Machine Learning

یادگیری پیزی

یادگیری ماشین
بخش سوم

دانشگاه شهید بهشتی
پژوهشکده‌ی فضای مجازی
پاییز ۱۴۰۱
احمد محمودی ازناوه

فهرست مطالب

- یادگیری بیزی
  - معیارهای تصمیم‌گیری
- تابع درست‌نمایی
- MAP
- Bayesian inference
- دسته‌بندی
- رگرسیون چند جمله‌ای تک متغیره
- Regularization

احتمال و استنتاج

- داده‌هایی که مورد استفاده قرار می‌دهیم، حاصل فرآیندی است که (کاملا) شناخته شده نیست.
- در پدیده‌های تصادفی، متغیرهای غیرقابل مشاهده، موجب پیدایش عدم قطعیت می‌شود.
- $x=f(z)$
- با توجه به این که چنین فرآیندهایی بدین‌شیوه قابل مدل کردن نیستند، خروجی را به صورت یک متغیر تصادفی تعریف می‌کنیم:
- $P(X=x)$
- بر اساس نمونه‌های ورودی می‌توان این توزیع را تخمین زد، به عنوان مثال برای سکه

$$p_o = \#\{Heads\}/\#\{Tosses\} = \sum_t x^t / N$$

دسته‌بندی بهینه

- مسأله‌ی دسته‌بندی اعتبار مشتریان:
  - ورودی: درآمد و پس‌انداز
  - خروجی: مشتری low risk و High risk
  - Input: $\boldsymbol{x} = [x_1, x_2]^T$,Output: $C \in \{0,1\}$
  - پیش‌بینی:
  - high risk(C=1) or low risk(C=0)

$$\text{choose} \begin{cases} C=1 \text{ if } P(C=1 \mid x_1, x_2) > 0.5 \\ C=0 \text{ otherwise} \end{cases}$$

or

$$\text{choose} \begin{cases} C=1 \text{ if } P(C=1 \mid x_1, x_2) > P(C=0 \mid x_1, x_2) \\ C=0 \text{ otherwise} \end{cases}$$

احتمال شرطی

Bayesian inference is a method of statistical inference in which Bayes' theorem is used to update the probability for a hypothesis as more evidence or information becomes available[wiki].

دسته‌بندی (ادامه...)

- با فرض این ورودی x، متغیر مشاهده شده است، مسأله یافتن احتمال $P(C|x)$ است.

Bayes' Rule

$$P(C|\mathbf{x}) = \frac{P(C)\,p(\mathbf{x}|C)}{p(\mathbf{x})}$$

- $P(C|\mathbf{x})$: posterior — احتمال پسین — با چه احتمالی C، دسته‌ی مربوط به x است.
- $P(C)$: prior — احتمال پیشین (احتمال اولیه)
- $p(\mathbf{x}|C)$: Class likelihood — درست‌نمایی دسته — با چه احتمالی در دسته‌ی C نمونه‌ای با مشخصات x وجود دارد.
- $p(\mathbf{x})$: evidence

$$P(C=0)+P(C=1)=1$$

$$p(\mathbf{x}) = p(\mathbf{x}|C=1)P(C=1) + p(\mathbf{x}|C=0)P(C=0)$$

دسته‌بندی (ادامه...)

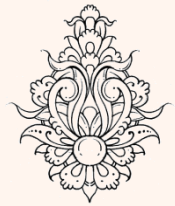
$$P(C_1|\mathrm{x}) \underset{2}{\overset{1}{>\!<}} P(C_2|\mathrm{x})$$

$$\frac{P(C_1|\mathrm{x})}{P(C_2|\mathrm{x})} \underset{2}{\overset{1}{>\!<}} 1$$

$$\frac{\frac{P(x|\mathrm{C}_1)P(C_1)}{P(x)}}{\frac{P(x|\mathrm{C}_2)P(C_2)}{P(x)}} \lessgtr 1$$

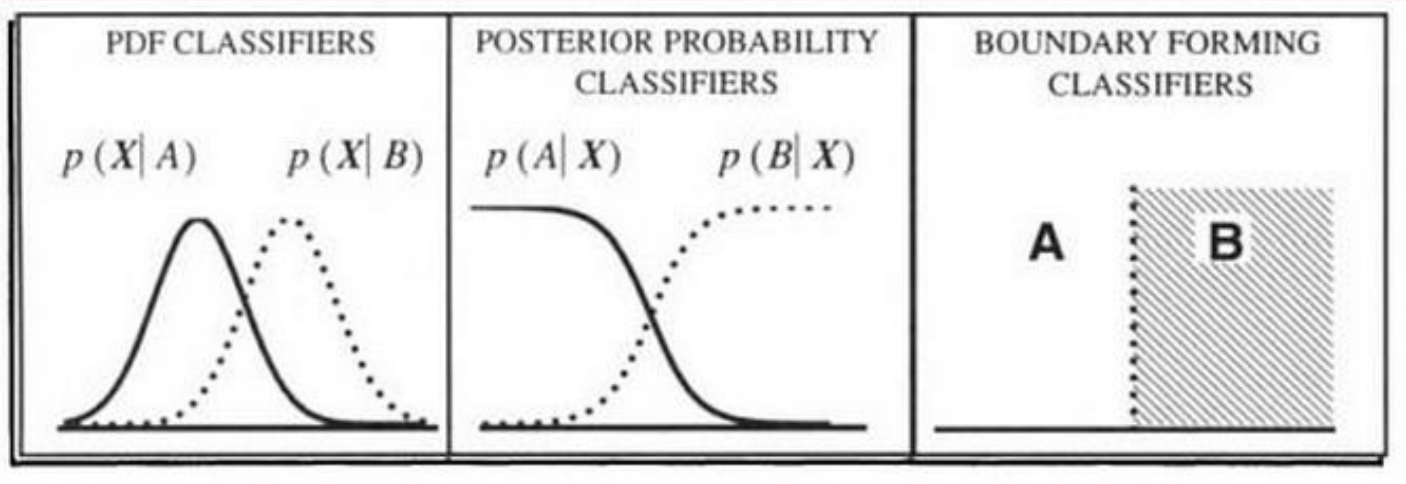
$$\frac{P(x|\mathrm{C}_1)P(C_1)}{P(x|\mathrm{C}_2)P(C_2)} \lessgtr 1$$

$$\frac{P(x|\mathrm{C}_1)}{P(x|\mathrm{C}_2)} \lessgtr \frac{P(C_2)}{P(C_1)}$$



posterior ∝ likelihood × prior,

دسته‌بندی (ادامه...)



pattern recognition using neural networks theory and algorithms for engineers and scientists, by Carl G. Looney

$$P(C|\mathbf{x}) = \frac{P(C)\,p(\mathbf{x}|C)}{p(\mathbf{x})}$$

دسته‌بندی چندکلاسی

احتمال رخداد x هنگامی که می‌دانیم به دسته‌ی C_i تعلق دارد
Class likelihood

$$P(C_i \mid \mathbf{x}) = \frac{P(C_i)\, p(\mathbf{x} \mid C_i)}{p(\mathbf{x})}$$

$$P(C_i \mid \mathbf{x}) = \frac{P(C_i)\, p(\mathbf{x} \mid C_i)}{p(\mathbf{x})} = \frac{P(C_i)\, p(\mathbf{x} \mid C_i)}{\sum_{k=1}^{K} P(C_i)\, p(\mathbf{x} \mid C_i)}$$

$$P(C_i \mid \mathbf{x}) = \max_k P(C_k \mid \mathbf{x})$$

در این صورت دسته‌ی C_i انتخاب می‌شود.

