

❑ **تابع خطا (loss function)** – تابع خطا تابعی است به صورت $L : (z, y) \in \mathbb{R} \times Y \mapsto L(z, y) \in \mathbb{R}$ که به عنوان ورودی مقدار پیش‌بینی‌شده‌ی z متناظر با مقدار داده‌ی حقیقی y را می‌گیرد و اختلاف این دو را خروجی می‌دهد. توابع خطای معمول در جدول زیر آمده‌اند:

خطای کمترین مربعات	خطای لجستیک	خطای Hinge	آنتروپی متقاطع
$\frac{1}{2}(y - z)^2$	$\log(1 + \exp(-yz))$	$\max(0, 1 - yz)$	$-[y \log(z) + (1 - y) \log(1 - z)]$
وایازش خطی	وایازش لجستیک	ماشین بردار پشتیبان	شبکه‌ی عصبی

راهنمای کوتاه یادگیری با نظارت

اثنین عیدی و شروین عیدی

۱۵ شهریور ۱۳۹۸

ترجمه به فارسی توسط امیرحسین کاظم نژاد، بازبینی توسط عرفان نوری و محمد کریمی.

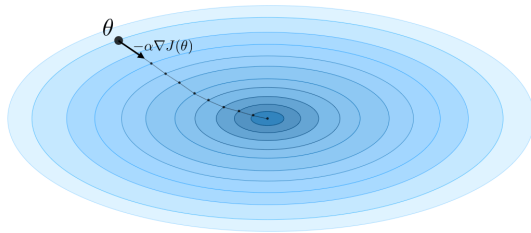
مبانی یادگیری با نظارت

❑ **تابع هزینه (cost function)** – تابع هزینه‌ی J معمولاً برای ارزیابی عملکرد یک مدل استفاده می‌شود و با توجه به تابع خطای L به صورت زیر تعریف می‌شود:

$$J(\theta) = \sum_{i=1}^m L(h_{\theta}(x^{(i)}), y^{(i)})$$

❑ **گرادیان کاهشی (gradient descent)** – با نمایش نرخ یادگیری به صورت $\alpha \in \mathbb{R}$ ، رویه‌ی به‌روزرسانی گرادیان کاهشی که با نرخ‌یادگیری و تابع هزینه‌ی J بیان می‌شود به شرح زیر است:

$$\theta \leftarrow \theta - \alpha \nabla J(\theta)$$



نکته: گرادیان کاهشی تصادفی (SGD) عوامل را بر اساس تک‌تک نمونه‌های آموزش به‌روزرسانی می‌کند، در حالی که گرادیان کاهشی دسته‌ای این کار را بر اساس دسته‌ای از نمونه‌های آموزش انجام می‌دهد.

❑ **درست‌نمایی (likelihood)** – از مقدار درست‌نمایی یک مدل $L(\theta)$ با پارامترهای θ در پیدا کردن عوامل بهینه θ از طریق روش بیشینه‌سازی درست‌نمایی مدل استفاده می‌شود. البته در عمل از لگاریتم درست‌نمایی $\ell(\theta) = \log(L(\theta))$ به‌روزرسانی آن ساده‌تر است استفاده می‌شود. داریم:

$$\theta^{\text{opt}} = \arg \max_{\theta} L(\theta)$$

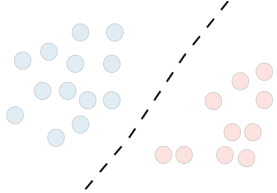
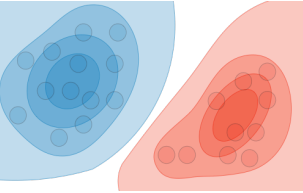
❑ **الگوریتم نیوتن (Newton's algorithm)** – الگوریتم نیوتن یک روش عددی است که θ را به گونه‌ای پیدا می‌کند که $\ell'(\theta) = 0$ باشد. رویه‌ی به‌روزرسانی آن به صورت زیر است:

با در نظر گرفتن مجموعه‌ای از نمونه‌های داده‌ی $\{x^{(1)}, \dots, x^{(m)}\}$ متناظر با مجموعه‌ی خروجی‌های $\{y^{(1)}, \dots, y^{(m)}\}$ هدف ساخت دسته‌بندی است که پیش‌بینی y از روی x را یاد می‌گیرد.

