

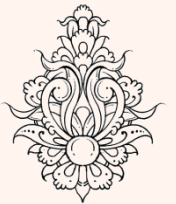
شناسایی آماری الگو
بخش دوم
(۰۱-۷۱۱-۱۰-۱۴۱)



دانشگاه شهید بهشتی
پژوهشکده‌ی فضای مجازی
زمستان ۱۳۹۵
احمد محمودی ازناوه

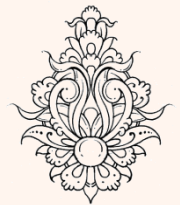
فهرست مطالب

- یادگیری بیزی
 - معیارهای تصمیم‌گیری
- تخمین تابع چگالی احتمال
- تابع درست‌نمایی
- برآورد درست‌نمایی بیشینه
 - مثال
- ارزیابی برآورد
- برآورد بیشینه‌گر احتمال پسین
- دسته‌بندی پارامتری



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 دسته‌بندی اعتبار مشتریان:

– ورودی: درآمد و پس‌انداز

– خروجی: مشتری High risk و low risk

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

– پیش‌بینی:

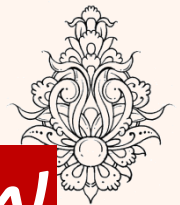
– high risk ($C=1$) or low risk ($C=0$)

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

امتثال شرطی

or

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



دسته‌بندی (ادامه...)

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

Bayes' Rule

posterior

احتمال پسین

با چه احتمالی C ، کلاس مربوط به x است.

احتمال پیشین

prior

درست‌نمایی کلاس

Class likelihood

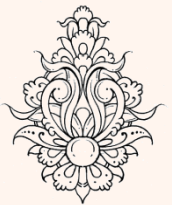
با چه احتمالی x توسط کلاس C تولید می‌شود.

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

evidence

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

$$p(x) = p(x|C=1)P(C=1) + p(x|C=0)P(C=0)$$

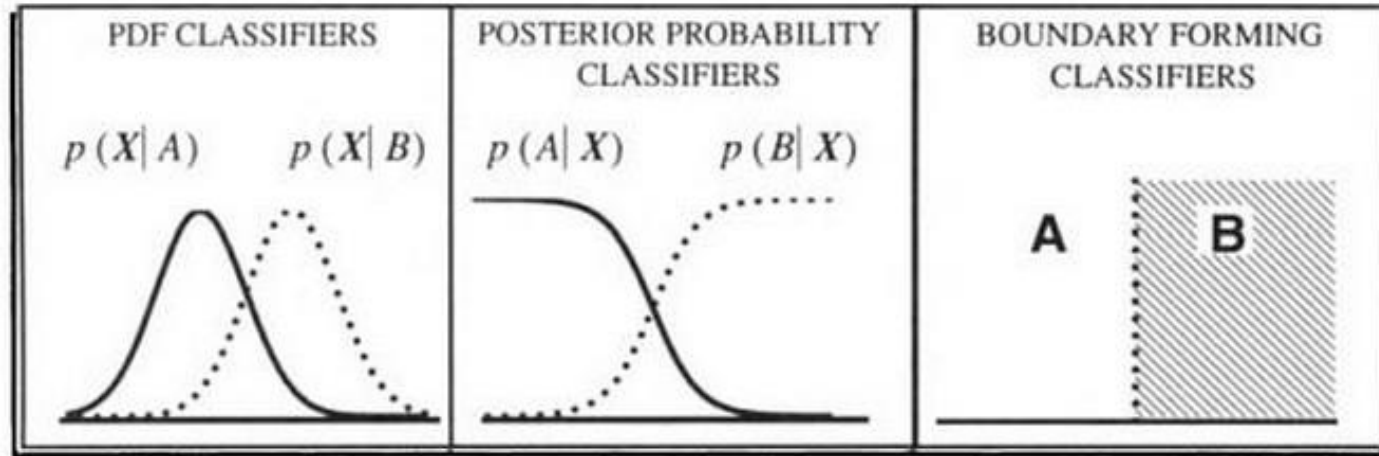


دسته بندی (ادامه...)

Class likelihood

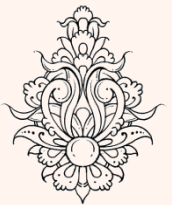
posterior

Class Boundary



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

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



دسته بندی چندکلاسی

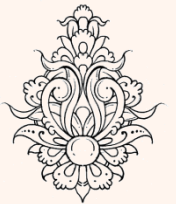
امتمال رخداد x هنگامی که می دانیم به
کلاس C_i تعلق دارد
Class likelihood

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

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

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

در این صورت کلاس C_i انتخاب می شود.

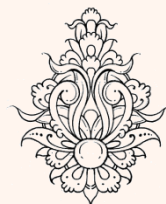


Losses and Risks

- در برخی موارد، تصمیم‌ها پی‌آمد یکسانی ندارند.
– «کنش α_i » به عنوان انتخاب کلاس C_i تعریف شده است.
- λ_{ik} به عنوان میزان ریسک انتخاب کلاس i در زمانی که ورودی به این کلاس k تعلق دارد.
- در این صورت، **expected risk** به صورت زیر محاسبه می‌شود:

$$R(\alpha_i|\mathbf{x}) = \sum_{k=1}^K \lambda_{ik} P(C_k|\mathbf{x})$$

$$\text{choose } \alpha_i \text{ if } R(\alpha_i|\mathbf{x}) = \min_k R(\alpha_k|\mathbf{x})$$

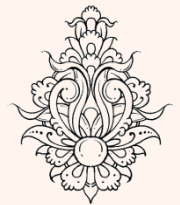


بررسی 0/1 Loss

$$\lambda_{ik} = \begin{cases} 0 & \text{if } i = k \\ 1 & \text{if } i \neq k \end{cases}$$

$$\begin{aligned} R(\alpha_i | \mathbf{x}) &= \sum_{k=1}^K \lambda_{ik} P(C_k | \mathbf{x}) \\ &= \sum_{k \neq i} P(C_k | \mathbf{x}) \\ &= 1 - P(C_i | \mathbf{x}) \end{aligned}$$

برای داشتن کم‌ترین ریسک **محتمل‌ترین** حالت
را انتخاب می‌کنیم



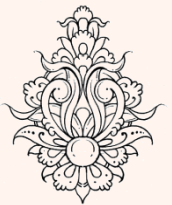
هزینه‌ی بالای انتخاب اشتباه

- در برخی کاربردها، انتخاب اشتباه کلاس هزینه‌ی بالایی دارد، به نحوی که بهتر است هیچ انتخابی توسط سیستم خودکار صورت نپذیرد. در این حالت نمونه به عنوان «مشکوک» تلقی شده و «رد» می‌شود.

– «کنش» جدیدی تعریف می‌شود: رد (reject) : α_{k+1}

*choose C_i if $R(\alpha_i|\mathbf{x}) < R(\alpha_k|\mathbf{x}) \quad \forall k \neq i$ and
 $R(\alpha_i|\mathbf{x}) < R(\alpha_{k+1}|\mathbf{x})$*

reject $R(\alpha_{k+1}|\mathbf{x}) < R(\alpha_i|\mathbf{x}) \quad i = 1, 2, \dots, k$



هزینه‌ی بالای انتخاب اشتباه (ادامه...)

- به عنوان مثال تابع ریسک به صورت زیر تعریف می‌شود:

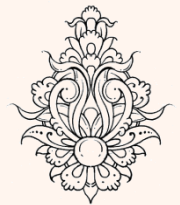
$$\lambda_{ik} = \begin{cases} 0 & \text{if } i = k \\ \lambda & \text{if } i = K + 1, \quad 0 < \lambda < 1 \\ 1 & \text{otherwise} \end{cases}$$

$$R(\alpha_{K+1} | \mathbf{x}) = \sum_{k=1}^K \lambda P(C_k | \mathbf{x}) = \lambda$$

$$R(\alpha_i | \mathbf{x}) = \sum_{k \neq i} P(C_k | \mathbf{x}) = 1 - P(C_i | \mathbf{x})$$

choose C_i if $P(C_i | \mathbf{x}) > P(C_k | \mathbf{x}) \quad \forall k \neq i$ and $P(C_i | \mathbf{x}) > 1 - \lambda$

reject otherwise



Discriminant Functions

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

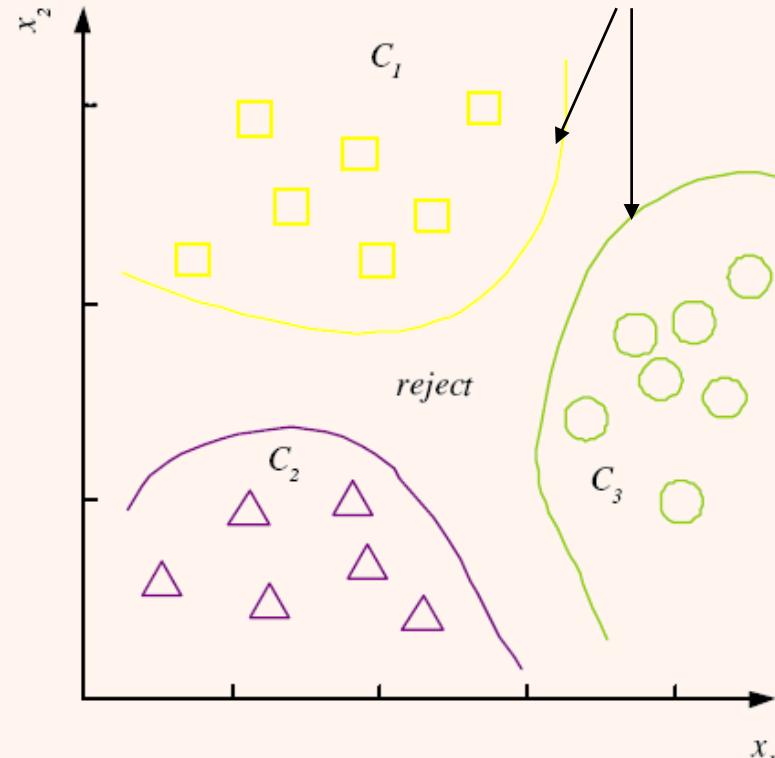
$$g_i(\mathbf{x}) = \begin{cases} -R(\alpha_i | \mathbf{x}) \\ P(C_i | \mathbf{x}) \\ p(\mathbf{x} | C_i)P(C_i) \end{cases}$$