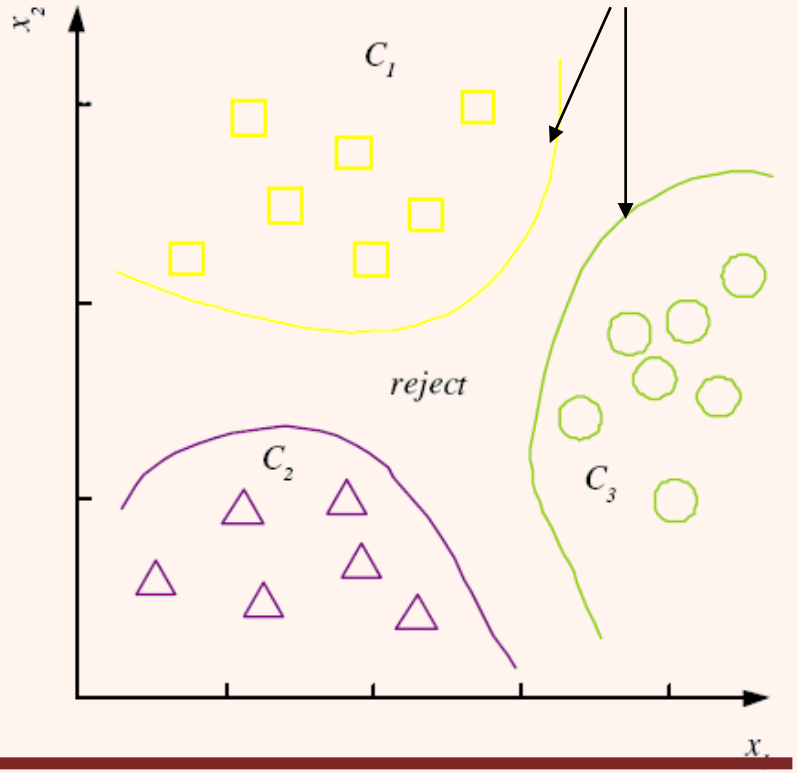
Discriminant Functions

توابع جداساز

choose C_i if $g_i(\mathbf{x}) = \max_k g_k(\mathbf{x})$

$g_i(\mathbf{x}), i = 1, \ldots, K$

$$g_i(\mathbf{x}) = \begin{cases} P(C_i|\mathbf{x}) \\ p(\mathbf{x}|C_i)P(C_i) \end{cases}$$



K decision regions $\mathcal{R}_1, \ldots, \mathcal{R}_K$

$$\mathcal{R}_i = \{\mathbf{x} \mid g_i(\mathbf{x}) = \max_k g_k(\mathbf{x})\}$$

Dichotomizer

$g(x) = g_1(x) - g_2(x)$

$$\text{choose} \begin{cases} C_1 \text{ if } g(\mathbf{x}) > 0 \\ C_2 \text{ otherwise} \end{cases}$$

برآورد پارامتری

- در فصل پیش در مورد «اتخاذ تصمیم بهینه» با در نظر گرفتن احتمال مشاهده‌ی ورودی با فرض دانستن دسته و احتمال وقوع دسته بحث شد.
  - با توجه به این فرض که توزیع داده‌ها، از توزیعی خاص پیروی می‌کند، این روش‌ها را «روش‌های پارامتری» می‌نامند.
    - $\mathcal{X} = \{ x^t \}_{t=1}^{N}$ where $x^t \sim p(x)$
- تخمین پارامتر:
  - تخمین پارامترهای θ از روی داده‌های آموزشی $\mathcal{X}$
  - برای داده‌ها یک مدل به صورت $p(x \mid \theta)$ در نظر گرفته می‌شود (θ «آماره‌ی بسنده» است؛ تمام اطلاعات در مورد توزیع را در بر دارد)

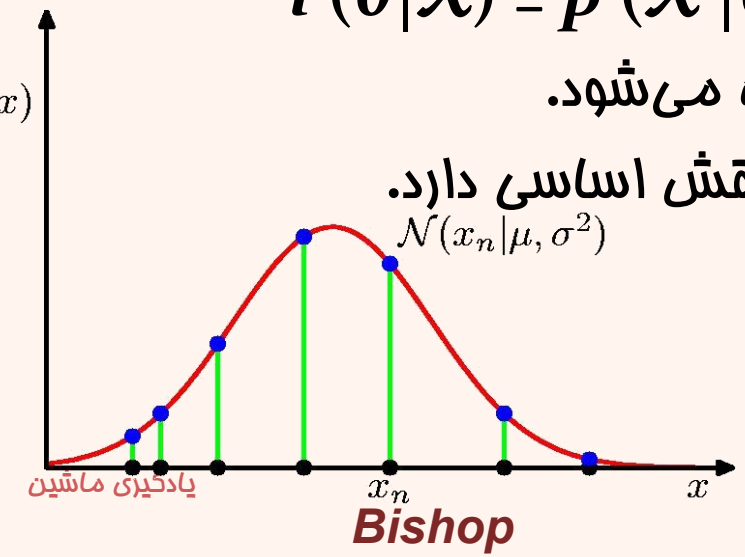
Sufficient statistic

تابع درست‌نمایی

- «تابع درست‌نمایی»، تابعی از پارامترهای مدل آماری است.

– درست‌نمایی یک مجموعه از پارامترها، θ، برای مقادیری معین ($\mathcal{X}$)؛ برابرست با احتمال رخداد $\mathcal{X}$ به ازای مجموعه پارامترها (احتمال درستی θ آن به شرط $\mathcal{X}$)

- $l(\theta|\mathcal{X}) \equiv p(\mathcal{X}|\theta)$



- $\mathcal{X}$ ثابت است و θ را تغییر داده می‌شود.
- این تابع در «استنباط آماری» نقش اساسی دارد.

