Machine Learning

یادگیری بیزی

مقدمه‌ای بر

یادگیری ماشین

بخش سوم

دانشگاه شهید بهشتی
پژوهشکده‌ی فضای مجازی
پاییز ۱۳۹۸
احمد محمودی ازناوه

فهرست مطالب

- یادگیری بیزی
  - معیارهای تصمیم‌گیری
- تابع درست‌نمایی

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

- 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)

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

احتمال شرطی

or

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

دسته‌بندی (ادامه...)

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

Bayes' Rule

posterior

احتمال پسین

با چه احتمالی C، کلاس مربوط به x است.

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

احتمال پیشین

prior

درست‌نمایی کلاس

Class likelihood

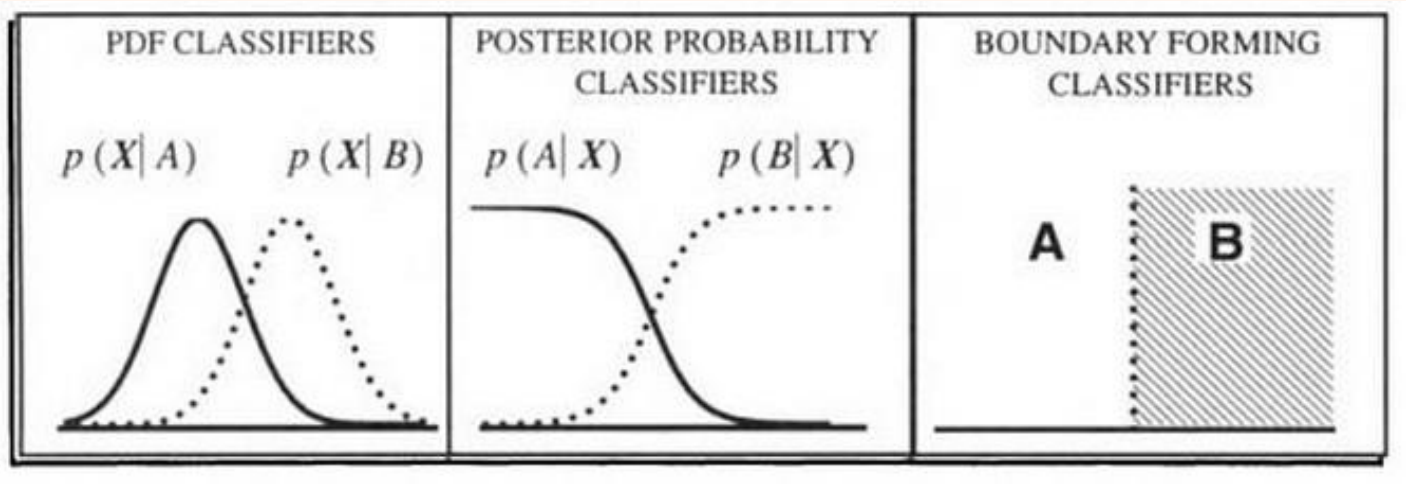
با چه احتمالی x توسط کلاس C تولید می‌شود.

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...)



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 کلاس C_i تعلق دارد
Class likelihood

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

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

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

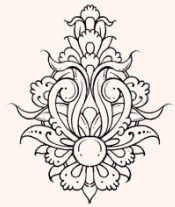
در این صورت کلاس C_i انتخاب می‌شود.

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:

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

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



بررسی 0/1 Loss

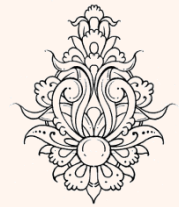
$$\lambda_{ik} = \begin{cases} 0 \text{ if } i = k \\ 1 \text{ if } i \neq k \end{cases}$$

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

$$= \sum_{k \neq i} P(C_k \mid \mathbf{x})$$

$$= 1 - P(C_i \mid \mathbf{x})$$

برای داشتن کم‌ترین ریسک محتمل‌ترین حالت را انتخاب می‌کنیم

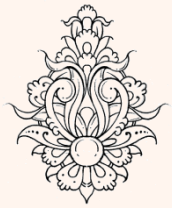


هزینه‌ی بالای انتخاب اشتباه

- 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}

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

$$reject \qquad R(\alpha_{k+1}/\boldsymbol{x}) < R(\alpha_i/\boldsymbol{x}) \quad i = 1,2,\ldots,k$$



هزینه‌ی بالای انتخاب اشتباه (ادامه...)