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

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

توابع جداساز

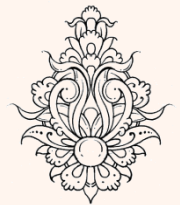
$g_i(\mathbf{x}), i = 1, \dots, K$



Dichotomizer

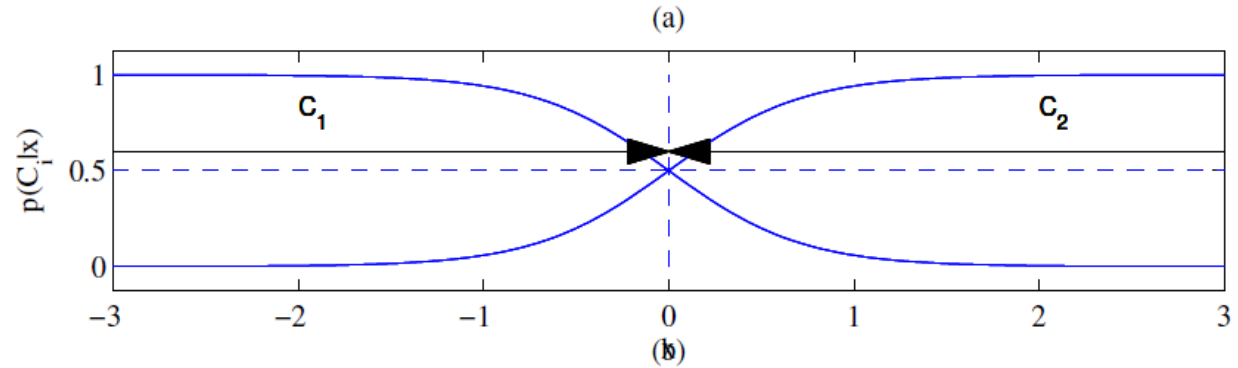
$$g(x) = g_1(x) - g_2(x)$$

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

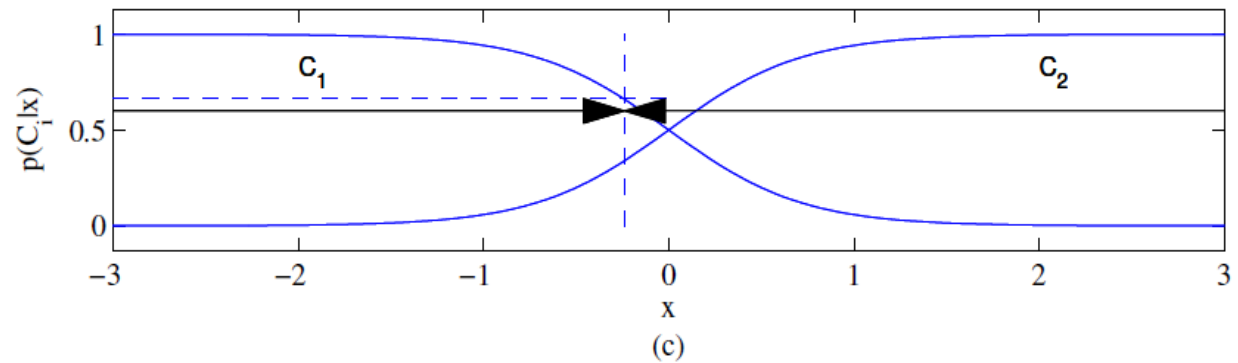


بررسی حالات مختلف

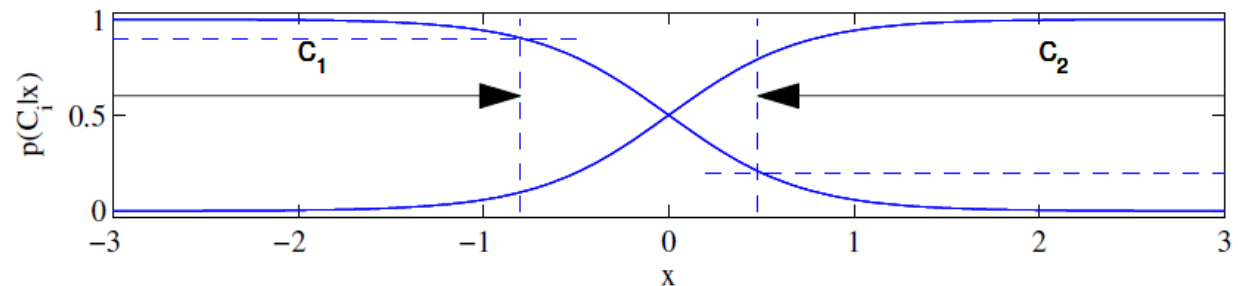
Equal losses



Unequal losses



With reject



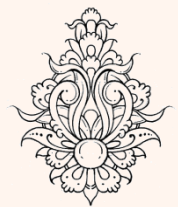
- احتمال این که با در اختیار داشتن ورودی x ، در حالت S_k باشیم: $P(S_k | x)$
- سودمندی کنش α_i وقتی در حالت k ام هستیم:

$$- U_{ik}$$

$$EU(\alpha_i | \mathbf{x}) = \sum_k U_{ik} P(S_k | \mathbf{x})$$

Choose α_i if $EU(\alpha_i | \mathbf{x}) = \max_j EU(\alpha_j | \mathbf{x})$

Expected utility



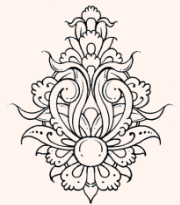
- در اسلایدهای پیش در مورد «اتخاذ تصمیم بهینه» با در نظر گرفتن احتمال مشاهدهی ورودی با فرض دانستن کلاس و احتمال وقوع کلاس بحث شد.
- با توجه به این فرض که توزیع داده‌ها، از توزیعی خاص پیروی می‌کند، این روش‌ها را «روش‌های پارامتری» می‌نامند.

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

- تخمین پارامتر:

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

Sufficient statistic



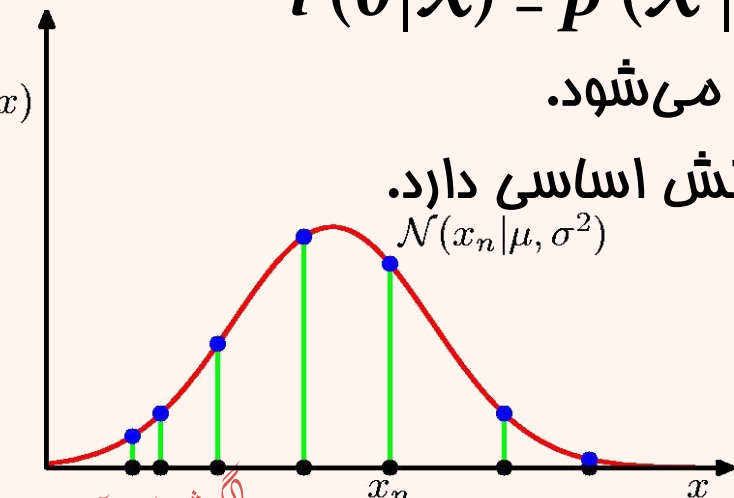
• «تابع درست‌نمایی»، تابعی از پارامترهای مدل آماری است.

– درست‌نمایی یک مجموعه از پارامترها، 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.

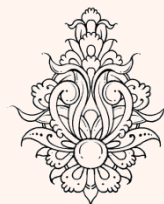
• این تابع در «استنباط آماری» نقش اساسی دارد.