Statistical inference

MLE

برآورد درست‌نمایی بیشینه

Maximum Likelihood Estimation

Make sampling x^t from $p(x^t|\theta)$ as likely as possible

- در صورتی که نمونه‌ها، $\mathcal{X} = \{ x^t \}$، «متغیرهای مستقل با توزیع یکسان(i.i.d.)» باشد:

independent and identically distributed

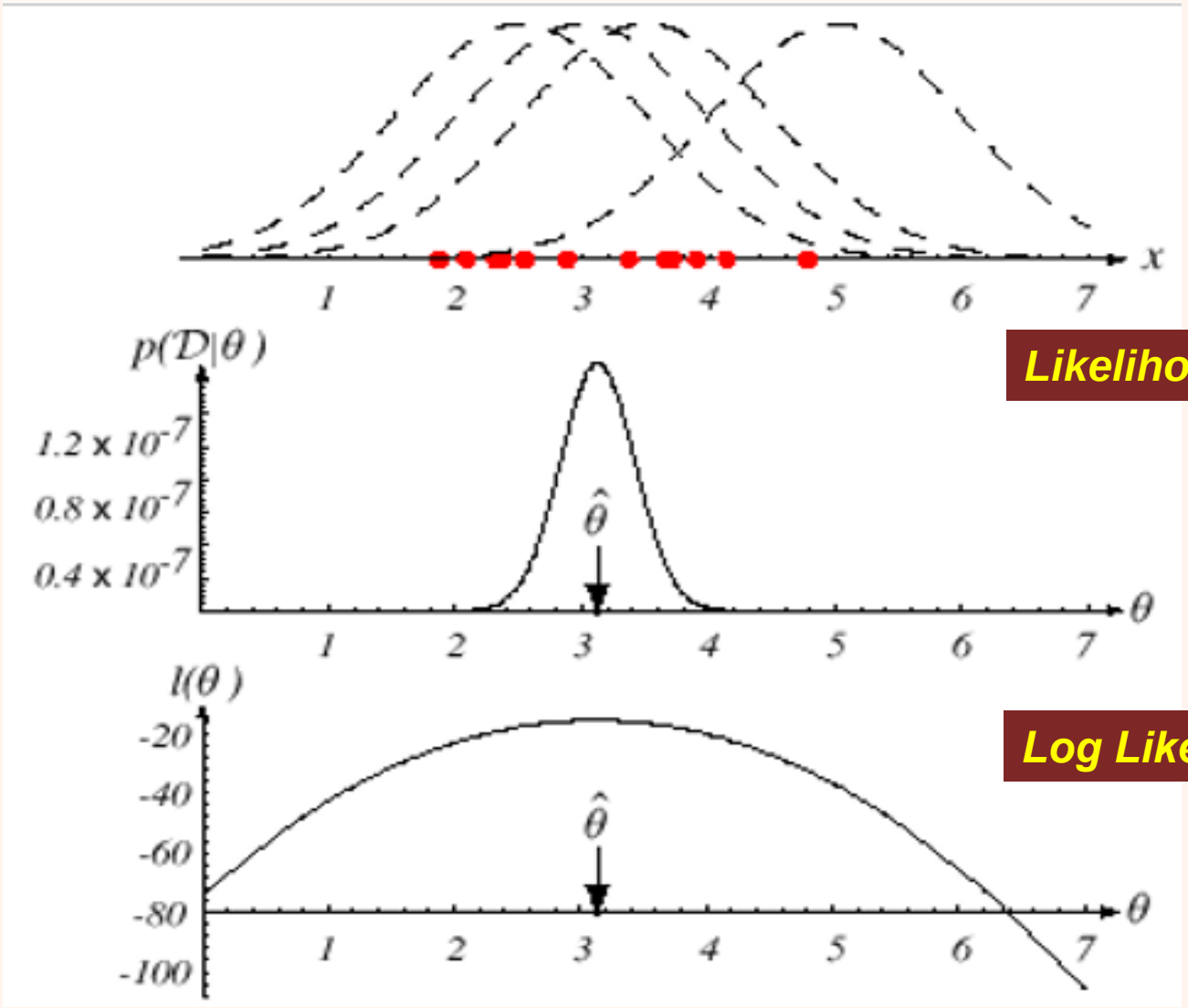
- $l(\theta|\mathcal{X}) = p(\mathcal{X}|\theta) = \prod_t p(x^t|\theta)$
- در برآورد درست‌نمایی بیشینه در پی یافتن θ هستیم به گونه‌ای که احتمال تعلق X به p حداکثر شود؛ درست‌نمایی بیشینه شود.
- برای سادگی محاسبات، به جای درست‌نمایی، از لگاریتم آن استفاده می‌شود:

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

Log likelihood

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

برآورد درست‌نمایی بیشینه



Likelihood

Log Likelihood

Pattern Classification, Chapter 3

Bernoulli /catagorical (generalized Bernoulli) Density

- توزیع برنولی

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: $\hat{p}_o = \sum_t x^t / N$

- توزیع برنولی تعمیم‌یافته

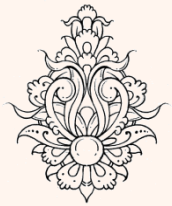
- $K>2$ states, x_i in {0,1}

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

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

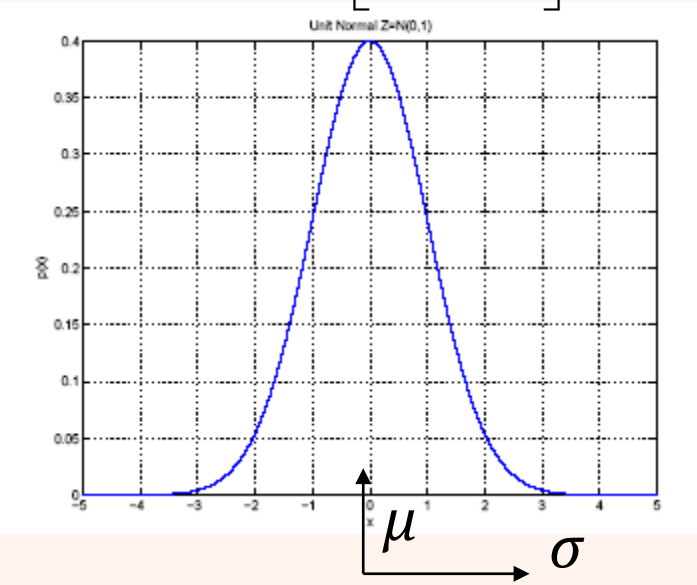
MLE: $\hat{p}_i = \sum_t x_i^t / N$

$$x_i^t = \begin{cases} 1 & \text{if exprimnet } t \text{ choose state } i \\ 0 & \text{otherwise} \end{cases}$$



Gaussian (Normal) Distribution

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



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

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

- MLE for μ and σ^2:

$$L(\mu,\sigma \mid X) = -\frac{N}{2}\log(2\pi) - N\log\sigma - \frac{\sum_t (x^t-\mu)^2}{2\sigma^2}$$

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

دسته‌بندی پارامتری

$$g_i(x) = p(x \mid C_i) P(C_i)$$

تابع جداساز

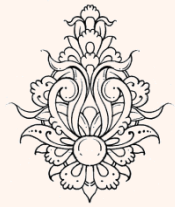
or

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

در صورتی که چگالی دسته را گاوسی در نظر بگیریم:

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

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



دسته‌بندی پارامتری (ادامه...)