❑ **انواع پیش‌بینی** – انواع مختلف مدل‌های پیش‌بینی‌کننده در جدول زیر به اختصار آمده‌اند:

وایازش (رگرسیون)	دسته‌بندی
اعداد پیوسته	دسته
وایازش خطی	وایازش لجستیک، ماشین بردار پشتیبان، بی‌ز ساده

❑ **نوع مدل** – انواع مختلف مدل‌ها در جدول زیر به اختصار آمده‌اند.

هدف	مدل مستقیم $P(y x)$	مدل متمایزکننده	مدل مولد
چیزی که یاد گرفته می‌شود	مرز تصمیم‌گیری	توزیع احتمال داده‌ها	$P(x y)$ و سپس نتیجه‌گیری $P(y x)$
تصویر			
نمونه‌ها	وایازش‌ها، ماشین‌های بردار پشتیبان	بی‌ز ساده، GDA	

نمادها و مفاهیم کلی

❑ **فرضیه (hypothesis)** – فرضیه که با h_{θ} نمایش داده می‌شود، همان مدلی است که ما انتخاب می‌کنیم. به ازای هر نمونه داده ورودی $x^{(i)}$ ، حاصل پیش‌بینی مدل $h_{\theta}(x^{(i)})$ می‌باشد.

$$\phi_i = \frac{\exp(\theta_i^T x)}{\sum_{j=1}^K \exp(\theta_j^T x)}$$

مدل‌های خطی تعمیم‌یافته

□ **خانواده‌ی نمایی (exponential family)** – به گروهی از توزیع‌ها خانواده‌ی نمایی گوییم اگر بتوان آن‌ها را با استفاده از عامل طبیعی η ، که معمولاً عامل متعارف یا تابع پیوند نیز گفته می‌شود، آماره‌ی کافی $T(y)$ ، و تابع دیواره‌بندی لگاریتمی $a(\eta)$ به صورت زیر نوشت :

$$p(y; \eta) = b(y) \exp(\eta T(y) - a(\eta))$$

نکته : معمولاً داریم $y = T(y)$. هم‌چنین می‌توان به $\exp(-a(\eta))$ به عنوان یک عامل نرمال‌کننده نگاه کرد که باعث می‌شود جمع احتمال‌ها حتماً برابر با یک شود.

رایج‌ترین توزیع‌های نمایی در جدول زیر به اختصار آمده‌اند :

توزیع	η	$T(y)$	$a(\eta)$	$b(y)$
برنولی	$\log\left(\frac{\phi}{1-\phi}\right)$	y	$\log(1 + \exp(\eta))$	1
گوسی	μ	y	$\frac{\eta^2}{2}$	$\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{y^2}{2}\right)$
پواسون	$\log(\lambda)$	y	e^η	$\frac{1}{y!}$
هندسی	$\log(1 - \phi)$	y	$\log\left(\frac{e^\eta}{1 - e^\eta}\right)$	1

□ **فرضیه‌های مدل‌های خطی تعمیم‌یافته** – مدل‌های خطی تعمیم‌یافته (GLM) Generalized Linear Models به دنبال پیش‌بینی متغیر تصادفی y به عنوان تابعی از $x \in \mathbb{R}^{n+1}$ هستند و بر سه فرض زیر استوارند :

$$(1) \quad y|x; \theta \sim \text{ExpFamily}(\eta) \quad (2) \quad h_\theta(x) = E[y|x; \theta] \quad (3) \quad \eta = \theta^T x$$

نکته : کمینه‌ی مربعات و وایزش لجستیک حالت‌های خاصی از مدل‌های خطی تعمیم‌یافته هستند.

ماشین‌های بردار پشتیبان

هدف ماشین‌های بردار پشتیبان (Support Vector Machines) پیدا کردن خطی هست که حداقل فاصله تا خط را بیشینه می‌کند.