- 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 مدها کمتر شود؛ درست‌نمایی بیشینه شود.
- برای سادگی محاسبات، به جای درست‌نمایی، از لگاریتم آن استفاده می‌شود:



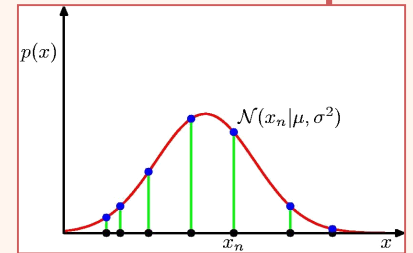
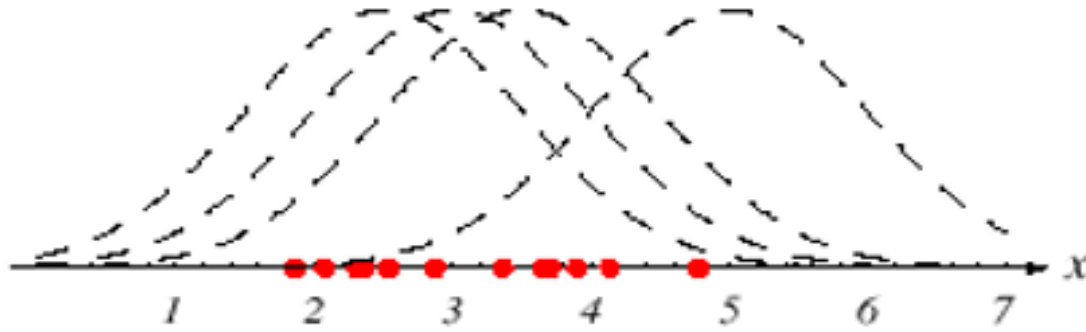
$$\mathcal{L}(\theta|\mathcal{X}) = \log l(\theta|\mathcal{X}) = \sum_t \log p(x^t|\theta)$$

Log likelihood

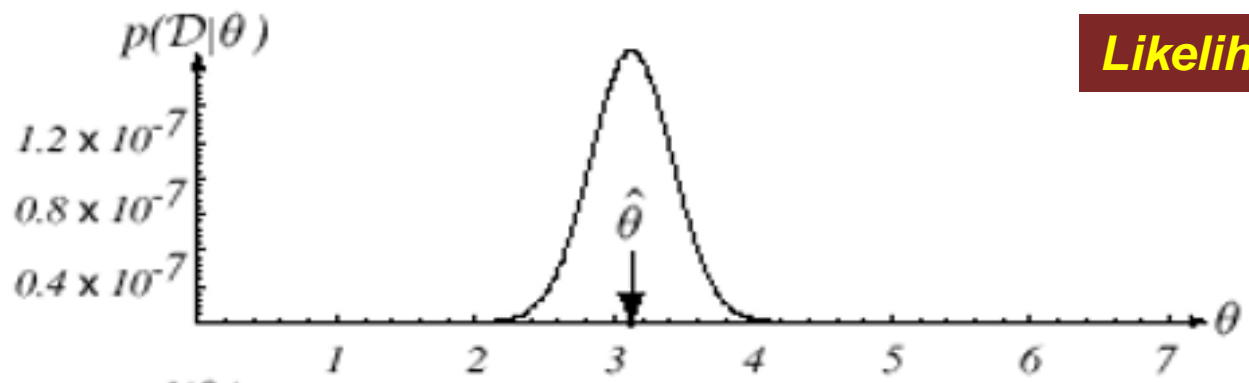
$$\theta^* = \operatorname{argmax}_{\theta} \mathcal{L}(\theta|\mathcal{X})$$



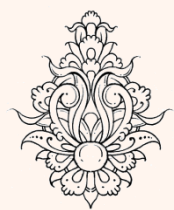
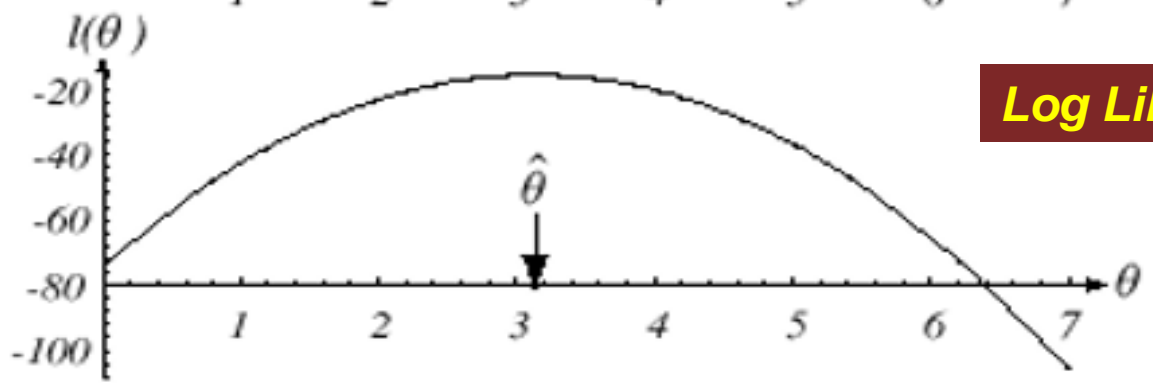
برآورد درست‌نمایی بیشینه



Likelihood



Log Likelihood



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

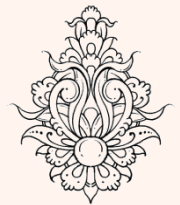
- $K > 2$ states, x_i in $\{0,1\}$

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

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

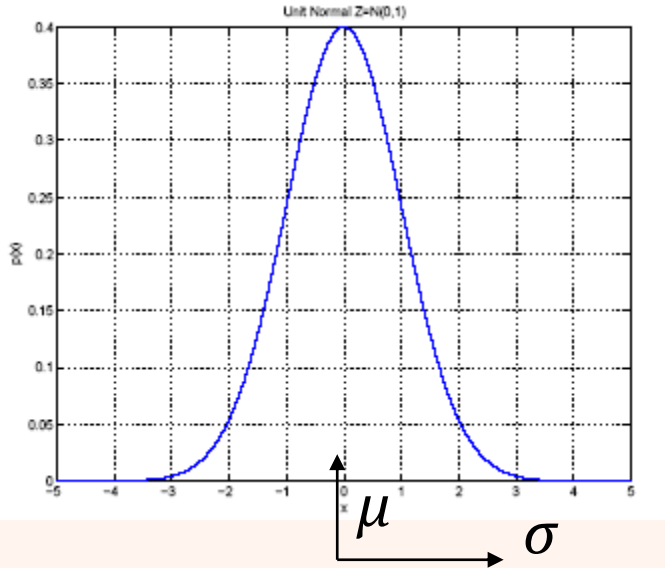
$$\text{MLE: } \hat{p}_i = \sum_t x_i^t / N$$

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



Gaussian (Normal) Distribution

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



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

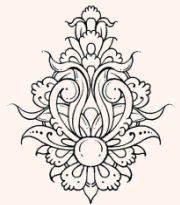
$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]$$

- MLE for μ and σ^2 :

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

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

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



- یک نمونه از داده‌ها: \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:

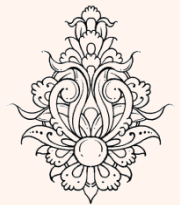
$$r(d, \theta) = E[(d(\mathcal{X}) - \theta)^2]$$

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،

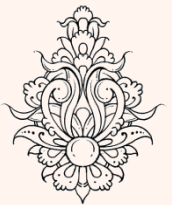
$$E[m] = E\left[\frac{\sum_t x^t}{N}\right] = \frac{1}{N} \sum_t E[x^t] = \frac{N\mu}{N} = \mu$$

- 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

$$\text{Var}(m) \rightarrow 0 \text{ as } N \rightarrow \infty$$

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



مثال - تخمین واریانس

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

$$E[s^2] = \frac{\sum E[(x^t)^2] - N \cdot E[m^2]}{N}$$

یادآوری

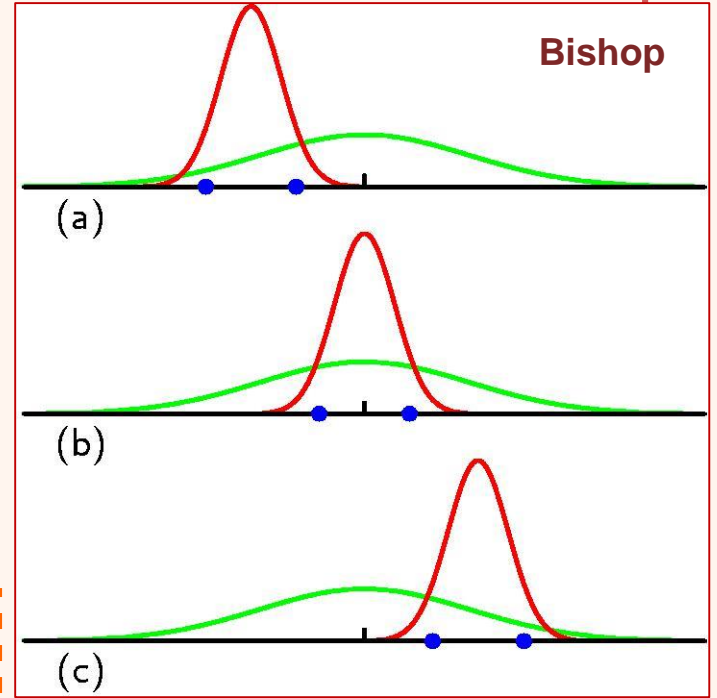
$$Var(X) = E[X^2] - E[X]^2$$

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

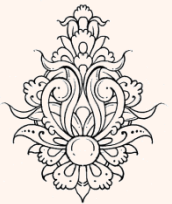
$$E[s^2] = \frac{N(\sigma^2 + \mu^2) - N(\sigma^2/N + \mu^2)}{N} = \frac{N-1}{N} \sigma^2 \neq \sigma^2$$

asymptotically unbiased estimator

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



Bishop

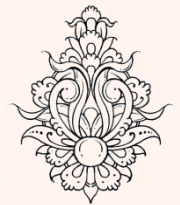
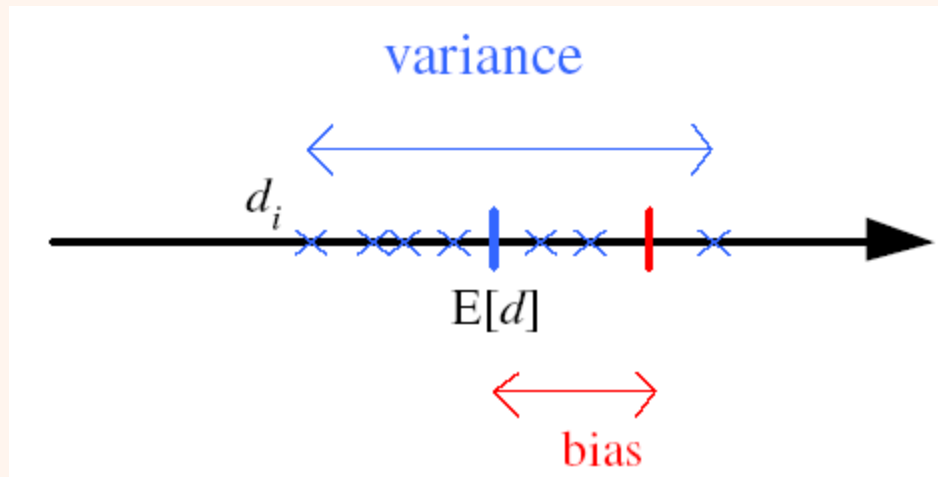


