CSE 405: Machine Learning

Parametric Methods

Dr Muhammad Abul Hasan

Department of Computer Science and Engineering
Green University of Bangladesh
muhammad.hasan@cse.green.edu.bd

Fall 2023

Outline

“There are no secrets to success. It is the result of preparation, hard work, and learning from failure.

– Colin Powell”

Parametric Estimation

Parametric Estimation

- $X = \{x^t\}_t$ where $x^t \sim p(x)$
- Parametric estimation: Assume a form for $p(x|q)$ and estimate q, its sufficient statistics, using X e.g., $N(\mu, \sigma^2)$ where $q = \{\mu, \sigma^2\}$

Maximum Likelihood Estimation

- Likelihood of q given the sample $\mathcal{X}$

$$l(\theta|\mathcal{X}) = p(\mathcal{X}|\theta) = \Pi_t p(x^t|\theta)$$

- Log-likelihood:

$$\mathcal{L}(\theta|\mathcal{X}) = \log l(\theta|\mathcal{X}) = \Sigma_t \log p(x^t|\theta)$$

- Maximum likelihood estimator (MLE):

$$\theta^* = \arg\max_{\theta} \mathcal{L}(\theta|\mathcal{X})$$

Examples: Bernoulli/Multinomial

- Bernoulli: Two states, failure/success: $x \in \{0, 1\}$

$$p(x) = (p_o)^x \cdot (1 - p_o)^{(1-x)}$$

$$\mathcal{L}(p_o|\mathcal{X}) = \log \Pi_t (p_o)^{x^t} \cdot (1 - p_o)^{\left(1-x^t\right)}$$

$$MLE : p_o = \sum_t \frac{x^t}{N}$$

- Multinomial: $K > 2$ states, $x_i \in \{0, 1\}$

$$P(x_1, x_2, \ldots, x_K) = \Pi_i (p_i)^{x_i}$$

$$\mathcal{L}(p_1, p_2, \ldots, p_K|\mathcal{X}) = \log \Pi_t \Pi_i (p_i)^{x_i^t}$$

$$MLE : p_i = \sum_t \frac{x_i^t}{N}$$

Gaussian (Normal) Distribution

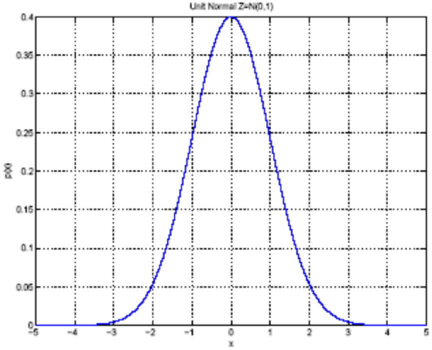
- $p(x) = \mathcal{N}(\mu, \sigma^2)$

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

- MLE for μ and σ^2:

$$\mu = \frac{\sum_t x^t}{N}$$

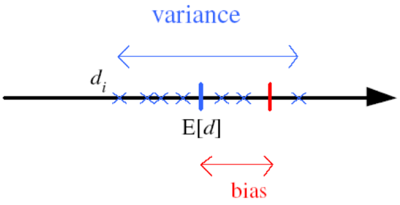
$$\sigma^2 = \frac{\sum_t (x^t - \mu)^2}{N}$$



Bias and Variance

- Unknown parameter θ Estimator $d_i = d(X_i)$ on sample X_i
- Bias: $b_\theta(d) = E[d] - \theta$
- Variance: $E[(d - E[d])^2]$
- Mean square error:

$$
\begin{aligned}
r(d, \theta) &= E[(d - \theta)^2] \\
&= (E[d] - \theta)^2 + E\left[(d - E[d])^2\right] \\
&= \text{Bias}^2 + \text{Variance}
\end{aligned}
$$



Bayes' Estimator

Bayes' Estimator

- Treat θ as a random variable with prior $p(\theta)$
- Bayes' rule: $p(\theta|X) = \frac{p(X|\theta)p(\theta)}{p(X)}$
- Full: $p(x|\mathcal{X}) = \int p(x|\theta)p(\theta|\mathcal{X})d\theta$
- Maximum a posteriori (MAP): $\theta_{MAP} = \arg\max_\theta p(\theta|\mathcal{X})$
- Maximum Likelihood (ML): $\theta_{ML} = \arg\max_\theta p(\mathcal{X}|\theta)$
- Bayes': $\theta_{Bayes'} = E[\theta|\mathcal{X}] = \int \theta p(\theta|\mathcal{X})d\theta$