□ **دسته‌بند حاشیه‌ی بهینه** – دسته‌بند حاشیه‌ی بهینه‌ی h به گونه‌ای است که :

$$h(x) = \text{sign}(w^T x - b)$$

که $(w, b) \in \mathbb{R}^n \times \mathbb{R}$ راه‌حلی برای مسأله‌ی بهینه‌سازی زیر باشد :

$$\min \frac{1}{2} \|w\|^2 \quad \text{و} \quad y^{(i)}(w^T x^{(i)} - b) \geq 1$$

$$\theta \leftarrow \theta - \frac{\ell'(\theta)}{\ell''(\theta)}$$

نکته : تعمیم چندبُعدی این روش، که به روش نیوتون-رافسون معروف است، قانون به‌روزرسانی زیر را دارد :

$$\theta \leftarrow \theta - \left(\nabla_\theta^2 \ell(\theta) \right)^{-1} \nabla_\theta \ell(\theta)$$

وایزش خطی

در این‌جا فرض می‌کنیم $y|x; \theta \sim \mathcal{N}(\mu, \sigma^2)$

□ **معادلات نرمال (normal equations)** – اگر X یک ماتریس باشد، مقداری از θ که تابع هزینه را کمینه می‌کند یک راه‌حل به فرم بسته دارد به طوری که :

$$\theta = (X^T X)^{-1} X^T y$$

□ **الگوریتم LMS** – با نمایش نرخ یادگیری با α ، رویه‌ی به‌روزرسانی الگوریتم کمینه‌ی میانگین مربعات Least Mean Squares (LMS) برای یک مجموعه‌ی آموزش با m نمونه داده، که به رویه‌ی به‌روزرسانی Widrow-Hoff نیز معروف است، به صورت زیر خواهد بود :

$$\forall j, \quad \theta_j \leftarrow \theta_j + \alpha \sum_{i=1}^m [y^{(i)} - h_\theta(x^{(i)})] x_j^{(i)}$$

نکته : این رویه‌ی به‌روزرسانی، حالت خاصی از الگوریتم گرادین کاهشی است.

□ **LWR** – وایزش محلی‌وزن‌دار یا Locally Weighted Regression نوعی دیگر از انواع وایزش‌های خطی است که در محاسبه‌ی تابع هزینه‌ی خود هر کدام از نمونه‌های آموزش را وزن $w^{(i)}(x)$ می‌دهد، که این وزن با عامل $\tau \in \mathbb{R}$ به شکل زیر تعریف می‌شود :

$$w^{(i)}(x) = \exp\left(-\frac{(x^{(i)} - x)^2}{2\tau^2}\right)$$

دسته‌بندی و وایزش لجستیک

□ **تابع سیگموئید (sigmoid)** – تابع سیگموئید g که به تابع لجستیک هم معروف است به صورت زیر تعریف می‌شود :

$$\forall z \in \mathbb{R}, \quad g(z) = \frac{1}{1 + e^{-z}} \in]0, 1[$$

□ **وایزش لجستیک (logistic regression)** – فرض می‌کنیم که $y|x; \theta \sim \text{Bernoulli}(\phi)$ داریم :

$$\phi = p(y = 1|x; \theta) = \frac{1}{1 + \exp(-\theta^T x)} = g(\theta^T x)$$

نکته : هیچ راه‌حل بسته‌ای برای وایزش لجستیک وجود ندارد.

□ **وایزش softmax** – وایزش softmax یا وایزش چنددسته‌ای، در مواقعی که بیش از ۲ کلاس خروجی داریم برای تعمیم وایزش لجستیک استفاده می‌شود. طبق قرارداد داریم $\theta_K = 0$. در نتیجه عامل برنولی ϕ_i برای هر کلاس i به صورت زیر خواهد بود :

یادگیری مولد

یک مدل مولد (generative model) ابتدا با تخمین زدن $P(x|y)$ سعی می‌کند یاد بگیرد چگونه می‌توان داده را تولید کرد، سپس با استفاده از $P(x|y)$ و همچنین قضایهی بیز، $P(y|x)$ را تخمین می‌زند.

تحلیل متمایزکنندهی گاوسی

❑ **فرضیات** – در تحلیل متمایزکنندهی گاوسی فرض می‌کنیم y و $x|y = 0$ و $x|y = 1$ به طوری که :

$$y \sim \text{Bernoulli}(\phi)$$

$$x|y = 0 \sim \mathcal{N}(\mu_0, \Sigma) \quad \text{و} \quad x|y = 1 \sim \mathcal{N}(\mu_1, \Sigma)$$

❑ **تخمین** – جدول زیر تخمین‌هایی که هنگام بیشینه‌کردن تابع درست‌نمایی به آن می‌رسیم را به اختصار آورده‌است :