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

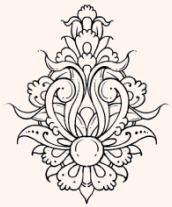
$$\lambda_{ik} = \begin{cases} 0 & \text{if } i = k \\ \lambda & \text{if } i = K+1, \quad 0 < \lambda < 1 \\ 1 & \text{otherwise} \end{cases}$$

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

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

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

reject otherwise



Discriminant Functions

توابع جداساز

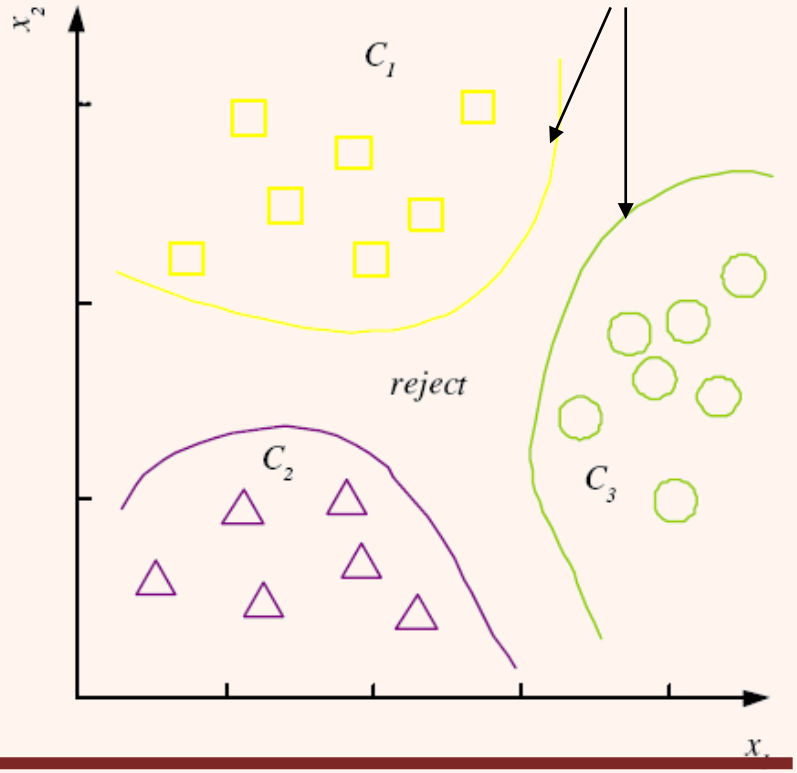
choose C_i if $g_i(\mathbf{x}) = \max_k g_k(\mathbf{x})$

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

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

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

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



Dichotomizer

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

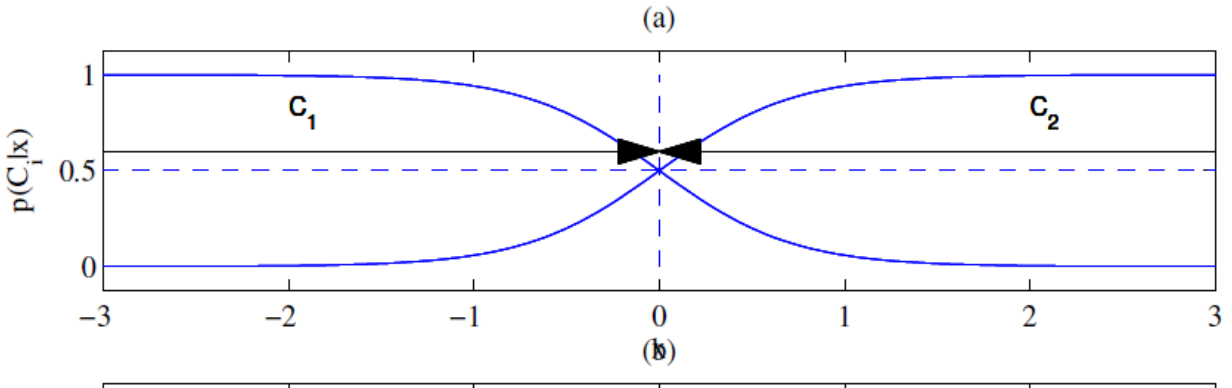
$$\text{choose} \begin{cases} C_1 \text{ if } g(\mathbf{x}) > 0 \\ C_2 \text{ otherwise} \end{cases}$$

