Machine Learning

روش‌های پارامتری

یادگیری ماشینی

(۱۳-۱۱-۸۰۵-۰۱)

فصل چهارم

دانشگاه شهید بهشتی

پژوهشکده‌ی فضای مجازی

پاییز ۱۳۹۷

احمد محمودی ازناوه

فهرست مطالب

- تخمین تابع چگالی احتمال
- تابع درست‌نمایی
- برآورد درست‌نمایی بیشینه
  - مثال
- ارزیابی برآورد
- برآورد بیشینه‌گر احتمال پسین
- دسته‌بندی پارامتری
- رگرسیون

Parametric Estimation

برآورد پارامتری

- در فصل پیش در مورد «اتخاذ تصمیم بهینه» با در نظر گرفتن احتمال مشاهده‌ی ورودی با فرض دانستن کلاس و احتمال وقوع کلاس 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)

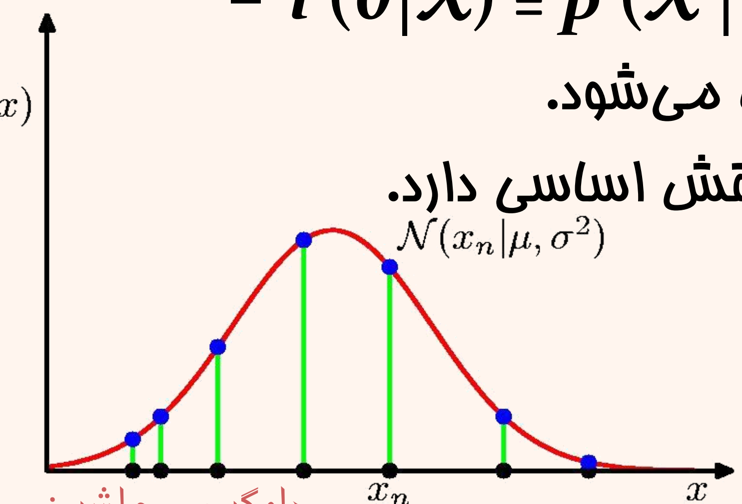
Sufficient statistic

Likelihood Function

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



یادگیری ماشین

Bishop

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

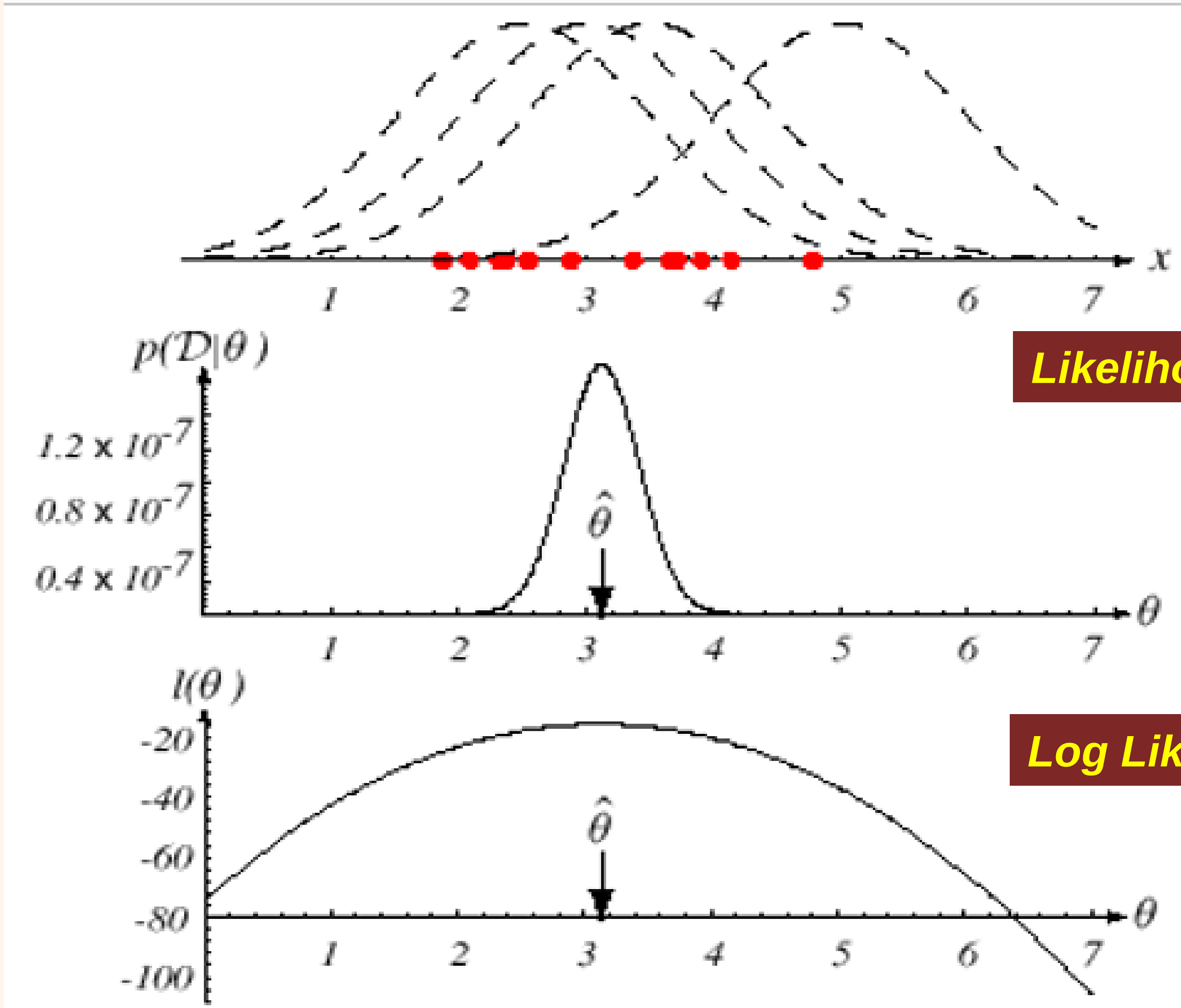
- $l\,(\theta|\mathcal{X}) = p\,(\mathcal{X}\,|\theta) = \prod_t p\,(x^t|\theta)$
- در برآورد درست‌نمایی بیشینه در پی یافتن θ هستیم به گونه‌ای که احتمال تعلق X به p حداکثر شود؛ درست‌نمایی بیشینه شود.
- برای سادگی محاسبات، به جای درست‌نمایی، از لگاریتم آن استفاده می‌شود:

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

Log likelihood

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

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 $\{0,1\}$

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

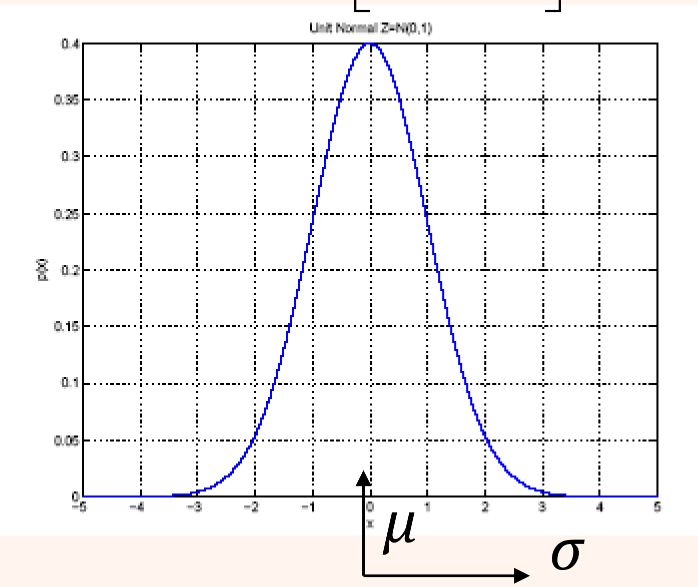
$\mathcal{L}(p_1,p_2,\ldots,p_K|\mathcal{X}) = \log \prod_t \prod_i p_i^{x_i^t} = \log \prod_i p_i^{\Sigma t(x_i^t)}$