تاشکانه
تعمیراتی
بهشتی

ارزیابی برآورد

Mean square error:

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



برآورد بیشینه‌گر احتمال پسین

Maximum a Posteriori (MAP)

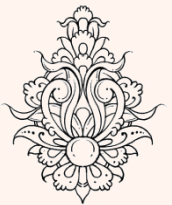
• در MLE، پارامتر مورد نظر به عنوان مجهول در نظر گرفته می‌شود، ممکن است در مورد پارامتر مورد نظر از پیش اطلاعاتی (prior information) داشته باشیم. این اطلاعات می‌توانند به تخمین دقیق‌تر کمک کنند، به ویژه زمانی که داده‌های آموزش کم‌تعداد باشند.

- 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 $p(\theta)$ وجود دارد. با ترکیب این اطلاعات با آنچه داده‌ها به ما می‌گویند (likelihood density)، خواهیم داشت:

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

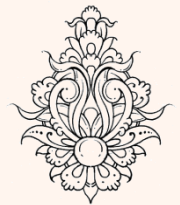
Maximum a Posteriori (MAP)

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

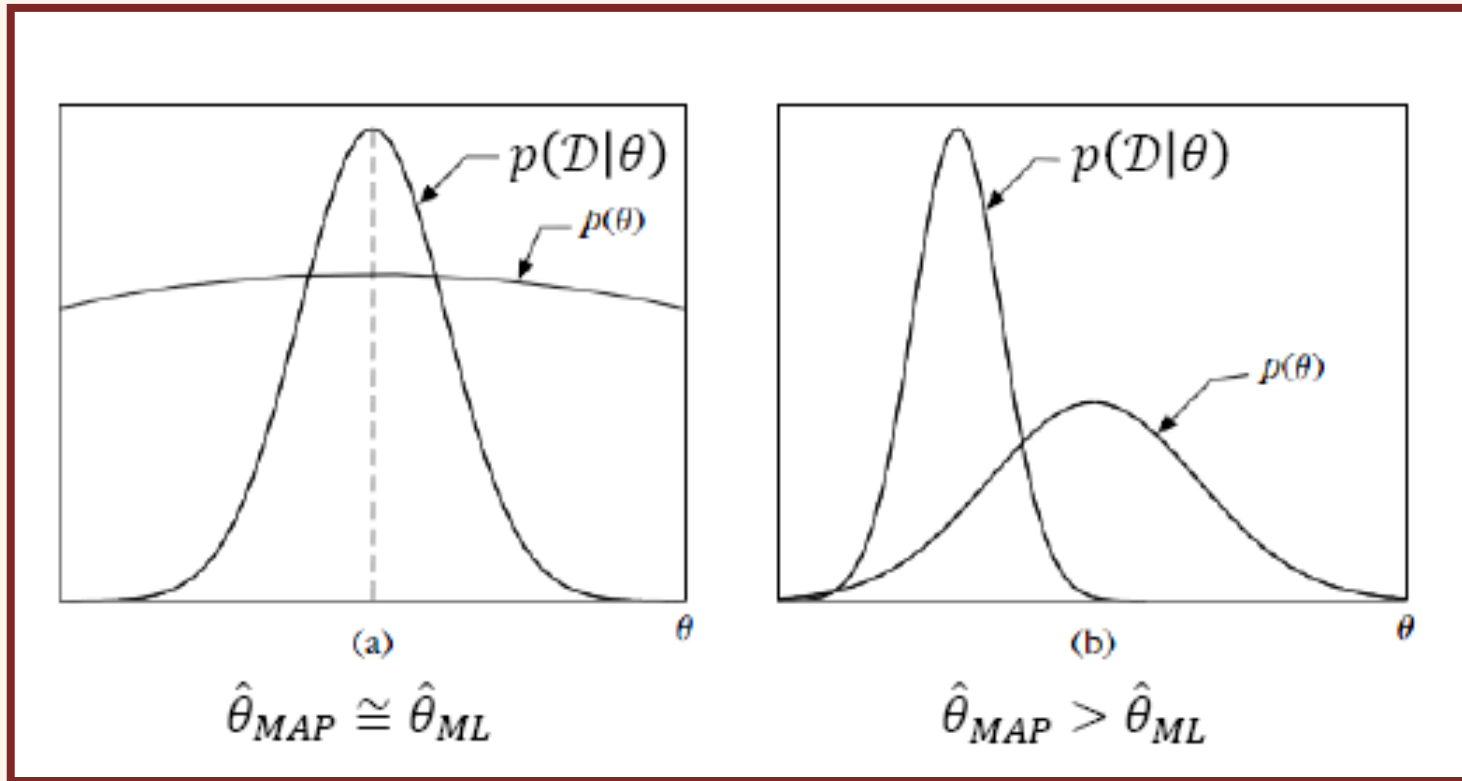
تفاوت با ML در نظر گرفتن $p(\theta)$ است.

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

Maximum Likelihood (ML)

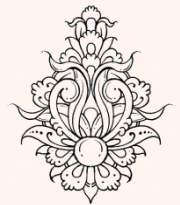


بر آورد بیشینه‌گر احتمال پسین (ادامه...)



Pattern recognition, Sergios Theodoridis

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



مثال

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

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

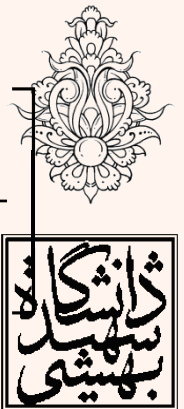
$$p(\mu | X) \propto p(\mu)p(X | \mu)$$

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

برای تخمین MAP باید رابطه‌ی زیر مناسبه بشود:

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

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



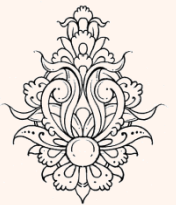
مثال - ادامه

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

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

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

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



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 وزن دار 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

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

دسته بندی پارامتری

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

or

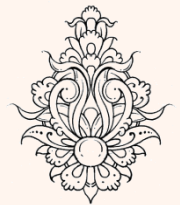
$$g_i(x) = \log p(x | C_i) + \log P(C_i)$$

تابع جدا ساز

در صورتی که چگالی کلاس 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

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

تابع جداساز

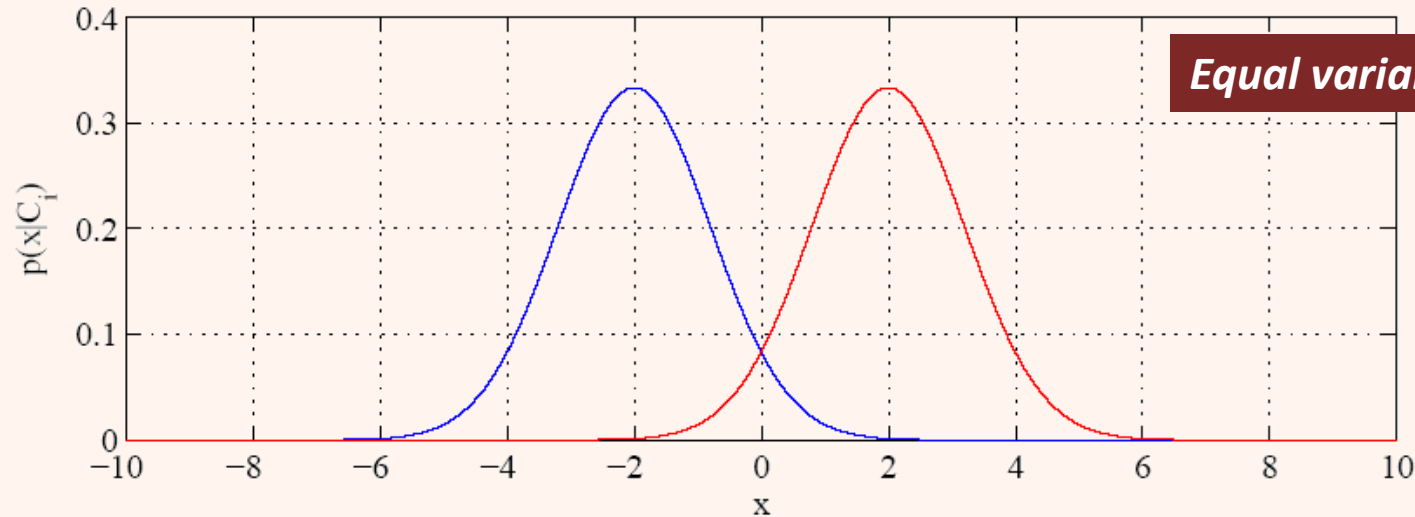
$$g_i(x) = -\frac{1}{2} \log 2\pi - \log s_i - \frac{(x - m_i)^2}{2s_i^2} + \log \hat{P}(C_i)$$