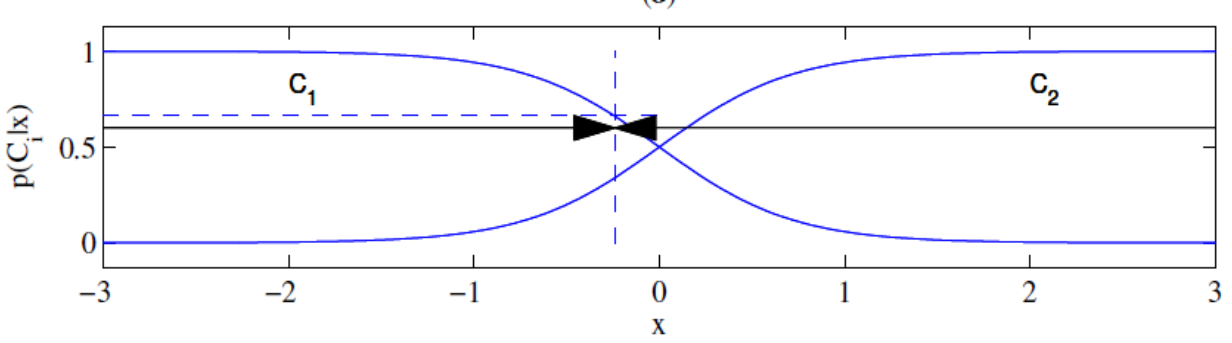
بررسی حالات مختلف

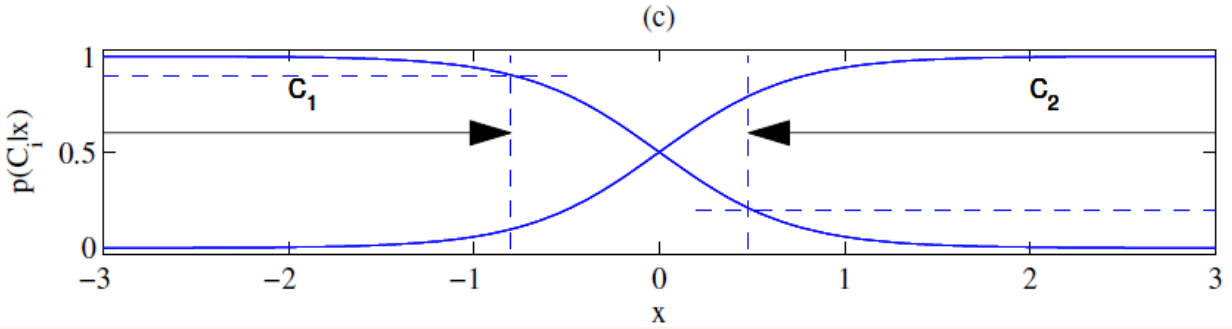
Equal losses

Unequal losses

With reject







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

تابع درست‌نمایی

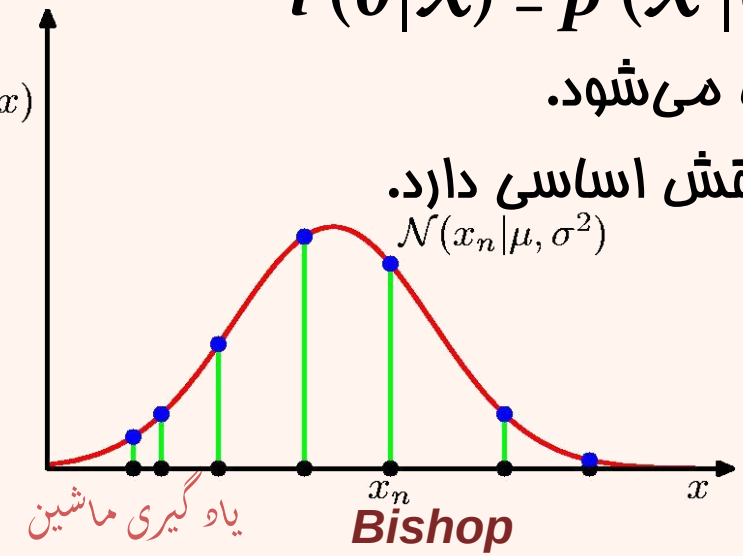
- «تابع درست‌نمایی»، تابعی از پارامترهای مدل آماری است.

– 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.



