Parametric Methods (cont.)

CAI 5107: Machine Learning

Instructor: Anowarul Kabir

Email: akabir@usf.edu

Fall 2025

Parametric classification

- From Baye's rule: $P(C_i|x) = \frac{p(x|C_i)P(C_i)}{p(x)}$
- The discriminant function: $g_i(x) = p(x|C_i)P(C_i)$
- Assume, $p(x|C_i)$ is Gaussian: $p(x|C_i) = \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left[-\frac{(x-\mu_i)^2}{2\sigma_i^2}\right]$
- We can write our discriminant as:

$$g_i(x) = -\frac{1}{2}\log 2\pi - \log \sigma_i - \frac{(x-\mu_i)^2}{2\sigma_i^2} + \log P(C_i)$$

An example

- Problem: Selling K different cars (C_1, …, C_K) given the yearly income (x) of the customers
- $P(C_i)$ ~ #-of customers bought car type *i*
- If such customers yearly income distribution can be approximated by Gaussian, $p(x|C_i) \sim \mathcal{N}(\mu_i, \sigma_i^2)$
  - Where μ_i is the mean income and σ_i^2 is the income variance

- We can plug-in those values in the discriminant function to do K-class classification

Extending previous example

- If we do not know $P(C_i)$ and $p(x|C_i)$, we will estimate from sample.
- Given a sample: $\mathcal{X} = \{x^t, \boldsymbol{r}^t\}_{t=1}^N$
  - Where r^t_i is defined as $$r_i^t = \begin{cases} 1 & \text{if } \boldsymbol{x}^t \in C_i \\ 0 & \text{if } \boldsymbol{x}^t \in C_k, k \neq i \end{cases}$$
- The means and variances for each class will be:

$$m_i = \frac{\sum_t x^t r_i^t}{\sum_t r_i^t} \qquad s_i^2 = \frac{\sum_t (x^t - m_i)^2 r_i^t}{\sum_t r_i^t}$$

- The prior estimate is: $\hat{P}(C_i) = \frac{\sum_t r_i^t}{N}$

Extending previous example (cont.)

- The discriminant function with the estimates will be:

$$g_i(x) = -\frac{1}{2}\log 2\pi - \log s_i - \frac{(x - m_i)^2}{2s_i^2} + \log \hat{P}(C_i)$$

- We can simplify $g_i(x) = -(x - m_i)^2$
  - When class priors and variances are equal
- So, our classifier will be:

$$\text{Choose } C_i \text{ if } |x - m_i| = \min_k |x - m_k|$$

- For 2 classes:

$$\begin{aligned} (x - m_1)^2 &= (x - m_2)^2 \\ x &= \frac{m_1 + m_2}{2} \end{aligned}$$

What ifs?

- When priors and variances are unequal for 2-classes:

$$p(x|C_1) \sim \mathcal{N}(\mu_1, \sigma_1^2) \text{ and } p(x|C_2) \sim \mathcal{N}(\mu_2, \sigma_2^2)$$

- The discriminant point is *x* that satisfies:

$$P(C_1|x) = P(C_2|x)$$

- Solve this at home. You will find something in the format of
  - $ax^2+bx+c = 0$
  - Then, we compute for the roots as $x_1, x_2 = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$

Regression

- Generic regression function: $r = f(x) + \epsilon$
- We will estimate f(x) with an estimator $g(x|\theta)$
  - Where θ is the set of parameters

- Assuming, $\epsilon \sim \mathcal{N}(0, \sigma^2)$, our estimator will be

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

Regression (cont.)

- Using MLE to learn parameters, we can write
  - Where $\mathcal{X} = \{x^t, r^t\}_{t=1}^N$ is IID

$$\begin{aligned}\mathcal{L}(\theta|\mathcal{X}) &= \log \prod_{t=1}^N p(x^t, r^t) \\ &= \log \prod_{t=1}^N p(r^t|x^t) + \log \prod_{t=1}^N p(x^t)\end{aligned}$$

- Doing further simplification:

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

Least squares estimate:

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

Linear regression

- For linear regression, we have a linear model:

$$g(x^t | w_1, w_0) = w_1 x^t + w_0$$

- Take derivatives w.r.t. w_0 and w_1

$$\sum_t r^t = N w_0 + w_1 \sum_t x^t$$

$$\sum_t r^t x^t = w_0 \sum_t x_t + w_1 \sum_t (x^t)^2$$

- We can write this in vector-matrix format and solve for ***w=A^{-1}y***

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

Polynomial regression

- In polynomial regression, the model is polynomial in x of order k

$$g(x^t|w_k,\ldots,w_2,w_1,w_0) = w_k(x^t)^k + \cdots + w_2(x^t)^2 + w_1x^t + w_0$$

- Note: still linear w.r.t parameters

$$\mathbf{A} = \begin{bmatrix} N & \sum_t x^t & \sum_t (x^t)^2 & \cdots & \sum_t (x^t)^k \\ \sum_t x^t & \sum_t (x^t)^2 & \sum_t (x^t)^3 & \cdots & \sum_t (x^t)^{k+1} \\ \vdots & & & & \\ \sum_t (x^t)^k & \sum_t (x^t)^{k+1} & \sum_t (x^t)^{k+2} & \cdots & \sum_t (x^t)^{2k} \end{bmatrix}$$

$$\boldsymbol{w} = \begin{bmatrix} w_0 \\ w_1 \\ w_2 \\ \vdots \\ w_k \end{bmatrix}, \boldsymbol{y} = \begin{bmatrix} \sum_t r^t \\ \sum_t r^t x^t \\ \sum_t r^t (x^t)^2 \\ \vdots \\ \sum_t r^t (x^t)^k \end{bmatrix}$$

We can write $\mathbf{A} = \mathbf{D}^T\mathbf{D}$ and $\boldsymbol{y} = \mathbf{D}^T\boldsymbol{r}$ where

$$\mathbf{D} = \begin{bmatrix} 1 & x^1 & (x^1)^2 & \cdots & (x^1)^k \\ 1 & x^2 & (x^2)^2 & \cdots & (x^2)^k \\ \vdots & & & & \\ 1 & x^N & (x^N)^2 & \cdots & (x^N)^k \end{bmatrix}, \boldsymbol{r} = \begin{bmatrix} r^1 \\ r^2 \\ \vdots \\ r^N \end{bmatrix}$$

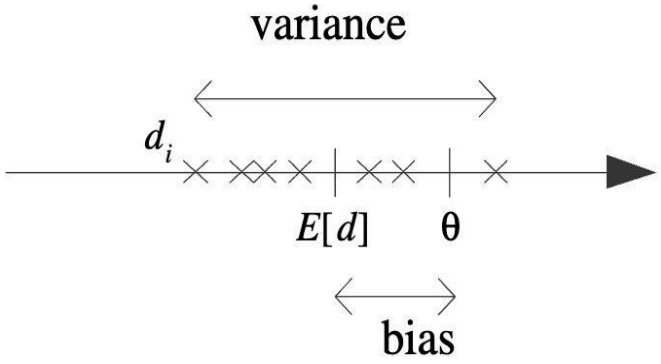
and we can then solve for the parameters as