نمونه‌های آموزشی

$$\mathcal{X}=\{x^t,r^t\}_{t=1}^N$$

$$x\in\Re \qquad r_i^t=\begin{cases}1 \text{ if } x^t\in C_i\\ 0 \text{ if } x^t\in C_j, j\neq i\end{cases}$$

برآورد درست‌نمایی بیشینه

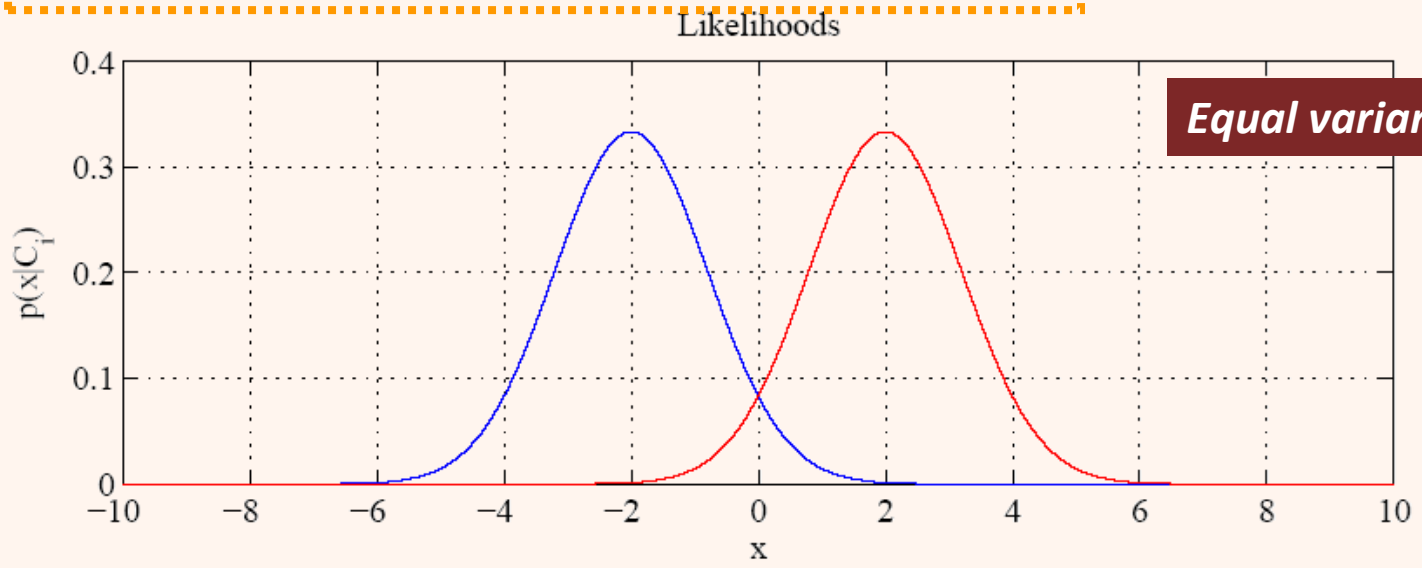
$$\hat{P}(C_i)=\frac{\sum_t r_i^t}{N} \quad m_i=\frac{\sum_t x^t r_i^t}{\sum_t r_i^t} \quad s_i^2=\frac{\sum_t (x^t-m_i)^2 r_i^t}{\sum_t r_i^t}$$

توابع جداساز

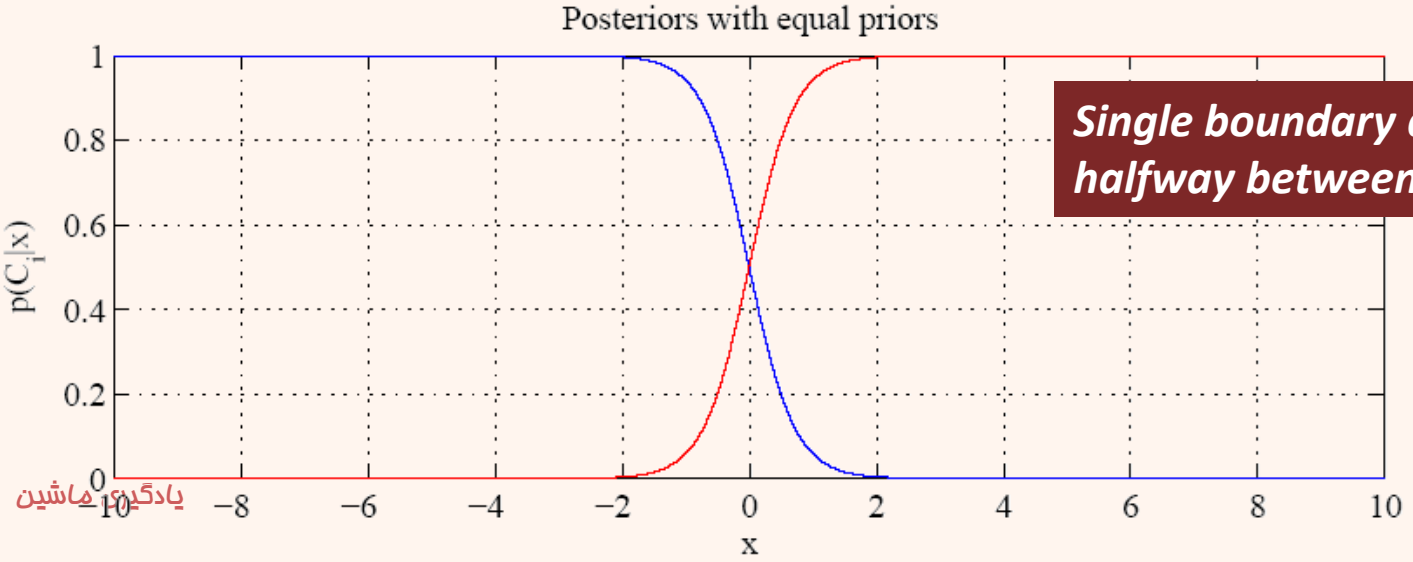
$$g_i(x)=-\frac{1}{2}\log 2\pi-\log s_i-\frac{(x-m_i)^2}{2s_i^2}+\log\hat{P}(C_i)$$

