

# راهنمای کوتاه ویژه : یادگیری ماشین

افشین عمیدی و شروین عمیدی

۱۵ شهریور ۱۳۹۸

جدول محتوا

۸	یادگیری عمیق	۳
۸	شبکه‌های عمیق	۳/۱
۸	شبکه‌های عمیق پیچشی	۳/۲
۸	شبکه‌های عمیق بازگشتی	۳/۳
۹	یادگیری تقویتی و کنترل	۳/۴
۱۰	نکات و ترفندهای یادگیری ماشین	۴
۱۰	معیارهای دسته‌بندی	۴/۰/۱
۱۰	معیارهای وایزش	۴/۰/۲
۱۱	انتخاب مدل	۴/۱
۱۱	عیب‌شناسی	۴/۲
۱۲	آمار و احتمالات	۵
۱۲	مقدمه‌ای بر احتمالات و ترکیبیات	۵/۱
۱۲	احتمال شرطی	۵/۲
۱۳	متغیرهای تصادفی	۵/۳
۱۳	متغیرهای تصادفی با توزیع مشترک	۵/۴
۱۴	تخمین پارامتر	۵/۵
۱۵	جبر خطی و حسابان	۶
۱۵	نمادها	۶/۱
۱۵	عملیات ماتریسی	۶/۲
۱۵	ضرب	۶/۲/۱
۱۵	دیگر عملیات	۶/۲/۲
۱۶	ویژگی‌های ماتریس‌ها	۶/۳
۱۶	تعاریف	۶/۳/۱
۱۶	قضیه	۶/۳/۲
۱۶	حسابان ماتریسی	۶/۴

۱ یادگیری با نظارت

۱/۱	مبانی یادگیری با نظارت
۱/۲	نمادها و مفاهیم کلی
۱/۳	مدل‌های خطی
۱/۳/۱	وایزش خطی
۱/۳/۲	دسته‌بندی و وایزش لجستیک
۱/۳/۳	مدل‌های خطی تعمیم‌یافته
۱/۴	ماشین‌های بردار پشتیبان
۱/۵	یادگیری مولد
۱/۵/۱	تحلیل متمایزکننده‌ی گاوسی
۱/۵/۲	دسته‌بند بیز ساده
۱/۶	روش‌های مبتنی بر درخت و گروه
۱/۷	سایر رویکردهای غیر عاملی
۱/۸	نظریه یادگیری

۲ یادگیری بدون نظارت

۲/۱	مبانی یادگیری بدون نظارت
۲/۲	خوشه‌بندی
۲/۲/۱	بیشینه‌سازی امید ریاضی
۲/۲/۲	خوشه‌بندی $k$ - میانگین
۲/۲/۳	خوشه‌بندی سلسله‌مراتبی
۲/۲/۴	معیارهای ارزیابی خوشه‌بندی
۲/۳	کاهش ابعاد
۲/۳/۱	تحلیل مولفه‌های اصلی
۲/۳/۲	تحلیل مولفه‌های مستقل

## ۱ یادگیری با نظارت

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

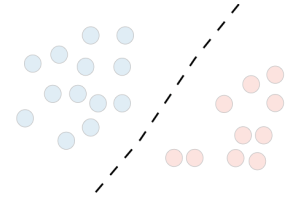
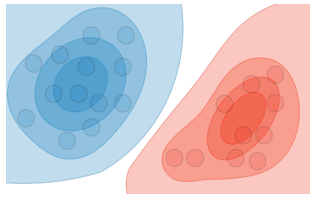
## ۱/۱ مبانی یادگیری با نظارت

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 پیش‌بینی کننده در جدول زیر به اختصار آمده‌اند :

خروجی	وایازش (رگرسیون)	دسته‌بندی
نمونه‌ها	وایازش خطی	دسته
	وایازش لجستیک، ماشین بردار پشتیبان، بیز ساده	

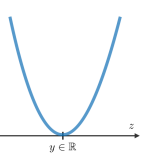
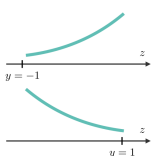
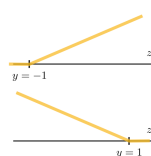
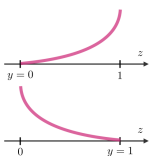
□ نوع مدل – انواع مختلف مدل‌ها در جدول زیر به اختصار آمده‌اند.

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

## ۱/۲ نمادها و مفاهیم کلی

□ فرضیه (hypothesis) – فرضیه که با  $h_\theta$  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.

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

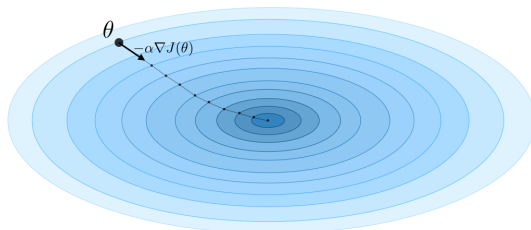
خطای کمترین مربعات	خطای لجستیک	خطای Hinge	آنتروپی متقاطع
$\frac{1}{2}(y - z)^2$	$\log(1 + \exp(-yz))$	$\max(0, 1 - yz)$	$-\left[y \log(z) + (1 - y) \log(1 - z)\right]$
			
وایازش خطی	وایازش لجستیک	ماشین بردار پشتیبان	شبکه‌ی عصبی

□ 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  $\theta$  در پیدا کردن عوامل بهینه  $\theta$  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  $\theta$  را به گونه‌ای پیدا می‌کند که  $\ell'(\theta) = 0$  باشد. رویه‌ی به‌روزرسانی آن به صورت زیر است :

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

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

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

۱/۳ مدل‌های خطی

۱/۳/۳ مدل‌های خطی تعمیم‌یافته

۱/۳/۱ وایزش خطی

□ **خانواده‌ی نمایی (exponential family)** – به گروهی از توزیع‌ها خانواده‌ی نمایی گوییم اگر بتوان آن‌ها را با استفاده از عامل طبیعی  $\eta$ ، 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	$\eta$	$T(y)$	$a(\eta)$	$b(y)$
برنولی	$\log\left(\frac{\phi}{1-\phi}\right)$	$y$	$\log(1 + \exp(\eta))$	1
گاوسی	$\mu$	$y$	$\frac{\eta^2}{2}$	$\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{y^2}{2}\right)$
پواسون	$\log(\lambda)$	$y$	$e^\eta$	$\frac{1}{y!}$
هندسی	$\log(1 - \phi)$	$y$	$\log\left(\frac{e^\eta}{1 - e^\eta}\right)$	1

□ **فرضیه‌های مدل‌های خطی تعمیم‌یافته** – مدل‌های خطی تعمیم‌یافته Generalized Linear Models (GLM) 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.

۱/۴ 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 :

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

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

□ **معادلات نرمال (normal equations)** – اگر  $X$  یک ماتریس باشد، مقداری از  $\theta$  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  $\alpha$ ، 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 :

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

۱/۳/۲ 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  $\phi_i$  برای هر کلاس  $i$  به صورت زیر خواهد بود :

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

## ۱/۵ 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 قضیه بیز،  $P(y|x)$  را تخمین می‌زند.

## ۱/۵/۱ 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:

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

## ۱/۵/۲ 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** – پیشنهاد کردن 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 مسأله‌های دسته‌بندی متن و تشخیص هرزنامه به صورت گسترده استفاده می‌شود.