یاد گیری ماشین

Bishop

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

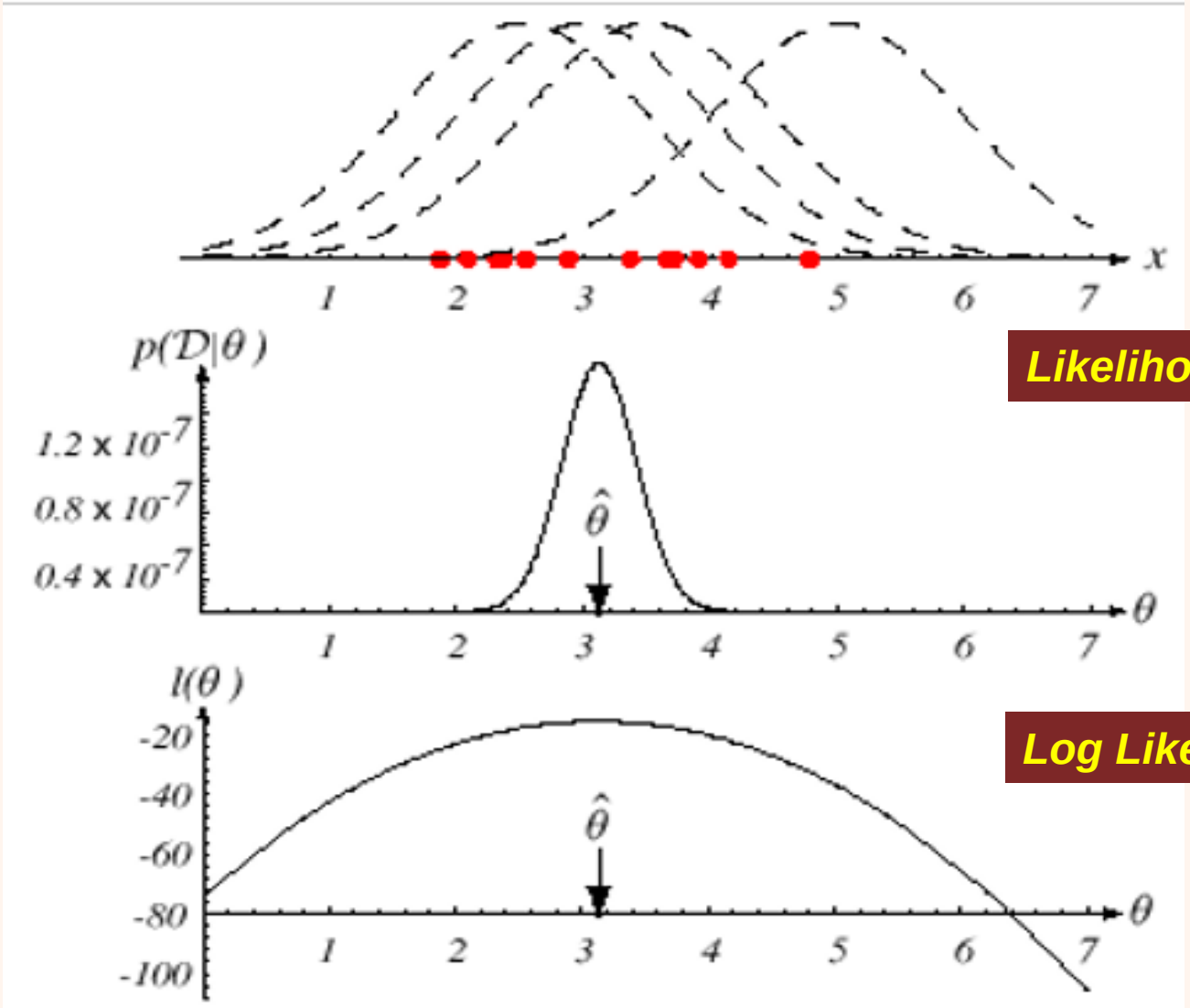
- $l(\theta|\mathcal{X}) = p(\mathcal{X}|\theta) = \prod_t p(x^t|\theta)$
- در برآورد درست‌نمایی بیشینه در پی یافتن θ هستیم به گونه‌ای که احتمال تعلق X به p حداکثر شود؛ درست‌نمایی بیشینه شود.
- برای سادگی محاسبات، به جای درست‌نمایی، از لگاریتم آن استفاده می‌شود:

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

Log likelihood

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

برآورد درست‌نمایی بیشینه



Likelihood

Log Likelihood

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

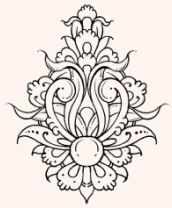
- $K>2$ states, x_i in {0,1}

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

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

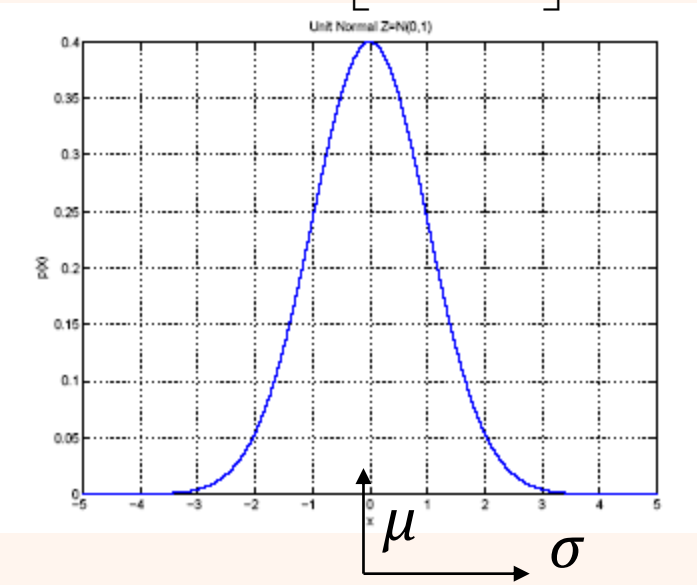
MLE: $\hat{p}_i = \sum_t x_i^t / N$

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



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:

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

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

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:

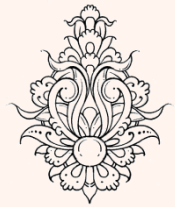
$$r(d,\theta)=E[(d(\mathcal{X})-\theta)^2]$$

