

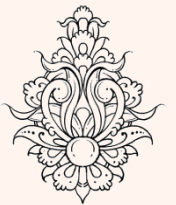
یادگیری ماشین  
بخش سوم



دانشگاه شهید بهشتی  
پژوهشکده‌ی فضای مجازی  
پاییز ۱۴۰۲  
احمد محمودی ازناوه

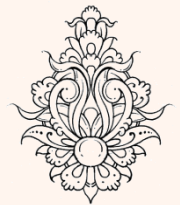
# فهرست مطالب

- 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 (کاملاً) شناخته شده نیست.
- در پدیده‌های تصادفی، متغیرهای غیرقابل مشاهده، موجب پیدایش عدم قطعیت می‌شود.
- $x=f(z)$
- با توجه به این که چنین فرآیندهایی بدین شیوه قابل مدل کردن نیستند، فروجی 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



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

• مسأله‌ی دسته بندی اعتبار مشتریان:

– ورودی: درآمد و پس انداز

– خروجی: مشتری low risk و high risk

– Input:  $\mathbf{x} = [x_1, x_2]^T$ , Output:  $C \in \{0, 1\}$

– پیش بینی:

– high risk ( $C=1$ ) or low risk ( $C=0$ )

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

or

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

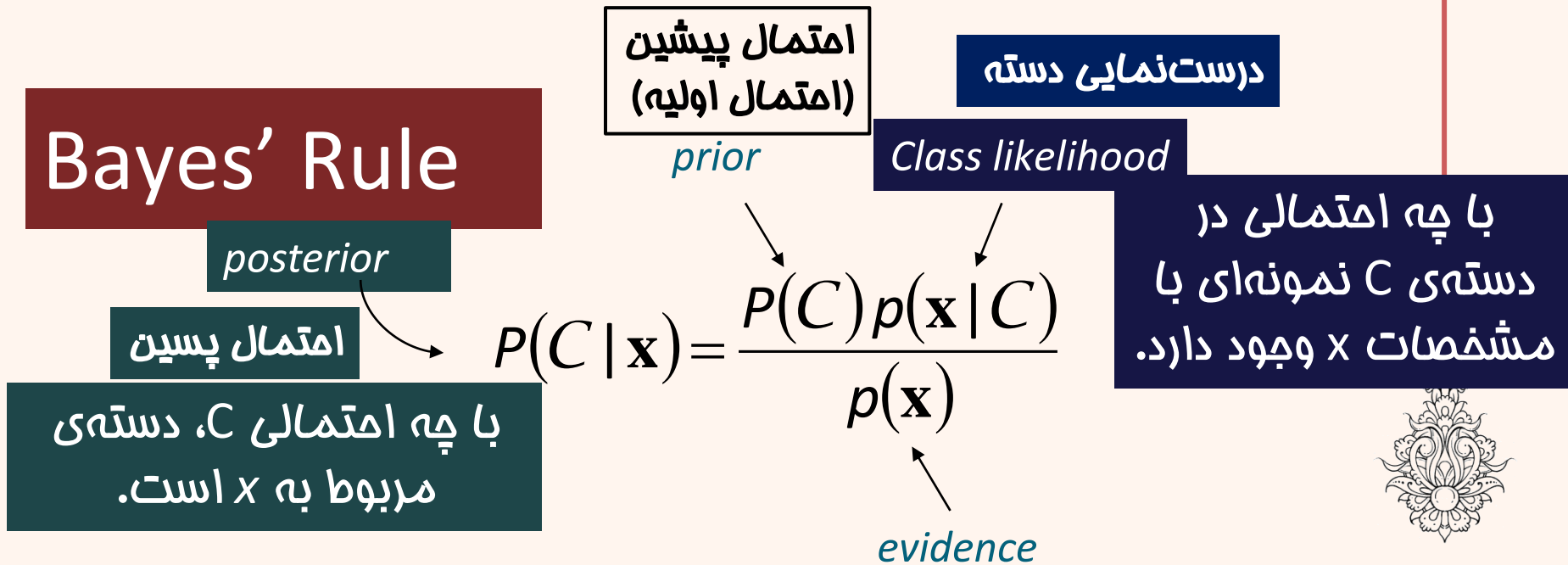
احتمال شرطی



**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].

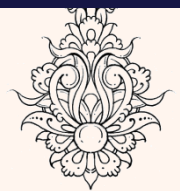
# دسته‌بندی (ادامه...)

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



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

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

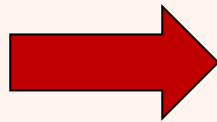


# دسته بندی (ادامه...)

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

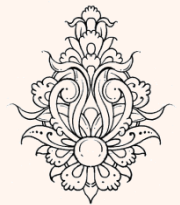
$$\frac{P(C_1|x)}{P(C_2|x)} > \frac{1}{2} < 1$$

$$\frac{\frac{P(x|C_1)P(C_1)}{P(x)}}{\frac{P(x|C_2)P(C_2)}{P(x)}} \approx 1$$

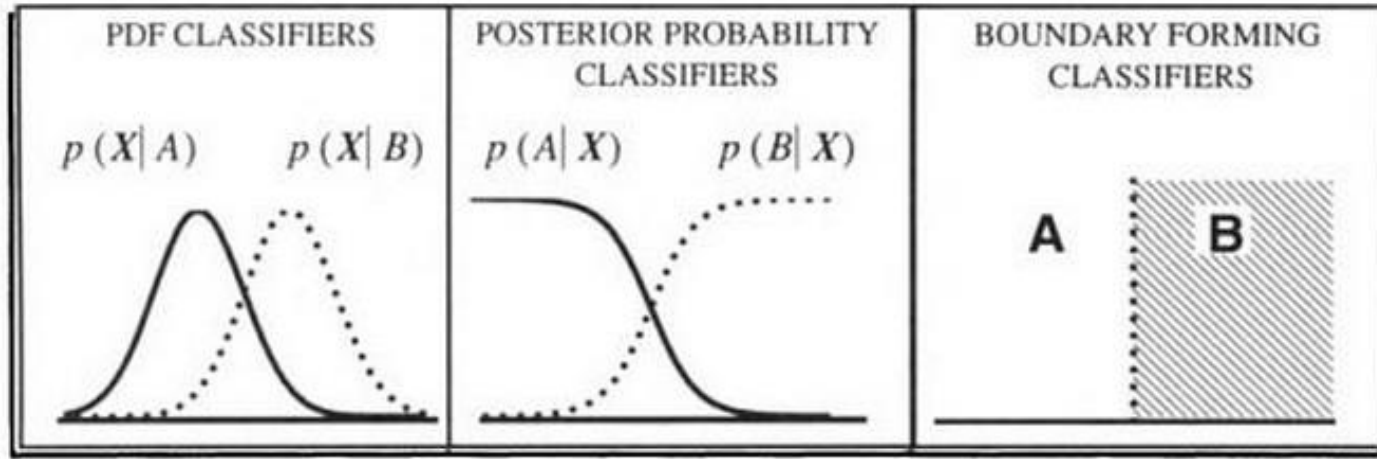


$$\frac{P(x|C_1)P(C_1)}{P(x|C_2)P(C_2)} \approx 1$$

$$\frac{P(x|C_1)}{P(x|C_2)} \approx \frac{P(C_2)}{P(C_1)}$$

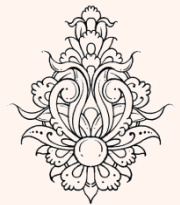


# دسته‌بندی (ادامه...)



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

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



# دسته بندی چندکلاسی

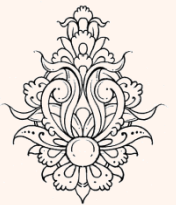
امتمال رخداد  $x$  هنگامی که می دانیم به  
دسته  $C_i$  تعلق دارد  
*Class likelihood*

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

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

$$P(C_i | \mathbf{x}) = \max_k P(C_{\hat{k}} | \mathbf{x})$$

در این صورت دسته  $C_i$  انتخاب می شود.