## ۱/۶ روش‌های مبتنی بر درخت و گروه

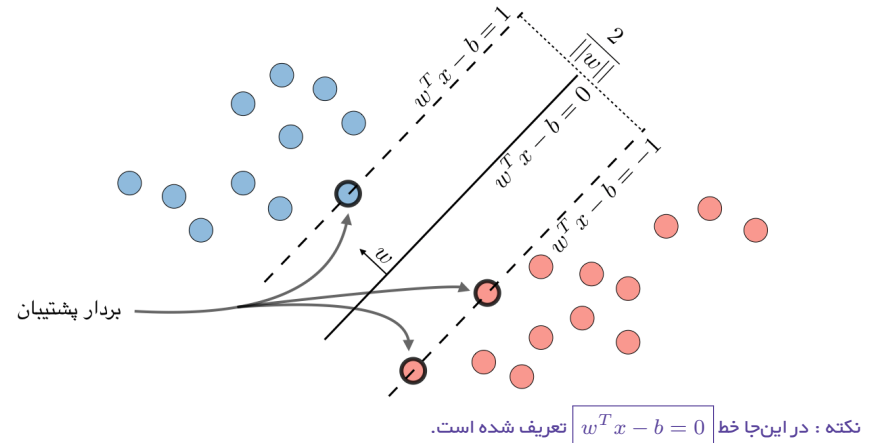
این روش‌ها هم در مسائل وایازش و هم در مسائل دسته‌بندی می‌توانند استفاده شوند.

❑ **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:



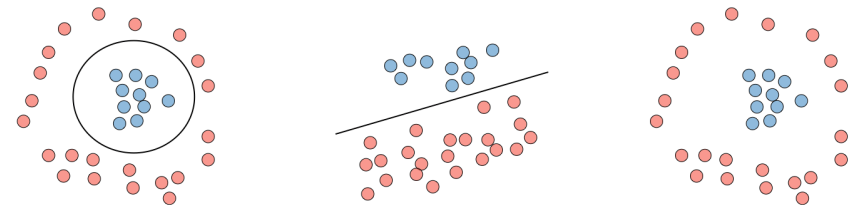
❑ **خطای Hinge** – در ماشین‌های بردار پشتیبان از تابع خطای Hinge استفاده می‌شود و تعریف آن به صورت زیر است:

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

❑ **هسته (kernel)** – برای هر تابع نگاشت ویژگی‌های  $\phi$ ، 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 پر استفاده محسوب می‌شود.



جداپذیری غیر خطی ← به کارگیری نگاشت هسته  $\phi$  ← 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  $\phi$  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  $\beta_i$  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 دسته‌بندهای  $\mathcal{H}$  می‌گوییم، 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 (دسته‌بندها) باشد به طوری که  $|\mathcal{H}| = k$  باشد و  $\delta$  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 (VC - Vapnik-Chervonenkis) برای هر مجموعه‌ی نامتناهی از فرضیه‌ها (دسته‌بندها)  $\mathcal{H}$  که با  $VC(\mathcal{H})$  نمایش داده می‌شود، برابر است با اندازه‌ی بزرگ‌ترین مجموعه‌ای که می‌توان با استفاده از آن را خرد کرد.

نکته : بُعد VC مجموعه‌ی {همه‌ی دسته‌بندهای خطی در ۲ بعد}  $\mathcal{H}$  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 :

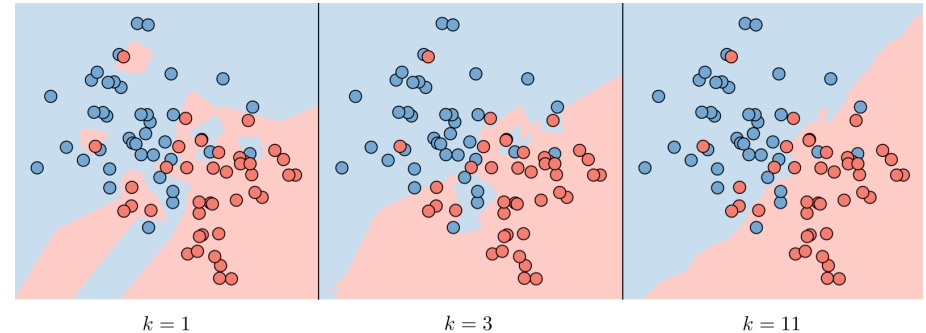
$$\epsilon(\hat{h}) \leq \left( \min_{h \in \mathcal{H}} \epsilon(h) \right) + O \left( \sqrt{\frac{d}{m} \log \left( \frac{m}{d} \right)} + \frac{1}{m} \log \left( \frac{1}{\delta} \right) \right)$$

ترقی‌دادن سازگار شونده (Adaptive boosting)	ترقی‌دادن گرادایانی (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

۱/۷ سایر رویکردهای غیر عاملی

□ **k-همسایه‌ی نزدیک (k-nearest neighbors)** – الگوریتم  $k$  – همسایه‌ی نزدیک که عموماً با  $k$ -nearest -  $k$ -NN 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 وایازش استفاده می‌شود.

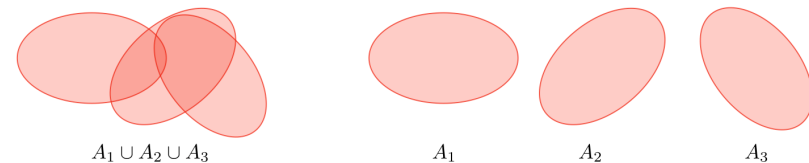
نکته : هرچه پارامتر  $k$  بزرگ‌تر باشد پیش‌قدر مدل بیشتر خواهد بود، و هر چه کوچک‌تر باشد واریانس مدل بیشتر خواهد شد.



۱/۸ 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  $\phi$  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 :

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

## ۲ یادگیری بدون نظارت

ترجمه به فارسی توسط عرفان نوری. بازبینی توسط محمد کریمی.

## ۲/۱ مبانی یادگیری بدون نظارت

انگیزه – هدف از یادگیری بدون نظارت unsupervised learning کشف الگوهای پنهان در داده‌های بدون برچسب  $\{x^{(1)}, \dots, x^{(m)}\}$  است.

نابرابری یسن (Jensen's inequality) – فرض کنید  $f$  تابعی محدب و  $X$  یک متغیر تصادفی باشد. در این صورت نابرابری زیر را داریم:

$$E[f(X)] \geq f(E[X])$$

## ۲/۲ خوشه‌بندی

## ۲/۲/۱ پیشینه‌سازی امید ریاضی

متغیرهای نهفته (latent variables) – متغیرهای نهفته متغیرهای پنهان یا مشاهده‌نشده‌ای هستند که مسائل تخمین را دشوار می‌کنند، و معمولاً با  $z$  نمایش داده می‌شوند. شرایط معمول که در آن‌ها متغیرهای نهفته وجود دارند در زیر آمده‌اند:

موقعیت	متغیر نهفته $z$	$x z$	توضیحات
ترکیب $k$ توزیع گاوسی	Multinomial( $\phi$ )	$\mathcal{N}(\mu_j, \Sigma_j)$	$\mu_j \in \mathbb{R}^n, \phi \in \mathbb{R}^k$
تحلیل عامل	$\mathcal{N}(0, I)$	$\mathcal{N}(\mu + \Lambda z, \psi)$	$\mu_j \in \mathbb{R}^n$