دسته‌بندی دو دسته با واریانس یکسان و احتمال اولیه مساوی

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



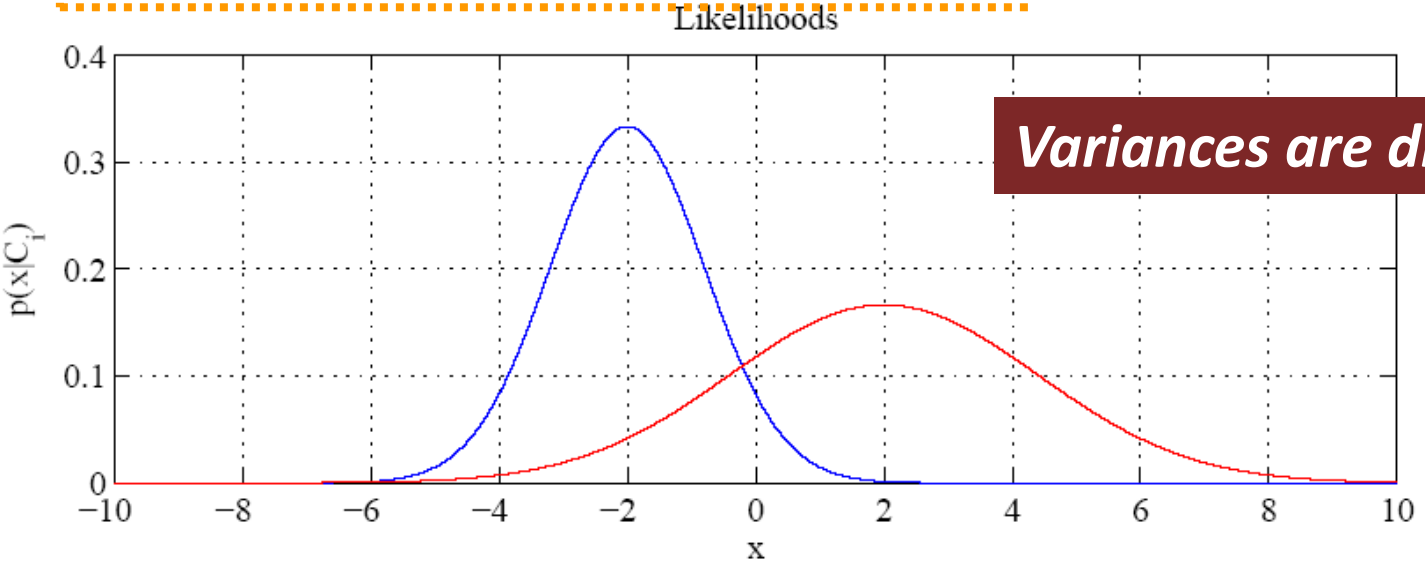
Equal variances



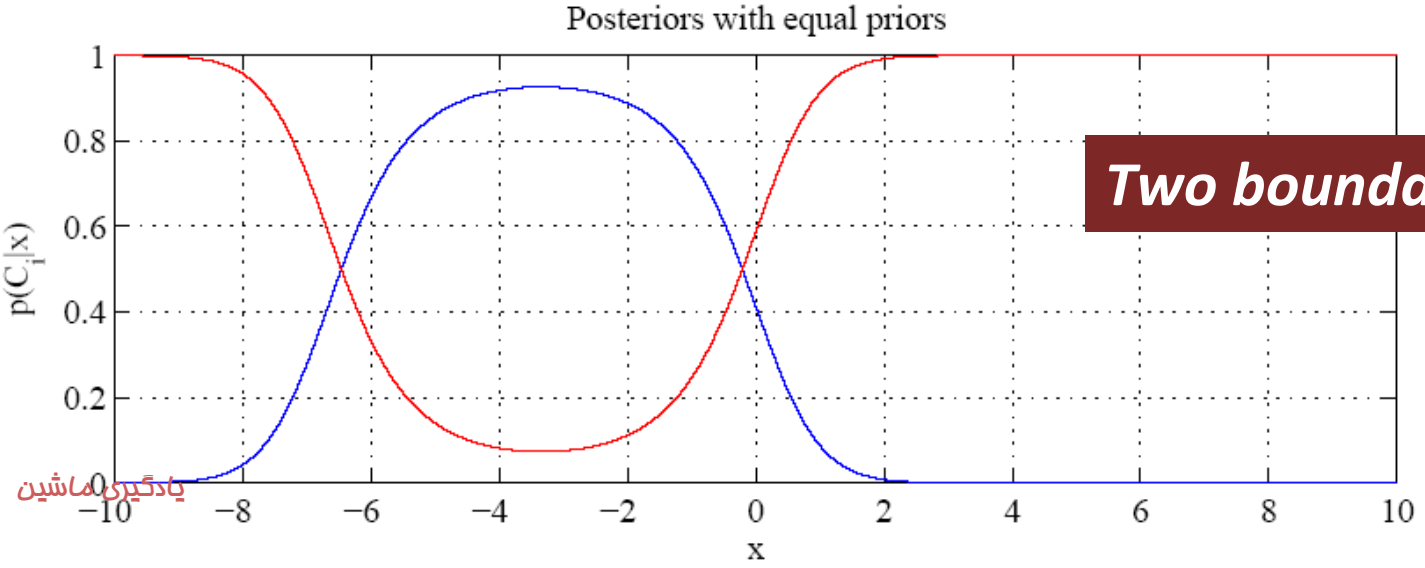
Single boundary at halfway between means

