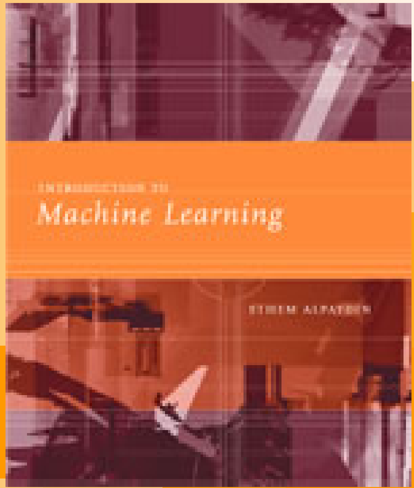


Lecture Slides for

INTRODUCTION TO

Machine Learning

ETHEM ALPAYDIN

alpaydin@boun.edu.tr
http://www.cmpe.boun.edu.tr/~ethem/i2ml

CHAPTER 4:

Parametric Methods

Parametric Estimation

- $\mathcal{X} = \{ x^t \}_t$ where $x^t \sim p(x)$
- Parametric estimation:

  Assume a form for $p(x \mid \theta)$ and estimate θ, its sufficient statistics, using $\mathcal{X}$

  e.g., $\mathcal{N}(\mu, \sigma^2)$ where $\theta = \{ \mu, \sigma^2 \}$

Maximum Likelihood Estimation

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

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

- Log likelihood

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

- Maximum likelihood estimator (MLE)