Bayes' Estimator: Example

- $x^t \sim \mathcal{N}(\theta, \sigma_o^2)$ and $\theta \sim \mathcal{N}(\mu, \sigma^2)$
- $\theta_{ML} = m$
- $\theta_{MAP} = \theta_{\text{Bayes'}} = E[\theta|\mathcal{X}] = \frac{N/\sigma_o^2}{N/\sigma_o^2 + 1/\sigma^2} m + \frac{1/\sigma^2}{N/\sigma_o^2 + 1/\sigma^2} \mu$

Parametric Classification

Parametric Classification

$$g_i(x) = p(x|C_i)p(C_i)$$

or

$$g_i(x) = \log p(x|C_i) + \log p(C_i)$$

$$p(x|C_i) = \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left[-\frac{(x-\mu_i)^2}{2\sigma_i^2}\right]$$

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

- Given the sample $\mathcal{X} = \{x^t, r^t\}_{t=1}^N$

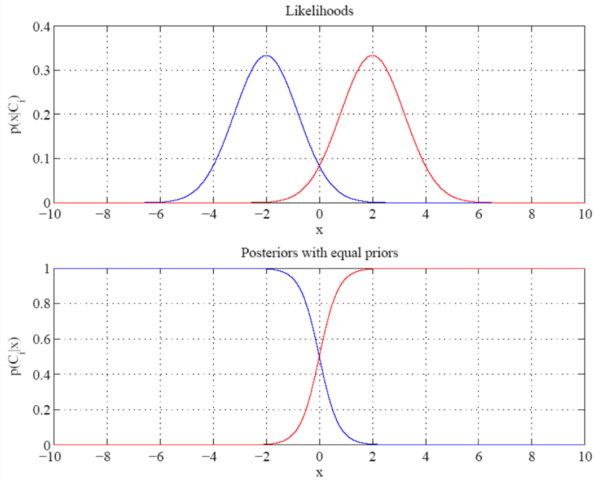
$$x_i^t \in \mathcal{R}, \quad r_i^t = \begin{cases} 1 & \text{if } x_i^t \in C_i \\ 0 & \text{if } x_i^t \in C_j, j \neq i \end{cases}$$

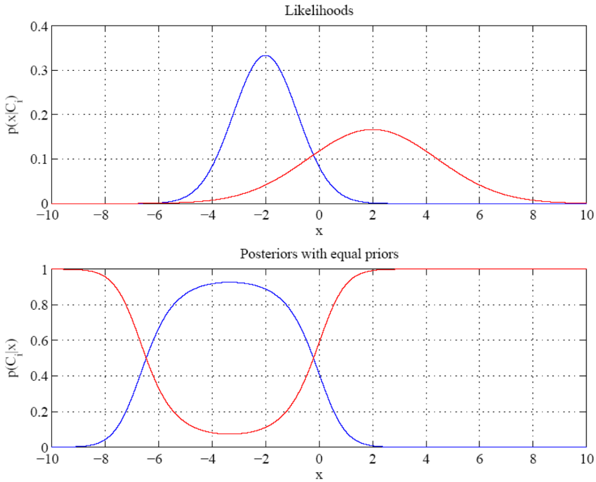
- ML estimates are:

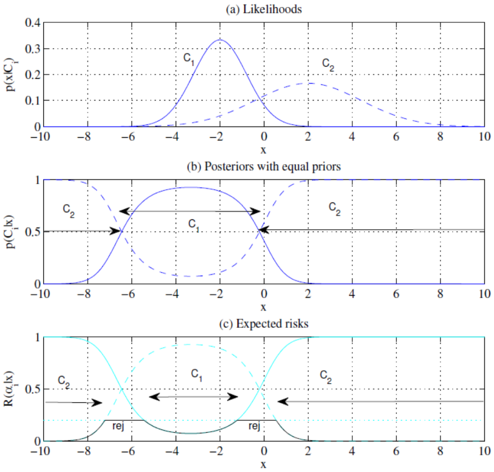
$$\hat{p}(C_i) = \frac{\Sigma_t r_i^t}{N}, \quad m_i = \frac{\Sigma_t x^t r_i^t}{\Sigma_t r_i^t}, \quad s_i^2 = \frac{\sigma_t (x^t - m_i)^2 \cdot r_i^t}{\Sigma_t r_i^t}$$

- Discriminant

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


Likelihoods
p(x|C_i)
0.4
0.3
0.2
0.1
0
−10 −8 −6 −4 −2 0 2 4 6 8 10
x
Posteriors with equal priors
p(C_i|x)
1
0.8
0.6
0.4
0.2
0
−10 −8 −6 −4 −2 0 2 4 6 8 10
x


Likelihoods
p(x|C_i)
0.4
0.3
0.2
0.1
0
−10 −8 −6 −4 −2 0 2 4 6 8 10
x
Posteriors with equal priors
p(C_i|x)
1
0.8
0.6
0.4
0.2
0
−10 −8 −6 −4 −2 0 2 4 6 8 10
x