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:

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

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

Evaluating an Estimator

ارزیابی برآورد

- یک نمونه از داده‌ها(جمعیت): $\mathcal{X}$
- پارامتر مجهول: θ
- برآورد پارامتر از روی داده‌ها $d = d(\mathcal{X})$
- معیار کیفیت تخمین: $(d(\mathcal{X})-\theta)^2$
  - با توجه به این که این معیار به نمونه‌ها وابسته است، از میانگین استفاده می‌کنیم:

$$r(d,\theta)=E[(d(\mathcal{X})-\theta)^2]$$

Mean square error

  - همچنین «بایاس تخمین» به صورت زیر تعریف می‌شود:

$$b_\theta(d)=E[d(\mathcal{X})]-\theta$$

- چنان‌چه این مقدار برابر صفر باشد، d را unbiased estimator می‌گویند.

مثال- تخمین میانگین

- در صورتی که x^t نمونه‌های از یک توزیع با میانگین μ باشد،

$$E[m]=E\left[\frac{\sum_t x^t}{N}\right]=\frac{1}{N}\sum_t E\left[x^t\right]=\frac{N\mu}{N}=\mu$$

- میانگین نمونه‌ها unbiased است.
- در صورتی که واریانس تخمین، با افزایش تعداد نمونه‌ها به صفر میل کند، به برآورد انجام شده «سازگار» گفته می‌شود.

Consistent estimator

Var(m) →0 as N→∞

$$\operatorname{var}(m)=\operatorname{var}\left(\frac{\sum_t x^t}{N}\right)=\frac{1}{N^2}\sum_t \operatorname{var}\left(x^t\right)=\frac{N\sigma^2}{N^2}=\frac{\sigma^2}{N}$$

مثال- تخمین واریانس

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

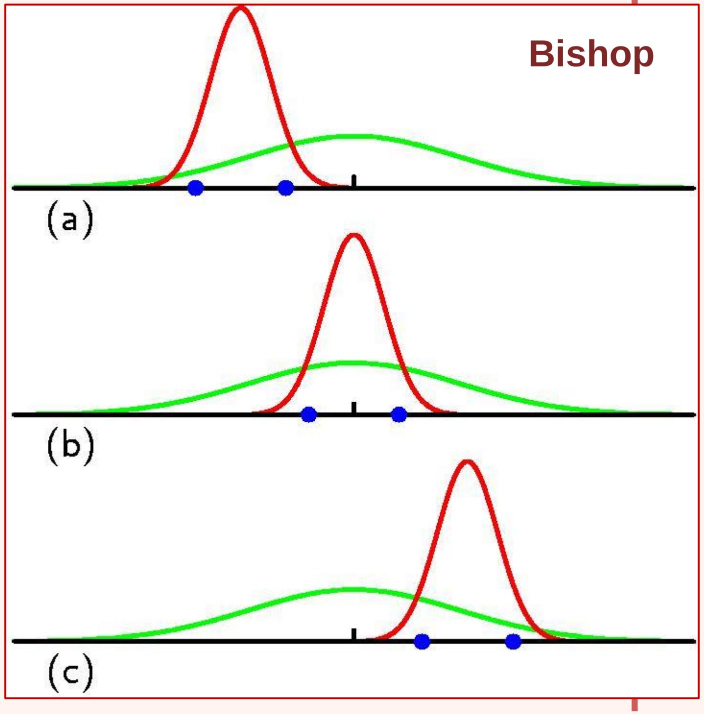
$$E\left[s^2\right] = \frac{\sum_t E\left[\left(x^t\right)^2\right] - N.E\left[m^2\right]}{N}$$

یادآوری

$$Var(X) = E\left[X^2\right] - E[X]^2$$

$$E\left[\left(x^t\right)^2\right] = \sigma^2 + \mu^2 \quad E\left[m^2\right] = \frac{\sigma^2}{N} + \mu^2$$

$$E\left[s^2\right] = \frac{N\left(\sigma^2 + \mu^2\right) - N\left(\sigma^2/N + \mu^2\right)}{N} = \frac{N-1}{N}\sigma^2 \neq \sigma^2$$

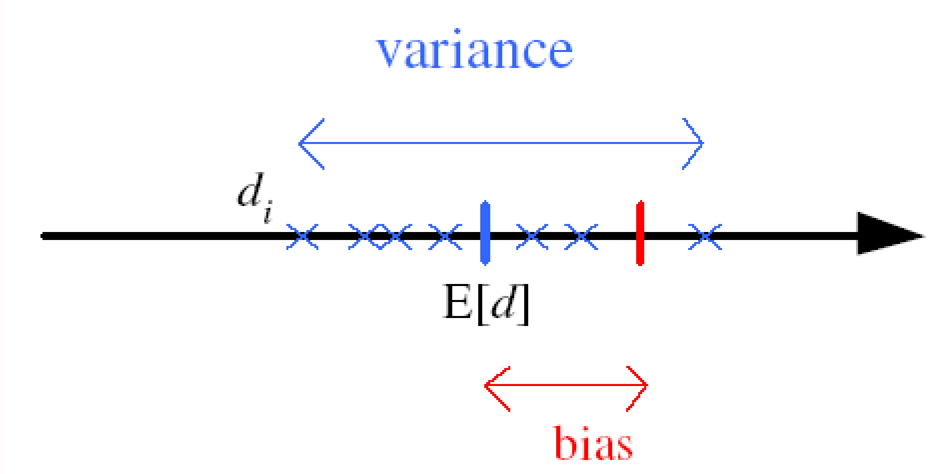


asymptotically unbiased estimator

$b_\theta(s) \rightarrow 0$ as $N \rightarrow \infty$

Mean square error:

$$r(d,\theta) = E[(d-\theta)^2]$$
$$= (E[d]-\theta)^2 + E[(d-E[d])^2]$$
$$= \text{Bias}^2 + \text{Variance} \quad (4.11)$$



برآورد بیشینه‌گر احتمال پسین

Maximum a Posteriori (MAP)

- در MLE، پارامتر مورد نظر به عنوان مجهول در نظر گرفته می‌شود، ممکن است در مورد پارامتر مورد نظر از پیش اطلاعاتی (prior information) داشته باشیم. این اطلاعات می‌توانند به تخمین دقیق‌تر کمک کنند، به ویژه زمانی که داده‌های آموزش کم‌تعداد باشند.
  - در این حالت به θ به صورت یک متغیر تصادفی نگاه می‌کنیم.
  - به عنوان مثال می‌دانیم، θ با احتمال نوددرصد، با توزیع گاوسی بین ۵ و ۹ به (صورت متقارن) قرار دارد.