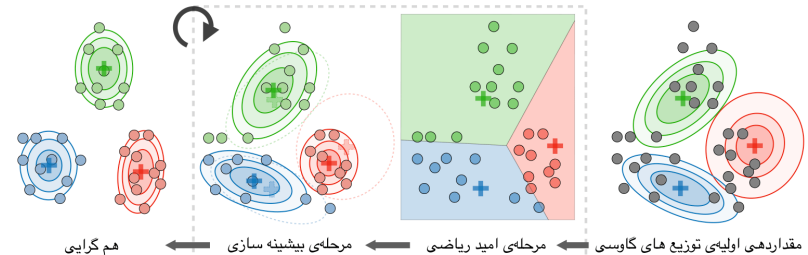
الگوریتم – الگوریتم پیشینه‌سازی امید ریاضی (Expectation-Maximization - EM) روشی بهینه برای تخمین پارامتر  $\theta$  تخمین درستی پیشینه در اختیار قرار می‌دهد. این کار با تکرار مرحله‌ای به دست آوردن یک کران پایین برای درستی (مرحله امید ریاضی) و همچنین بهینه‌سازی آن کران پایین (مرحله پیشینه‌سازی) طبق توضیح زیر انجام می‌شود:

مرحله امید ریاضی: احتمال پسین  $Q_i(z^{(i)})$  که هر نمونه داده  $x^{(i)}$  متعلق به خوشه‌ی  $z^{(i)}$  باشد به صورت زیر محاسبه می‌شود:

$$Q_i(z^{(i)}) = P(z^{(i)} | x^{(i)}; \theta)$$

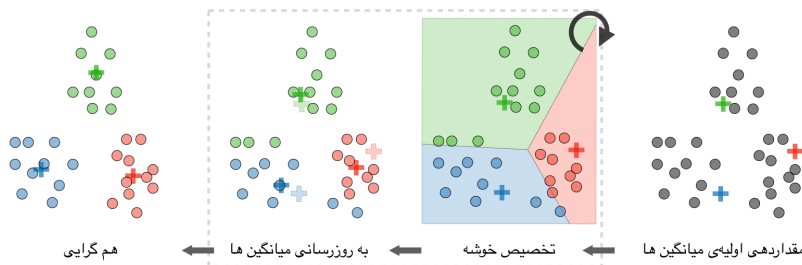
مرحله پیشینه‌سازی: با استفاده از احتمالات پسین  $Q_i(z^{(i)})$  به عنوان وزن‌های وابسته به خوشه‌ها برای نمونه‌های داده‌ی  $x^{(i)}$ ، مدل مربوط به هر کدام از خوشه‌ها، طبق توضیح زیر، دوباره تخمین زده می‌شود:

$$\theta_i = \operatorname{argmax}_{\theta} \sum_i \int_{z^{(i)}} Q_i(z^{(i)}) \log \left( \frac{P(x^{(i)}, z^{(i)}; \theta)}{Q_i(z^{(i)})} \right) dz^{(i)}$$

۲/۲/۲ خوشه‌بندی  $k$  – میانگینتوجه کنید که  $c^{(i)}$  خوشه‌ی نمونه داده‌ی  $i$  و  $\mu_j$  مرکز خوشه‌ی  $j$  است.

الگوریتم – بعد از مقداردهی اولیه‌ی تصادفی مراکز خوشه  $\mu_1, \mu_2, \dots, \mu_k \in \mathbb{R}^n$ ، الگوریتم  $k$  – میانگین مراحل زیر را تا همگرایی تکرار می‌کند:

$$c^{(i)} = \operatorname{argmin}_j \|x^{(i)} - \mu_j\|^2 \quad \text{و} \quad \mu_j = \frac{\sum_{i=1}^m 1_{\{c^{(i)}=j\}} x^{(i)}}{\sum_{i=1}^m 1_{\{c^{(i)}=j\}}}$$



تابع اعوجاج – برای تشخیص اینکه الگوریتم به همگرایی رسیده است، به تابع اعوجاج (distortion function) که به صورت زیر تعریف می‌شود رجوع می‌کنیم:

$$J(c, \mu) = \sum_{i=1}^m \|x^{(i)} - \mu_{c^{(i)}}\|^2$$

## ۲/۲/۳ خوشه‌بندی سلسله‌مراتبی

الگوریتم – یک الگوریتم خوشه‌بندی سلسله‌مراتبی تجمعی است که خوشه‌های تودرتو را به صورت پی‌درپی ایجاد می‌کند.

انواع – انواع مختلفی الگوریتم خوشه‌بندی سلسله‌مراتبی وجود دارند که هر کدام به دنبال بهینه‌سازی توابع هدف مختلفی هستند، که در جدول زیر به اختصار آمده‌اند:

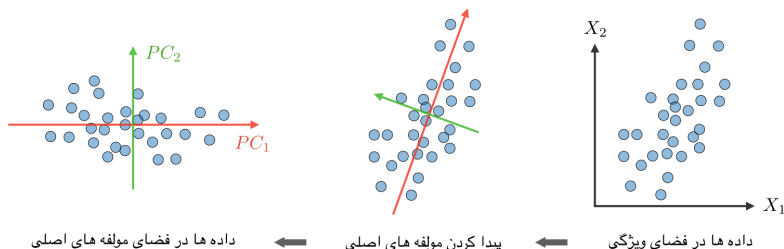
پیوند بخشی (Ward)	پیوند میانگین (Average)	پیوند کامل (Complete)
کمینه‌کردن فاصله‌ی درون خوشه	کمینه‌کردن فاصله‌ی میانگین بین هر دو جفت خوشه	کمینه‌کردن حداکثر فاصله بین هر دو جفت خوشه

## ۲/۲/۴ معیارهای ارزیابی خوشه‌بندی

در یک وضعیت یادگیری بدون نظارت، معمولاً ارزیابی یک مدل کار دشواری است، زیرا برخلاف حالت یادگیری نظارتی اطلاعاتی در مورد برچسب‌های حقیقی داده‌ها نداریم.

ضریب نیم‌رخ (Silhouette coefficient) – با نمایش  $a$  به عنوان میانگین فاصله‌ی یک نمونه با همه‌ی نمونه‌های دیگر در همان کلاس، و با نمایش  $b$  به عنوان میانگین فاصله‌ی یک نمونه با همه‌ی نمونه‌های دیگر از نزدیک‌ترین خوشه، ضریب نیم‌رخ  $s$  به صورت زیر تعریف می‌شود:

- مرحله ۴ : داده‌ها بر روی فضای  $\text{span}_{\mathbb{R}}(u_1, \dots, u_k)$  تصویر می‌شوند. این رویه واریانس را در فضای  $k$ -بعدی به دست آمده پیشینه می‌کند.



داده‌ها در فضای مولفه‌های اصلی

پیدا کردن مولفه‌های اصلی

داده‌ها در فضای ویژگی

## ۲/۳/۲ تحلیل مولفه‌های مستقل

روشی است که برای پیدا کردن منابع مولد داده به کار می‌رود.

- فرضیه‌ها** – فرض می‌کنیم که داده‌ی  $x$  توسط بردار  $n$ -بعدی  $s = (s_1, \dots, s_n)$  تولید شده است، که  $s_i$  ها متغیرهای تصادفی مستقل هستند، و این تولید داده از طریق بردار منبع به وسیله‌ی یک ماتریس معکوس‌پذیر و ترکیب‌کننده‌ی  $A$  به صورت زیر انجام می‌گیرد :

$$x = As$$

هدف پیدا کردن ماتریس ضدترکیب  $W = A^{-1}$  است.