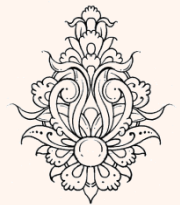
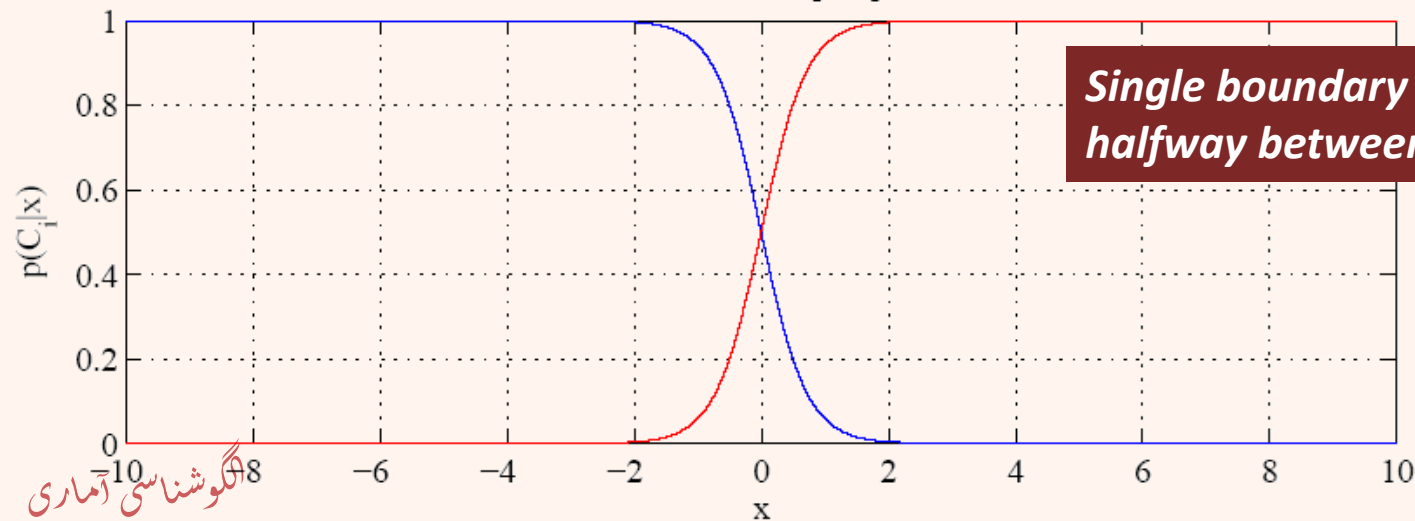
دسته‌بندی دو کلاس با واریانس یکسان

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

Likelihoods



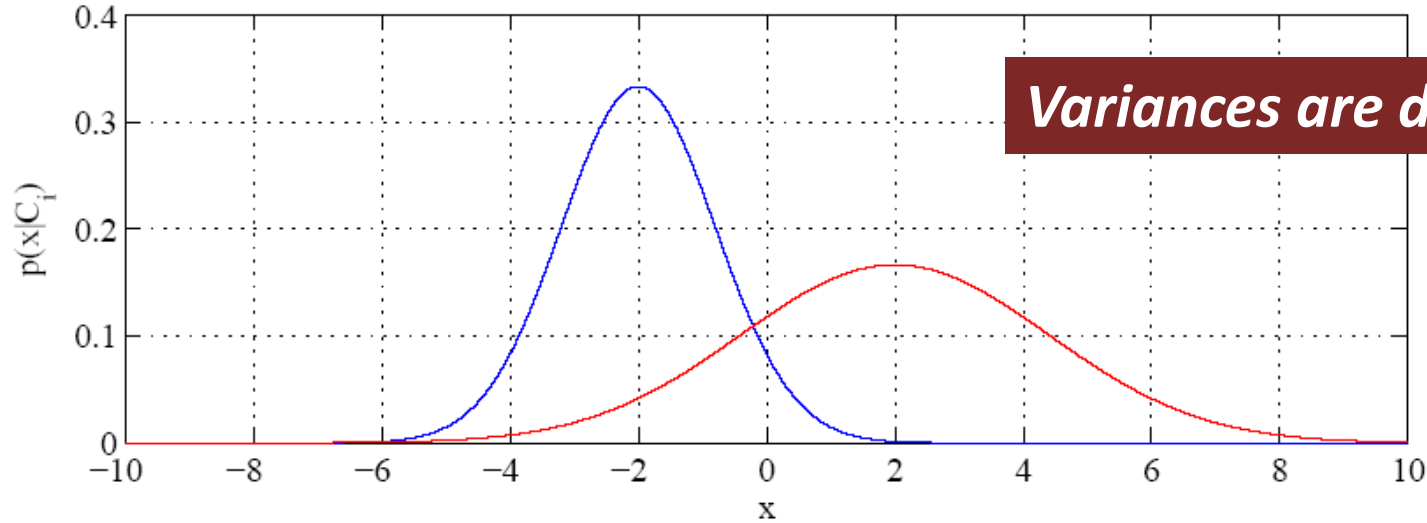
Posteriors with equal priors



دسته‌بندی دو کلاس با واریانس متفاوت

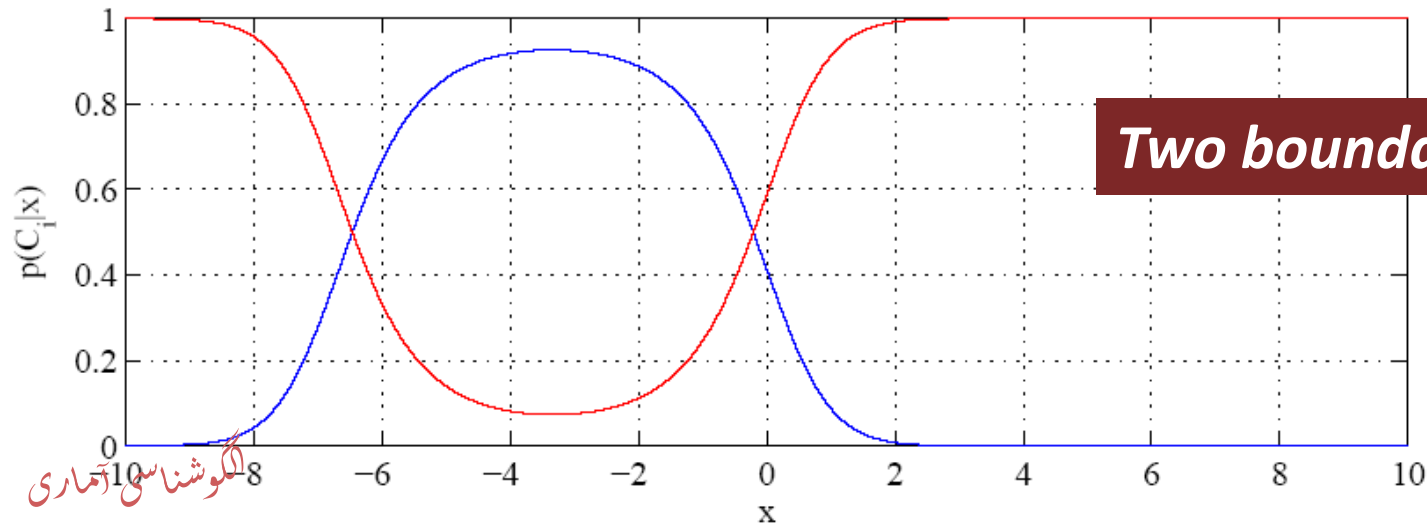
$$g_i(x) = -\frac{1}{2} \log 2\pi - \log s_i - \frac{(x - m_i)^2}{2s_i^2} + \log \hat{P}(C_i)$$