دسته‌بندی دو دسته با واریانس متفاوت و احتمال اولیه مساوی

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



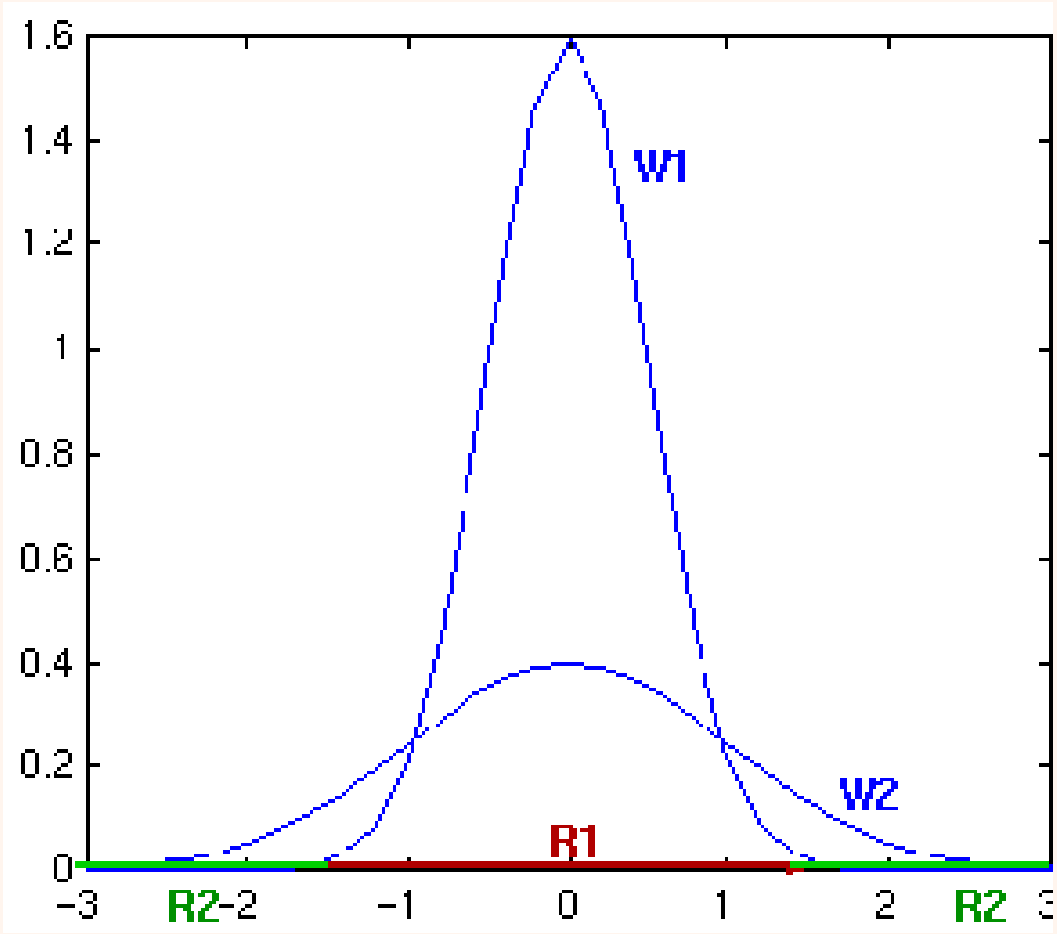
Variances are different



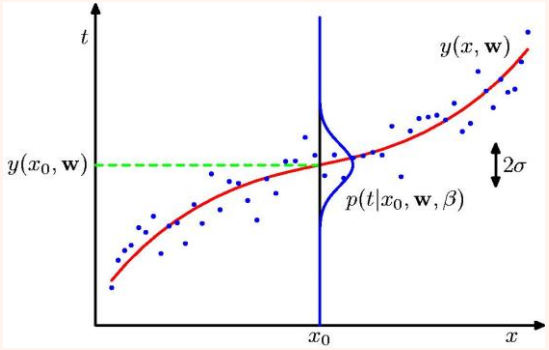
Two boundaries

مثال

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



رگرسیون چندجمله‌ای



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

Least Squares estimates

$$g\left(x^t \mid w_k,\ldots,w_2,w_1,w_0\right) = w_k\left(x^t\right)^k + \cdots + w_2\left(x^t\right)^2 + w_1 x^t + w_0$$

$$\frac{\partial E}{\partial w_i} = \sum_{t=1}^{N}\left[r^t - g\left(x^t \mid \theta\right)\right] \times \frac{\partial g}{\partial w_i} \qquad \forall i \qquad 0 \le i \le k$$

$$\Rightarrow \quad \sum_{t=1}^{N} g\left(x^t \mid \theta\right)\left(x^t\right)^i = \sum_{t=1}^{N} r^t\left(x^t\right)^i$$

رگرسیون چندجمله‌ای

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

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

$$\mathbf{w}=\begin{bmatrix} w_0 \\ w_1 \\ w_2 \\ \vdots \\ w_k \end{bmatrix}$$

$$\mathbf{Aw}=\mathbf{y}$$

$$\sum_{t=1}^{N} g\left(x^t|\theta\right)\left(x^t\right)^i=\sum_{t=1}^{N} r^t\left(x^t\right)^i$$

رگرسیون چندجمله‌ای

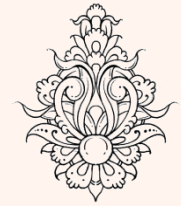
$$\mathbf{A} = \left(\mathbf{D}^T\mathbf{D}\right) \qquad \mathbf{y} = \mathbf{D}^T\mathbf{r}$$

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

$$\mathbf{y} = \begin{bmatrix} \sum_t r^t \\ \sum_t r^t x^t \\ \sum_t r^t \left(x^t\right)^2 \\ \vdots \\ \sum_t r^t \left(x^t\right)^k \end{bmatrix}$$

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

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



معیارهای خطا

- مجموع مربعات خطا

sum of squared error(SSE)

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

- خطای نسبی

relative square error(RSE)

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

the coefficient of determination

$$R^2 = 1 - E_{RSE}$$

در صورتی که «ضریب تعیین» نزدیک به ۱ باشد، می‌توان نتیجه گرفت که مدل به دست آمده مفید است.

Bias/Variance Dilemma

مثال

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

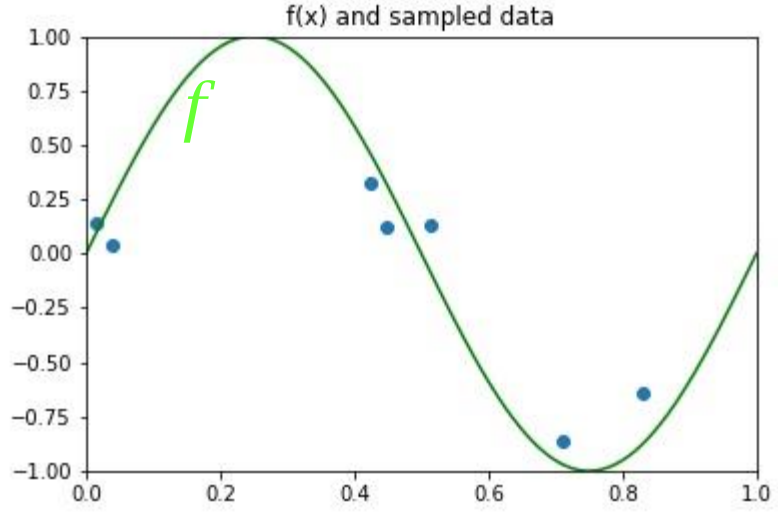
$$\text{Bias}^2(g)=\frac{1}{N}\sum_t\left[\bar{g}(x^t)-f(x^t)\right]^2$$

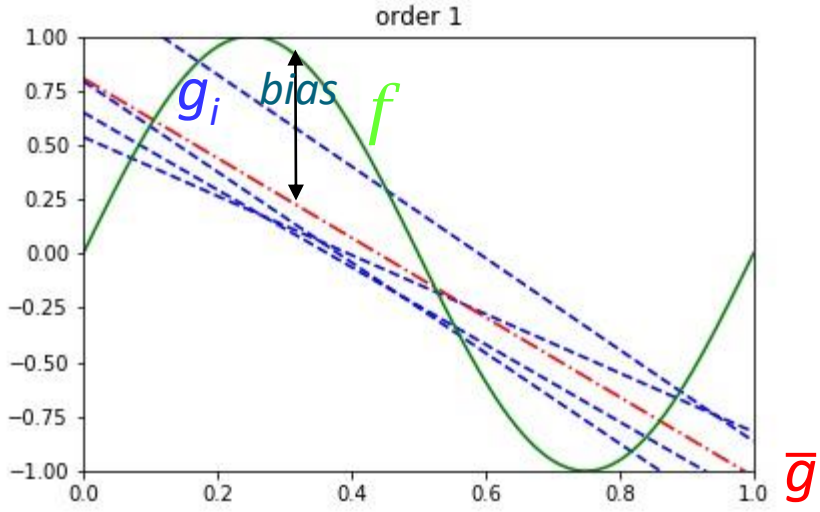
$$\text{Variance}(g)=\frac{1}{NM}\sum_t\sum_i\left[g_i(x^t)-\bar{g}(x^t)\right]^2$$