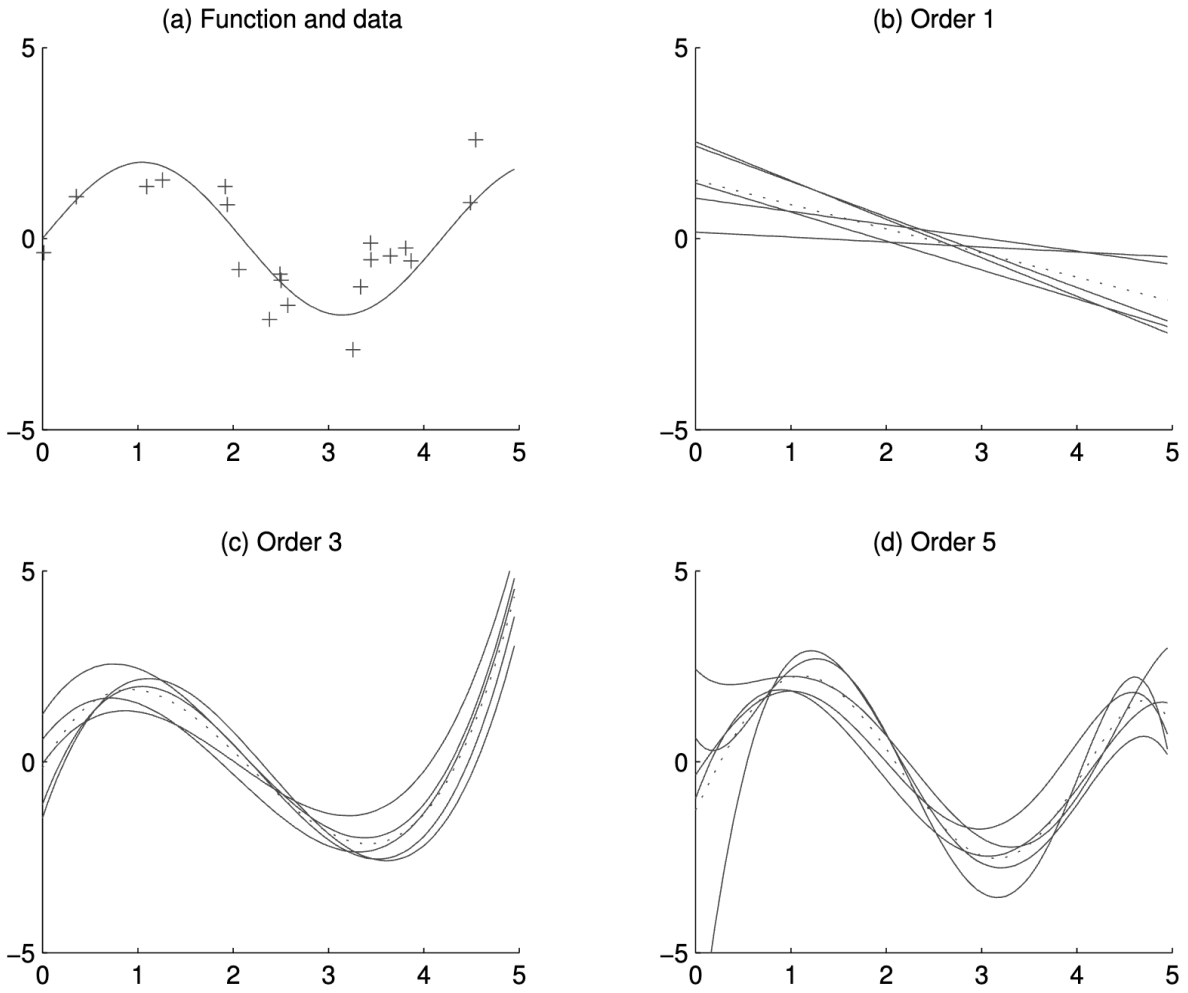
$$\boldsymbol{w} = (\mathbf{D}^T\mathbf{D})^{-1}\mathbf{D}^T\boldsymbol{r}$$