Likelihoods

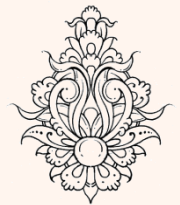


Variances are different

Posteriors with equal priors

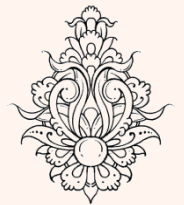
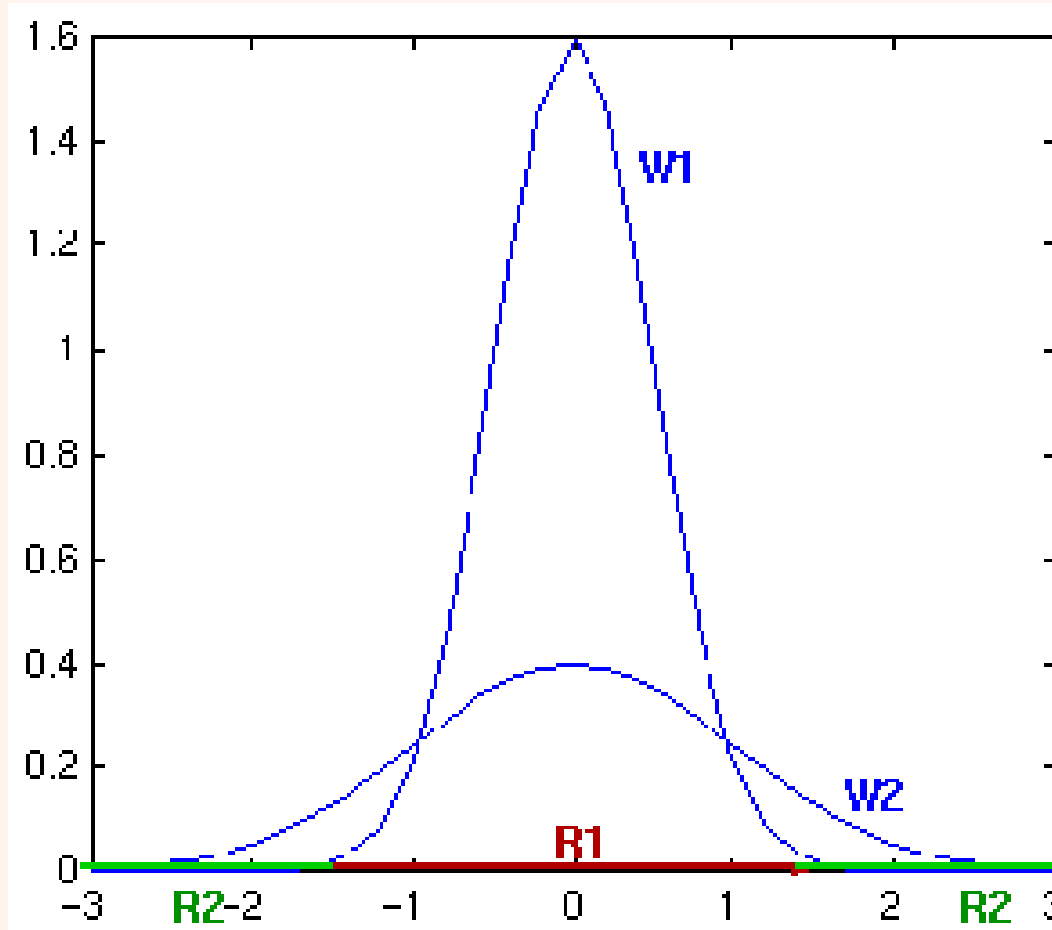


Two boundaries



مثال

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



Dependent variable

رگرسیون

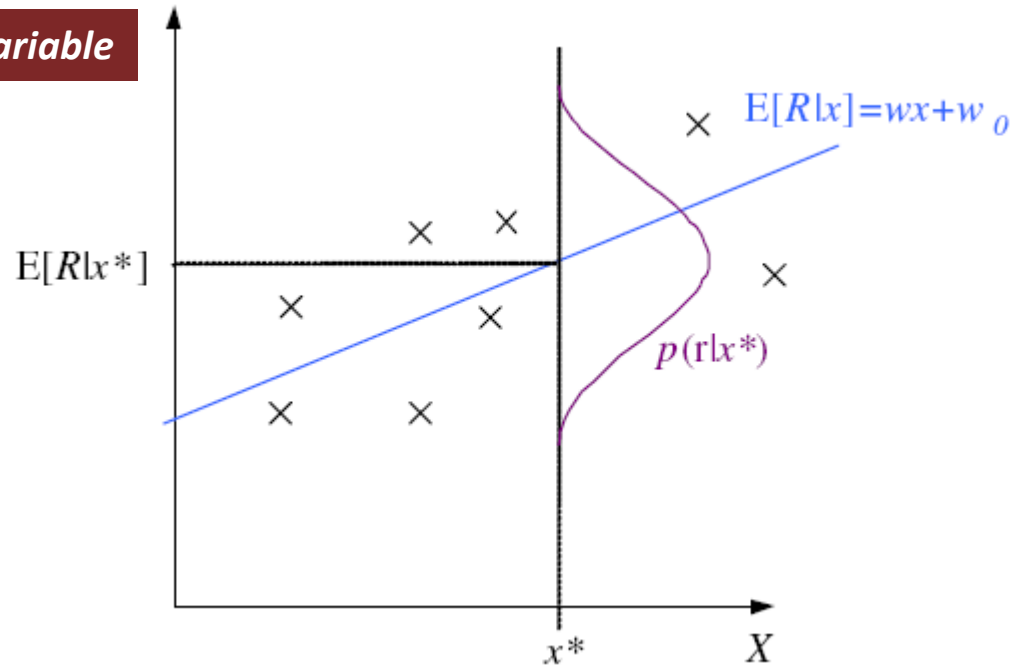
Independent variable

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


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

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

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



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

$$p(x, r) = p(r | x)p(x)$$


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



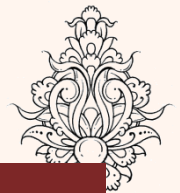
محاسبه‌ی تابع خطا

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

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

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

Least Squares estimates



رگرسیون خطی

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

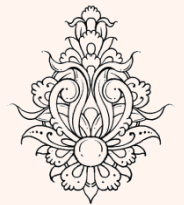
$$g(x^t | w_1, w_0) = w_1 x^t + w_0$$

$$\sum_t r^t = N w_0 + w_1 \sum_t x^t$$

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

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

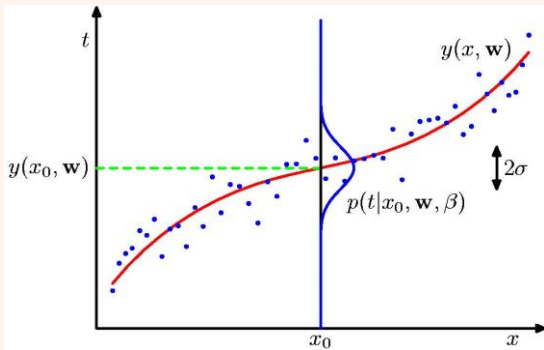
$$\mathbf{w} = \mathbf{A}^{-1} \mathbf{y}$$



رگرسیون چندجمله‌ای

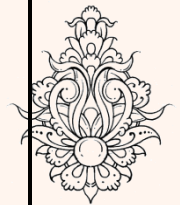
$$g(x^t | w_k, \dots, w_2, w_1, w_0) = w_k (x^t)^k + \dots + w_2 (x^t)^2 + w_1 x^t + w_0$$

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



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

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

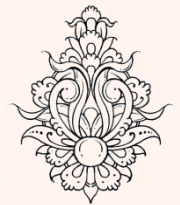


رگرسیون چندجمله‌ای

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

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

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



Bias and Variance

noise

squared error

Expected square error at x

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

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

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

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.



تاریخچه

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

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

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

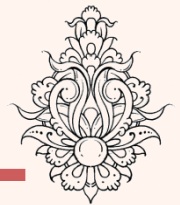
$$\bar{g}(x) = \frac{1}{M} \sum_i g_i(x)$$

$$g_i(x) = 2$$

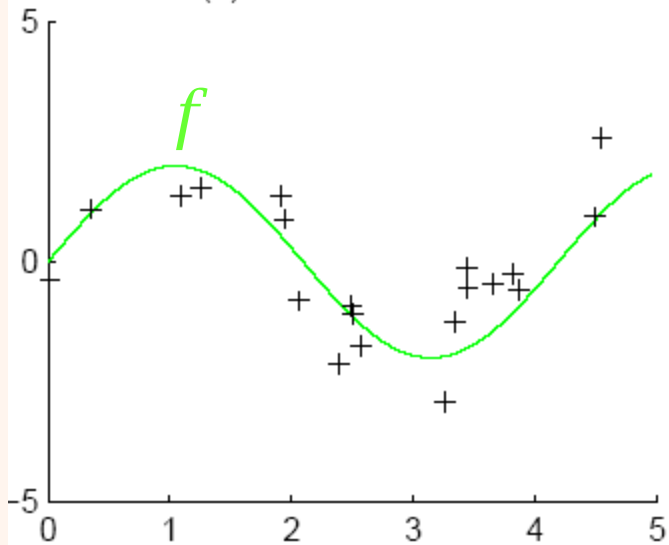
واریانس صفر است، اما بایاس بالایی دارد

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

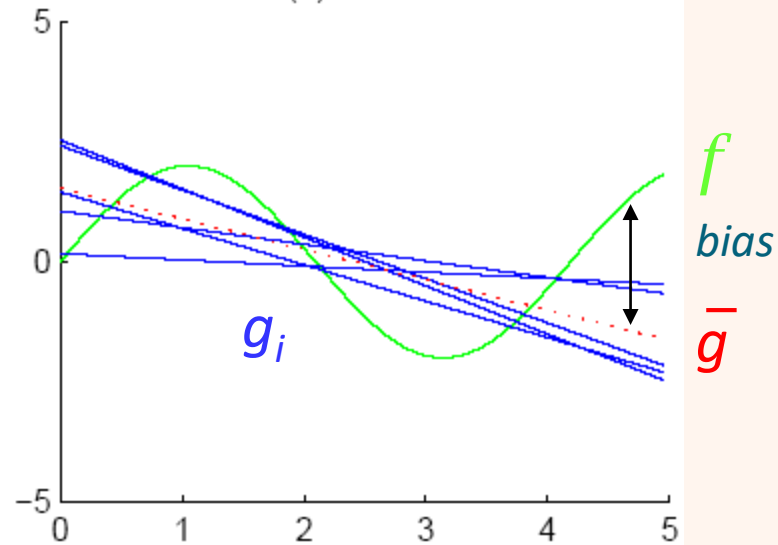
بایاس کاهش می‌یابد، اما واریانس افزایش می‌یابد