$$P\left\{-1.64<\frac{\theta-\mu}{\sigma}<1.64\right\}=0.9$$

$$P\{\mu-1.64\sigma<\theta<\mu+1.64\sigma\}=0.9$$

$$P(\theta)\sim N\left(7,(2/1.64)^2\right)$$

برآورد بیشینه‌گر احتمال پسین

- در چنین حالتی اطلاعاتی در مورد p(θ) وجود دارد. با ترکیب این اطلاعات با آن‌چه داده‌ها به ما می‌گویند(likelihood density)، خواهیم داشت:

$$p(\theta|\mathcal{X}) = p(\mathcal{X}|\theta)\, p(\theta) / p(\mathcal{X})$$

- و

Maximum a Posteriori (MAP)

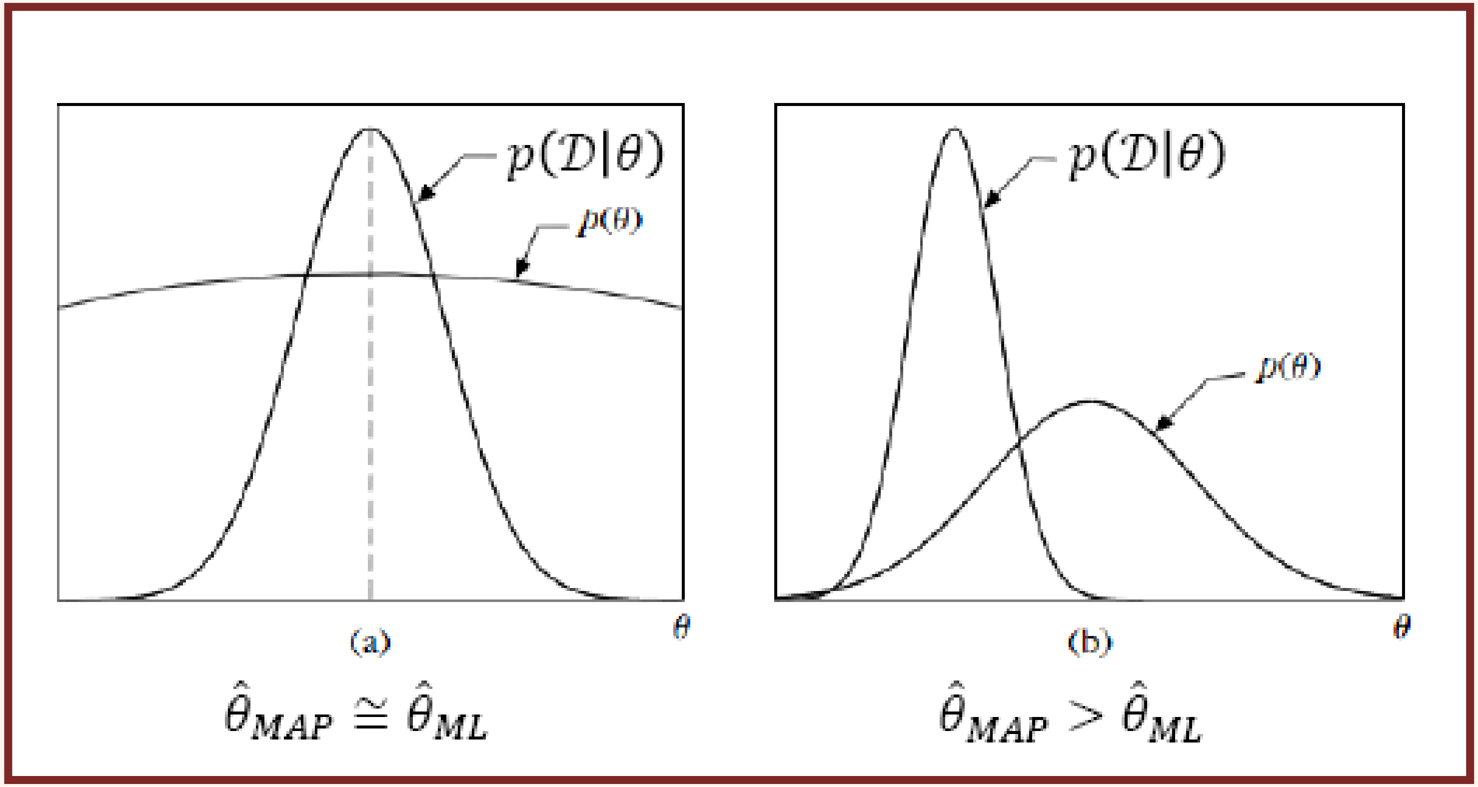
$$\theta_{\mathrm{MAP}} = \operatorname{argmax}_\theta\, p(\theta|\mathcal{X}) = \operatorname{argmax}_\theta\, p(\theta)\, p(\mathcal{X}|\theta)$$

تفاوت با ML در نظر گرفتن $p(\theta)$ است.

$$\theta_{\mathrm{ML}} = \operatorname{argmax}_\theta\, p(\mathcal{X}|\theta)$$

Maximum Likelihood (ML)

برآورد بیشینه‌گر احتمال پسین(ادامه...)



Pattern recognition, Sergios Theodoridis

در صورتی که $p(\theta)$ دارای توزیع یکنواخت باشد، دو روش پاسخ یکسانی به دست می‌آورند.

مثال

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

$$p(\mu) = \frac{1}{\sqrt{2\pi}\sigma_0} \exp\left[-\frac{(\mu-\mu_0)^2}{2\sigma_0^2}\right]$$

$$\boxed{p(\mu \mid X) \propto p(\mu) p(X \mid \mu)}$$

$$\prod_t p(\mu \mid x^t) = p(\mu) \prod_t p(x^t \mid \mu)$$

برای تخمین MAP باید رابطه‌ی زیر محاسبه شود:

$$\frac{\partial}{\partial \mu} \ln\left[p(\mu) \prod_t p(x^t \mid \mu)\right] = \frac{\partial}{\partial \mu}\left[\ln p(\mu) + \sum_t \ln p(x^t \mid \mu)\right] = 0$$

$$\frac{\partial}{\partial \mu}\left[-\frac{1}{2}\ln 2\pi - \ln \sigma_0 - \frac{(\mu-\mu_0)^2}{2\sigma_0^2} - \frac{N}{2}\ln 2\pi - \ln \sigma - \frac{\sum_t (x^t-\mu)^2}{2\sigma^2}\right]$$

مثال- ادامه

$$\mu_N = \frac{N\sigma_0^2\overline{x} + \sigma^2\mu_0}{N\sigma_0^2 + \sigma^2}$$

$$\mu_N = \frac{\mu_0 + \frac{\sigma_0^2}{\sigma^2}\sum_t x^t}{1 + \frac{\sigma_0^2}{\sigma^2} N}$$

$\mu_N \rightarrow \frac{1}{N}\sum_t x^t$ as $N \rightarrow \infty$

$\mu_N \rightarrow \frac{1}{N}\sum_t x^t$ as $\sigma_0 >> \sigma$

برای واریانس هم به صورت مشابه خواهیم داشت:

$$\sigma_N^2 = \frac{\sigma^2\sigma_0^2}{N\sigma_0^2 + \sigma^2}$$

استنباط بیزی

- رویکرد دیگر محاسبه‌ی $P(x|X)$ است، در شرایطی که $p(\theta)$ را می‌دانیم.

$$
\begin{aligned}
p(x|X) &= \int p(x,\theta|X)d\theta \\
&= \int p(x|\theta,X)p(\theta|X)d\theta \\
&= \int p(x|\theta)p(\theta|X)d\theta
\end{aligned}
$$

اگر پارامتر θ را بدانیم، کل توزیع مشخص است

میانگین وزن‌دار تخمین را بر اساس احتمال مقادیر مدل

- عیب عمده‌ی این روش حجم محاسبات بالاست، و محاسبات تحلیلی تنها در حالات خاصی امکان‌پذیر است.