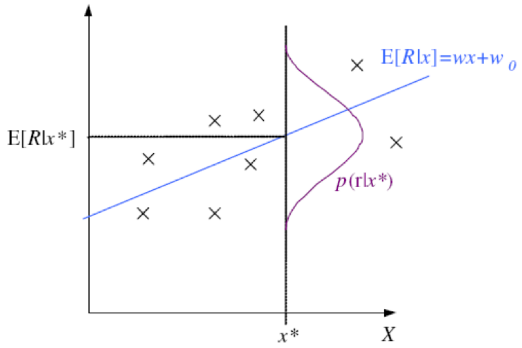

(a) Likelihoods
p(x|C_i)
C_1
C_2
x
(b) Posteriors with equal priors
p(C_i|x)
C_2
C_1
C_2
x
(c) Expected risks
R(α_i|x)
C_2
C_1
C_2
rej
rej
x

Parametric Regression

Regression

- $r = f(x) + \epsilon$
- estimator $g(x|\theta)$
- $\epsilon \sim \mathcal{N}(0, \sigma^2)$
- $p(r|x) \sim \mathcal{N}(g(x|\theta), \sigma^2)$

$$\mathcal{L}(\theta|\mathcal{X}) = \log \Pi_{t=1}^{N} p(x^t, r^t)$$
$$= \log \Pi_{t=1}^{N} p(r^t|x^t) + \log \Pi_{t=1}^{N} p(x^t)$$



Regression: From $\log \mathcal{L}$ to Error

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

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

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

Linear Regression

$$g(x^t|\omega_1, \omega_2) = \omega_1 x^t + \omega_0$$
$$\Sigma_t r^t = N\omega_0 + \omega_1 \Sigma_t x^t$$
$$\Sigma_t r^t x^t = \omega_0 \Sigma_t x^t + \omega_1 \Sigma_t (x^t)^2$$

$$\mathbf{A} = \begin{bmatrix} N & \Sigma_t x^t \\ \Sigma_t x^t & \Sigma_t (x^t)^2 \end{bmatrix} \quad \mathbf{w} = \begin{bmatrix} \omega_0 \\ \omega_1 \end{bmatrix} \quad \mathbf{y} = \begin{bmatrix} \Sigma_t r^t \\ \Sigma_t r^t x^t \end{bmatrix}$$

$$\mathbf{w} = \mathbf{A}^{-1}\mathbf{y}$$

Polynomial Regression

$$g(x^t|\omega_k, \ldots, \omega_2, \omega_1, \omega_0) = \omega_k (x^t)^k + \cdots + \omega_2 (x^t)^2 + \omega_1 x^t + \omega_0$$

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

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

Other Error Measures

- Square Error: $E(\theta|\mathcal{X}) = \frac{1}{2}\Sigma_{t=1}^{N}[r^t - g(x^t|\theta)]^2$
- Relative Square Error: $E(\theta|\mathcal{X}) = \frac{\Sigma_{t=1}^{N}[r^t - g(x^t|\theta)]^2}{\Sigma_{t=1}^{N}[r^t - \bar{r}]^2}$
- Absolute Error: $E(\theta|\mathcal{X}) = \Sigma_t|r^t - g(x^t|\theta)|$
- ϵ-sensitive Error: $E(\theta|\mathcal{X}) = \Sigma_t 1(|r^t - g(x^t|\theta)| > \epsilon)(|r^t - g(x^t|\theta)| - \epsilon)$

Bias and Variance

$$E\big[(r-g(x))^2|x\big] = E\big[(r-E[r|c])^2|x\big] + (E[r|x]-g(x))^2$$

$$E_x\big[(E[r|x]-g(x))^2|x\big] = (E[r|x]-E_x[g(x)])^2 + E_x\big[(g(x)-E_x[g(x)])^2\big]$$

Estimating Bias and Variance

- M samples $X_i = \{x_i^t, r_i^t\}, i = 1, \ldots, M$ are used to fit $g_i(x), i = 1, \ldots, M$