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،

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

- 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

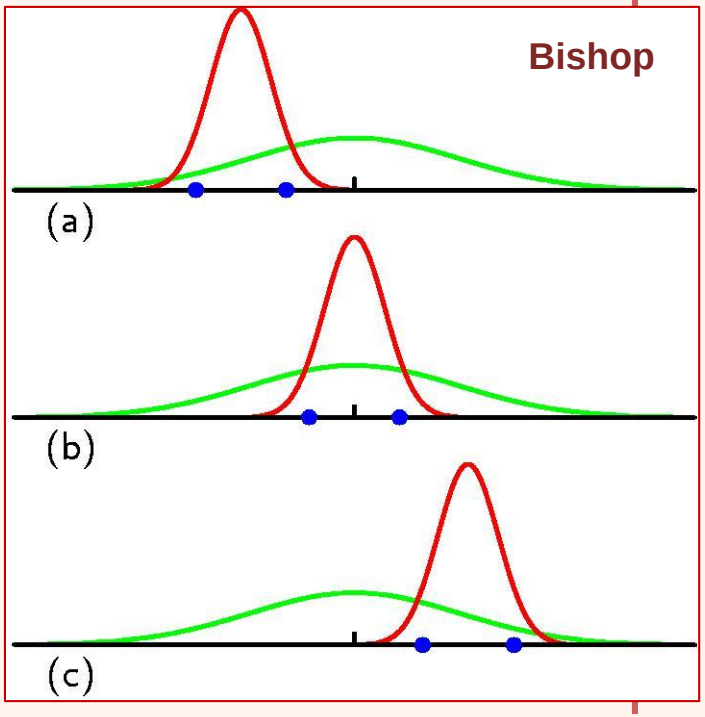
Var(m) $\rightarrow$0 as N$\rightarrow\infty$

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

مثال- تخمین واریانس

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

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



یادآوری

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

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

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

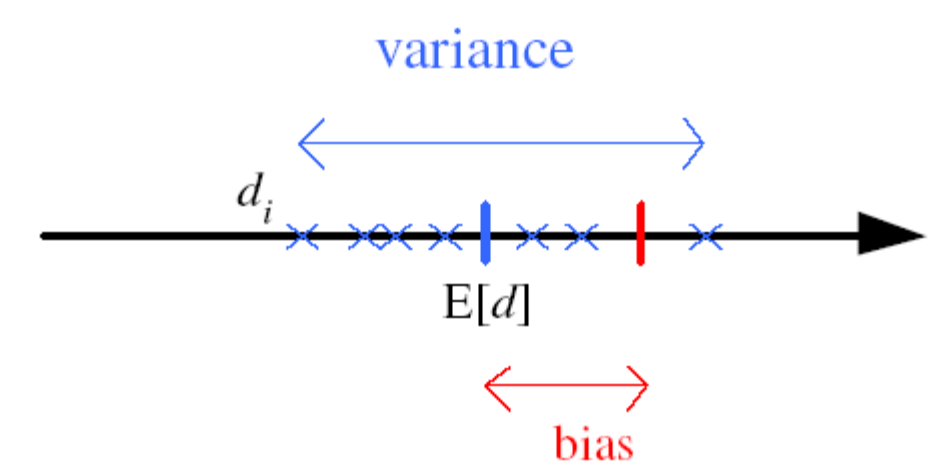
asymptotically unbiased estimator

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

ارزیابی برآورد

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 $p(\theta)$ 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)