برای سادگی می‌توان فرض کرد که $P(\theta|X)$ شبیه تابع ضربه است، در این صورت

$P(x|X)=P(x|\theta_{MAP})$

دسته‌بندی پارامتری

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

or

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

تابع جداساز

در صورتی که چگالی کلاس 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

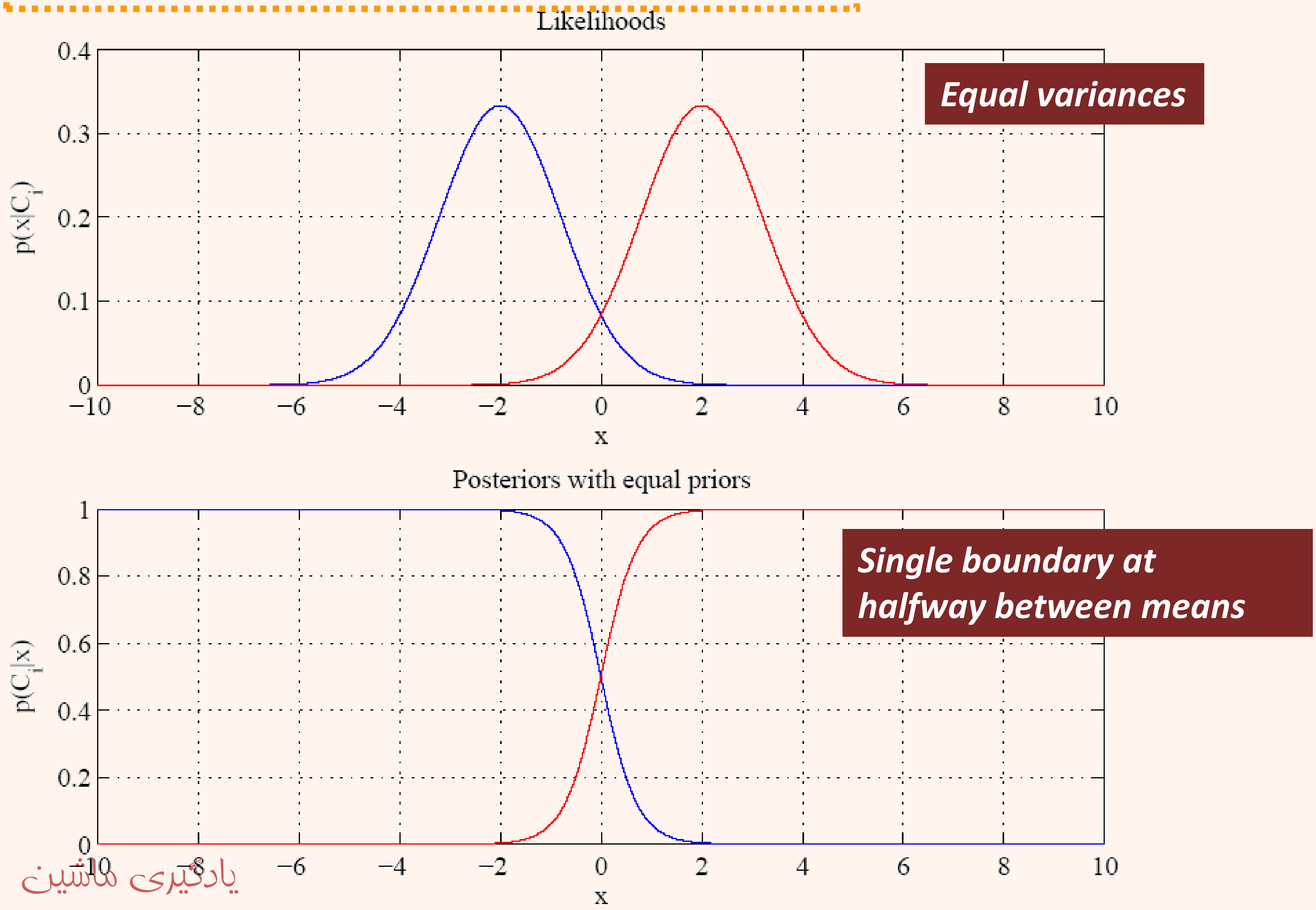
$$\hat{P}(C_i) = \frac{\sum_t r_i^t}{N} \quad m_i = \frac{\sum_t x^t r_i^t}{\sum_t r_i^t} \quad s_i^2 = \frac{\sum_t \left(x^t - m_i\right)^2 r_i^t}{\sum_t r_i^t}$$