# Discriminant Functions

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

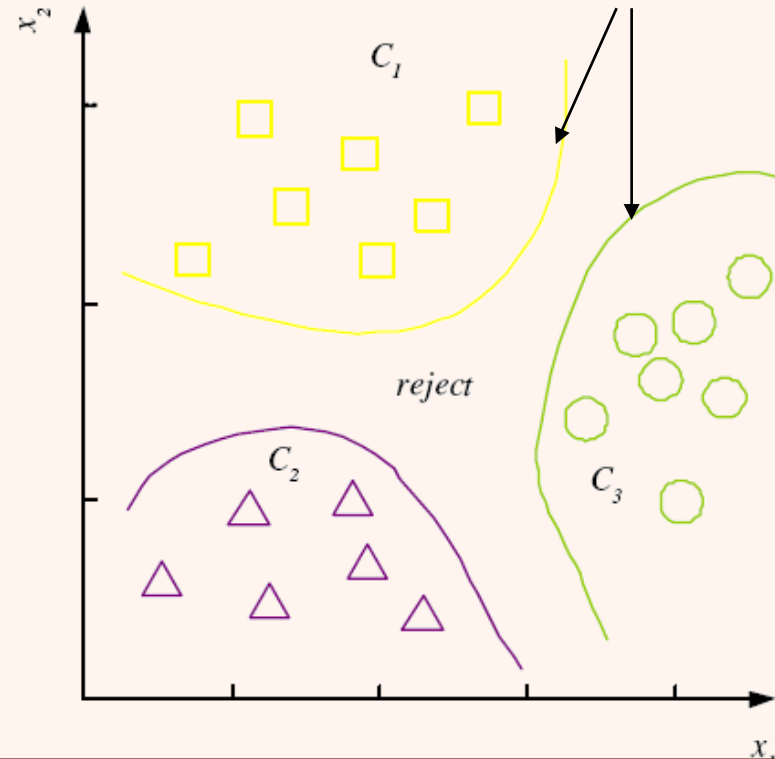
$$g_i(\mathbf{x}) = \begin{cases} P(C_i|\mathbf{x}) \\ p(\mathbf{x}|C_i)P(C_i) \end{cases}$$

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

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

# توابع جداساز

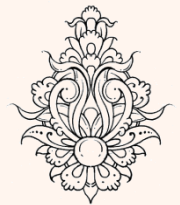
$g_i(\mathbf{x}), i = 1, \dots, K$



# Dichotomizer

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

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



- 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.

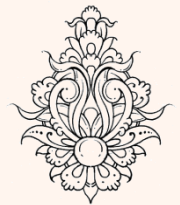
- با توجه به این فرض که توزیع داده‌ها، از توزیعی خاص پیروی می‌کند، این روش‌ها را «روش‌های پارامتری» می‌نامند.

- $\mathcal{X} = \{x^t\}_{t=1}^N$  where  $x^t \sim p(x)$

- تخمین پارامتر:

- تخمین پارامترهای  $\theta$  از روی داده‌های آموزشی  $\mathcal{X}$
- برای داده‌ها یک مدل به صورت  $p(x | \theta)$  در نظر گرفته می‌شود ( $\theta$  «آماره‌ی بسنده» است؛ تمام اطلاعات در مورد توزیع را در بر دارد)

## Sufficient statistic



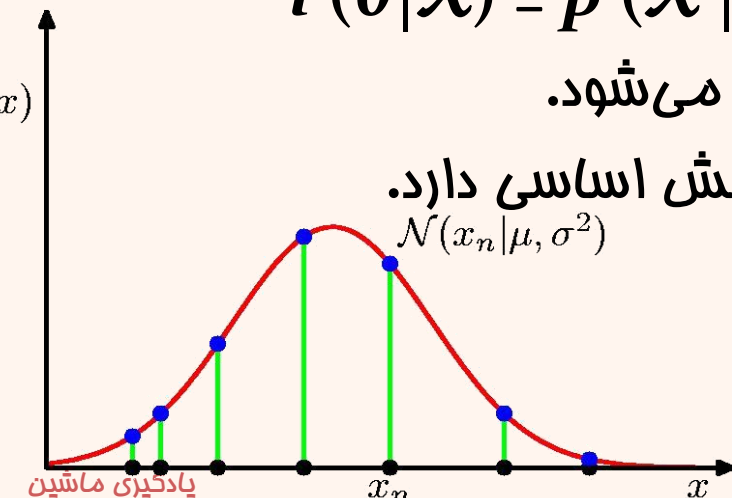
• «تابع درست‌نمایی»، تابعی از پارامترهای مدل آماری است.

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

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

•  $\mathcal{X}$  ثابت است و  $\theta$  را تغییر داده می‌شود.

• این تابع در «استنباط آماری» نقش اساسی دارد.



Bishop

Statistical inference

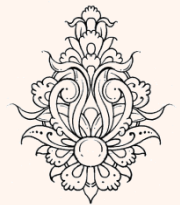
دانشگاه  
تهران  
پیشین

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  $\theta$  هستیم به گونه‌ای که احتمال تعلق  $X$  به  $p$  مدها کمتر شود؛ درست‌نمایی بیشینه شود.
- برای سادگی محاسبات، به جای درست‌نمایی، از لگاریتم آن استفاده می‌شود:



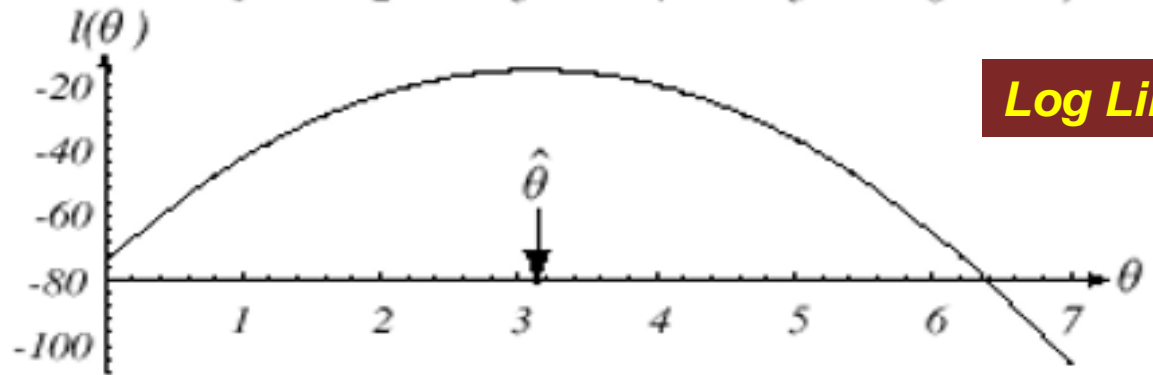
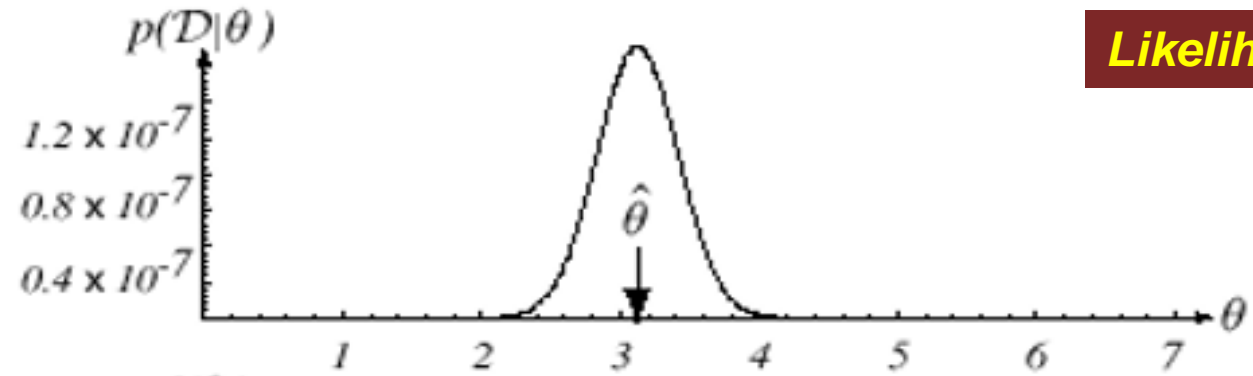
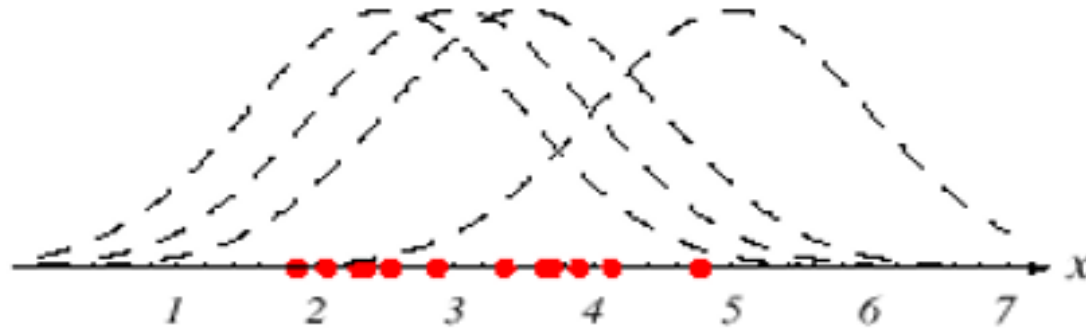
$$\mathcal{L}(\theta|\mathcal{X}) = \log l(\theta|\mathcal{X}) = \sum_t \log p(x^t|\theta)$$

Log likelihood

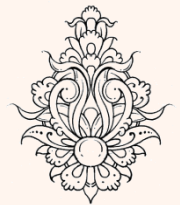
$$\theta^* = \operatorname{argmax}_{\theta} \mathcal{L}(\theta|\mathcal{X})$$



# برآورد درست‌نمایی بیشینه



Pattern Classification, Chapter 3



# Bernoulli /categorical (generalized Bernoulli) Density

$x$  in  $\{0,1\}$

• توزیع برنولی

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

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

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

• توزیع برنولی تعدیم یافته

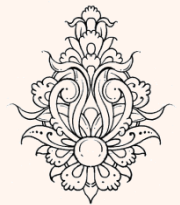
- $K > 2$  states,  $x_i$  in  $\{0,1\}$

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

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

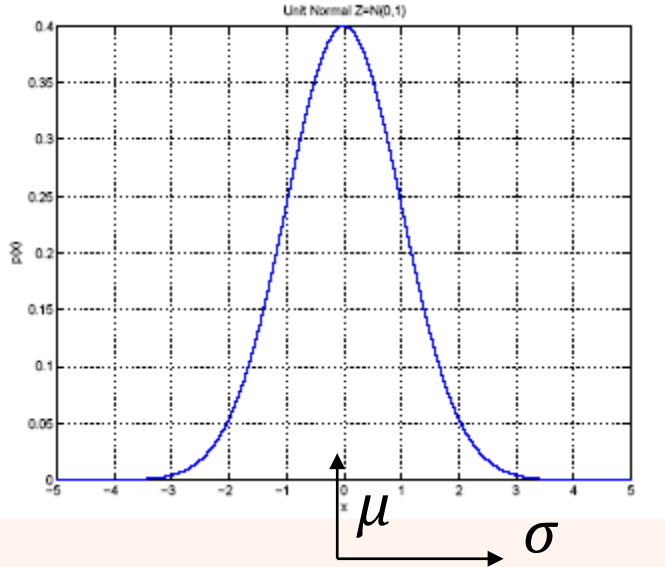
$$\text{MLE: } \hat{p}_i = \sum_t x_i^t / N$$

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



# Gaussian (Normal) Distribution

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



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

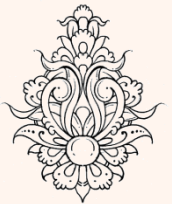
$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]$$

- MLE for  $\mu$  and  $\sigma^2$ :

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

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

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



# دسته بندی پارامتری

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

or

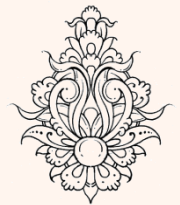
$$g_i(x) = \log p(x | C_i) + \log P(C_i)$$

تابع جدا ساز

در صورتی که چگالی دسته را گاوسی در نظر بگیریم:

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

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





# دسته‌بندی پارامتری (ادامه...)

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

نمونه‌های آموزشی

$$x \in \mathcal{R}$$

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

برآورد درست‌نمایی پیشینه

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

توابع جداساز

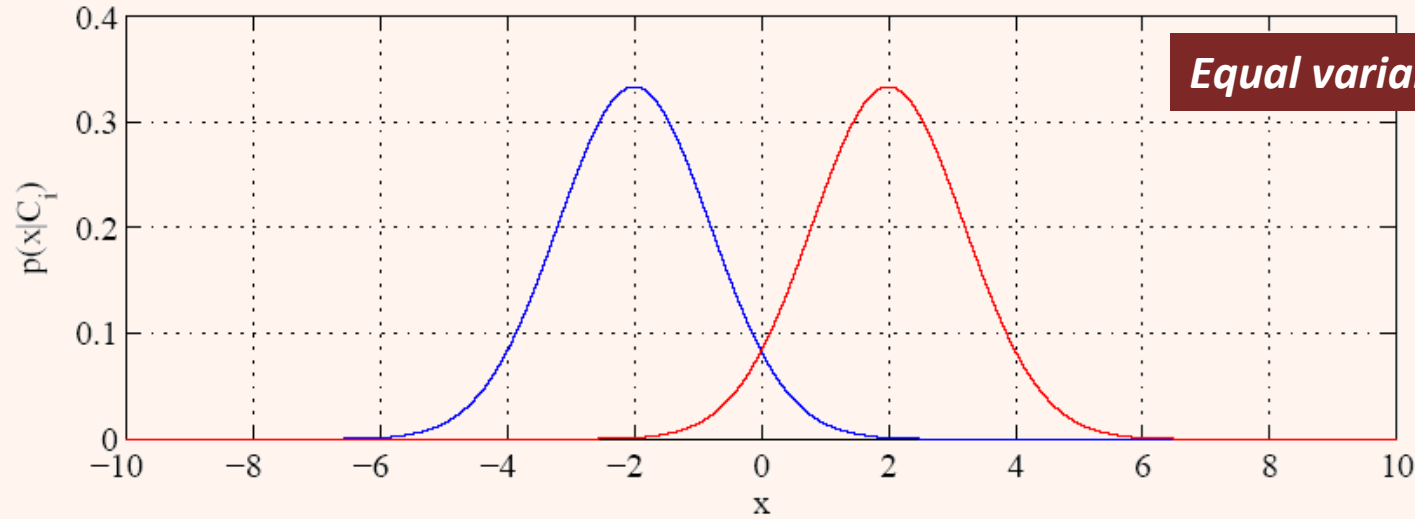
$$g_i(x) = -\frac{1}{2} \log 2\pi - \log s_i - \frac{(x - m_i)^2}{2s_i^2} + \log \hat{P}(C_i)$$



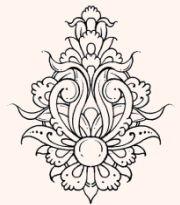
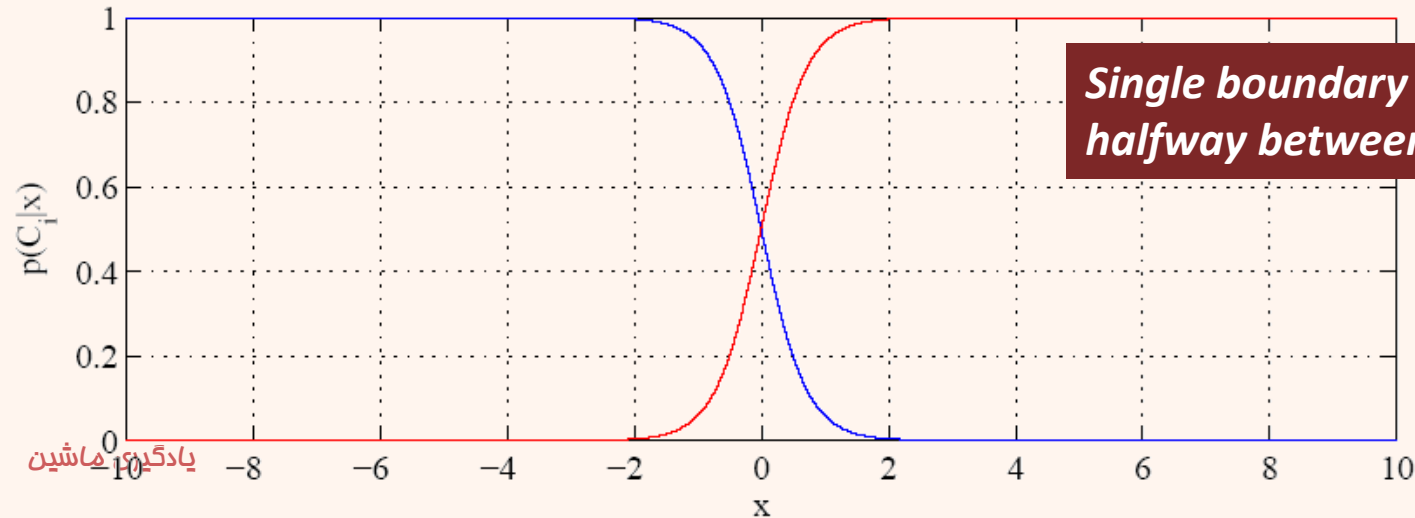
# دسته‌بندی دو دسته با واریانس یکسان و احتمال اولیه مساوی