- الگوریتم تحلیل مولفه‌های مستقل Bell و Sejnowski** – این الگوریتم ماتریس ضدترکیب  $W$  را در مراحل زیر پیدا می‌کند :
- احتمال  $x = As = W^{-1}s$  به صورت زیر نوشته می‌شود :

$$p(x) = \prod_{i=1}^n p_s(w_i^T x) \cdot |W|$$

- با نمایش تابع سیگموئید  $g$ ، لگاریتم درست‌نمایی با توجه به داده‌های  $\{x^{(i)}, i \in [1, m]\}$  به صورت زیر نوشته می‌شود :

$$l(W) = \sum_{i=1}^m \left( \sum_{j=1}^n \log \left( g'(w_j^T x^{(i)}) \right) + \log |W| \right)$$

- بنابراین، رویه‌ی یادگیری گرادینان تصادفی افزایشی برای هر نمونه از داده‌های آموزش  $x^{(i)}$  به گونه‌ای است که برای به‌روزرسانی  $W$  داریم :

$$W \leftarrow W + \alpha \left( \begin{pmatrix} 1 - 2g(w_1^T x^{(i)}) \\ 1 - 2g(w_2^T x^{(i)}) \\ \vdots \\ 1 - 2g(w_n^T x^{(i)}) \end{pmatrix} x^{(i)T} + (W^T)^{-1} \right)$$

$$s = \frac{b-a}{\max(a,b)}$$

- شاخص Calinski-Harabaz** – با در نظر گرفتن  $k$  به عنوان تعداد خوشه‌ها، ماتریس پراکندگی درون خوشه‌ای  $B_k$  و ماتریس پراکندگی میان‌خوشه‌ای  $W_k$  به صورت زیر تعریف می‌شوند :

$$B_k = \sum_{j=1}^k n_{c(i)} (\mu_{c(i)} - \mu)(\mu_{c(i)} - \mu)^T, \quad W_k = \sum_{i=1}^m (x^{(i)} - \mu_{c(i)})(x^{(i)} - \mu_{c(i)})^T$$

- شاخص Calinski-Harabaz  $s(k)$  بیان می‌کند که یک مدل خوشه‌بندی چگونه خوشه‌های خود را مشخص می‌کند، به گونه‌ای که هر چقدر مقدار این شاخص بیشتر باشد، خوشه‌ها متراکم‌تر و از هم تفکیک‌یافته‌تر خواهند بود. این شاخص به صورت زیر تعریف می‌شود :

$$s(k) = \frac{\text{Tr}(B_k)}{\text{Tr}(W_k)} \times \frac{N-k}{k-1}$$

## ۲/۳ کاهش ابعاد

### ۲/۳/۱ تحلیل مولفه‌های اصلی

روشی برای کاهش ابعاد است که جهت‌هایی را با حداکثر واریانس پیدا می‌کند تا داده‌ها را در آن جهت‌ها تصویر کند.

- مقدار ویژه، بردار ویژه (eigenvalue, eigenvector)** – برای ماتریس دلخواه  $A \in \mathbb{R}^{n \times n}$ ،  $\lambda$  مقدار ویژه‌ی ماتریس  $A$  است اگر وجود داشته باشد بردار  $z \in \mathbb{R}^n \setminus \{0\}$  که به آن بردار ویژه می‌گویند، به طوری که :

$$Az = \lambda z$$

- قضیه‌ی طیفی (spectral theorem)** – فرض کنید  $A \in \mathbb{R}^{n \times n}$  باشد. اگر  $A$  متقارن باشد، در این صورت  $A$  توسط یک ماتریس حقیقی متعامد  $U \in \mathbb{R}^{n \times n}$  قطری‌پذیر است. با نمایش  $\Lambda = \text{diag}(\lambda_1, \dots, \lambda_n)$  داریم :

$$\exists \Lambda \text{ diagonal, } A = U \Lambda U^T$$

نکته : بردار ویژه‌ی متناظر با بزرگ‌ترین مقدار ویژه، بردار ویژه‌ی اصلی ماتریس  $A$  نام دارد.

- الگوریتم** – رویه‌ی تحلیل مولفه‌های اصلی یک روش کاهش ابعاد است که داده‌ها را در فضای  $k$ -بعدی با پیشینه کردن واریانس داده‌ها، به صورت زیر تصویر می‌کند :

- مرحله ۱ : داده‌ها به گونه‌ای نرمال‌سازی می‌شوند که میانگین ۰ و انحراف معیار ۱ داشته باشند.

$$x_j^{(i)} \leftarrow \frac{x_j^{(i)} - \mu_j}{\sigma_j} \quad \text{و} \quad \mu_j = \frac{1}{m} \sum_{i=1}^m x_j^{(i)} \quad \text{و} \quad \sigma_j^2 = \frac{1}{m} \sum_{i=1}^m (x_j^{(i)} - \mu_j)^2$$

- مرحله ۲ : مقدار  $\Sigma = \frac{1}{m} \sum_{i=1}^m x^{(i)} x^{(i)T} \in \mathbb{R}^{n \times n}$  که ماتریسی متقارن با مقادیر ویژه‌ی حقیقی است محاسبه می‌شود.

- مرحله ۳ : بردارهای  $u_1, \dots, u_k \in \mathbb{R}^n$  که  $k$  بردارهای ویژه‌ی اصلی متعامد  $\Sigma$  هستند محاسبه می‌شوند. این بردارهای ویژه متناظر با  $k$  مقدار ویژه با بزرگ‌ترین مقدار هستند.

## ۳ یادگیری عمیق

□ **انتشار معکوس (backpropagation)** – انتشار معکوس روشی برای بروزرسانی وزن‌ها با توجه به خروجی واقعی و خروجی مورد انتظار در شبکه‌ی عصبی است. مشتق نسبت به وزن  $w$  توسط قاعده‌ی زنجیری محاسبه می‌شود و به شکل زیر است :

$$\frac{\partial L(z,y)}{\partial w} = \frac{\partial L(z,y)}{\partial a} \times \frac{\partial a}{\partial z} \times \frac{\partial z}{\partial w}$$

در نتیجه، وزن به صورت زیر بروزرسانی می‌شود :

$$w \leftarrow w - \eta \frac{\partial L(z,y)}{\partial w}$$

□ **بروزرسانی وزن‌ها** – در یک شبکه‌ی عصبی، وزن‌ها به صورت زیر بروزرسانی می‌شوند :

- **گام ۱** : یک دسته از داده‌های آموزشی را تهیه می‌کنیم.
- **گام ۲** : الگوریتم انتشار مستقیم را برای بدست آوردن خطای مربوطه اجرا می‌کنیم.
- **گام ۳** : خطا را انتشار معکوس می‌دهیم تا گرادینت‌ها به دست بیایند.
- **گام ۴** : از گرادینت‌ها برای بروزرسانی وزن‌های شبکه استفاده می‌کنیم.

□ **برون‌اندازی (dropout)** – برون‌اندازی یک روش برای جلوگیری از بیش‌برازش بر روی داده‌های آموزشی با حذف تصادفی واحدها در یک شبکه‌ی عصبی است. در عمل، واحدها با احتمال  $p$  حذف یا با احتمال  $1-p$  حفظ می‌شوند.