  $$\theta^* = \text{argmax}_\theta \mathcal{L}(\theta|\mathcal{X})$$

Examples: Bernoulli/Multinomial

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

  $$P(x) = p_o^{x}(1 - p_o)^{(1 - x)}$$

  $$\mathcal{L}(p_o|\mathcal{X}) = \log \prod_t p_o^{x^t}(1 - p_o)^{(1 - x^t)}$$

  MLE: $p_o = \sum_t x^t / N$

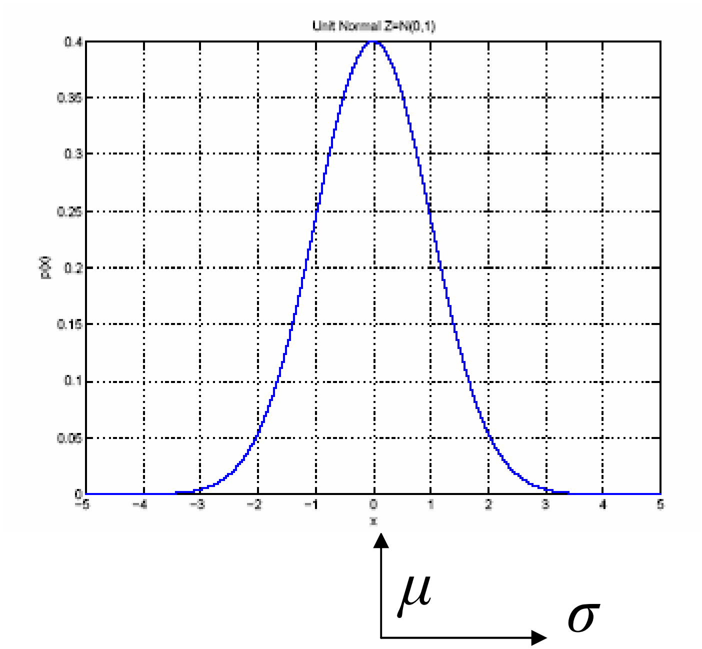
- Multinomial: K>2 states, x_i in {0,1}

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

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

  MLE: $p_i = \sum_t x_i^t / N$

Gaussian (Normal) Distribution



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

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

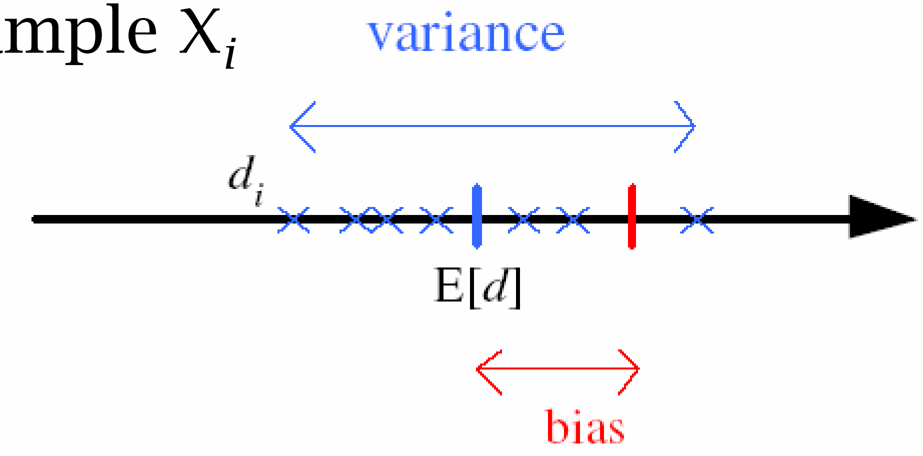
- MLE for μ and σ^2:

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

$$s^2 = \frac{\sum_t (x^t - m)^2}{N}$$

Bias and Variance

Unknown parameter θ
Estimator $d_i = d(\mathcal{X}_i)$ on sample $\mathcal{X}_i$



Bias: $b_\theta(d) = E[d] - \theta$
Variance: $E[(d - E[d])^2]$

Mean square error:

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

Bayes' Estimator

- Treat θ as a random var with prior $p(\theta)$
- Bayes' rule: $p(\theta|\mathcal{X}) = p(\mathcal{X}|\theta)\, p(\theta) / p(\mathcal{X})$
- Full: $p(x|\mathcal{X}) = \int p(x|\theta)\, p(\theta|\mathcal{X})\, d\theta$
- Maximum a Posteriori (MAP): $\theta_{\mathrm{MAP}} = \mathrm{argmax}_\theta\, p(\theta|\mathcal{X})$
- Maximum Likelihood (ML): $\theta_{\mathrm{ML}} = \mathrm{argmax}_\theta\, p(\mathcal{X}|\theta)$
- Bayes': $\theta_{\mathrm{Bayes'}} = E[\theta|\mathcal{X}] = \int \theta\, p(\theta|\mathcal{X})\, d\theta$

Bayes' Estimator: Example

- $x^t \sim \mathcal{N}(\theta, \sigma_o^2)$ and $\theta \sim \mathcal{N}(\mu, \sigma^2)$
- $\theta_{ML} = m$
- $\theta_{MAP} = \theta_{Bayes'} =$

$$E[\theta|\mathcal{X}] = \frac{N/\sigma_0^2}{N/\sigma_0^2 + 1/\sigma^2} m + \frac{1/\sigma^2}{N/\sigma_0^2 + 1/\sigma^2} \mu$$