تابع جداساز

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

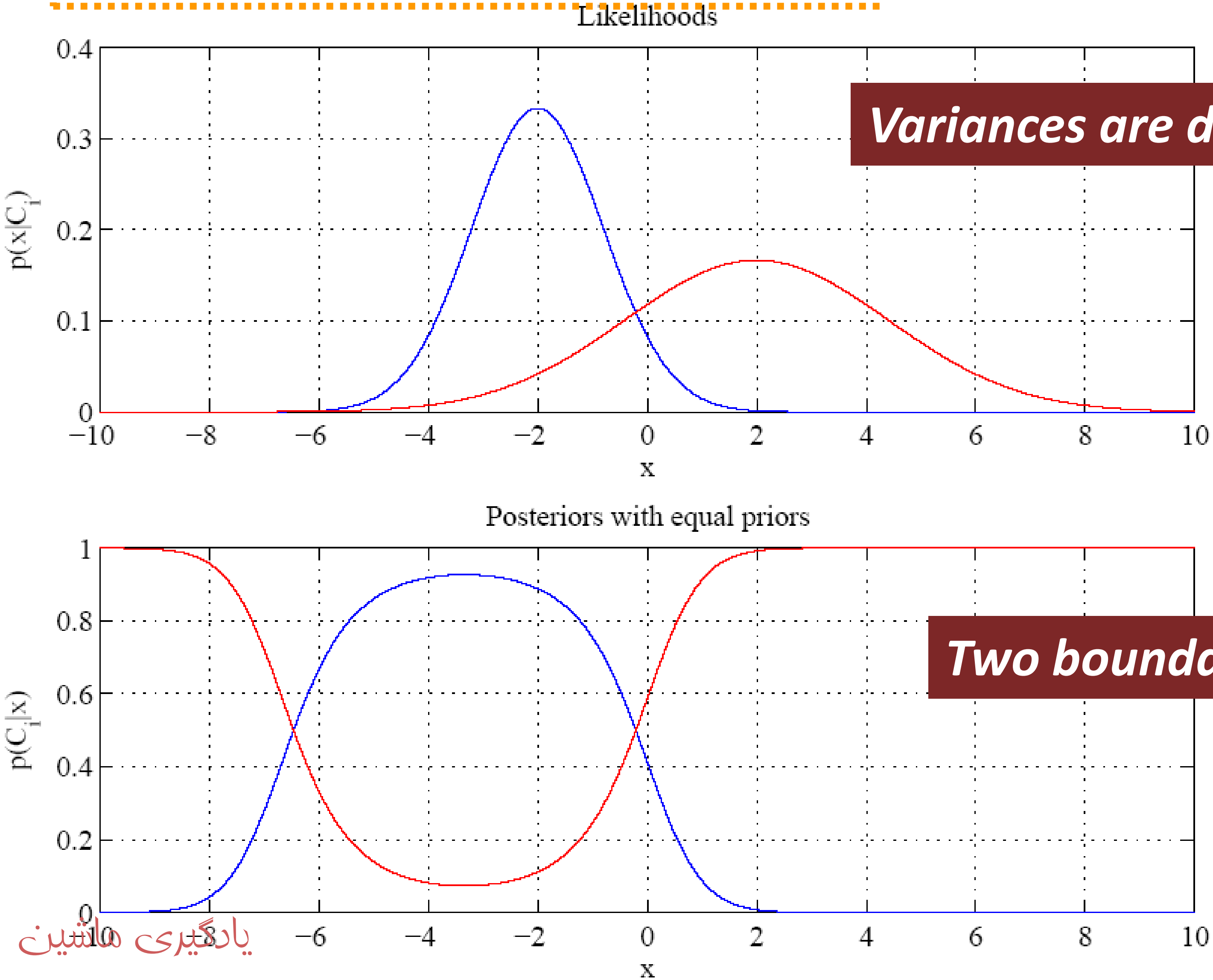
دسته‌بندی دو کلاس با واریانس یکسان

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



دسته‌بندی دو کلاس با واریانس متفاوت

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

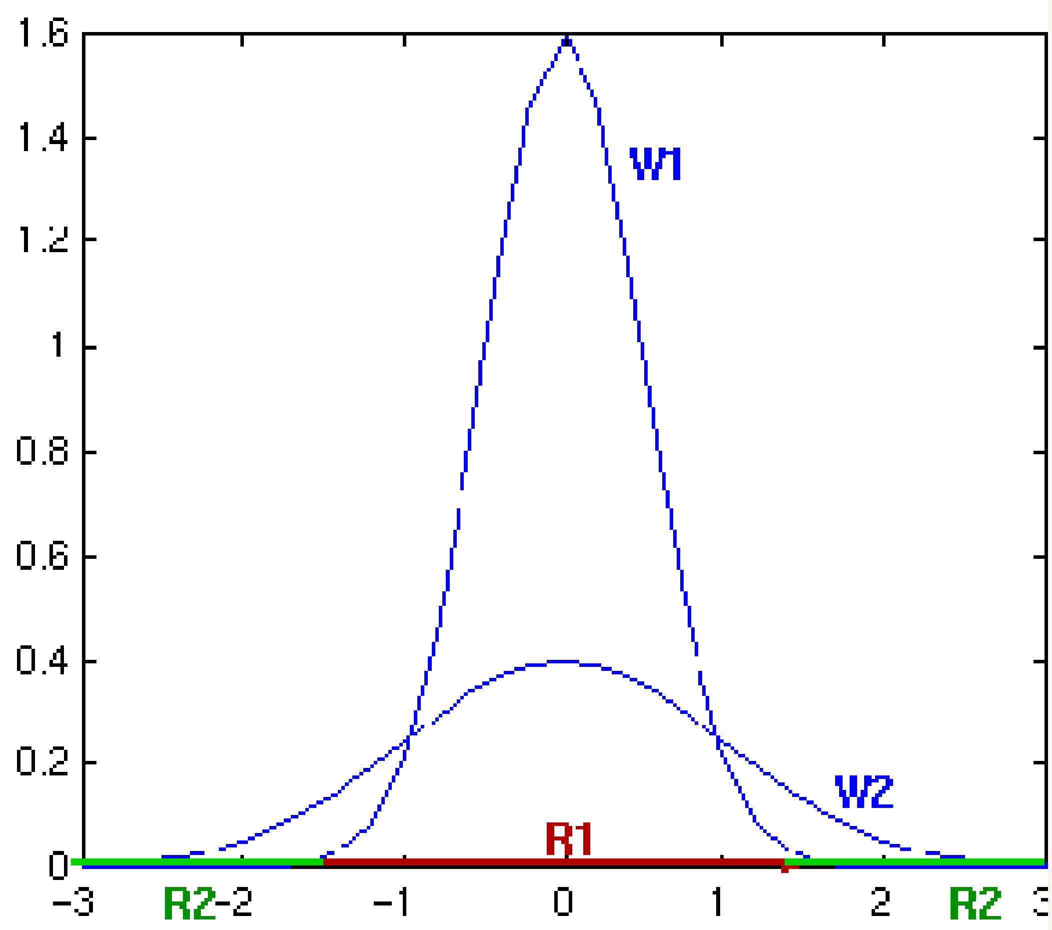


Variances are different

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

Dependent variable

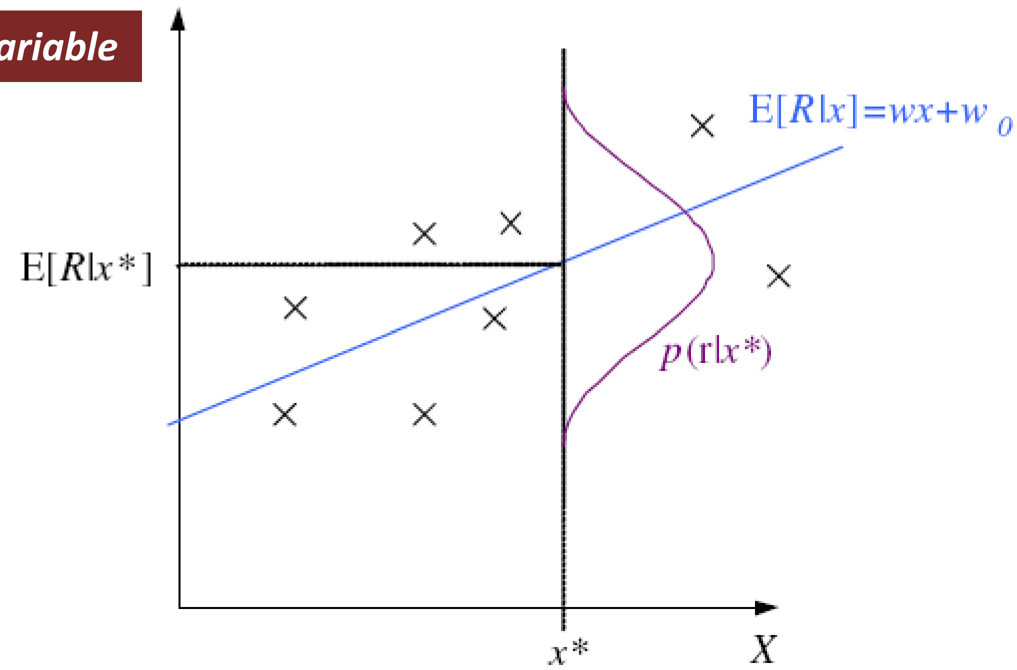
Independent variable

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

$$\text{estimator} : g(x \mid \theta)$$

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

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



$$\boxed{p(x,r) = p(r \mid x)\,p(x)}$$

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

$$= \log \prod_{t=1}^{N} p(r^t \mid x^t) + \log \prod_{t=1}^{N} p(x^t)$$

باتوجه به این که به تخمین ربطی ندارد، این بخش در نظر گرفته نمی‌شود

محاسبه‌ی تابع خطا

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

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

باتوجه به این که به تخمین ربطی ندارد، این بخش در نظر گرفته نمی‌شود

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

Least Squares estimates

رگرسیون خطی

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

$$g\left(x^t|w_1,w_0\right)=w_1x^t+w_0$$

$$\sum_t r^t=Nw_0+w_1\sum_t x^t$$

$$\sum_t r^tx^t=w_0\sum_t x^t+w_1\sum_t\left(x^t\right)^2$$

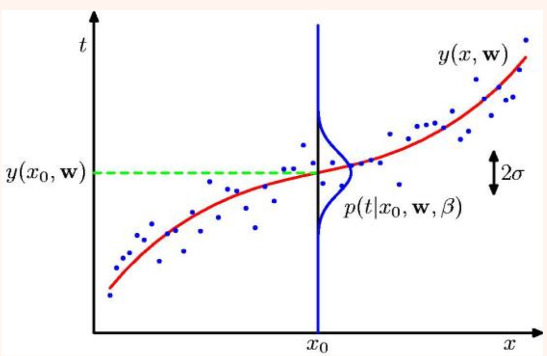
$$\mathbf{A}=\begin{bmatrix}N & \sum_t x^t\\ \sum_t x^t & \sum_t\left(x^t\right)^2\end{bmatrix}\quad \mathbf{w}=\begin{bmatrix}w_0\\ w_1\end{bmatrix}\quad \mathbf{y}=\begin{bmatrix}\sum_t r^t\\ \sum_t r^tx^t\end{bmatrix}$$

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

رگرسیون چندجمله‌ای

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