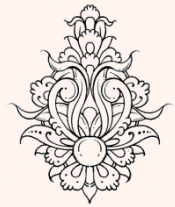
- و

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

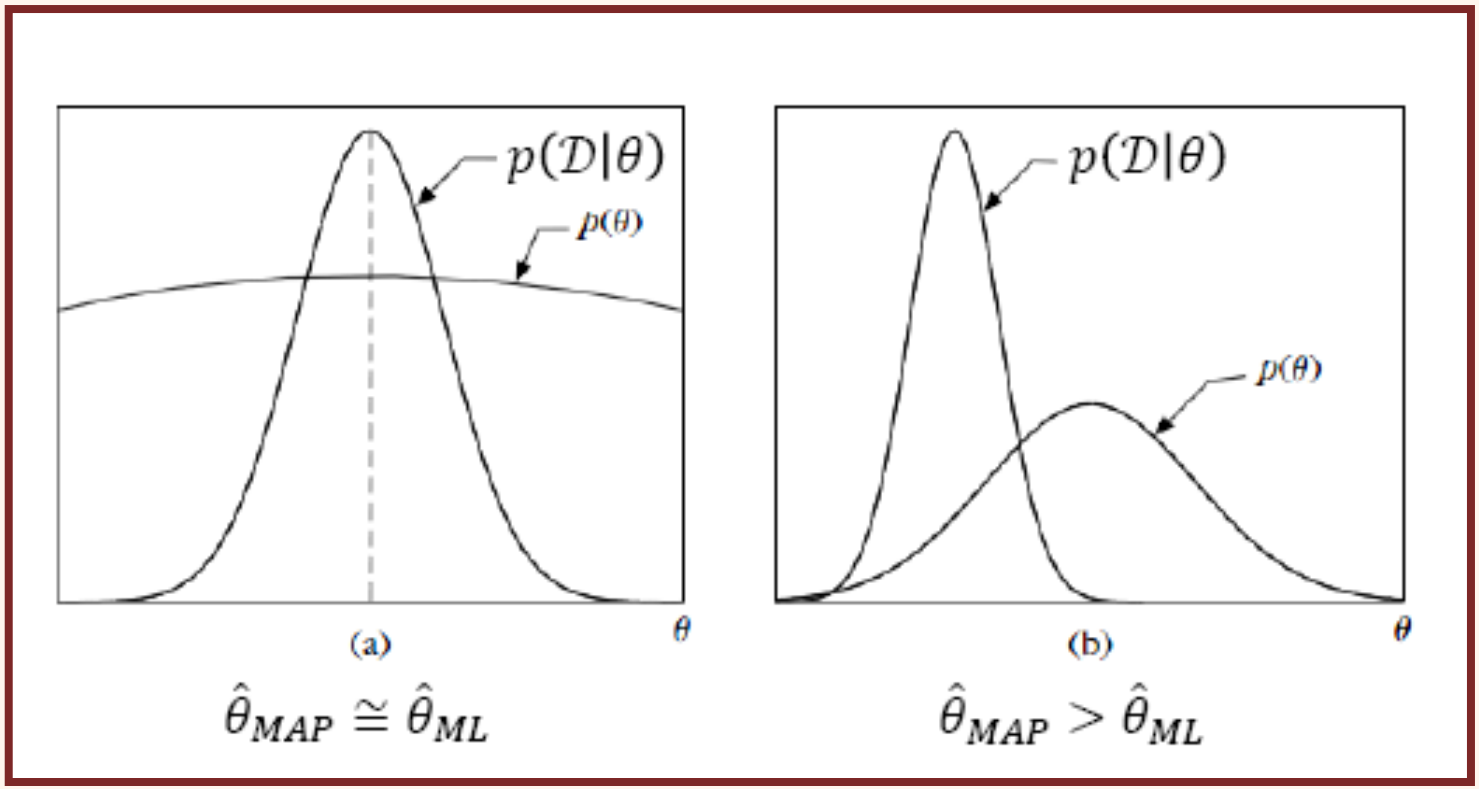
تفاوت با ML در نظر گرفتن $p(\theta)$ است.

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

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

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

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

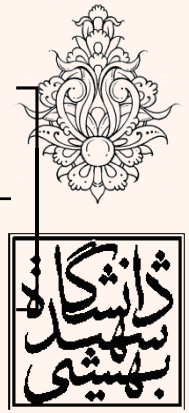
$$p(\mu \mid X) \propto p(\mu) p(X \mid \mu)$$

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

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

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

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

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

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

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.

$$p(x|\mathcal{X}) = \int p(x,\theta|\mathcal{X})d\theta$$
$$= \int p(x|\theta,\mathcal{X})p(\theta|\mathcal{X})d\theta$$
$$= \int p(x|\theta)p(\theta|\mathcal{X})d\theta$$

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

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

تابع جداساز

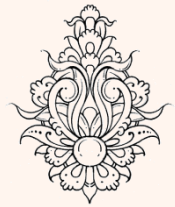
or

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

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

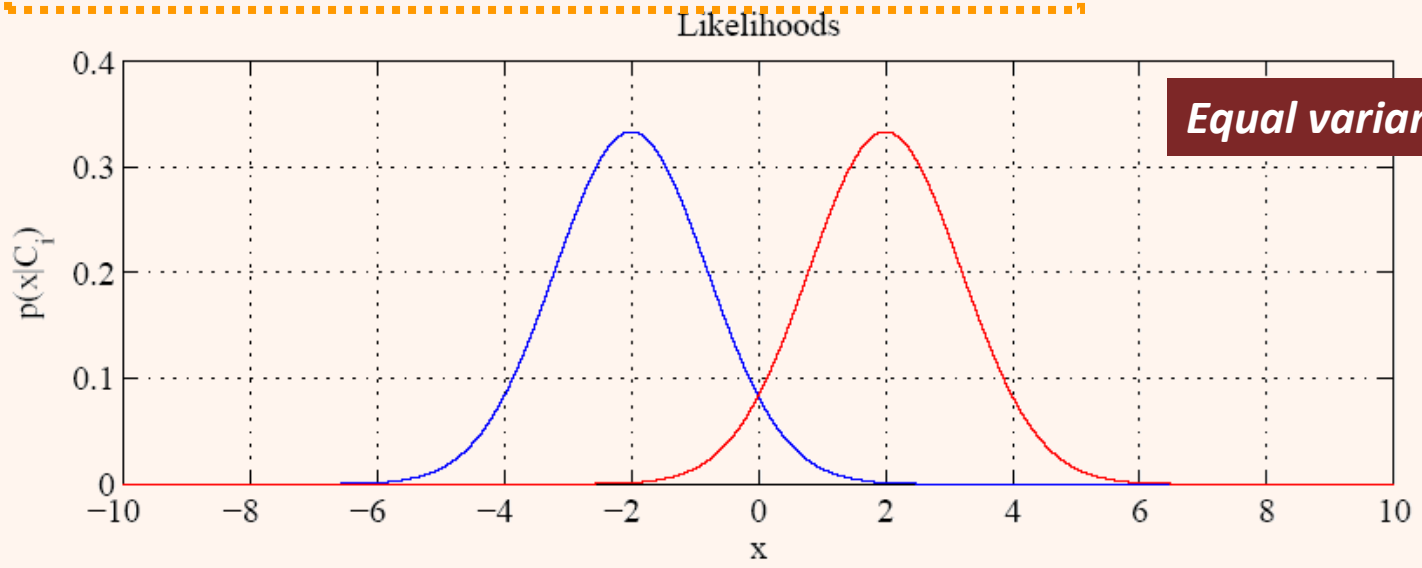
$$\hat{P}(C_i)=\frac{\sum_t r_i^t}{N} \quad m_i=\frac{\sum_t x^t r_i^t}{\sum_t r_i^t} \quad s_i^2=\frac{\sum_t (x^t-m_i)^2 r_i^t}{\sum_t r_i^t}$$