Parametric Classification

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

or equivalently

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

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

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

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

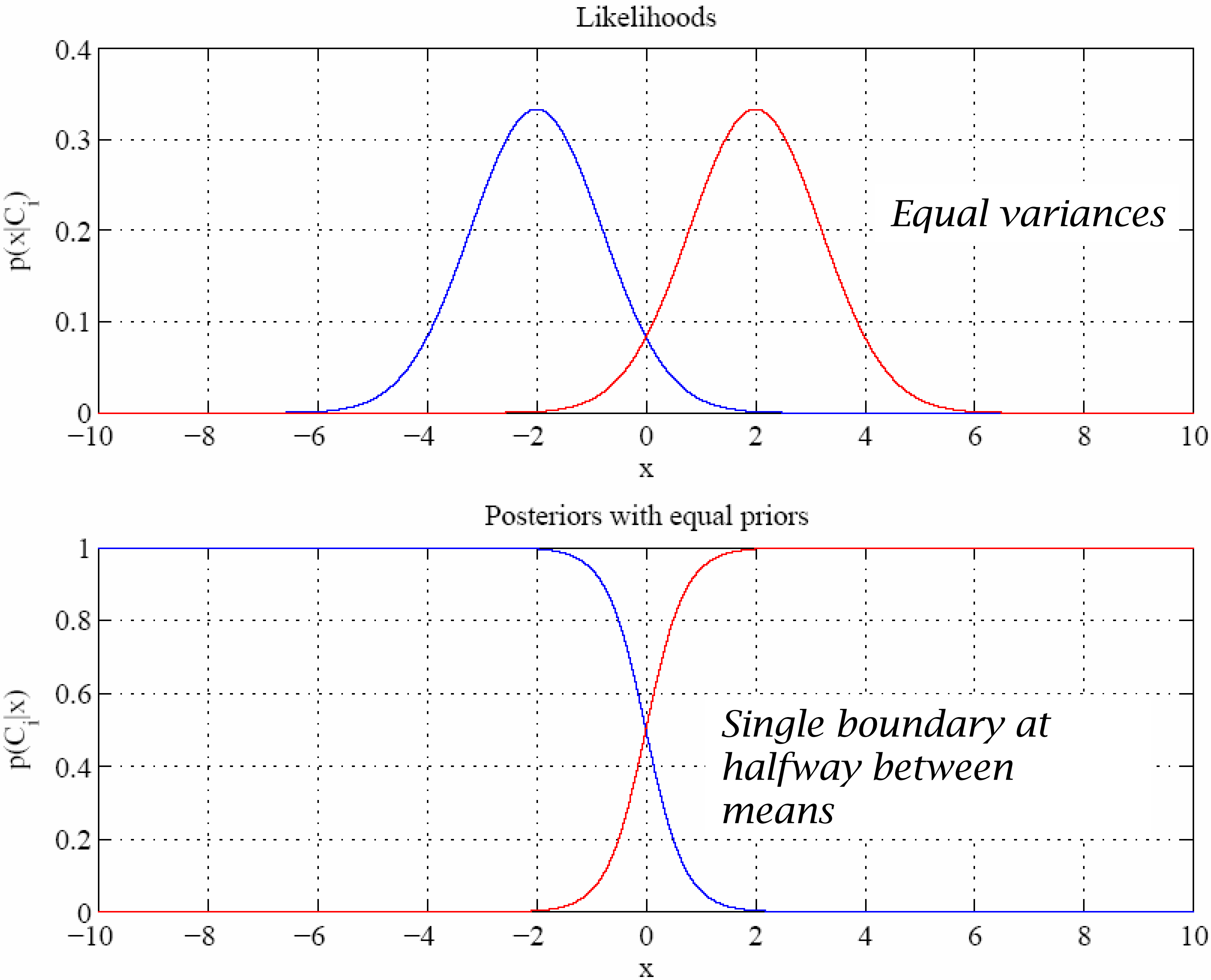
$$x \in \Re \qquad r_i^t = \begin{cases} 1 & \text{if } \boldsymbol{x}^t \in C_i \\ 0 & \text{if } \boldsymbol{x}^t \in C_k, k \neq i \end{cases}$$