## ۳/۲ شبکه‌های عصبی پیچشی

□ **الزامات لایه کانولوشنی** – با نمایش  $W$  اندازه توده‌ی ورودی،  $F$  اندازه نورون‌های لایه‌ی کانولوشنی،  $P$  اندازه‌ی حاشیه‌ی صفر، تعداد نورون‌های  $N$  که در توده‌ی داده شده قرار می‌گیرند برابر است با :

$$N = \frac{W - F + 2P}{S} + 1$$

□ **نرمال‌سازی دسته‌ای (batch normalization)** – یک مرحله از فرامعامل‌های  $\gamma$  و  $\beta$  که دسته‌ی  $\{x_i\}$  را نرمال می‌کند در زیر آمده است. نماد  $\mu_B, \sigma_B^2$  به میانگین و واریانس دسته‌ای که می‌خواهیم آن را اصلاح کنیم اشاره دارد که به صورت زیر است :

$$x_i \leftarrow \gamma \frac{x_i - \mu_B}{\sqrt{\sigma_B^2 + \epsilon}} + \beta$$

معمولاً بعد از یک لایه‌ی تمام‌متصل یا لایه‌ی کانولوشنی و قبل از یک لایه‌ی غیرخطی اعمال می‌شود و امکان استفاده از نرخ یادگیری بالاتر را می‌دهد و همچنین باعث می‌شود که وابستگی شدید مدل به مقداردهی اولیه کاهش یابد.

## ۳/۳ شبکه‌های عصبی بازگشتی

□ **انواع دروازه‌ها** – انواع مختلف دروازه‌هایی که در یک شبکه‌ی عصبی بازگشتی معمولی به آنها برمی‌خوریم در زیر آمده‌اند :

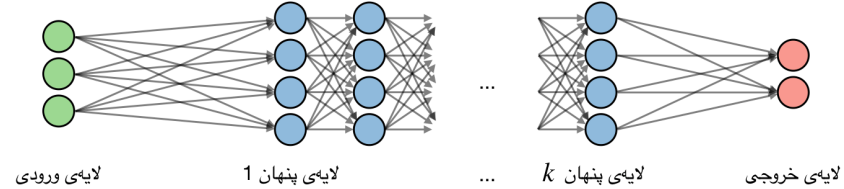
دروازه‌ی ورودی	دروازه‌ی فراموشی	دروازه	دروازه‌ی خروجی
در سلول بنویسد یا خیر؟	سلول را پاک کند یا خیر؟	چه مقدار در سلول بنویسد؟	چه مقدار برای سلول آشکار کند؟

□ **LSTM** – یک شبکه‌ی حافظه‌ی کوتاه-مدت طولانی (LSTM) یک نوع از مدل‌های RNN است که مشکل ناپدید شدن (صفر شدن) گرادینت را با اضافه کردن «دروازه‌ی فراموشی» حل می‌کند.

## ۳/۱ شبکه‌های عصبی

شبکه‌های عصبی دسته‌ای از مدل‌هایی هستند که با لایه‌بندی ساخته می‌شوند (ساختاری چند لایه دارند). شبکه‌های عصبی پیچشی (کانولوشنی (CNN)) و شبکه‌های عصبی بازگشتی (RNN) انواع رایج شبکه‌های عصبی هستند.

□ **معماری** – واژه معماری در شبکه‌های عصبی در شکل زیر توصیف شده است :



با نمایش  $i$  به عنوان لایه  $i$ ام و  $j$  به عنوان واحد  $j$ ام پنهان آن لایه، داریم :

$$z_j^{[i]} = w_j^{[i]T} x + b_j^{[i]}$$

که به ترتیب  $w, b, z$  و وزن، پیش‌قدر، و خروجی لایه هستند.

□ **تابع فعال‌سازی (activation function)** – توابع فعال‌سازی در انتهای واحد پنهان برای معرفی پیچیدگی غیر خطی به مدل استفاده می‌شوند. در اینجا رایج‌ترین آنها نمایش داده شده است :

Leaky ReLU	ReLU	Tanh	Sigmoid
$g(z) = \max(\epsilon z, z)$ $\epsilon \ll 1$ با	$g(z) = \max(0, z)$	$g(z) = \frac{e^z - e^{-z}}{e^z + e^{-z}}$	$g(z) = \frac{1}{1 + e^{-z}}$

□ **خطای آنتروپی متقاطع (cross-entropy loss)** – در مضمون شبکه‌های عصبی، عموماً از تابع خطای آنتروپی متقاطع  $L(z,y)$  استفاده می‌شود که به صورت زیر تعریف می‌شود :

$$L(z,y) = - \left[ y \log(z) + (1-y) \log(1-z) \right]$$

□ **نرخ یادگیری (learning rate)** – نرخ یادگیری اغلب با نماد  $\alpha$  و گاهی اوقات با نماد  $\eta$  نمایش داده می‌شود و بیانگر سرعت (گام) بروزرسانی وزن‌ها است که می‌تواند مقداری ثابت یا به سازگار شونده تغییر کند. محبوب‌ترین روش حال حاضر Adam نام دارد، متدی است که نرخ یادگیری را در حین فرآیند آموزش تنظیم می‌کند.



## ۳/۴ یادگیری تقویتی و کنترل

هدف یادگیری تقویتی برای یک عامل این است که یاد بگیرد در یک محیط چگونه تکامل یابد.

❑ **فرایندهای تصمیم‌گیری مارکوف (Markov Decision Processes)** – یک فرآیند تصمیم‌گیری مارکوف (به اختصار MDP) شامل پنج تایی  $(S, A, \{P_{sa}\}, \gamma, R)$  است به طوری که :

- $S$  مجموعه‌ای حالات است
- $A$  مجموعه‌ای از کنش‌ها است
- $\{P_{sa}\}$  احتمالات انتقال وضعیت برای هر  $s \in S$  و  $a \in A$  هستند.
- $\gamma \in [0, 1]$  ضریب تخفیف است.
- $R : S \times A \rightarrow \mathbb{R}$  یا  $R : S \rightarrow \mathbb{R}$  تابع پاداشی است که الگوریتم سعی دارد آن را بیشینه کند.

❑ **خطمشی (policy)** – یک خطمشی  $\pi$  تابعی است  $\pi : S \rightarrow A$  که حالات را به کنش‌ها نگاشت می‌کند.

نکته : می‌گوییم ما در حال اجرای خطمشی  $\pi$  هستیم اگر به ازای وضعیت  $s$  کنش  $a = \pi(s)$  را اجرا کنیم.

❑ **تابع ارزش (value function)** – برای سیاست  $\pi$  و وضعیت  $s$ ، تابع ارزش  $V^\pi$  را به صورت زیر تعریف می‌کنیم :

$$V^\pi(s) = E \left[ R(s_0) + \gamma R(s_1) + \gamma^2 R(s_2) + \dots | s_0 = s, \pi \right]$$

❑ **معادله‌ی بلمن (Bellman equation)** – معادله‌ی بلمن بهینه‌ی تابع ارزش  $V^{\pi^*}$  مربوط به خطمشی بهینه‌ی  $\pi^*$  مشخص می‌کند :

$$V^{\pi^*}(s) = R(s) + \max_{a \in A} \gamma \sum_{s' \in S} P_{sa}(s') V^{\pi^*}(s')$$

نکته : سیاست بهینه‌ی  $\pi^*$  برای وضعیت  $s$  این صورت است که :

$$\pi^*(s) = \operatorname{argmax}_{a \in A} \sum_{s' \in S} P_{sa}(s') V^*(s')$$

❑ **الگوریتم تکرار ارزش** – الگوریتم تکرار ارزش دو گام دارد :

- ارزش را مقداردهی اولیه می‌کنیم :

$$V_0(s) = 0$$

- ارزش را با توجه به ارزش‌های قبلی تکرار می‌کنیم :

$$V_{i+1}(s) = R(s) + \max_{a \in A} \left[ \sum_{s' \in S} \gamma P_{sa}(s') V_i(s') \right]$$

❑ **تخمین درست‌نمایی بیشینه** – تخمین‌های درست‌نمایی بیشینه برای احتمالات انتقال وضعیت به صورت زیر است :

$$P_{sa}(s') = \frac{\text{دفعاتی که کنش } a \text{ در وضعیت } s' \text{ رخ دهد}}{\text{دفعاتی که کنش } a \text{ در وضعیت } s \text{ اجرا شد.}}$$

❑ **یادگیری Q (Q-learning)** – یادگیری  $Q$  نوعی از یادگیری تقویتی بدون مدل برای تخمین  $Q$  است که به صورت زیر انجام می‌شود :

## ۴ نکات و ترندهای یادگیری ماشین

ترجمه به فارسی توسط الیستر و محمد رضا. بازبینی توسط عرفان نوری و محمد کریمی.