$\hat{\Sigma}$	$\hat{\mu}_j \quad (j = 0, 1)$	$\hat{\phi}$
$\frac{1}{m} \sum_{i=1}^m (x^{(i)} - \mu_{y^{(i)}})(x^{(i)} - \mu_{y^{(i)}})^T$	$\frac{\sum_{i=1}^m 1_{\{y^{(i)}=j\}} x^{(i)}}{\sum_{i=1}^m 1_{\{y^{(i)}=j\}}}$	$\frac{1}{m} \sum_{i=1}^m 1_{\{y^{(i)}=1\}}$

دسته‌بند بیز ساده

❑ **فرض** – مدل بیز ساده (Naive Bayes) فرض می‌کند تمام خصوصیات هر نمونهی داده از هم‌دیگر مستقل است.

$$P(x|y) = P(x_1, x_2, \dots | y) = P(x_1|y)P(x_2|y) \dots = \prod_{i=1}^n P(x_i|y)$$

❑ **راه‌حل‌ها** – بیشینه‌کردن لگاریتم درست‌نمایی به پاسخ‌های زیر می‌رسد، که $k \in \{0, 1\}$ و $l \in [1, L]$

$$P(y = k) = \frac{1}{m} \times \#\{j | y^{(j)} = k\} \quad \text{و} \quad P(x_i = l | y = k) = \frac{\#\{j | y^{(j)} = k \text{ و } x_i^{(j)} = l\}}{\#\{j | y^{(j)} = k\}}$$

نکته : دسته‌بند بیز ساده در مساله‌های دسته‌بندی متن و تشخیص هرزنامه به صورت گسترده استفاده می‌شود.

روش‌های مبتنی بر درخت و گروه

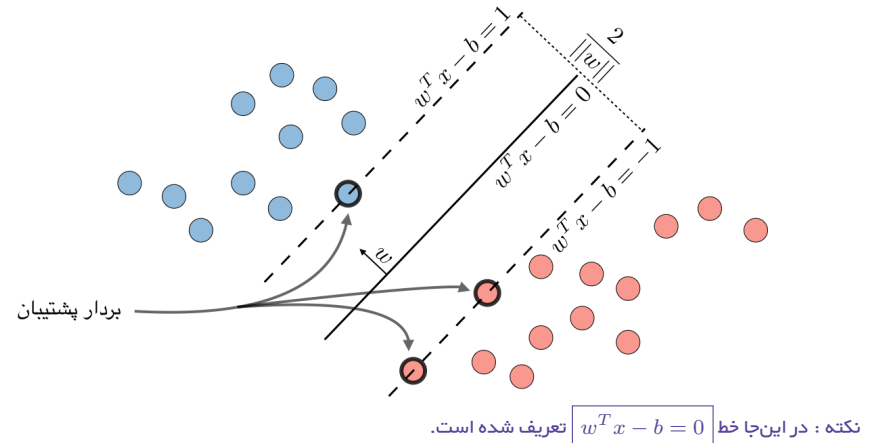
این روش‌ها هم در مسائل وایازش و هم در مسائل دسته‌بندی می‌توانند استفاده شوند.

❑ **CART** – درخت‌های وایازش و دسته‌بندی (Classification and Regression Trees)، عموماً با نام درخت‌های تصمیم‌گیری شناخته می‌شوند. می‌توان آن‌ها را به صورت درخت‌هایی دودویی نمایش داد. مزیت آن‌ها قابل تفسیر بودنشان است.

❑ **جنگل تصادفی (random forest)** – یک تکنیک مبتنی بر درخت است، که تعداد زیادی درخت تصمیم‌گیری که روی مجموعه‌هایی تصادفی از خصوصیات ساخته شده‌اند، را به کار می‌گیرد. روش جنگل تصادفی برخلاف درخت تصمیم‌گیری ساده، بسیار غیر قابل تفسیر است البته عمکرد عموماً خوب آن باعث شده است به الگوریتم محبوبی تبدیل شود.

نکته : جنگل تصادفی یکی از انواع «روش‌های گروهی» است.