- ML estimates are

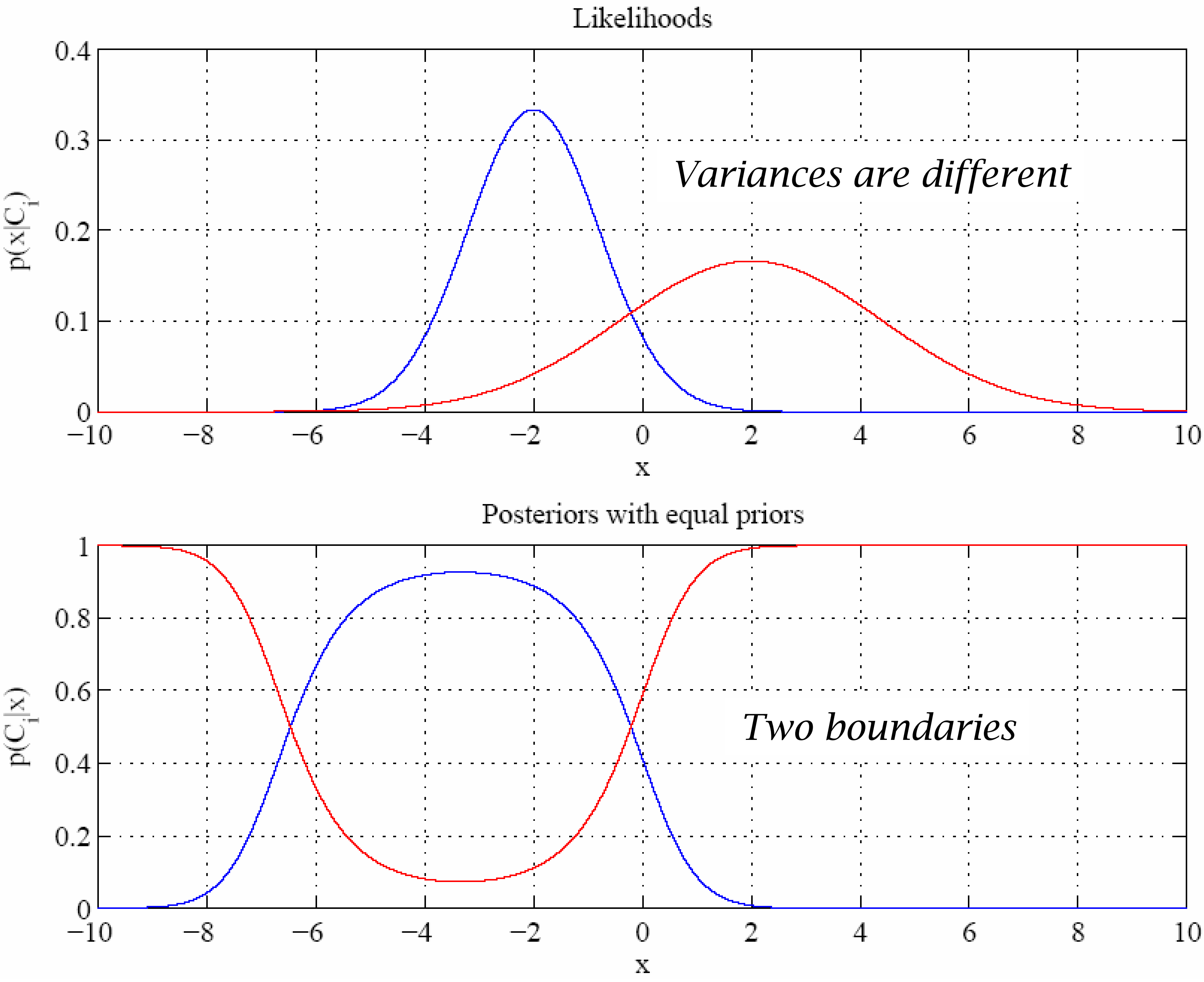
$$\hat{P}(C_i) = \frac{\sum_t r_i^t}{N} \qquad \begin{aligned} m_i &= \frac{\sum_t x^t r_i^t}{\sum_t r_i^t} \\ s_i^2 &= \frac{\sum_t (x^t - m_i)^2 r_i^t}{\sum_t r_i^t} \end{aligned}$$

- Discriminant becomes

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



Likelihoods
Equal variances
p(x|C_i)
x
Posteriors with equal priors
Single boundary at halfway between means
p(C_i|x)
x