## ۴/۰/۱ معیارهای دسته‌بندی

معیارهای اساسی و مهم برای پیگیری در زمینه‌ی دسته‌بندی دوتایی و به منظور ارزیابی عملکرد مدل در زیر آمده‌اند.

❑ **ماتریس درهم‌ریختگی (confusion matrix)** – از ماتریس درهم‌ریختگی برای دست یافتن به تصویری جامع‌تر در ارزیابی عملکرد مدل استفاده می‌شود. این ماتریس بصورت زیر تعریف می‌شود:

دسته پیش‌بینی‌شده			
	-	+	
دسته واقعی	FN False Negatives Type II error	TP True Positives	+
	TN True Negatives	FP False Positives Type I error	-

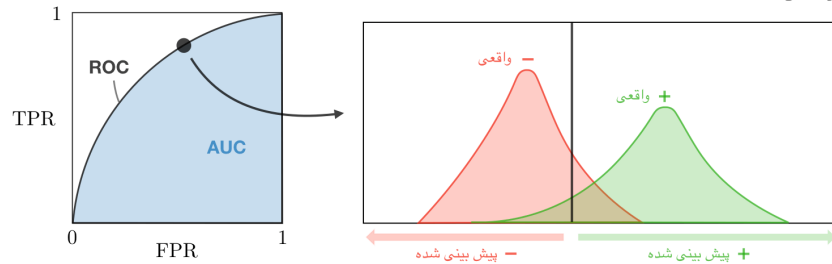
❑ **معیارهای اصلی** – معیارهای زیر معمولاً برای ارزیابی عملکرد مدل‌های دسته‌بندی بکار برده می‌شوند.

معیار	فرمول	تفسیر
صحت (Accuracy)	$\frac{TP + TN}{TP + TN + FP + FN}$	عملکرد کلی مدل
دقت (Precision)	$\frac{TP}{TP + FP}$	پیش‌بینی‌های مثبت چقدر دقیق هستند
فراخوانی (Recall)	$\frac{TP}{TP + FN}$	پوشش نمونه‌ی مثبت واقعی
ویژگی (Specificity)	$\frac{TN}{TN + FP}$	پوشش نمونه‌ی منفی واقعی
F1 score	$\frac{2TP}{2TP + FP + FN}$	معیار ترکیبی مفید برای دسته‌های نامتوازن

❑ **ROC** – منحنی عملیاتی گیرنده که تحت عنوان ROC نیز شناخته می‌شود تصویر TPR به ازای FPR و با تغییر مقادیر آستانه است. این معیارها بصورت خلاصه در جدول زیر آورده شده‌اند:

معیار	فرمول	معادل
True Positive Rate TPR	$\frac{TP}{TP + FN}$	فراخوانی
False Positive Rate FPR	$\frac{FP}{TN + FP}$	ویژگی-۱

❑ **AUC** – ناحیه‌ی زیر منحنی عملیاتی گیرنده، که با AUC یا AUROC نیز شناخته می‌شود، مساحت زیر منحنی ROC که در شکل زیر نشان داده شده است:



## ۴/۰/۲ معیارهای وایزش

❑ **معیارهای ابتدایی** – با توجه به مدل وایزش  $f$ ، معیارهای زیر برای ارزیابی عملکرد مدل مورد استفاده قرار می‌گیرند:

مجموع کل مربعات	مجموع مربعات توضیح داده شده	باقی‌مانده‌ی مجموع مربعات
$SS_{\text{tot}} = \sum_{i=1}^m (y_i - \bar{y})^2$	$SS_{\text{reg}} = \sum_{i=1}^m (f(x_i) - \bar{y})^2$	$SS_{\text{res}} = \sum_{i=1}^m (y_i - f(x_i))^2$

❑ **ضریب تعیین** – ضریب تعیین، که با  $R^2$  یا  $r^2$  هم نمایش داده می‌شود، معیاری برای سنجش این است که مدل به چه اندازه می‌تواند نتایج مشاهده‌شده را تکرار کند، و به صورت زیر تعریف می‌شود:

$$R^2 = 1 - \frac{SS_{\text{res}}}{SS_{\text{tot}}}$$

❑ **معیارهای اصلی** – از معیارهای زیر معمولاً برای ارزیابی عملکرد مدل‌های وایزش با در نظر گرفتن تعداد متغیرهای  $n$  که در نظر می‌گیرند، استفاده می‌شود:

Adjusted $R^2$	BIC	AIC	Mallow's Cp
$1 - \frac{(1 - R^2)(m - 1)}{m - n - 1}$	$\log(m)(n + 2) - 2 \log(L)$	$2[(n + 2) - \log(L)]$	$\frac{SS_{\text{res}} + 2(n + 1)\hat{\sigma}^2}{m}$

که  $L$  درست‌نمایی و  $\hat{\sigma}^2$  تخمینی از واریانس مربوط به هر یک از پاسخ‌ها است.

## ۴/۱ انتخاب مدل

□ **واژگان** – هنگام انتخاب مدل، سه بخش مختلف از داده‌ها را به صورت زیر مشخص می‌کنیم :

مجموعه آموزش (Training)	مجموعه اعتبارسنجی (Validation)	مجموعه آزمایش (Testing)
– مدل آموزش داده شده است – معمولاً ۸۰ درصد از مجموعه داده‌ها	– مدل ارزیابی شده است – معمولاً ۲۰ درصد از مجموعه داده‌ها – این مجموعه همچنین تحت عنوان مجموعه بیرون نگه‌داشته شده یا توسعه نیز شناخته می‌شود	– مدل پیش‌بینی می‌کند – داده‌های دیده نشده

بعد از اینکه مدل انتخاب شد، روی کل مجموعه داده‌ها آموزش داده می‌شود و بر روی مجموعه دادگان دیده نشده آزمایش می‌شود. این مراحل در شکل زیر آمده‌اند :



□ **اعتبارسنجی متقاطع (cross-validation)** – اعتبارسنجی متقاطع، که CV نیز نامیده می‌شود، عبارت است از روشی برای انتخاب مدلی که بیش از حد به مجموعه‌ی آموزش اولیه تکیه نمی‌کند. انواع مختلف بصورت خلاصه در جدول زیر ارائه شده‌اند :

## ۴/۲ عیب‌شناسی

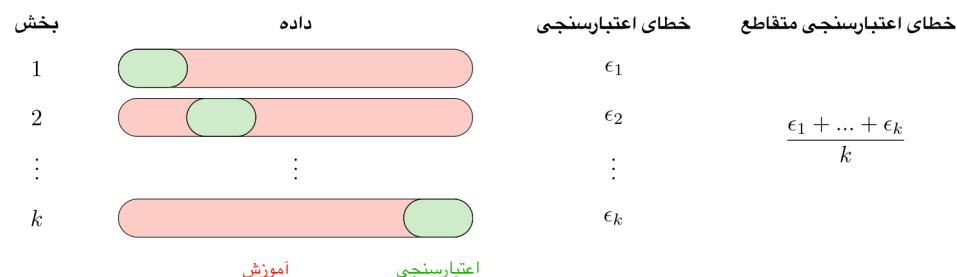
□ **پیش‌قدر (bias)** – پیش‌قدر مدل اختلاف بین پیش‌بینی مورد انتظار و مدل صحیح است که تلاش می‌کنیم برای نمونه داده‌های داده‌شده پیش‌بینی کنیم.