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



$$\mathrm{w}=\begin{bmatrix} w_0 \\ w_1 \\ w_2 \\ \vdots \\ w_k \end{bmatrix} \qquad \mathbf{y}=\begin{bmatrix} \sum_t r^t \\ \sum_t r^t x^t \\ \sum_t r^t \left(x^t\right)^2 \\ \vdots \\ \sum_t r^t \left(x^t\right)^k \end{bmatrix}$$

رگرسیون چندجمله‌ای

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

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

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

معیارهای خطا

- مجموع مربعات خطا

sum of squared error

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

- خطای نسبی

relative square error

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

- قدرمطلق خطا

Absolute Error

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

ε-sensitive Error

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

Bias and Variance

Expected square error at x

$$\underbrace{E\left[(r-g(x))^2 \mid x\right]}_{\text{Expected square error at x}} = \underbrace{\boxed{E\left[(r-E[r \mid x])^2 \mid x\right]}}_{\text{noise}} + \underbrace{\boxed{(E[r \mid x]-g(x))^2}}_{\text{squared error}}$$

noise

squared error

به مدل بستگی ندارد، واریانس نویز است؛ در واقع بخشی از خطاست که قابل حذف نیست

میزان خطا: وابسته به داده‌های آموزشی و مدل است

$$E_X\left[(E[r \mid x]-g(x))^2 \mid x\right] = \underbrace{\boxed{(E[r \mid x]-E_X[g(x)])^2}}_{\text{bias}} + \underbrace{\boxed{E_X\left[(g(x)-E_X[g(x)])^2\right]}}_{\text{variance}}$$

bias

variance

معیاری است که میزان خطا را صرفنظر از نمونه‌های آموزشی نشان می‌دهد

با تغییرات داده‌های آموزشی، مقدار g(x) به چه میزان تغییر می‌کند.

Bias/Variance Dilemma

مثال

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

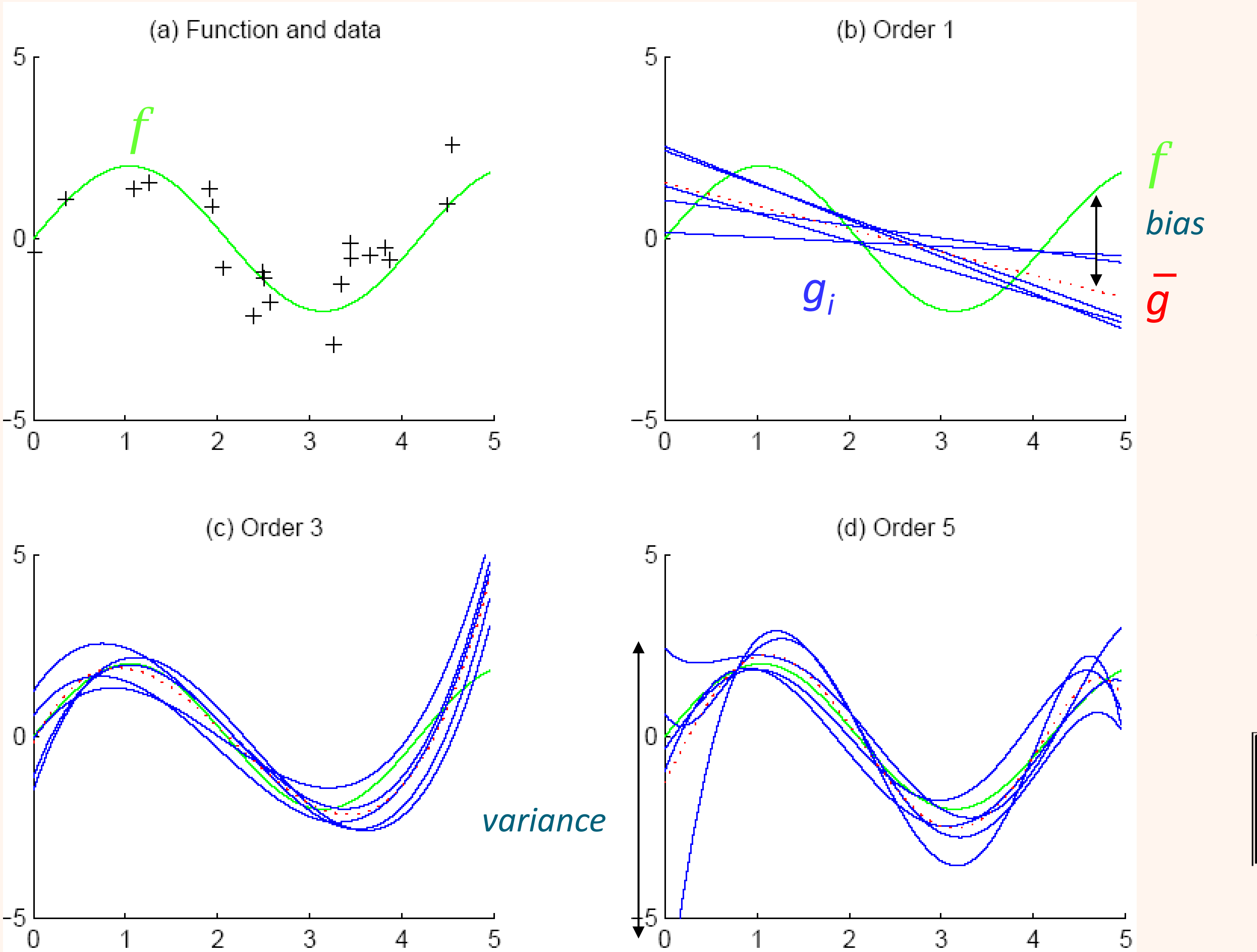
$$\text{Bias}^2(g)=\frac{1}{N}\sum_t\left[\bar{g}\left(x^t\right)-f\left(x^t\right)\right]^2$$