$$\text{Bias}^2(g) = \frac{1}{1}\Sigma_t[\bar{g}(x^t) - f(x^t)]^2$$

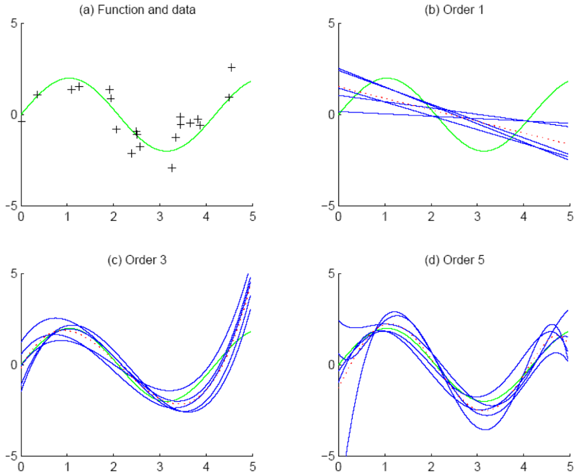
$$\text{Variance}(g) = \frac{1}{MN}\Sigma_t\Sigma_i[g_i(x^t) - (\bar{g}(x^t)]^2$$

$$(\bar{g})(x) = \frac{1}{M}\Sigma_t g_i(x)$$

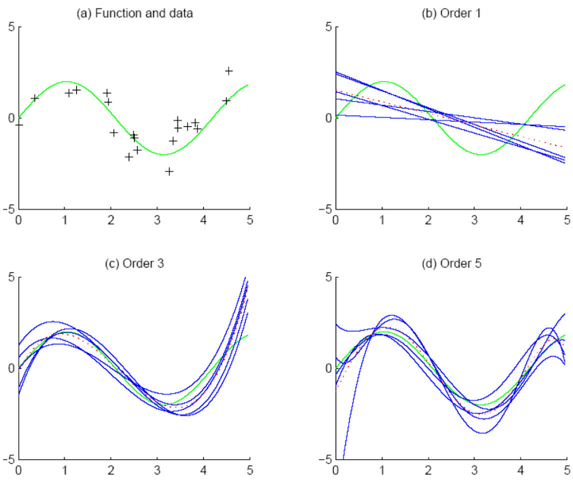
Bias/Variance Dilemma

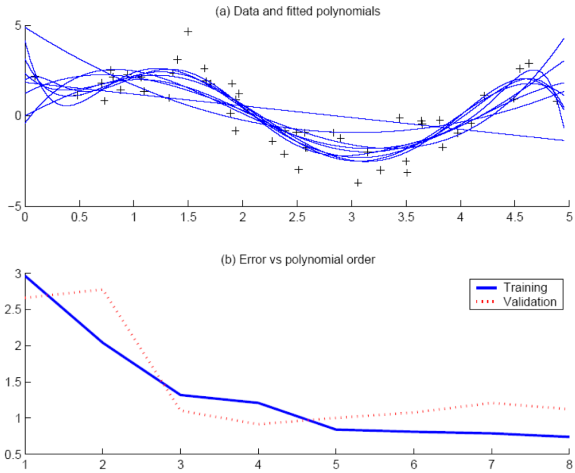
Bias/Variance Dilemma

- Example: $g_i(x) = 2$ has no variance and high bias $g_i(x) = \Sigma_t r_i^t / N$ 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)


(a) Function and data
(b) Order 1
(c) Order 3
(d) Order 5
5
0
−5
0
1
2
3
4
5

Polynomial Regression




(a) Data and fitted polynomials
5
0
−5
0
0.5
1
1.5
2
2.5
3
3.5
4
4.5
5
(b) Error vs polynomial order
Training
Validation
3
2.5
2
1.5
1
0.5
1
2
3
4
5
6
7
8

Model Selection

Model Selection

- Cross-validation: Measure generalization accuracy by testing on data unused during training
- Regularization: Penalize complex models

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

  Akaike's information criterion (AIC), Bayesian information criterion (BIC)
- Minimum description length (MDL): Kolmogorov complexity, shortest description of data
- Structural risk minimization (SRM)

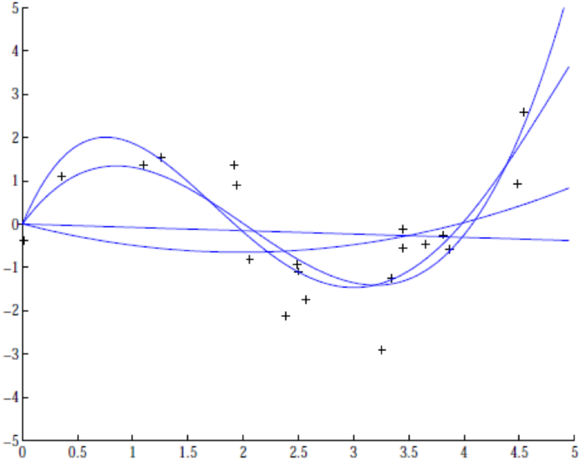
Bayesian Model Selection

- Prior on models, $p(\text{model})$

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

- Regularization, when prior favors simpler models
- Bayes, MAP of the posterior, $p(\text{model}|\text{data})$
- Average over a number of models with high posterior (voting, ensembles: Chapter 17)

Regression example



Thank You!