تابع جداساز

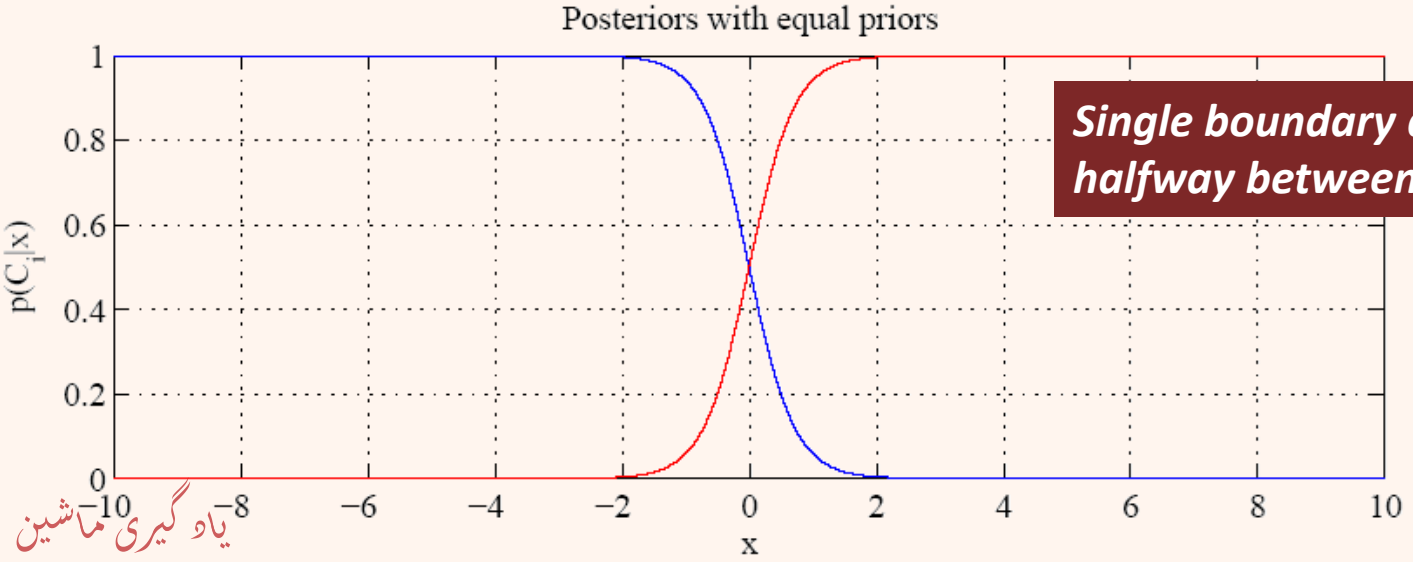
$$g_i(x)=-\frac{1}{2}\log 2\pi-\log s_i-\frac{(x-m_i)^2}{2s_i^2}+\log\hat{P}(C_i)$$