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

Likelihoods



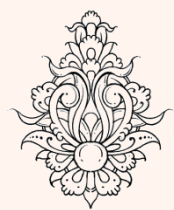
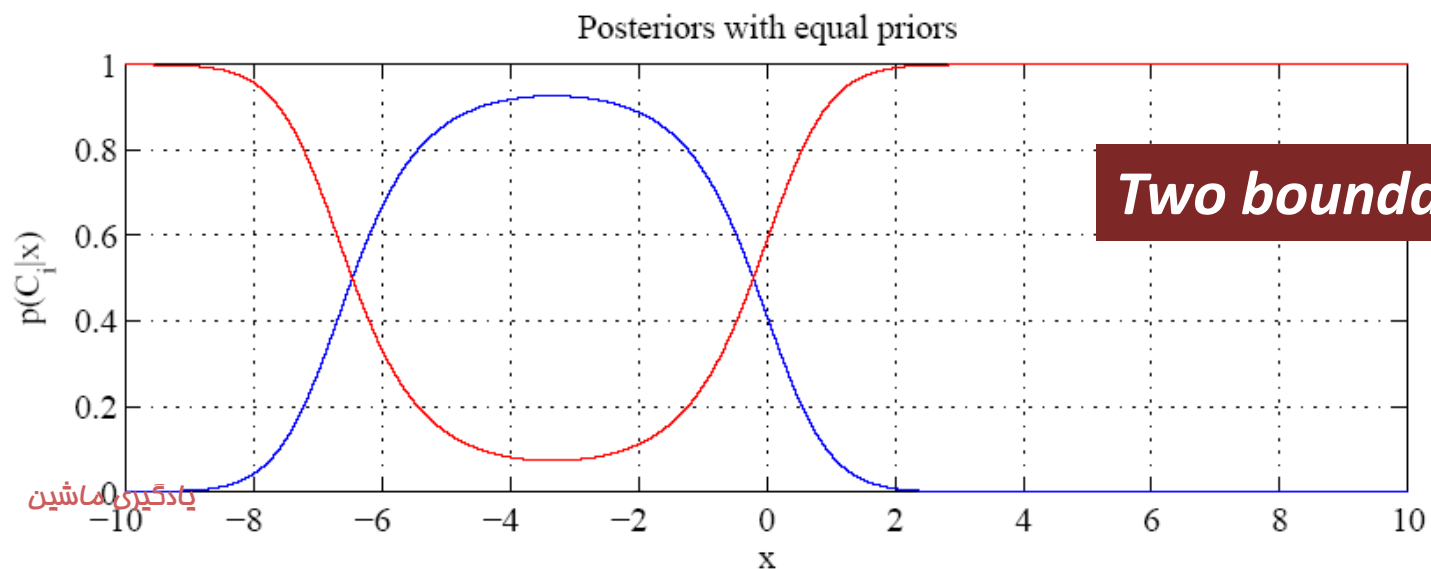
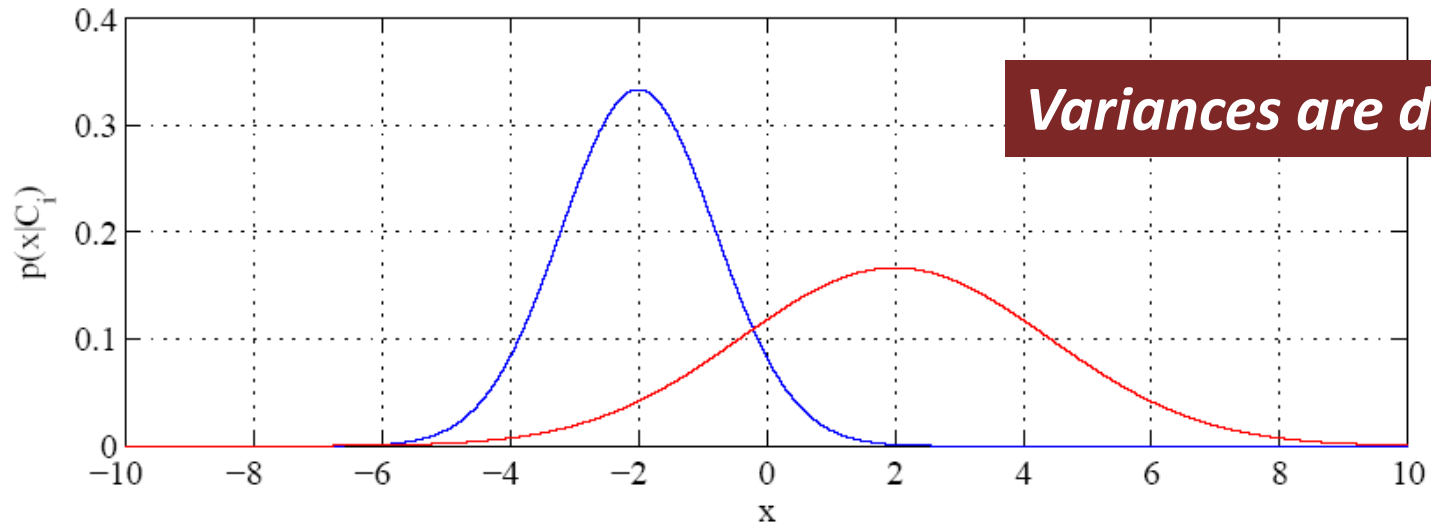
Posteriors with equal priors



# دسته‌بندی دو دسته با واریانس متفاوت و احتمال اولیه مساوی

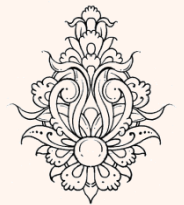
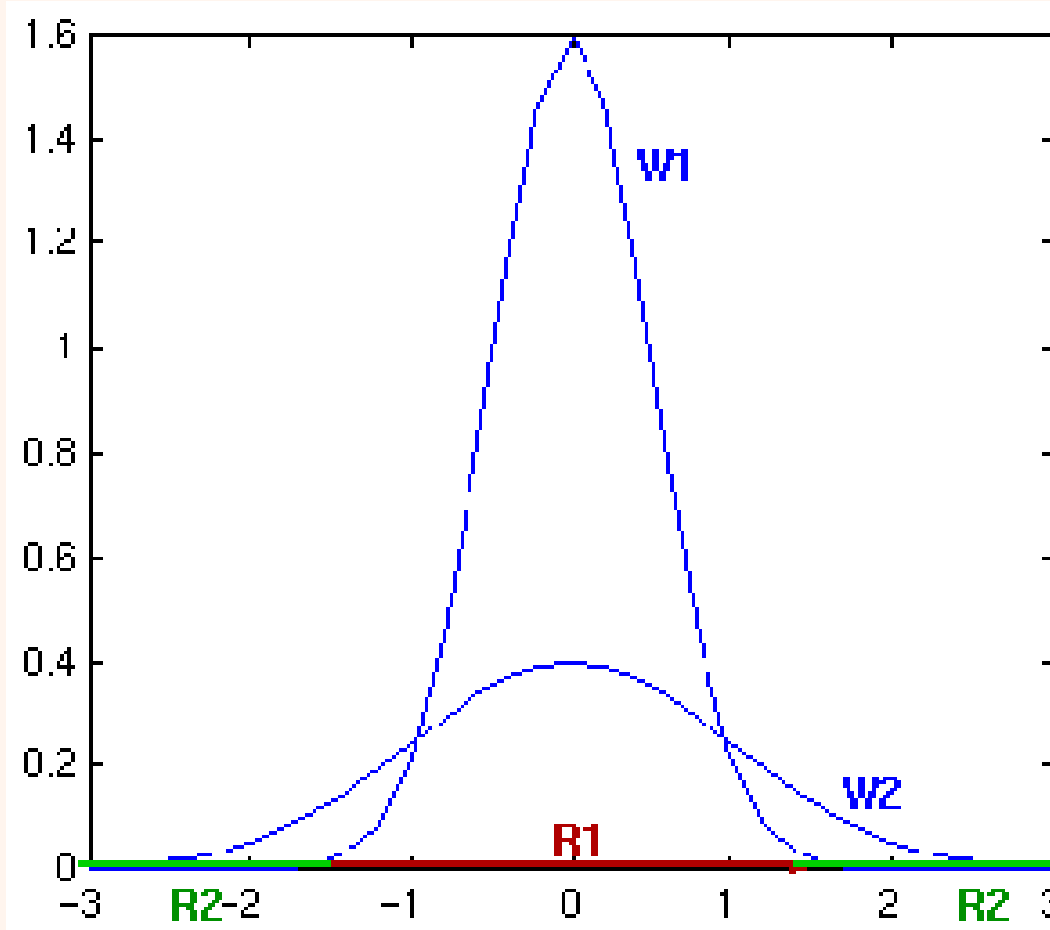
$$g_i(x) = -\frac{1}{2} \log 2\pi - \log s_i - \frac{(x - m_i)^2}{2s_i^2} + \log \hat{P}(C_i)$$