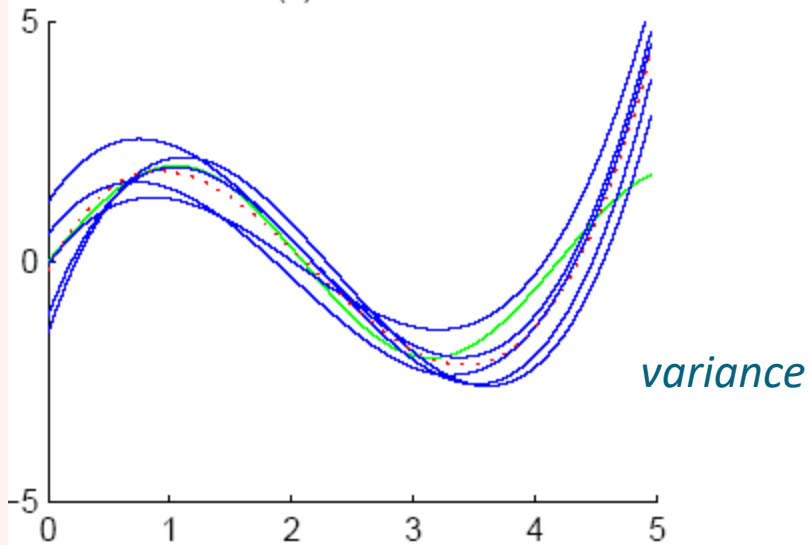
(a) Function and data



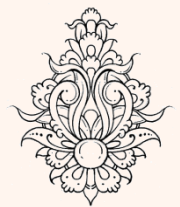
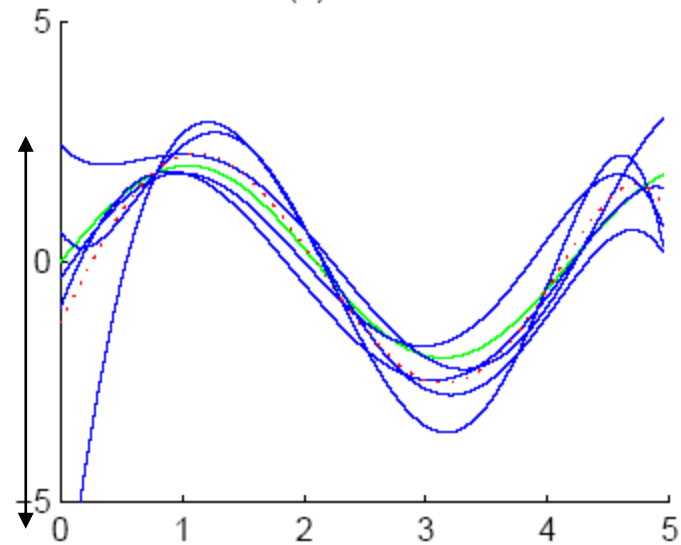
(b) Order 1



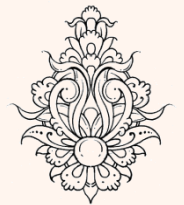
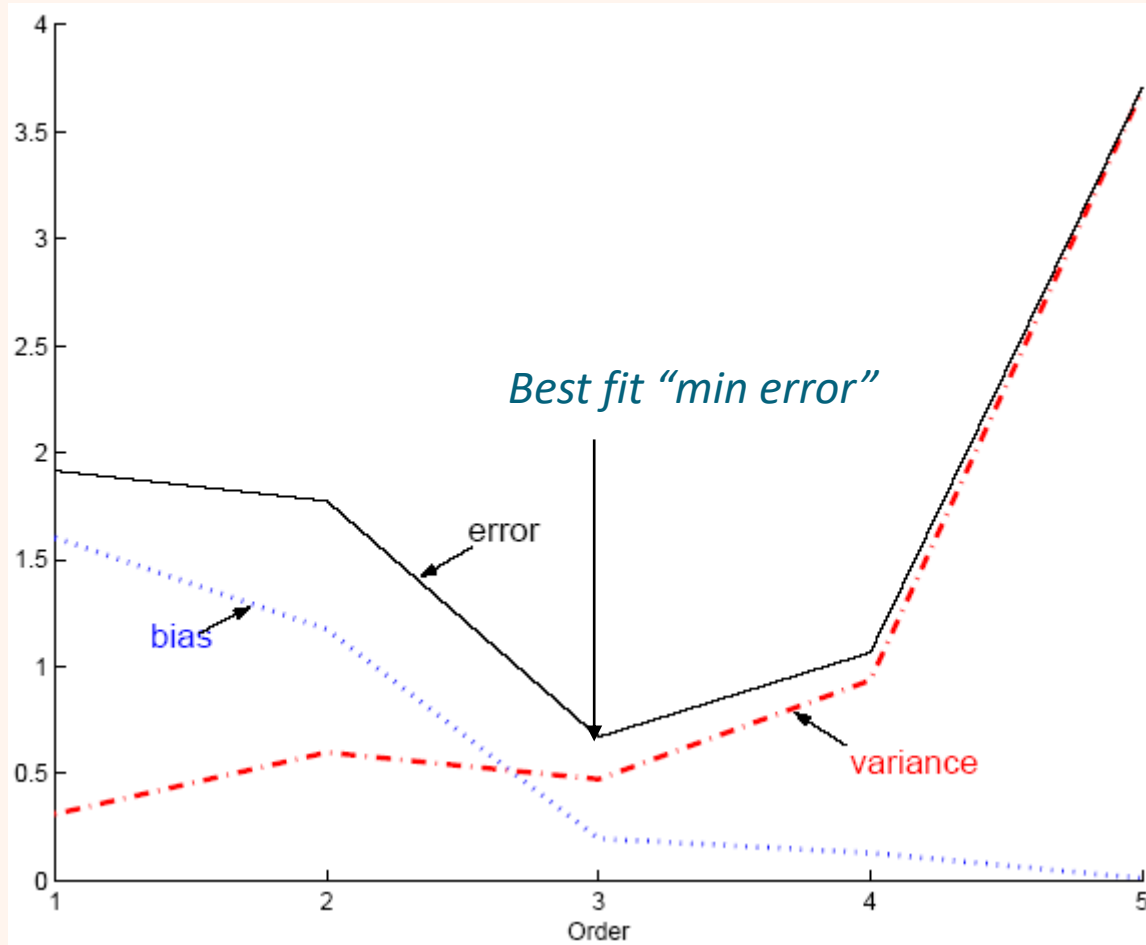
(c) Order 3