دسته‌بندی دو کلاس با واریانس یکسان

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



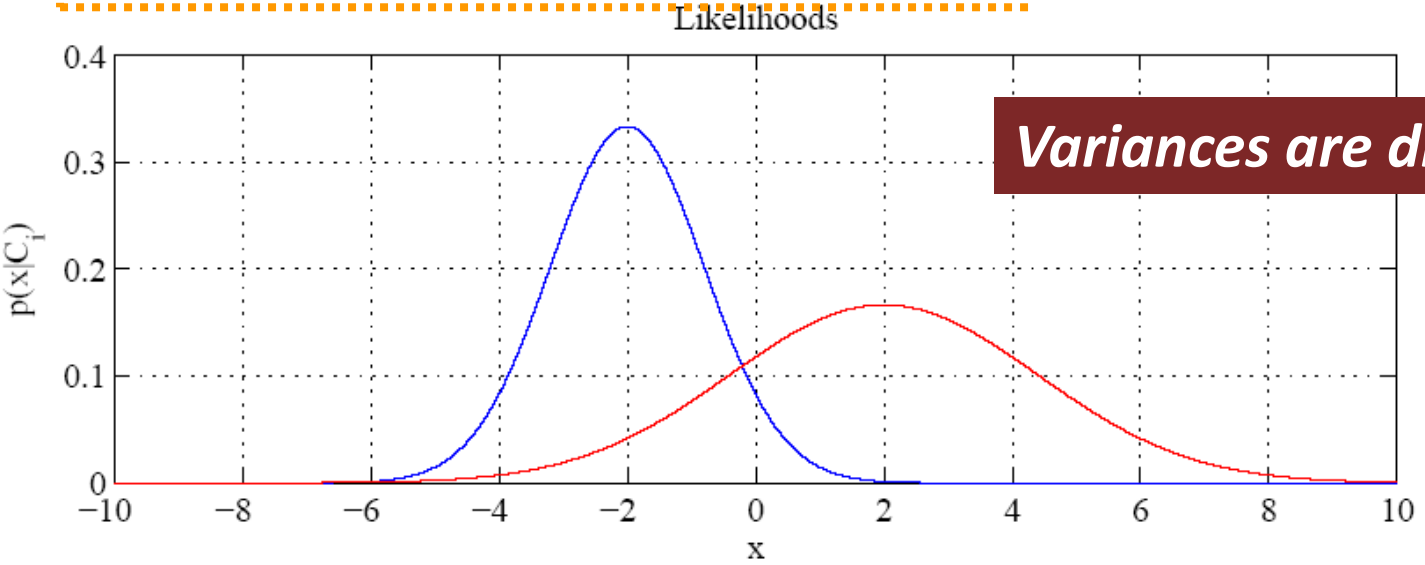
Equal variances



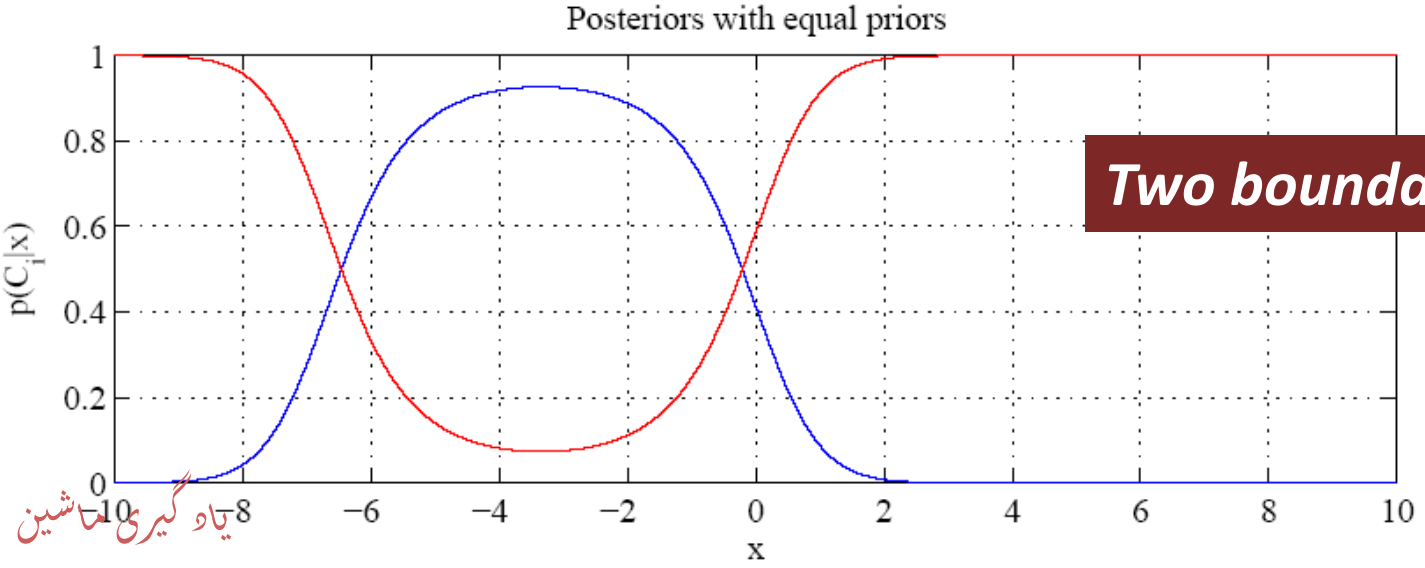
Single boundary at halfway between means

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

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