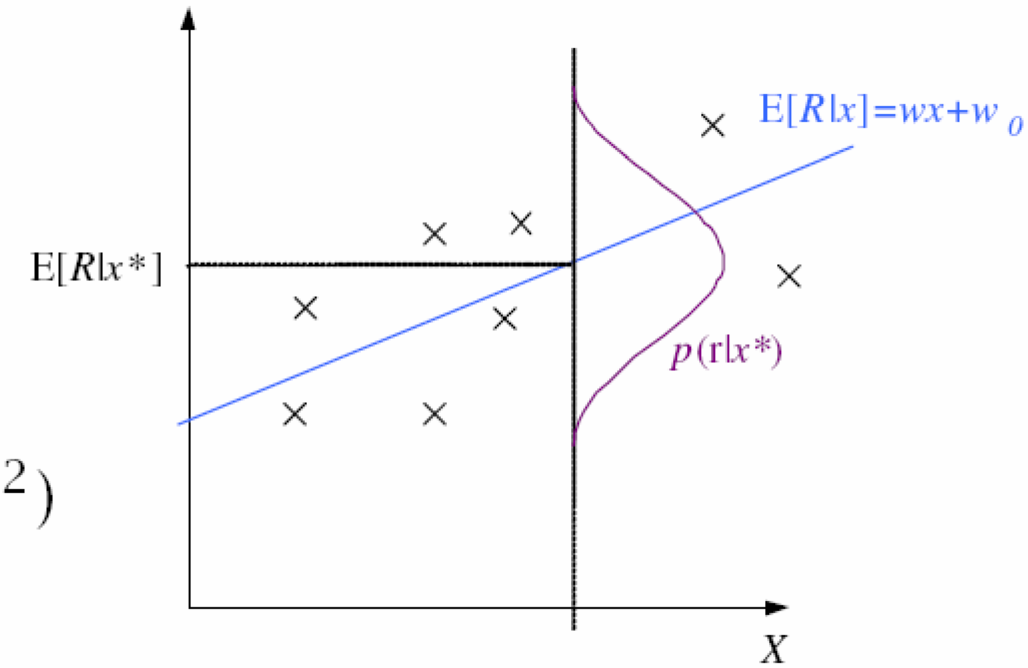

Likelihoods
Variances are different
p(x|C_i)
0.4
0.3
0.2
0.1
0
−10
−8
−6
−4
−2
0
2
4
6
8
10
x
Posteriors with equal priors
Two boundaries
p(C_i|x)
1
0.8
0.6
0.4
0.2
0
−10
−8
−6
−4
−2
0
2
4
6
8
10
x

Regression

$$r = f(x) + \epsilon$$

estimator, $g(x|\theta)$

$\epsilon \sim \mathcal{N}(0, \sigma^2)$



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

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

Regression: From LogL to Error

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

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

Linear Regression

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

$$\begin{aligned}
\sum_t r^t &= N w_0 + w_1 \sum_t x^t \\
\sum_t r^t x^t &= w_0 \sum_t x_t + w_1 \sum_t (x^t)^2
\end{aligned}$$

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

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

Polynomial Regression

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

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

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

Other Error Measures

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

- Relative Square Error: $E_{RSE} = \frac{\sum_t [r^t - g(x^t|\theta)]^2}{\sum_t (r^t - \overline{r})^2}$

- Absolute Error: $E(\theta|\mathcal{X}) = \sum_t |r^t - g(x^t|\theta)|$
- ε-sensitive Error:

$$E(\theta|\mathcal{X}) = \sum_t 1(|r^t - g(x^t|\theta)| > \varepsilon)\ (|r^t - g(x^t|\theta)| - \varepsilon)$$

Bias and Variance

$$E[(r-g(x))^2|x] = \underbrace{E[(r-E[r|x])^2|x]}_{noise} + \underbrace{(E[r|x]-g(x))^2}_{squared\ error}$$

$$E_X[(E[r|x]-g(x))^2|x] = \underbrace{(E[r|x]-E_X[g(x)])^2}_{bias} + \underbrace{E_X[(g(x)-E_X[g(x)])^2]}_{variance}$$