(d) Order 5

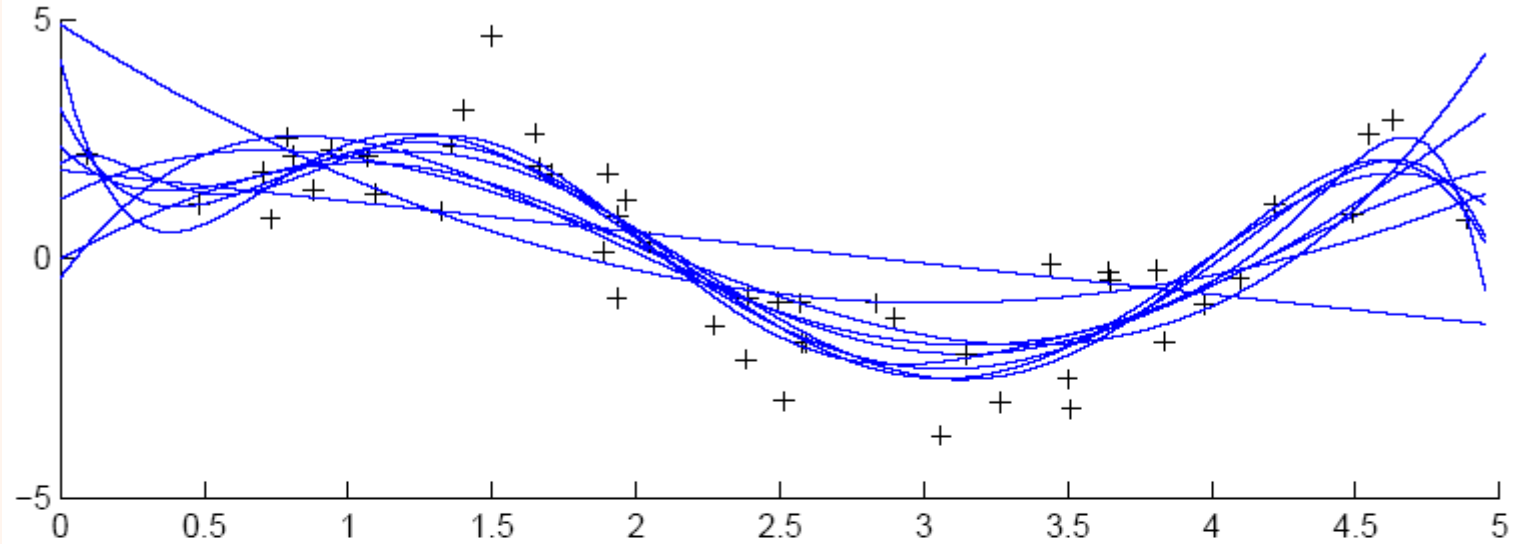


انتخاب مدل

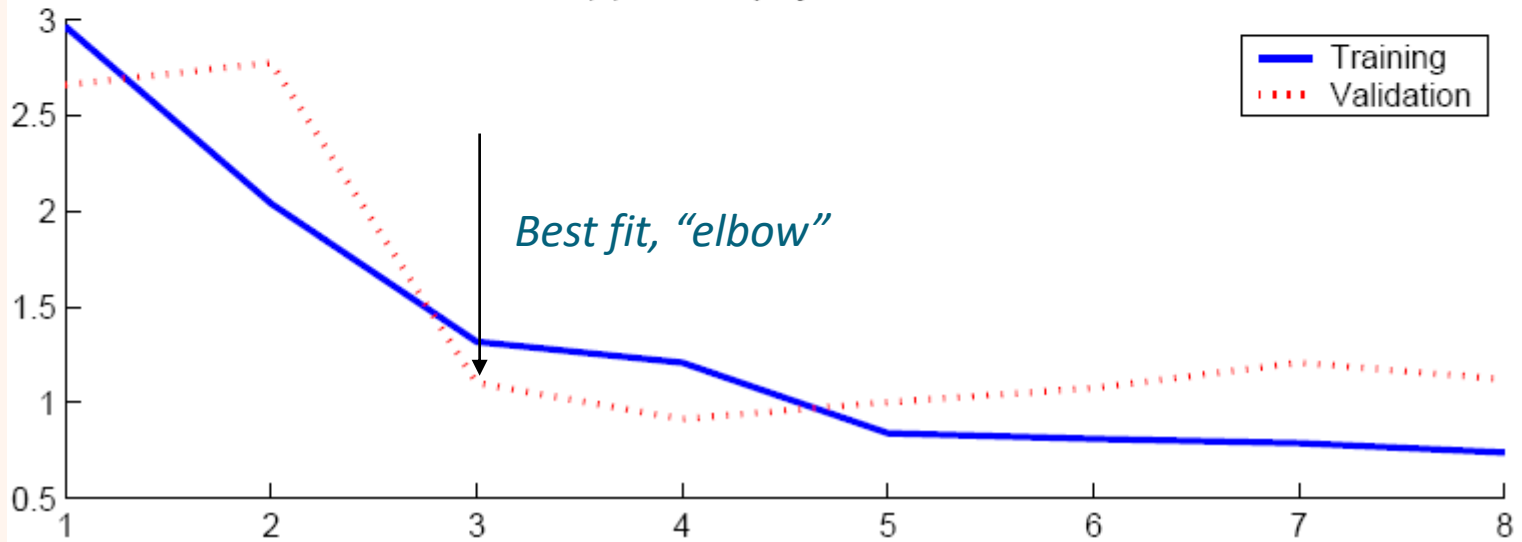


Cross validation

(a) Data and fitted polynomials



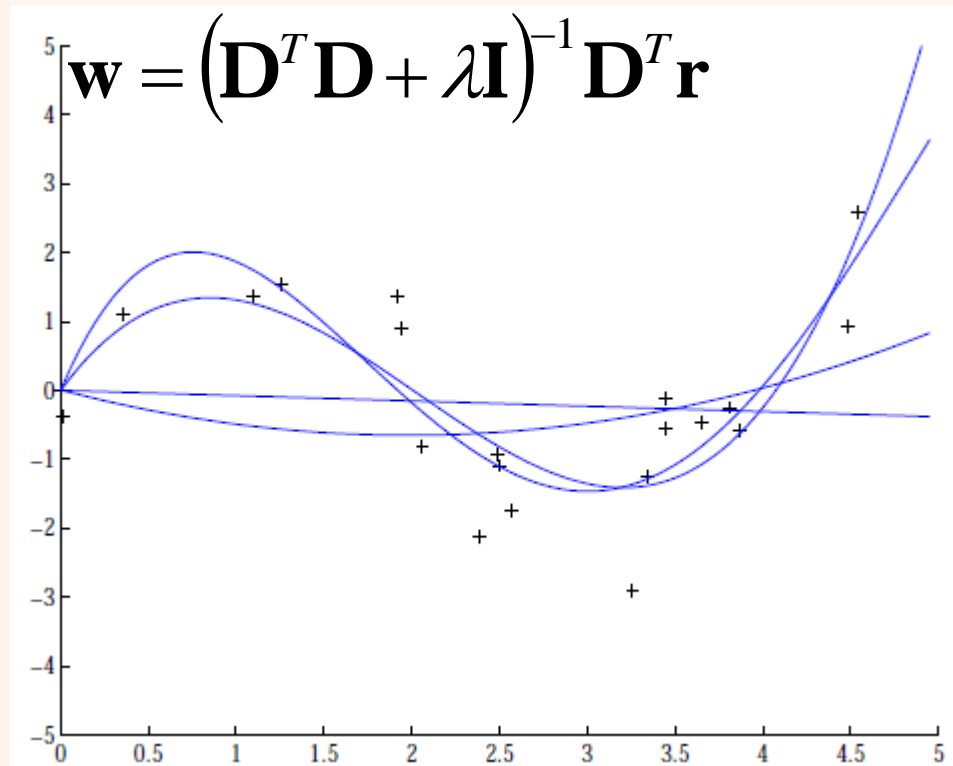
(b) Error vs polynomial order



Regularization

Penalize complex models

E' = error on data + λ model complexity



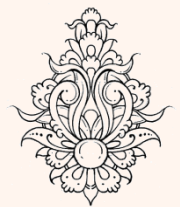
Coefficients increase in magnitude as order increases:

1: [-0.0769, 0.0016]

2: [0.1682, -0.6657, 0.0080]

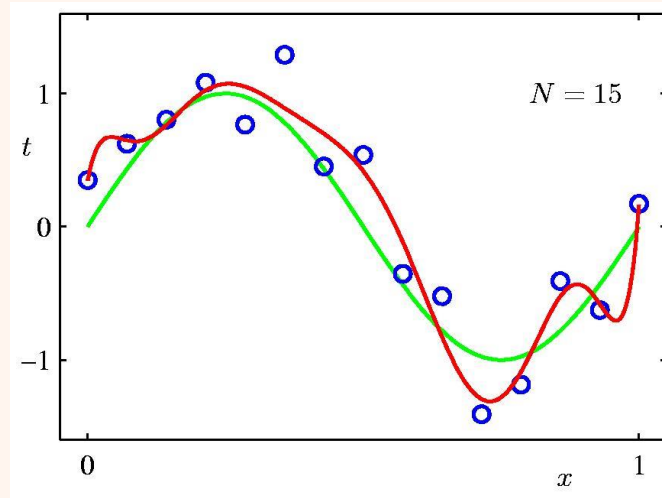
3: [0.4238, -2.5778, 3.4675, -0.0002]

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

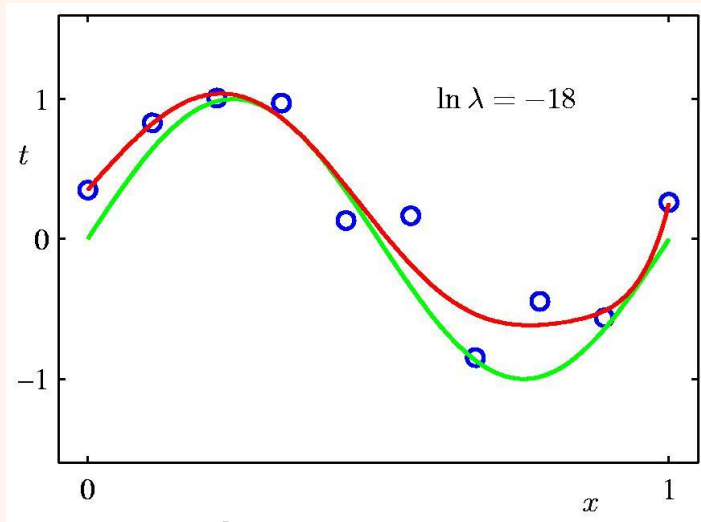


$$\text{regularization: } E(\mathbf{w} | \mathcal{X}) = \frac{1}{2} \sum_{t=1}^N [r^t - g(x^t | \mathbf{w})]^2 + \lambda \sum_i w_i^2$$

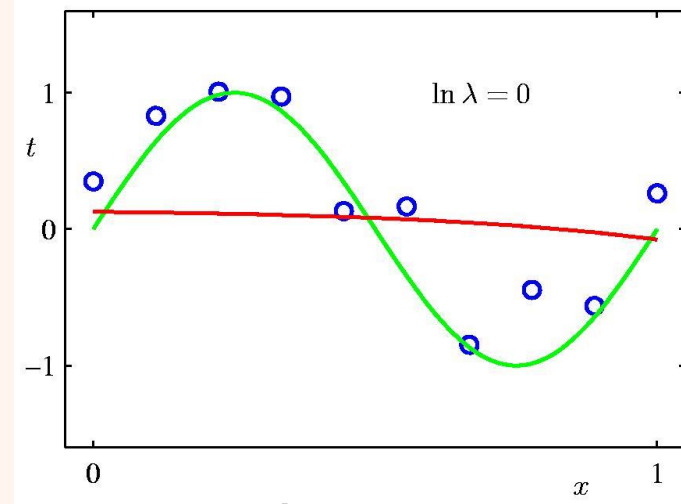
Regularization



9th Order Polynomial



$\ln \lambda = -18$



$\ln \lambda = 0$