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

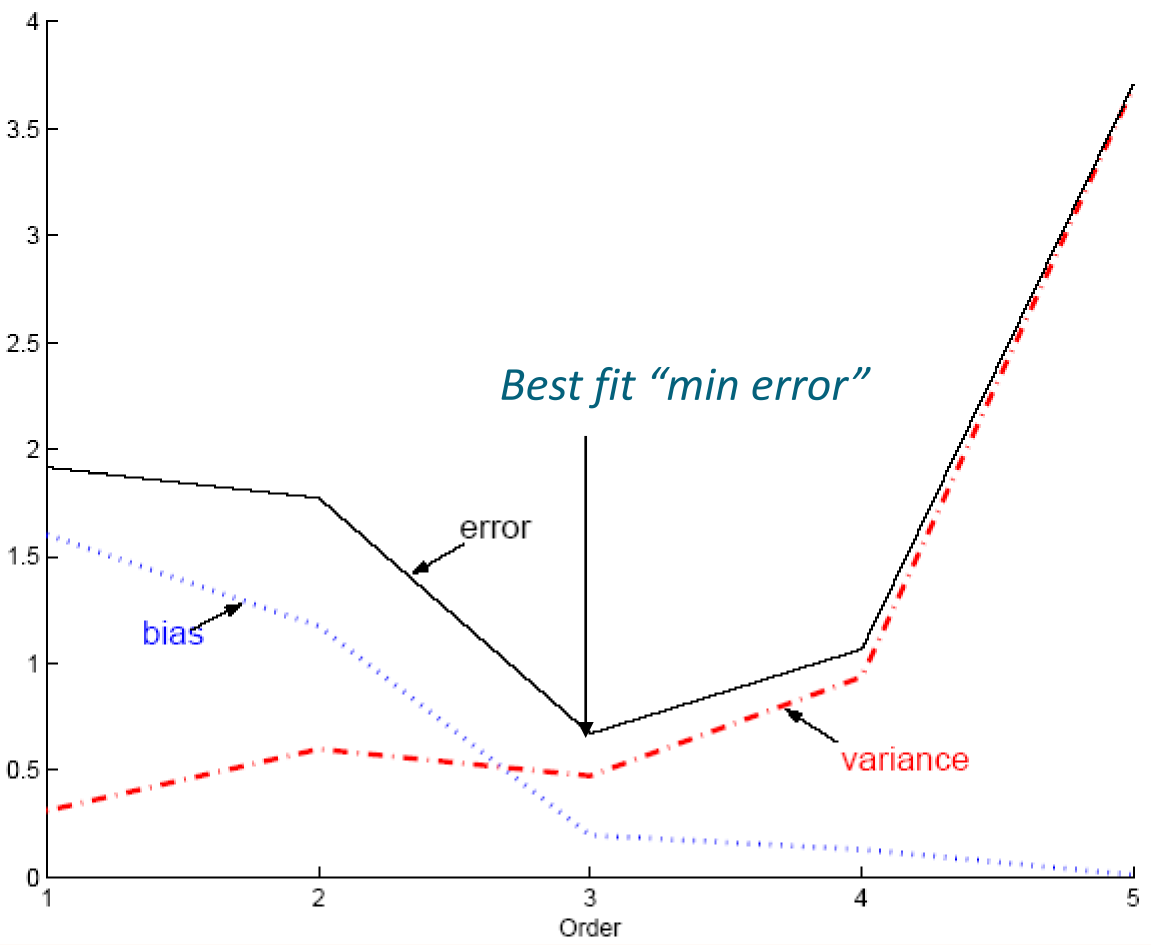
$$\bar{g}(x)=\frac{1}{M}\sum_i g_i(x)$$

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

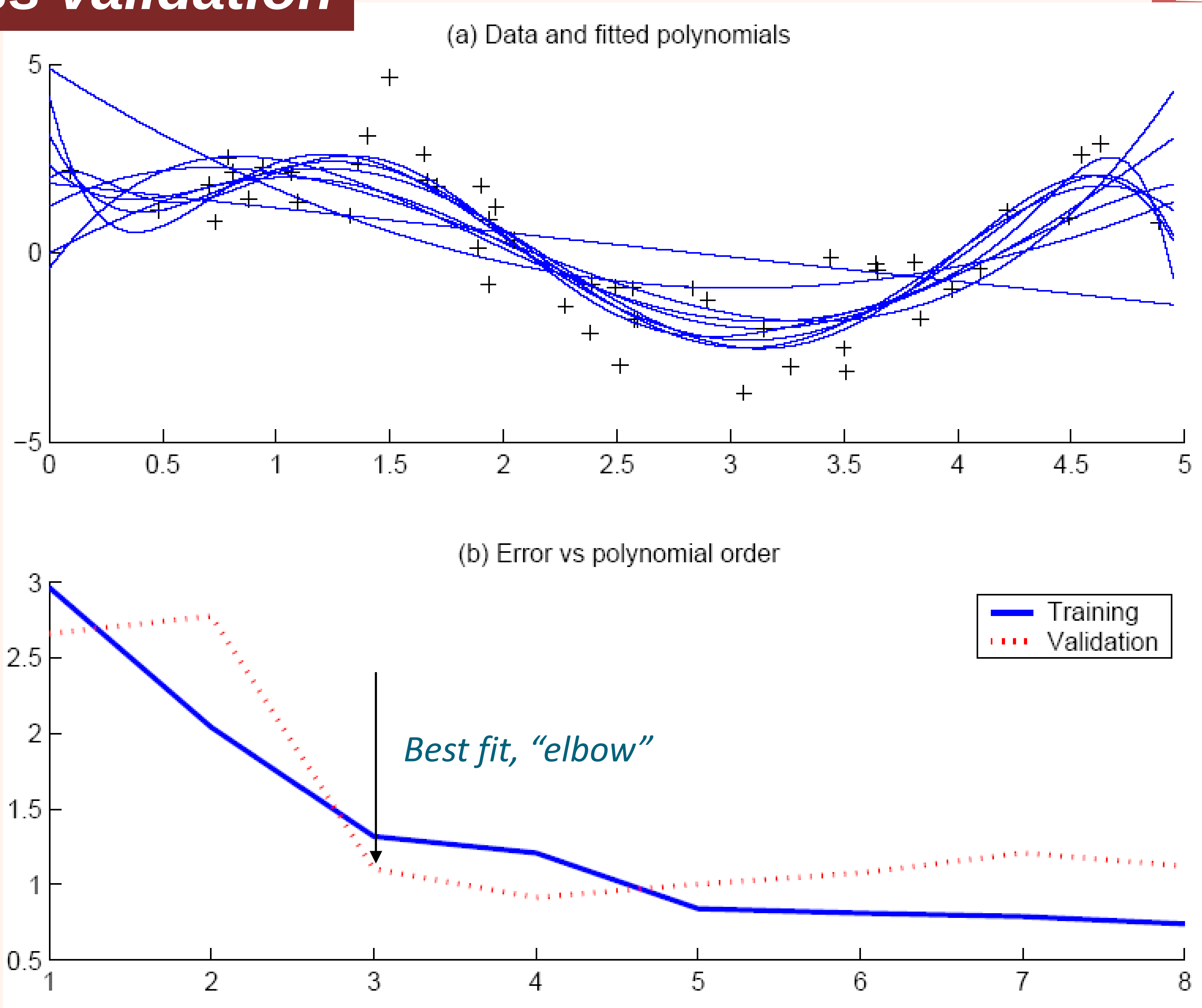
$g_i(x)=\sum_t r^t_i/N$ بایاس کاهش می‌یابد، اما واریانس افزایش می‌یابد