Estimating Bias and Variance

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

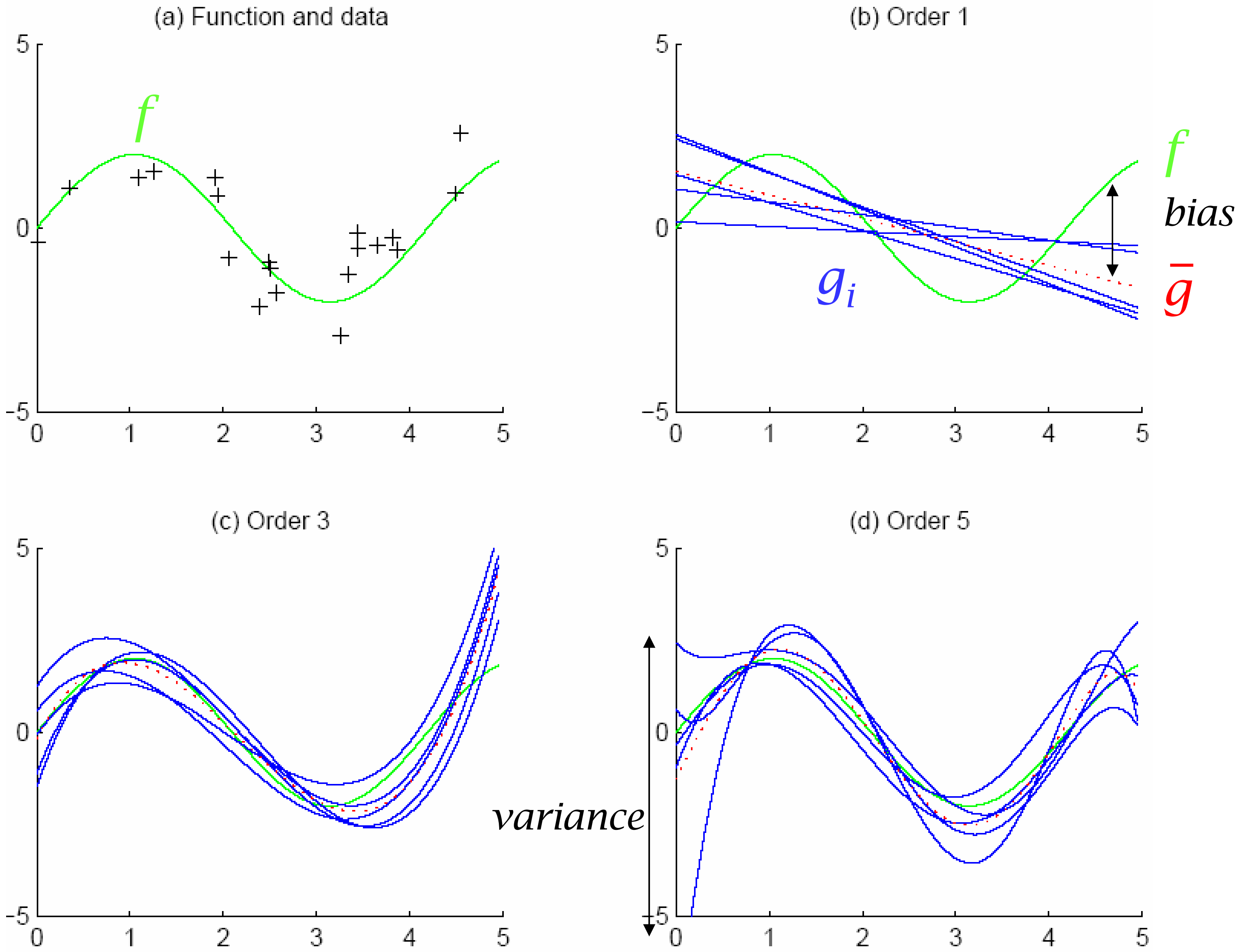
$$\text{Bias}^2(g) = \frac{1}{N}\sum_t [\overline{g}(x^t) - f(x^t)]^2$$