□ **واریانس (variance)** – واریانس یک مدل تنوع پیش‌بینی مدل برای نمونه داده‌های داده‌شده است.

□ **تعادل پیش‌قدر/واریانس** – هر چقدر مدل ساده‌تر باشد، پیش‌قدر بیشتر خواهد بود، و هر چه مدل پیچیده‌تر باشد واریانس بیشتر خواهد شد.

Leave- $p$ -out	$k$ -fold
– آموزش بر روی $n - p$ مشاهده و ارزیابی بر روی $p$ مشاهده‌ی باقی‌مانده – مورد $p = 1$ تحت عنوان حذف تک‌مورد گفته می‌شود	– آموزش بر روی $k - 1$ بخش دیگر و ارزیابی بر روی بخش باقی‌مانده – معمولاً $k = 5$ یا $10$

رایج‌ترین روش مورد استفاده، اعتبار سنجی متقاطع  $k$  – بخشی نامیده می‌شود که داده‌های آموزشی را به  $k$  بخش تقسیم می‌کند تا مدل روی یک بخش ارزیابی شود و در عین حال مدل را روی  $k - 1$  بخش دیگر آموزش دهد، و این عمل را  $k$  بار تکرار می‌کند. سپس میانگین خطا بر روی  $k$  بخش محاسبه می‌شود که خطای اعتبارسنجی متقاطع نامیده می‌شود.



□ **نظام‌بخشی (regularization)** – هدف از رویه‌ی نظام‌بخشی جلوگیری از بیش‌برازش به داده‌ها توسط مدل است و در نتیجه با مشکل واریانس بالا طرف است. جدول زیر خلاصه‌ای از انواع روش‌های متداول نظام‌بخشی را ارائه می‌دهد :

Elastic Net	Ridge	LASSO
بین انتخاب متغیر و ضرایب کوچک مصالحه می‌کند	ضرایب را کوچکتر می‌کند	– ضرایب را تا ۰ کاهش می‌دهد – برای انتخاب متغیر مناسب است
$\dots + \lambda \left[ (1 - \alpha) \ \theta\ _1 + \alpha \ \theta\ _2^2 \right]$ $\lambda \in \mathbb{R}, \alpha \in [0, 1]$	$\dots + \lambda \ \theta\ _2^2$ $\lambda \in \mathbb{R}$	$\dots + \lambda \ \theta\ _1$ $\lambda \in \mathbb{R}$

Overfitting	Just right	Underfitting	علامت
– خطای آموزش بسیار کم – خطای آموزش بسیار کمتر از خطای آزمایش – واریانس بالا	– خطای آموزش کمی کمتر از خطای آزمایش	– خطای بالای آموزش – خطای آموزش نزدیک به خطای آزمایش – پیش‌قدر زیاد	
			نمایش وایازش

## ۵ آمار و احتمالات

ترجمه به فارسی توسط عرفان نوری. بازبینی توسط محمد کریمی.

## ۵/۱ مقدمه‌ای بر احتمالات و ترکیبیات

□ **فضای نمونه** – مجموعه‌ی همگی پیشامدهای یک آزمایش را فضای نمونه‌ی آن آزمایش گویند که با  $S$  نمایش داده می‌شود.

□ **رخداد** – هر زیرمجموعه‌ی  $E$  از فضای نمونه یک رخداد در نظر گرفته می‌شود. به عبارت دیگر، یک رخداد مجموعه‌ای از پیشامدهای یک آزمایش است. اگر پیشامد یک آزمایش عضوی از مجموعه‌ی  $E$  باشد، در این حالت می‌گوییم که رخداد  $E$  اتفاق افتاده است.

□ **اصول موضوعی احتمالات** – برای هر رخداد  $E$ ،  $P(E)$  احتمال اتفاق افتادن رخداد  $E$  می‌باشد.

$$(1) \quad 0 \leq P(E) \leq 1 \quad (2) \quad P(S) = 1 \quad (3) \quad P\left(\bigcup_{i=1}^n E_i\right) = \sum_{i=1}^n P(E_i)$$

□ **جایگشت** – یک جایگشت چیدمانی از  $r$  شی از  $n$  شی با یک ترتیب خاص است. تعداد این چنین جایگشت‌ها  $P(n, r)$  است که به صورت زیر تعریف می‌شود:

$$P(n, r) = \frac{n!}{(n-r)!}$$

□ **ترکیب** – یک ترکیب چیدمانی از  $r$  شی از  $n$  شی است، به طوری که ترتیب اهمیتی نداشته باشد. تعداد این چنین ترکیب‌ها  $C(n, r)$  است که به صورت زیر تعریف می‌شود:

$$C(n, r) = \frac{P(n, r)}{r!} = \frac{n!}{r!(n-r)!}$$

نکته: برای  $n$ ،  $0 \leq r \leq n$  داریم  $P(n, r) \geq C(n, r)$ .

## ۵/۲ احتمال شرطی

□ **قضیه‌ی بیز** – برای رخداد‌های  $A$  و  $B$  به طوری که  $P(B) > 0$  داریم:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

نکته: داریم  $P(A \cap B) = P(A)P(B|A) = P(A|B)P(B)$ .

□ **افراز** – فرض می‌کنیم برای  $\{A_i, i \in [1, n]\}$  به ازای هر  $i$  داشته باشیم  $A_i \neq \emptyset$ . در این صورت می‌گوییم  $\{A_i\}$  یک افراز است اگر:

$$\forall i \neq j, A_i \cap A_j = \emptyset \quad \text{و} \quad \bigcup_{i=1}^n A_i = S$$

نکته: برای هر رخداد  $B$  در فضای نمونه داریم  $P(B) = \sum_{i=1}^n P(B|A_i)P(A_i)$ .

□ **تعمیم قضیه‌ی بیز** – فرض می‌کنیم  $\{A_i, i \in [1, n]\}$  یک افراز از فضای نمونه باشید. در این صورت داریم:

			نمایش دسته‌بندی
			نمایش یادگیری عمیق
– نظام‌بخشی انجام دهید – داده‌های بیشتری – گردآوری کنید		– مدل را پیچیده‌تر کنید – ویژگی‌های بیشتری اضافه کنید – مدت طولانی‌تری آموزش دهید	اصلاحات احتمالی

□ **تحلیل خطا (error analysis)** – تحلیل خطا به بررسی علت اصلی اختلاف در عملکرد بین مدل‌های کنونی و مدل‌های صحیح می‌پردازد.

□ **تحلیل تقطیعی (ablative analysis)** – تحلیل تقطیعی به بررسی علت اصلی اختلاف بین مدل‌های کنونی و مدل‌های پایه می‌پردازد.