Likelihoods

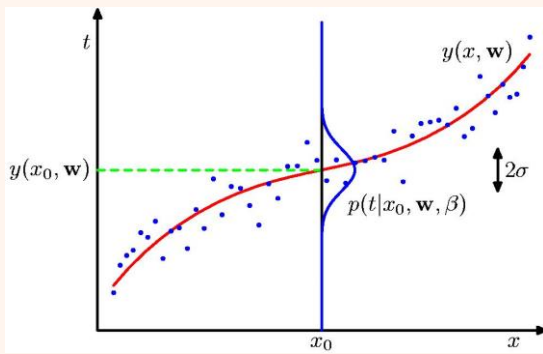


# مثال

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



# رگرسیون چندجمله‌ای

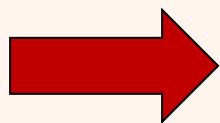
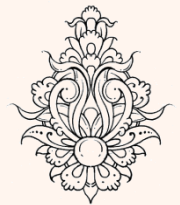


$$E(\theta | \mathcal{X}) = \frac{1}{2} \sum_{t=1}^N [r^t - g(x^t | \theta)]^2$$

*Least Squares estimates*

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

$$\frac{\partial E}{\partial w_i} = \sum_{t=1}^N [r^t - g(x^t | \theta)] \times \frac{\partial g}{\partial w_i} \quad \forall i \quad 0 \leq i \leq k$$



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



# رگرسیون چندجمله‌ای

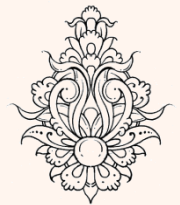
$$g(x^t | w_k, \dots, w_2, w_1, w_0) = w_k (x^t)^k + \dots + w_2 (x^t)^2 + w_1 x^t + w_0$$

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

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

$$\mathbf{A}\mathbf{w} = \mathbf{y}$$

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



# رگرسیون چندجمله‌ای

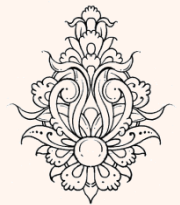
$$\mathbf{A} = (\mathbf{D}^T \mathbf{D}) \quad \mathbf{y} = \mathbf{D}^T \mathbf{r}$$

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

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

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

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



# معیارهای خطا

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

• مجموع مربعات خطا

*sum of squared error(SSE)*

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

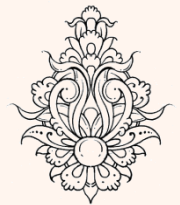
• خطای نسبی

*relative square error(RSE)*

*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.





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

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

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

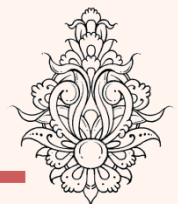
$$\bar{g}(x) = \frac{1}{M} \sum_i g_i(x)$$

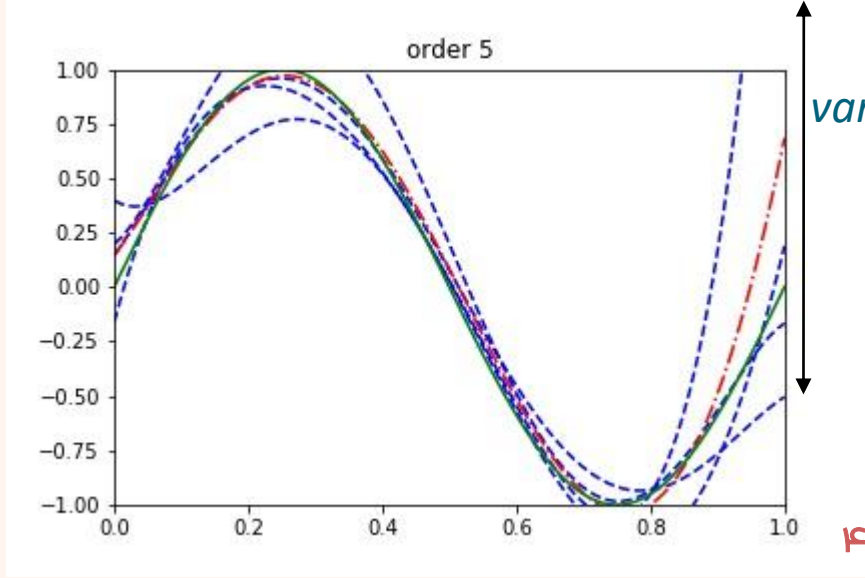
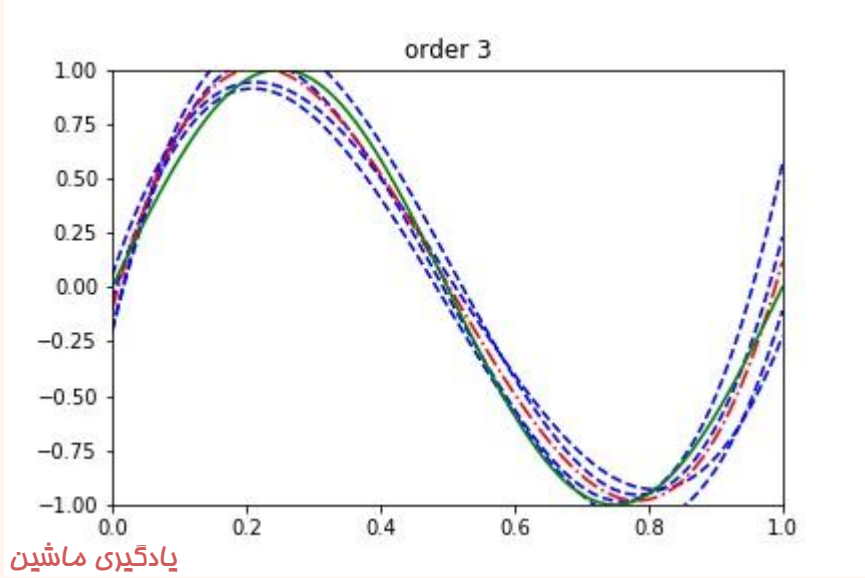
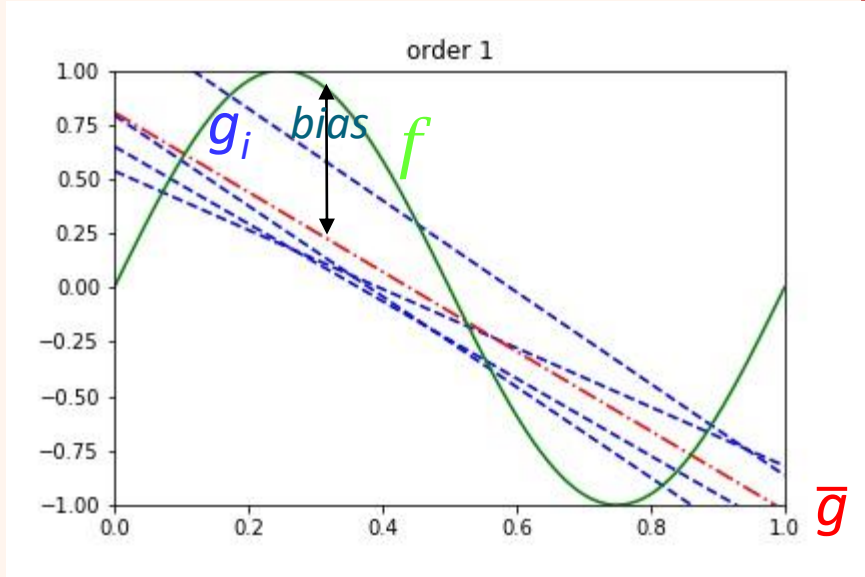
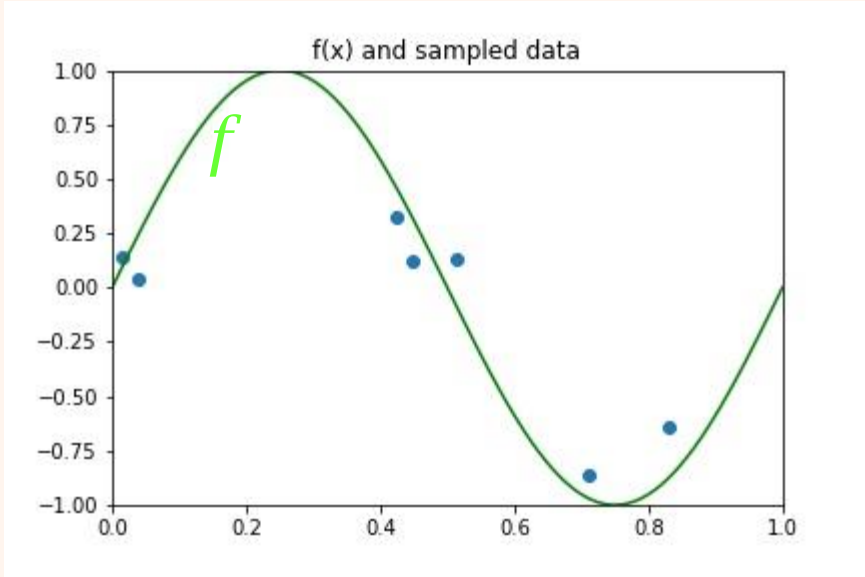
$$g_i(x) = 2$$