❑ **ترقی‌دادن (boosting)** – ایدهی اصلی روش‌های ترقی‌دادن ترکیب چند مدل ضعیف و ساخت یک مدل قوی از آن‌هاست. انواع اصلی آن به صورت خلاصه در جدول زیر آمده‌اند :



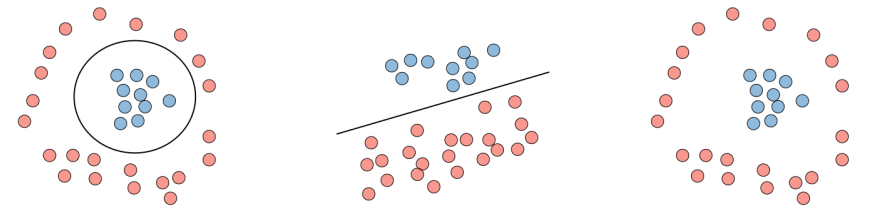
❑ **خطای Hinge** – در ماشین‌های بردار پشتیبان از تابع خطای Hinge استفاده می‌شود و تعریف آن به صورت زیر است :

$$L(z, y) = [1 - yz]_+ = \max(0, 1 - yz)$$

❑ **هسته (kernel)** – برای هر تابع نگاشت ویژگی‌های ϕ ، هسته‌ی K (kernel) به صورت زیر تعریف می‌شود :

$$K(x, z) = \phi(x)^T \phi(z)$$

در عمل، به هسته‌ی K که به صورت $K(x, z) = \exp\left(-\frac{\|x - z\|^2}{2\sigma^2}\right)$ تعریف شده باشد، هسته‌ی گاوسی می‌گوییم. این نوع هسته یکی از هسته‌های پراستفاده محسوب می‌شود.



جدانپذیری غیر خطی ← به کارگیری نگاشت هسته ϕ ← مرز تصمیم در فضای اصلی

نکته : می‌گوییم برای محاسبه‌ی تابع هزینه از «حقه‌ی هسته» استفاده می‌شود چرا که در واقع برای محاسبه‌ی آن، نیازی به دانستن دقیق نگاشت ϕ که بیشتر مواقع هم بسیار پیچیده‌ست، نداریم؛ تنها دانستن مقادیر $K(x, z)$ کافیست.

❑ **لاگرانژی (Lagrangian)** – لاگرانژی $\mathcal{L}(w, b)$ به صورت زیر تعریف می‌کنیم :

$$\mathcal{L}(w, b) = f(w) + \sum_{i=1}^l \beta_i h_i(w)$$

نکته : به ضرایب β_i ضرایب لاگرانژ هم می‌گوییم.

□ **احتمالاً تقریباً درست (Probably Approximately Correct - PAC)** – چارچوبی است که در ذیل آن نتایج متعددی در نظریه یادگیری اثبات شده است و فرض‌های زیر را در بر دارد :

- مجموعه‌ی آموزش و مجموعه‌ی آزمایش از یک توزیع هستند.
- نمونه‌های آموزشی مستقل از یکدیگر انتخاب شده‌اند.

□ **خرد شدن (shattering)** – برای مجموعه‌ی $S = \{x^{(1)}, \dots, x^{(d)}\}$ و مجموعه‌ای از دسته‌بندی‌های \mathcal{H} می‌گوییم، \mathcal{H} مجموعه‌ی S را اصطلاحاً خرد می‌کند اگر به ازای هر مجموعه‌ای از برچسب‌های $\{y^{(1)}, \dots, y^{(d)}\}$ داشته باشیم :

$$\exists h \in \mathcal{H}, \quad \forall i \in [1, d], \quad h(x^{(i)}) = y^{(i)}$$

□ **قضیه‌ی کران بالا** – اگر \mathcal{H} یک مجموعه‌ی متناهی از فرضیه‌ها (دسته‌بندی‌ها) باشد به طوری که $|\mathcal{H}| = k$ باشد و δ و m ثابت باشند، آنگاه با احتمال حداقل $1 - \delta$ داریم :

$$\epsilon(\hat{h}) \leq \left(\min_{h \in \mathcal{H}} \epsilon(h) \right) + 2\sqrt{\frac{1}{2m} \log \left(\frac{2k}{\delta} \right)}$$

□ **بُعد VC** – بُعد (Vapnik-Chervonenkis - VC) برای هر مجموعه‌ی نامتناهی از فرضیه‌ها (دسته‌بندی‌ها) \mathcal{H} که با $VC(\mathcal{H})$ نمایش داده می‌شود، برابر است با اندازه‌ی بزرگ‌ترین مجموعه‌ای که می‌توان با استفاده از آن \mathcal{H} را خرد کرد.

نکته : بُعد VC مجموعه‌ی {همه‌ی دسته‌بندی‌های خطی در ۲ بُعد} \mathcal{H} برابر با ۳ است.