(a) Function and data
f
(b) Order 1
f
bias
g_i
$\bar{g}$
(c) Order 3
variance
(d) Order 5
5
0
-5
0
1
2
3
4
5

انتخاب مدل



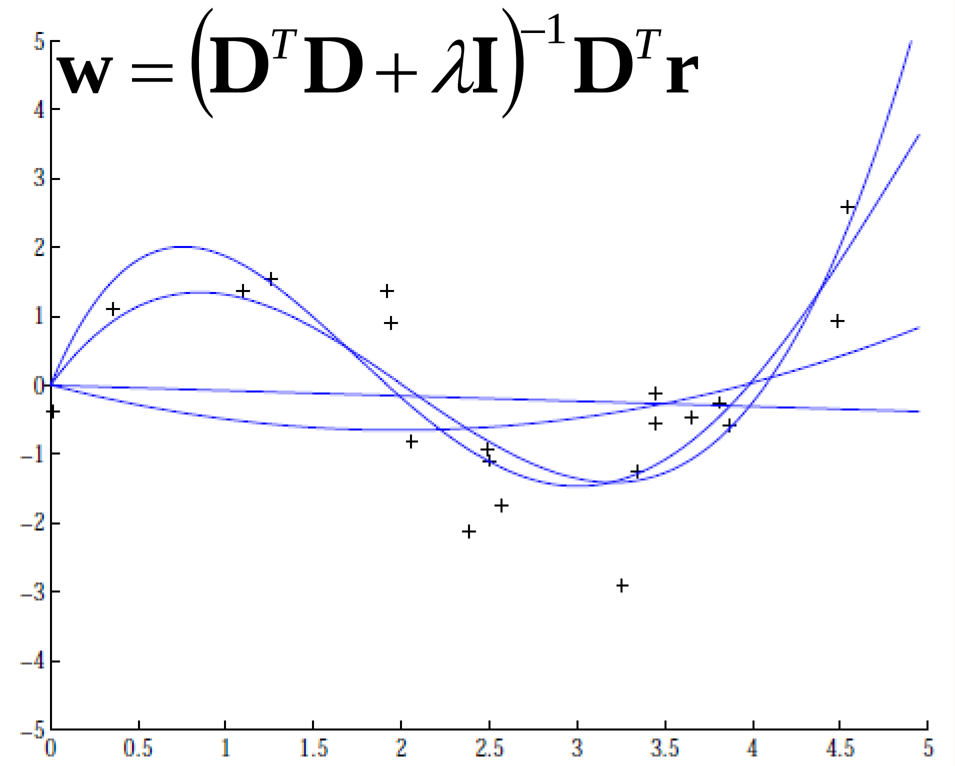
Cross validation



Regularization

Penalize complex models

E'=error on data + λ model complexity



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

Coefficients increase in magnitude as order increases:

1: [-0.0769, 0.0016]

2: [0.1682, -0.6657, 0.0080]

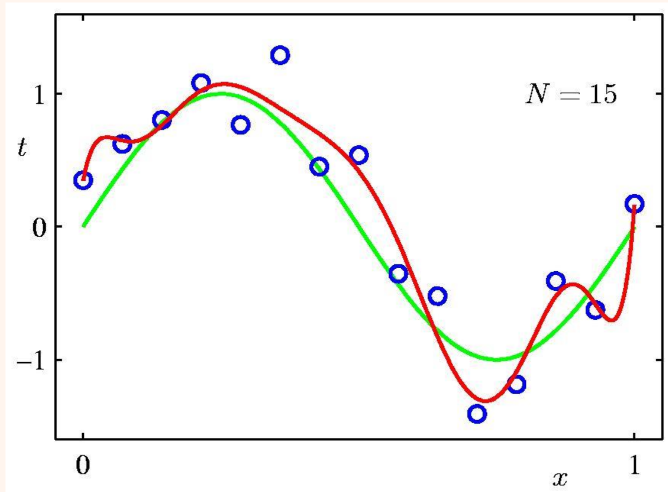
3: [0.4238, -2.5778, 3.4675, -0.0002

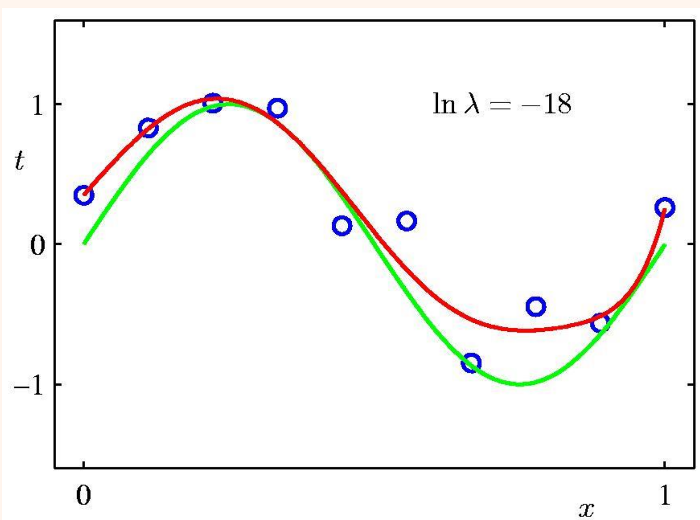
4: [-0.1093, 1.4356, -5.5007, 6.0454, -0.0019]

$$regularization: E(\mathbf{w} \mid \mathcal{X}) = \frac{1}{2}\sum_{t=1}^{N}\left[r^t - g(x^t \mid \mathbf{w})\right]^2 + \lambda \sum_i w_i^2$$

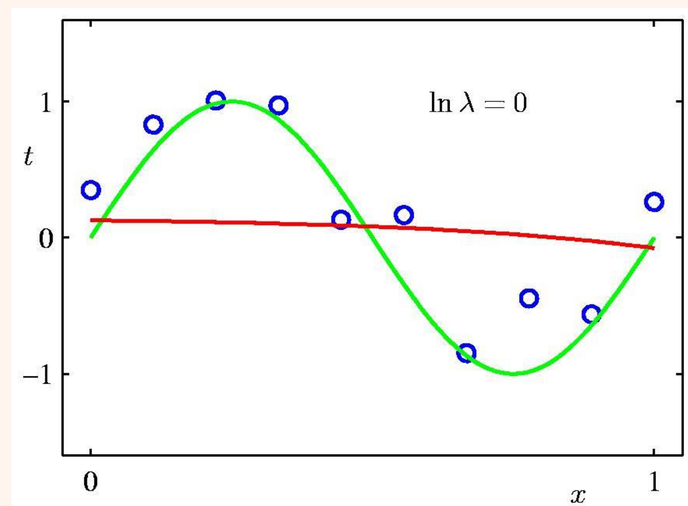
Regularization

9th Order Polynomial





$$\ln \lambda = -18$$



$$\ln \lambda = 0$$