\]
3: \([0.4238, -2.5778, 3.4675, -0.0002]\]
4: \([-0.1093, 1.4356, -5.5007, 6.0454, -0.0019]\]

\[
\text{regularization}: E(w \mid X) = \frac{1}{2} \sum_{t=1}^{N} \left[ r^t - g(x^t \mid w) \right]^2 + \lambda \sum_{i} w_i^2
\]
Model Selection

- Bayesian model selection

\[
p(\text{model} | \text{data}) = \frac{p(\text{data} | \text{model}) p(\text{model})}{p(\text{data})}
\]

- \[ \log p(\text{model} | \text{data}) = \log p(\text{data} | \text{model}) + \log p(\text{model}) - \log p(\text{data}) \]

- Average over a number of models with high posterior (voting, ensembles: Chapter 17)
Summary: Parametric Methods

- Select model
  - Tradeoff bias and variance
- Estimate model parameters from data
  - Regression