□ **قضیه (Vapnik)** – به ازای \mathcal{H} به طوری که $VC(\mathcal{H}) = d$ و همچنین m تعداد نمونه‌های آموزشی باشد، با احتمال حداقل $1 - \delta$ داریم :

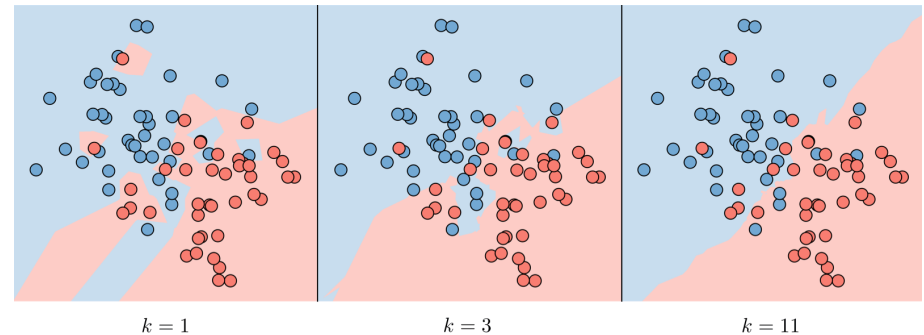
$$\epsilon(\hat{h}) \leq \left(\min_{h \in \mathcal{H}} \epsilon(h) \right) + O \left(\sqrt{\frac{d}{m} \log \left(\frac{m}{d} \right)} + \frac{1}{m} \log \left(\frac{1}{\delta} \right) \right)$$

ترقی‌دادن سازگار شونده (Adaptive boosting)	ترقی‌دادن گرادینانی (Gradient boosting)
برای خطاها وزن بالایی در نظر می‌گیرد تا در مرحله‌ی بعدی ترقی‌دادن، مدل بهبود یابد.	چند مدل ضعیف روی باقی خطاها آموزش می‌یابند

سایر رویکردهای غیر عاملی

□ **k -همسایه‌ی نزدیک (k -nearest neighbors)** – الگوریتم k -همسایه‌ی نزدیک که عموماً با k -nearest neighbors نیز شناخته می‌شود، یک الگوریتم غیرعاملی است که پاسخ مدل به هر نمونه داده از روی k همسایه‌ی آن در مجموعه دادگان آموزش تعیین می‌شود. این الگوریتم هم در دسته‌بندی و هم در واپازش استفاده می‌شود.

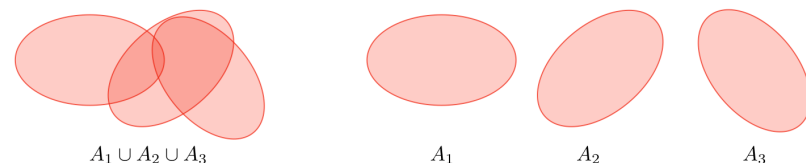
نکته : هرچه پارامتر k بزرگ‌تر باشد پیش‌قدر مدل بیشتر خواهد بود، و هر چه کوچک‌تر باشد واریانس مدل بیشتر خواهد شد.



نظریه یادگیری

□ **کران اجتماع** – اگر A_1, \dots, A_k عدد رخداد باشد، داریم :

$$P(A_1 \cup \dots \cup A_k) \leq P(A_1) + \dots + P(A_k)$$



□ **نامساوی هوفدینگ (Hoeffding inequality)** – اگر Z_1, \dots, Z_m عدد متغیر تصادفی مستقل با توزیع یکسان و نمونه‌برداری شده از توزیع برنولی با پارامتر ϕ باشند و همچنین $\hat{\phi}$ میانگین آن‌ها و $\gamma > 0$ ثابت باشد، داریم :

$$P(|\phi - \hat{\phi}| > \gamma) \leq 2 \exp(-2\gamma^2 m)$$

نکته : این نامساوی به کران چرنوف نیز معروف است.

□ **خطای آموزش** – به ازای هر دسته‌بند h ، خطای آموزش $\hat{\epsilon}(h)$ (یا همان خطای تجربی)، به صورت زیر تعریف می‌شود :

$$\hat{\epsilon}(h) = \frac{1}{m} \sum_{i=1}^m 1_{\{h(x^{(i)}) \neq y^{(i)}\}}$$