واریانس صفر است، اما بایاس بالایی دارد

$$g_i(x) = \sum_t r_i^t / N$$

بایاس کاهش می‌یابد، اما واریانس افزایش می‌یابد

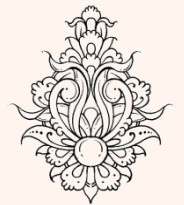
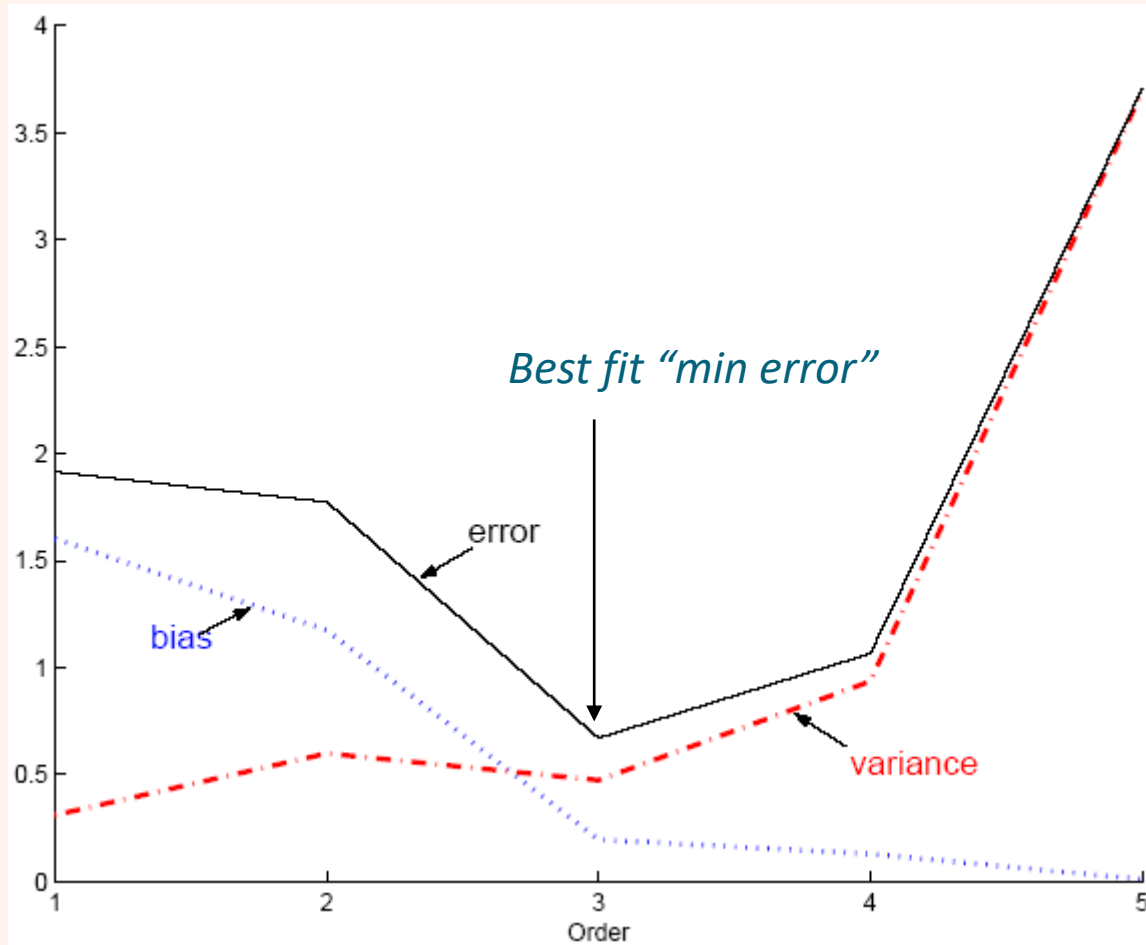