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

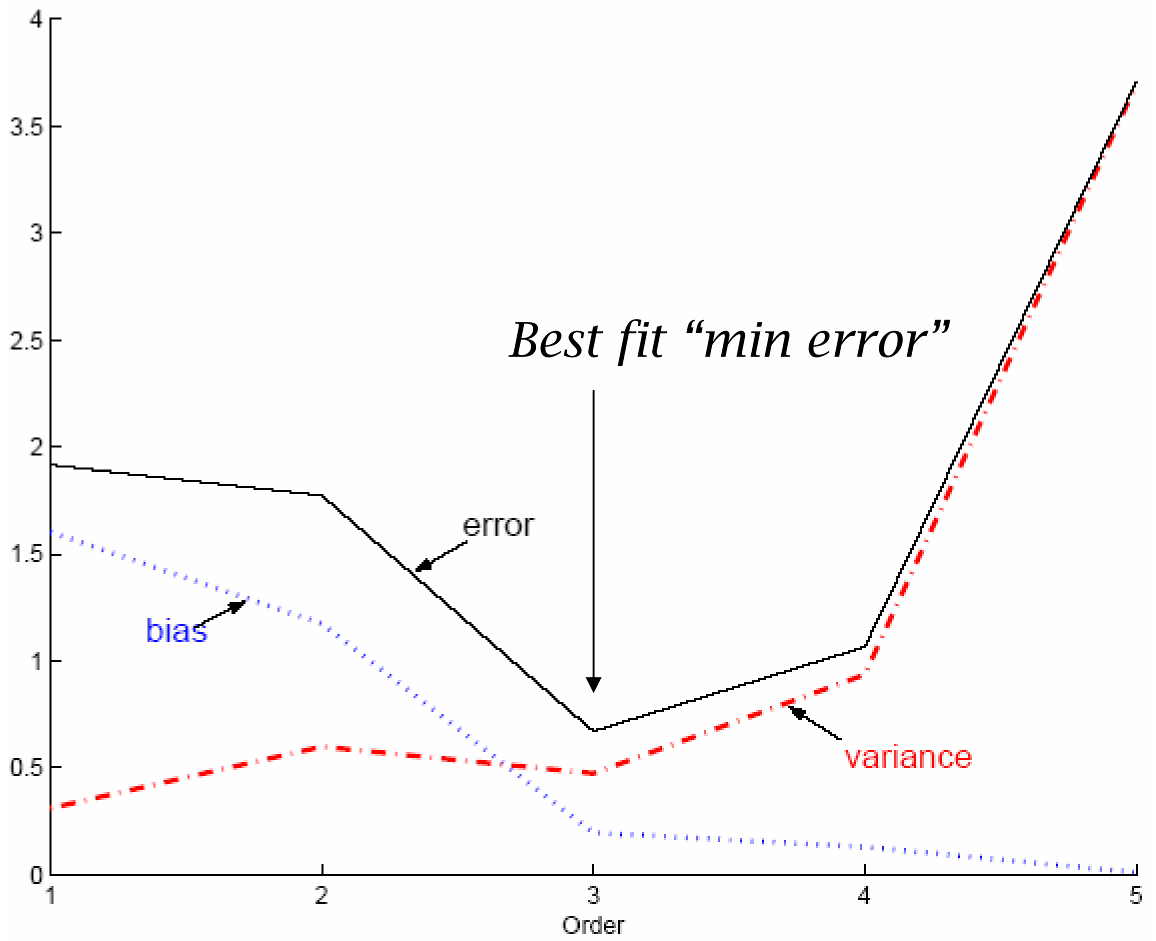
$$\overline{g}(x) = \frac{1}{M}\sum_{i=1}^{M} g_i(x)$$