Tuning model complexity: Bias/Variance Dilemma

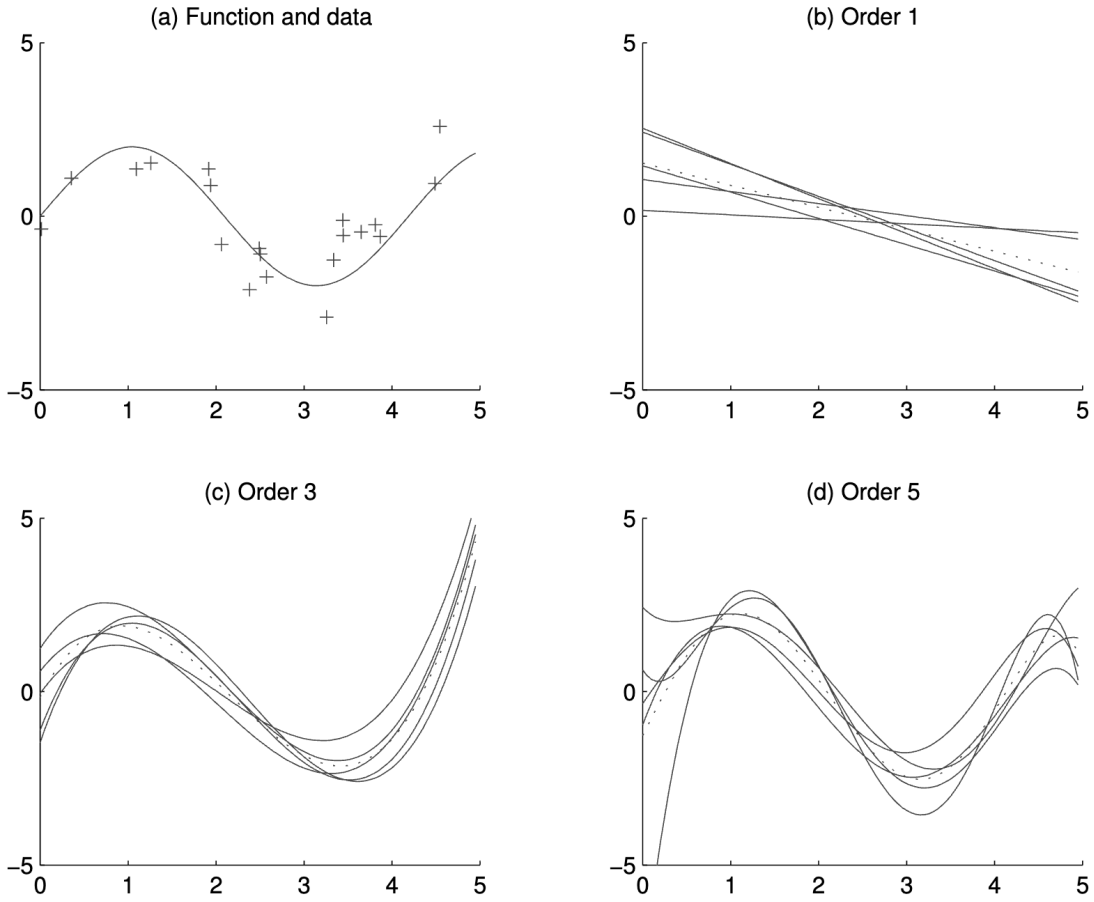
- **Variance**: measures how much, on average, d_i vary around the expected value (from one dataset to another)

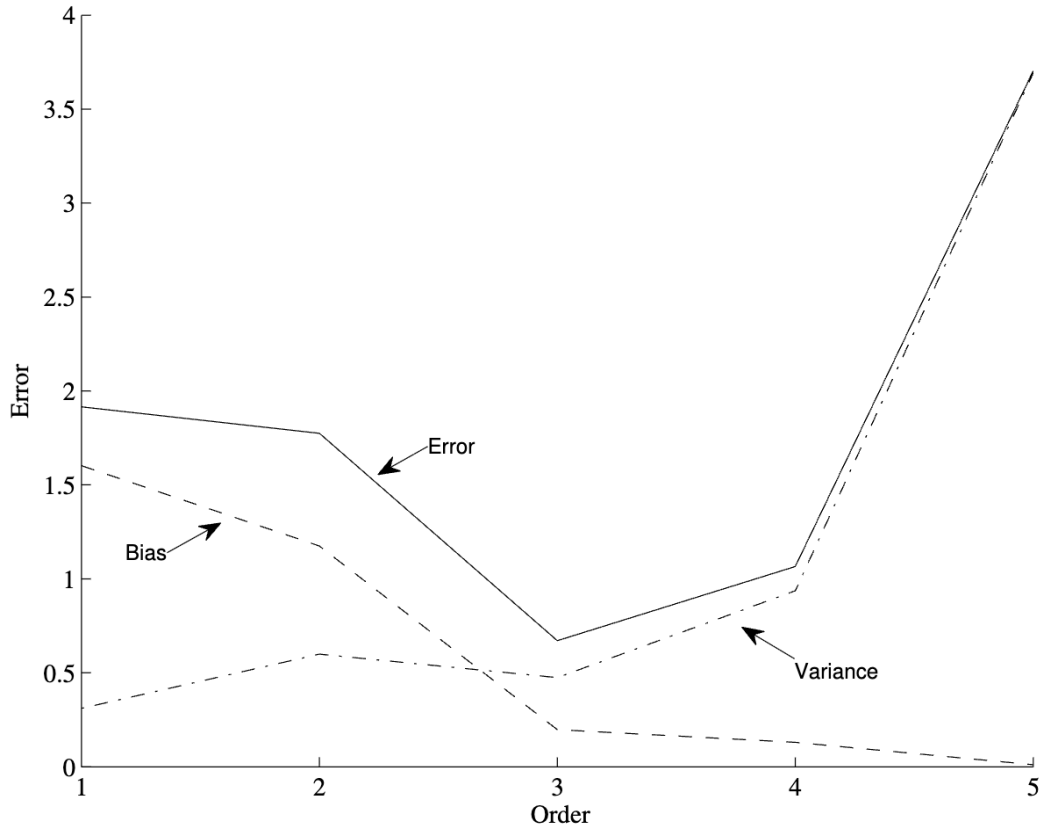
- **Bias**: measures how much the expected value varies from the correct value of θ