variance

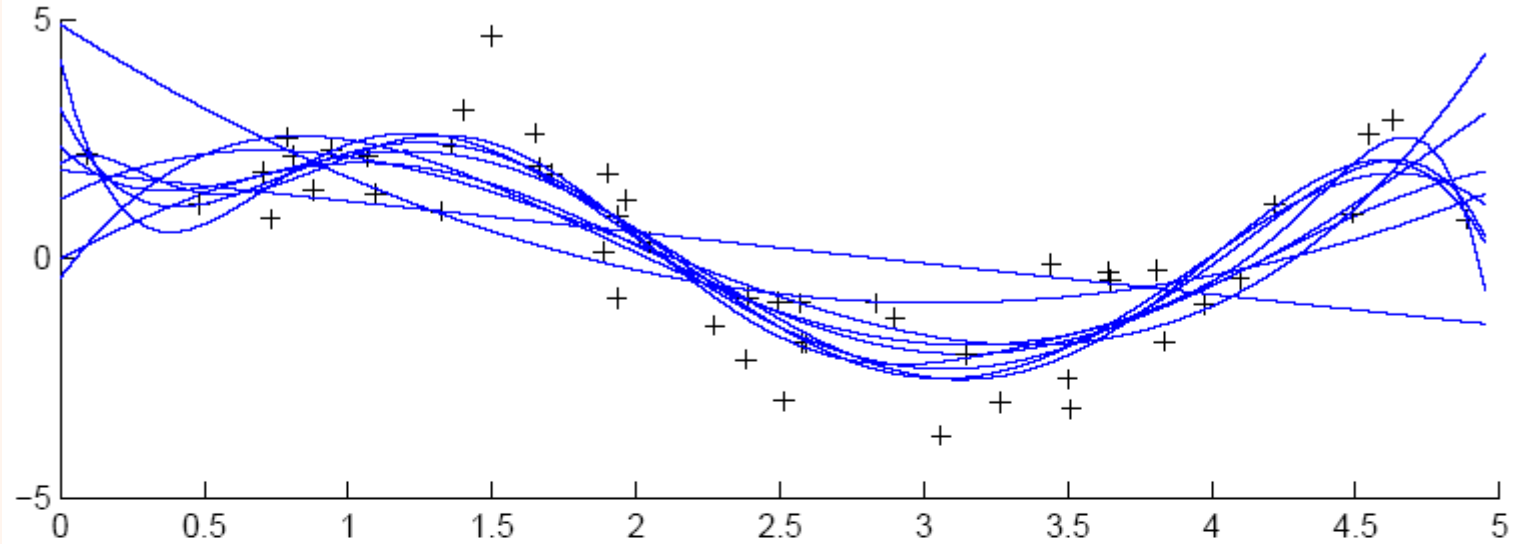


# انتخاب مدل

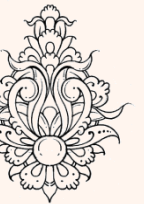
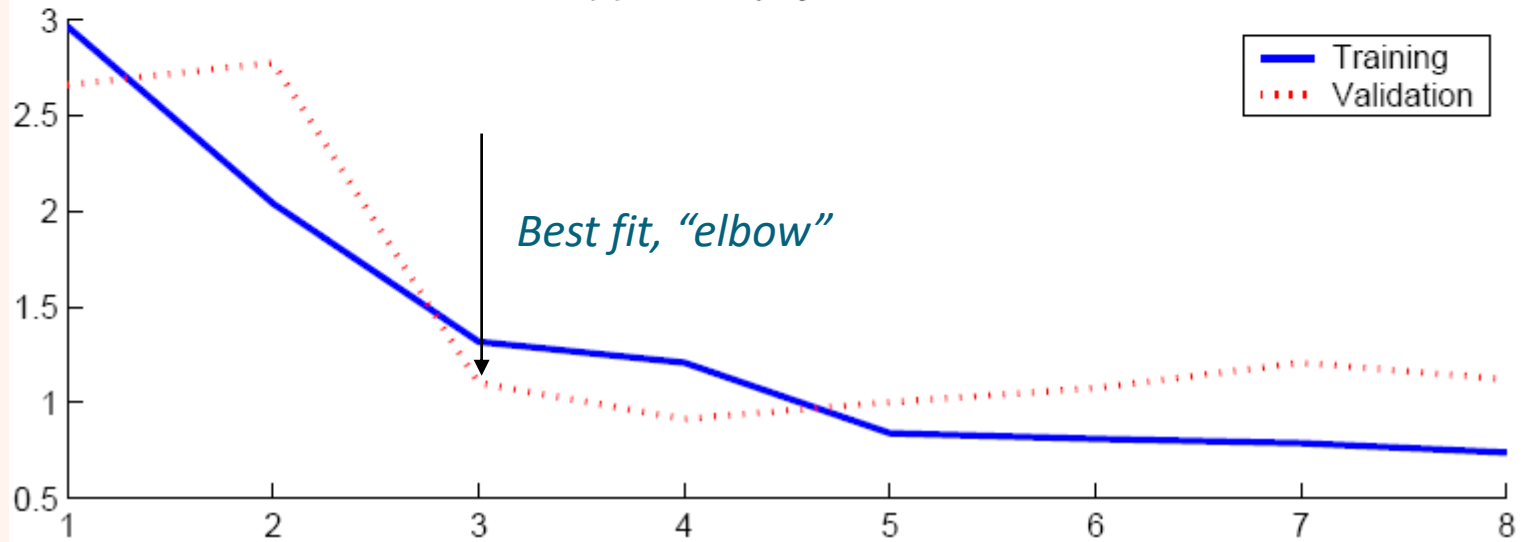


# Cross validation

(a) Data and fitted polynomials



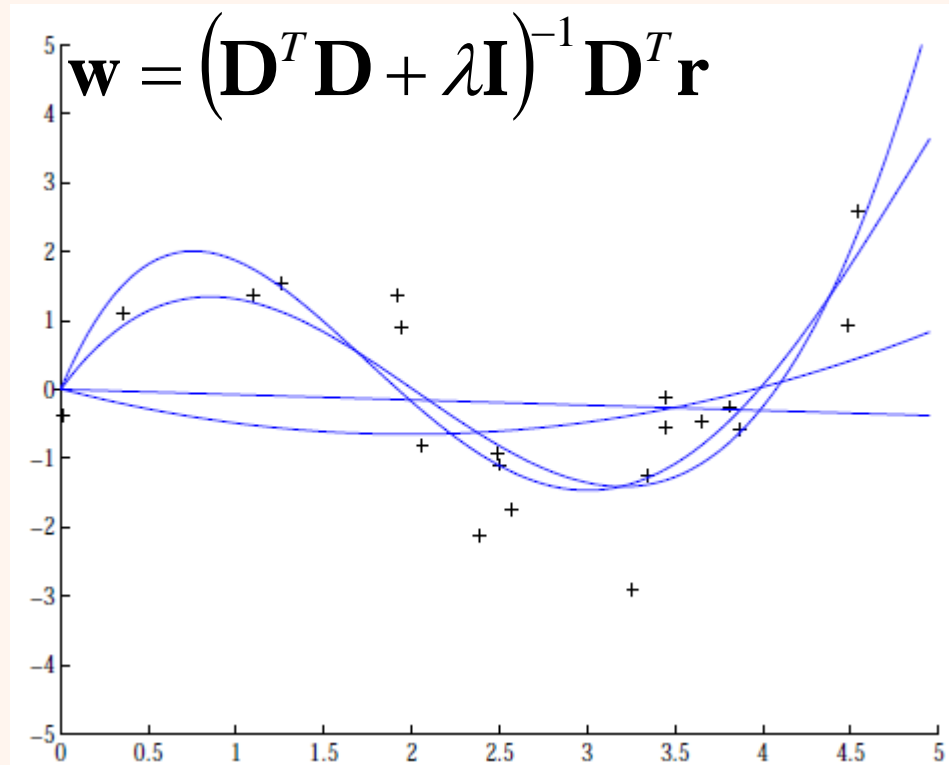
(b) Error vs polynomial order



# Regularization

## Penalize complex models

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



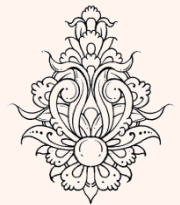
Coefficients increase in magnitude as order increases:

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]



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

# Regularization

9<sup>th</sup> Order Polynomial