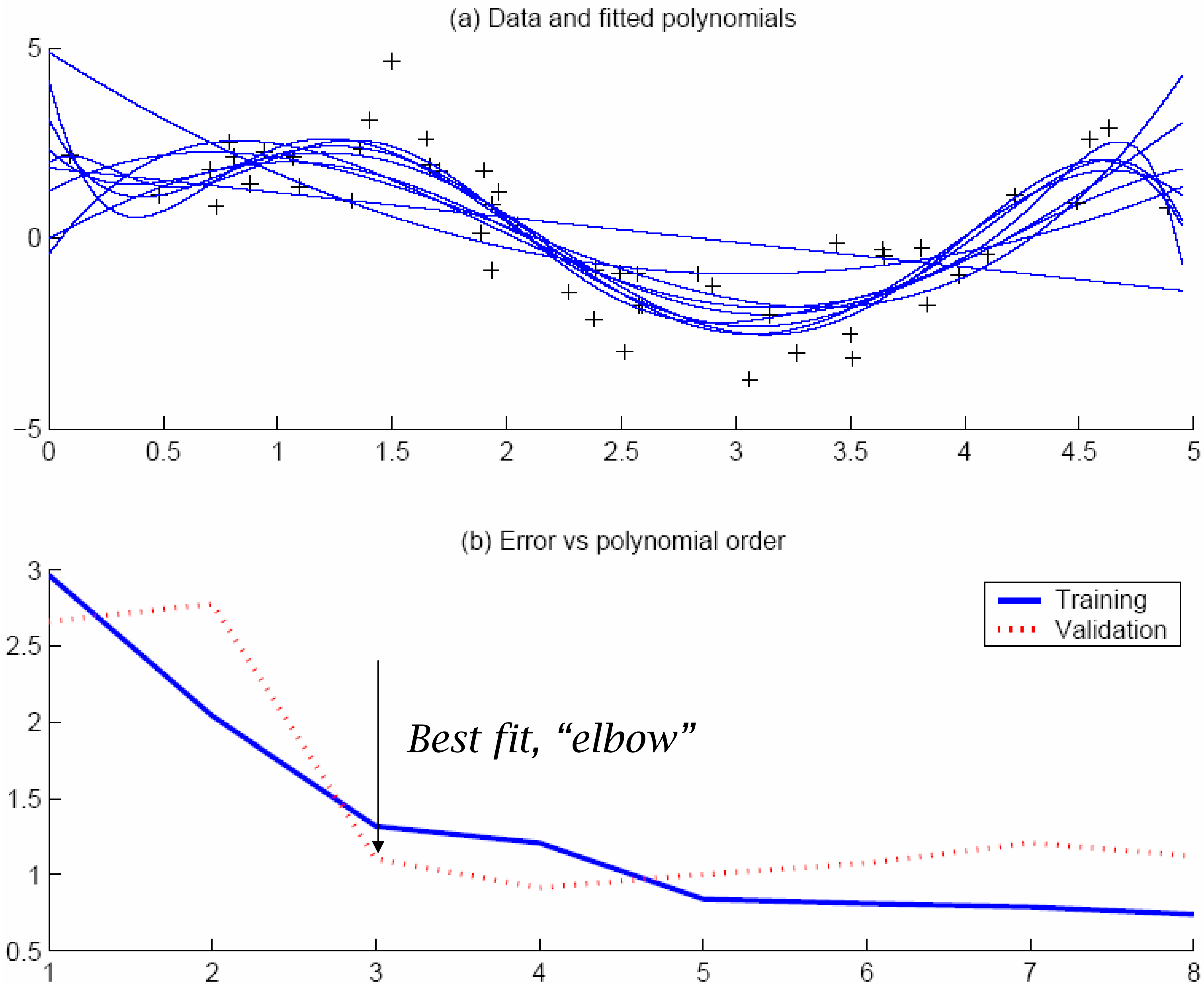
Bias/Variance Dilemma

- Example: $g_i(x) = 2$ has no variance and high bias
- $g_i(x) = \sum_t r_i^t / N$ 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
f
(b) Order 1
f
bias
g_i
$\bar{g}$
(c) Order 3
(d) Order 5
variance

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
Best fit, "elbow"
3
2.5
2
1.5
1
0.5
1
2
3
4
5
6
7
8

Model Selection

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

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

  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(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(model|data)
- Average over a number of models with high posterior (voting, ensembles: Chapter 15)