Tuning model complexity: Bias/Variance Dilemma

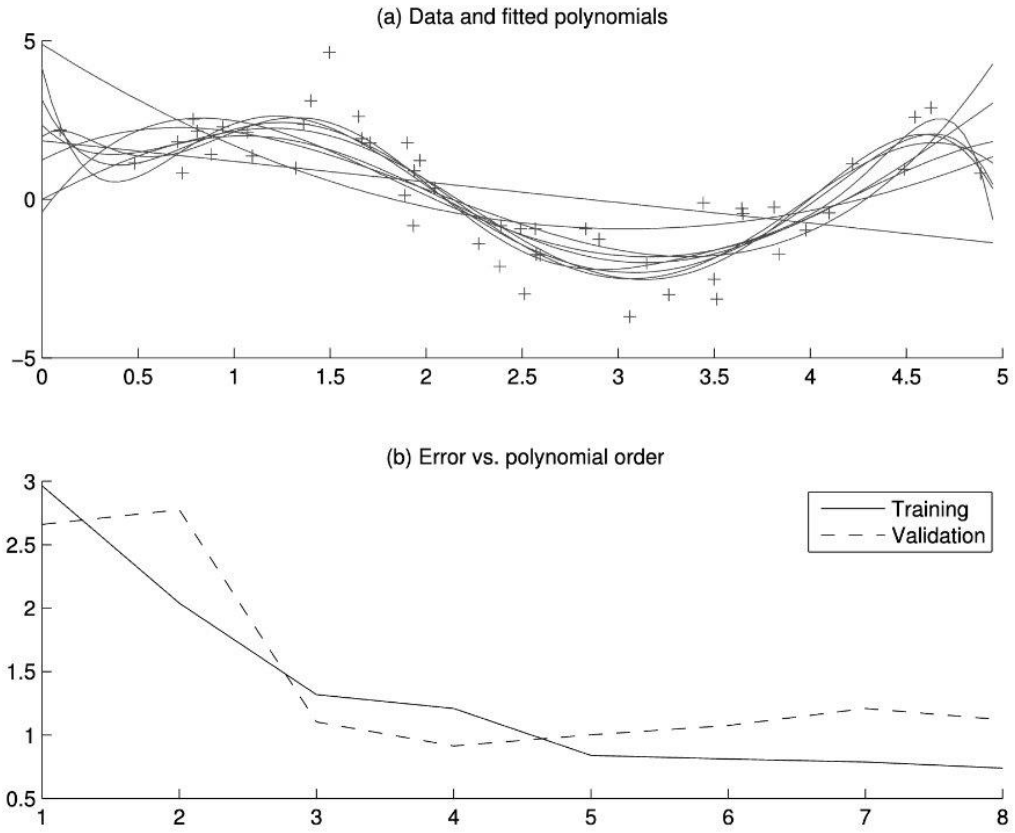




Model selection

- Cross validation
  - Elbow method for best model complexity
- Regularization

$$E' = \text{error on data} + \lambda \cdot \text{model complexity}$$



Reading materials

- Chapter 4: Parametric Methods
  - 4.5 - 4.8