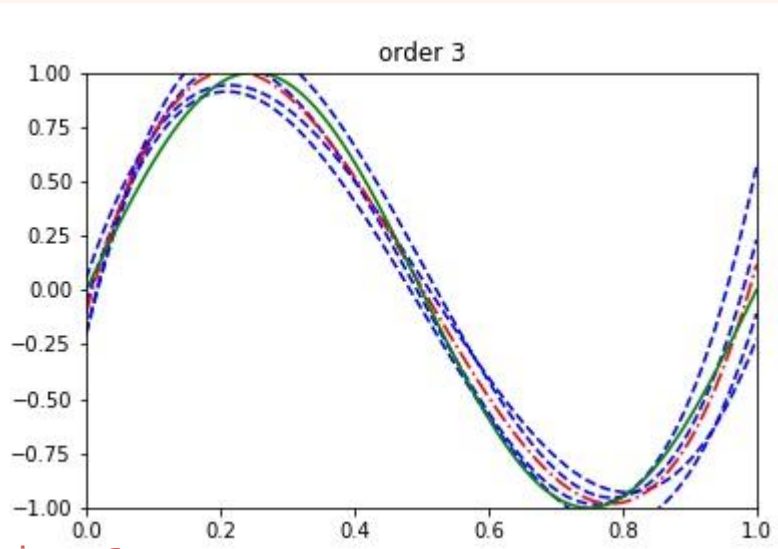
$$\bar{g}(x)=\frac{1}{M}\sum_i g_i(x)$$

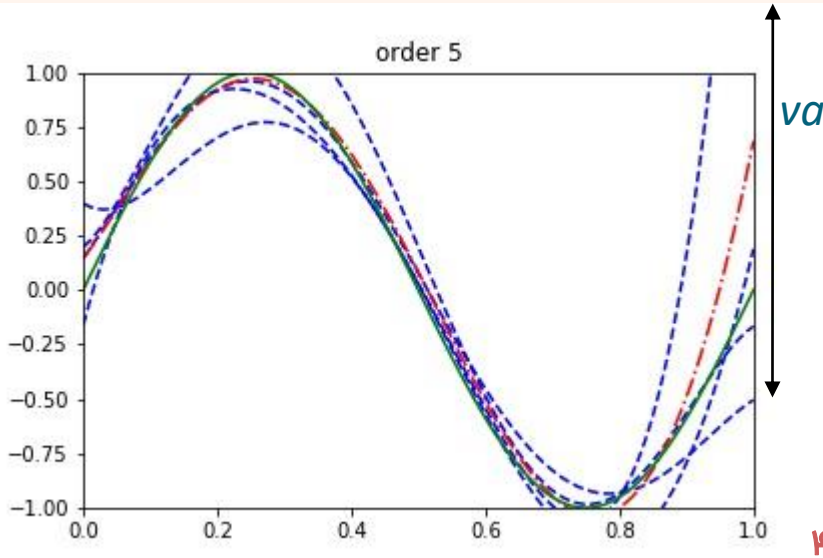
$g_i(x)=2$ واریانس صفر است، اما بایاس بالایی دارد

$g_i(x)=\sum_t r^t_i/N$ بایاس کاهش می‌یابد، اما واریانس افزایش می‌یابد

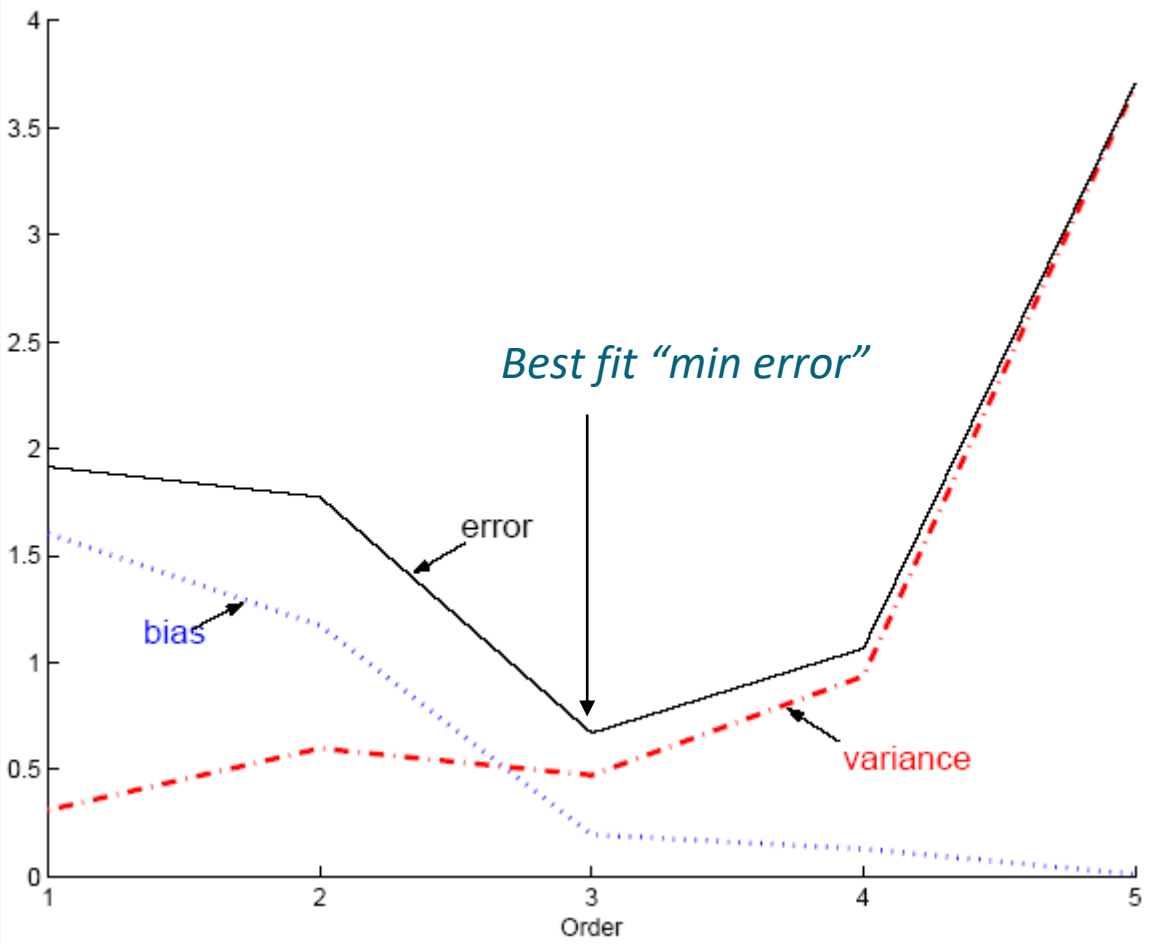

f(x) and sampled data
f


order 1
g_i
bias
f
$\bar{g}$

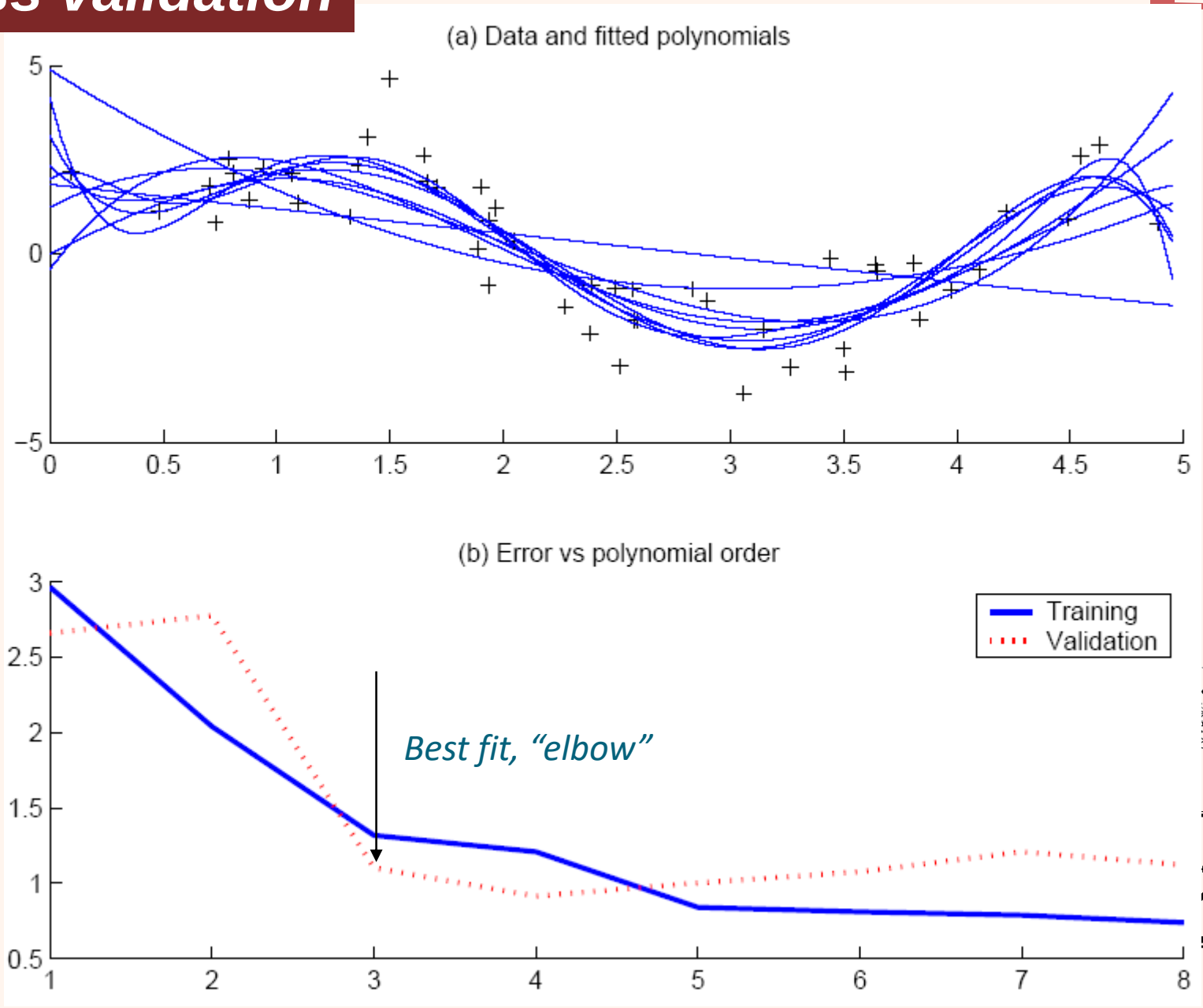

order 3


order 5
variance

انتخاب مدل


Best fit “min error”
error
bias
variance
Order
0
0.5
1
1.5
2
2.5
3
3.5
4
1
2
3
4
5

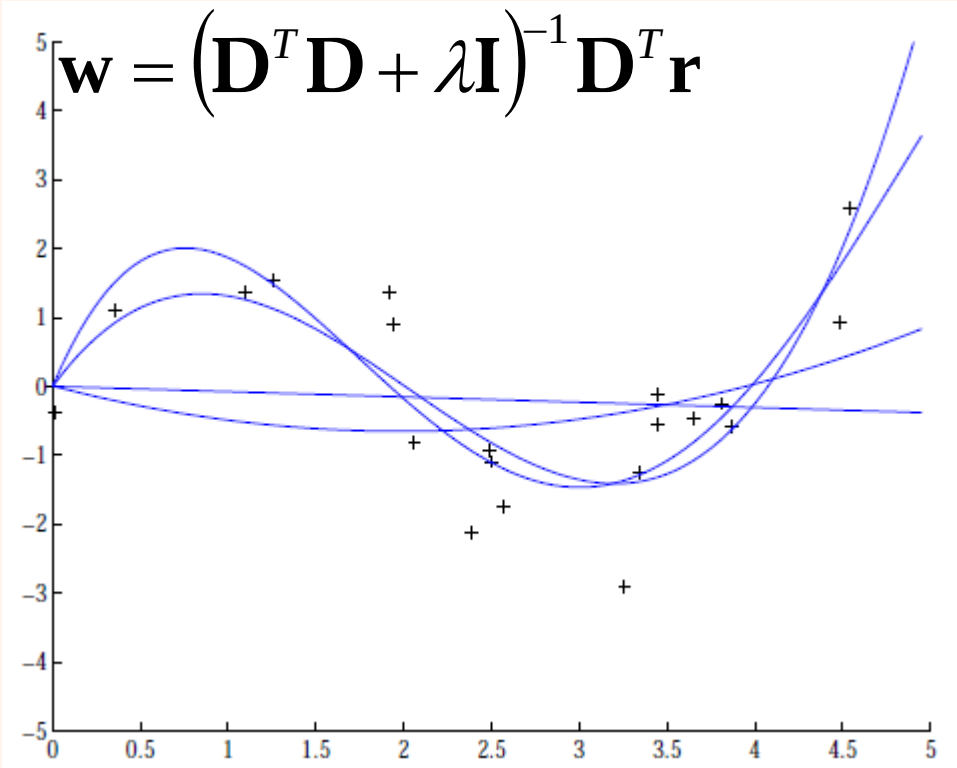
Cross validation



Regularization

Penalize complex models

E'=error on data + λ model complexity



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]

$$regularization: E'(\mathbf{w}|\mathcal{X}) = \frac{1}{2}\left\{\sum_{t=1}^{N}\left[r^t - g\left(x^t|\mathbf{w}\right)\right]^2 + \lambda\sum_i w_i^2\right\}$$

Regularization

9^{th} Order Polynomial