حالت	$E[X]$	$E[g(X)]$	$E[X^k]$	$\psi(\omega)$
(D)	$\sum_{i=1}^n x_i f(x_i)$	$\sum_{i=1}^n g(x_i) f(x_i)$	$\sum_{i=1}^n x_i^k f(x_i)$	$\sum_{i=1}^n f(x_i) e^{i\omega x_i}$
(C)	$\int_{-\infty}^{+\infty} x f(x) dx$	$\int_{-\infty}^{+\infty} g(x) f(x) dx$	$\int_{-\infty}^{+\infty} x^k f(x) dx$	$\int_{-\infty}^{+\infty} f(x) e^{i\omega x} dx$

نکته : داریم  $e^{i\omega x} = \cos(\omega x) + i \sin(\omega x)$ .

□ **تبدیلات متغیرهای تصادفی** – فرض کنید متغیرهای تصادفی  $X$  و  $Y$  توسط تابعی به هم مرتبط هستند. با نمایش تابع توزیع متغیرهای تصادفی  $X$  و  $Y$  با  $f_X$  و  $f_Y$  داریم :

$$f_Y(y) = f_X(x) \left| \frac{dx}{dy} \right|$$

□ **قضیه انتگرال لایبنیتس** – فرض کنید  $g$  تابعی از  $x$  و  $c$  باشد، و  $a$  و  $b$  کرانهایی باشند که مقدار آن‌ها وابسته به مقدار  $c$  باشد. داریم :

$$\frac{\partial}{\partial c} \left( \int_a^b g(x) dx \right) = \frac{\partial b}{\partial c} \cdot g(b) - \frac{\partial a}{\partial c} \cdot g(a) + \int_a^b \frac{\partial g}{\partial c}(x) dx$$

□ **نابرابری چبیشف** – فرض کنید  $X$  متغیری تصادفی با امید ریاضی  $\mu$  و انحراف معیار  $\sigma$ . برای هر  $k > 0$  نابرابری زیر را داریم :

$$P(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}$$

## ۵/۴ متغیرهای تصادفی با توزیع مشترک

□ **چگالی شرطی** – چگالی شرطی  $X$  نسبت به  $Y$ ، که معمولاً با  $f_{X|Y}$  نمایش داده می‌شود، به صورت زیر تعریف می‌شود :

$$f_{X|Y}(x) = \frac{f_{XY}(x,y)}{f_Y(y)}$$

□ **استقلال** – دو متغیر تصادفی  $X$  و  $Y$  مستقل هستند اگر داشته باشیم :

$$f_{XY}(x,y) = f_X(x)f_Y(y)$$

□ **چگالی حاشیه‌ای و توزیع تجمعی** – از تابع چگالی احتمالی مشترک  $f_{XY}$  داریم :

حالت	چگالی حاشیه‌ای	تابع تجمعی
(D)	$f_X(x_i) = \sum_j f_{XY}(x_i, y_j)$	$F_{XY}(x, y) = \sum_{x_i \leq x} \sum_{y_j \leq y} f_{XY}(x_i, y_j)$
(C)	$f_X(x) = \int_{-\infty}^{+\infty} f_{XY}(x, y) dy$	$F_{XY}(x, y) = \int_{-\infty}^x \int_{-\infty}^y f_{XY}(x', y') dx' dy'$

$$P(A_k|B) = \frac{P(B|A_k)P(A_k)}{\sum_{i=1}^n P(B|A_i)P(A_i)}$$

□ **استقلال** – دو رخداد  $A$  و  $B$  مستقل هستند اگر و فقط اگر داشته باشیم :

$$P(A \cap B) = P(A)P(B)$$

## ۵/۳ متغیرهای تصادفی

□ **متغیر تصادفی** – یک متغیر تصادفی، که معمولاً با  $X$  نمایش داده می‌شود، یک تابع است که هر عضو فضای نمونه را به اعداد حقیقی نگاشت می‌کند.

□ **تابع توزیع تجمعی (CDF)** – تابع توزیع تجمعی  $F$ ، که تابعی یکنوا و اکیدا غیرنزولی است و برای آن  $\lim_{x \rightarrow -\infty} F(x) = 0$  و

$\lim_{x \rightarrow +\infty} F(x) = 1$  صدق می‌کنید، به صورت زیر تعریف می‌شود :

$$F(x) = P(X \leq x)$$

نکته : داریم  $P(a < X \leq b) = F(b) - F(a)$ .

□ **تابع چگالی احتمال (PDF)** – تابع چگالی احتمال  $f$  احتمال آن است که متغیر تصادفی  $X$  مقداری بین دو تحقق همجوار این متغیر تصادفی را بگیرد.

□ **ارتباط بین PDF و CDF** – موارد زیر ویژگی‌های مهمی هستند که باید در مورد حالت گسسته و حالت پیوسته در نظر گرفت.

حالت	$F$ CDF	$f$ PDF	ویژگی‌های PDF
(D)	$F(x) = \sum_{x_i \leq x} P(X = x_i)$	$f(x_j) = P(X = x_j)$	$0 \leq f(x_j) \leq 1$ و $\sum_j f(x_j) = 1$
(C)	$F(x) = \int_{-\infty}^x f(y) dy$	$f(x) = \frac{dF}{dx}$	$f(x) \geq 0$ و $\int_{-\infty}^{+\infty} f(x) dx = 1$

□ **واریانس** – واریانس یک متغیر تصادفی، که معمولاً با  $\text{Var}(X)$  یا  $\sigma^2$  نمایش داده می‌شود، میزانی از پراکندگی یک تابع توزیع است. مقدار واریانس به صورت زیر به دست می‌آید :

$$\text{Var}(X) = E[(X - E[X])^2] = E[X^2] - E[X]^2$$

□ **انحراف معیار** – انحراف معیار یک متغیر تصادفی، که با  $\sigma$  نمایش داده می‌شود، میزانی از پراکندگی یک تابع توزیع است که با متغیر تصادفی هم‌واحد است. مقدار آن به صورت زیر به دست می‌آید :

$$\sigma = \sqrt{\text{Var}(X)}$$

□ **امید ریاضی و گشتاورهای یک توزیع** – عبارت‌های مربوط به امید ریاضی  $E[X]$ ، امید ریاضی تعمیم یافته  $E[g(X)]$ ،  $k$  – مین گشتاور  $E[X^k]$ ، و تابع ویژگی  $\psi(\omega)$  برای حالات پیوسته و گسسته به صورت زیر هستند :

□ **میانگین و واریانس نمونه** – میانگین نمونه‌ی یک نمونه‌ی تصادفی که برای تخمین مقدار واقعی میانگین  $\mu$  یک توزیع به کار می‌رود، معمولاً با  $\bar{X}$  نمایش داده می‌شود. واریانس نمونه‌ی یک نمونه‌ی تصادفی که برای تخمین مقدار واقعی واریانس  $\sigma^2$  یک توزیع به کار می‌رود، معمولاً با  $s^2$  یا  $\hat{\sigma}^2$  نمایش داده می‌شود و به صورت زیر تعریف می‌شود:

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i \quad \text{و} \quad s^2 = \hat{\sigma}^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$$

□ **قضیه‌ی حد مرکزی** – یک نمونه‌ی تصادفی  $X_1, \dots, X_n$  که از یک توزیع با میانگین  $\mu$  و واریانس  $\sigma^2$  به دست آمده‌اند را در نظر بگیرید؛ داریم:

$$\bar{X} \underset{n \rightarrow +\infty}{\sim} \mathcal{N}\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

□ **کواریانس** – کواریانس دو متغیر تصادفی  $X$  و  $Y$  که با  $\sigma_{XY}^2$  یا به صورت معمول‌تر با  $\text{Cov}(X, Y)$  نمایش داده می‌شود، به صورت زیر است:

$$\text{Cov}(X, Y) \triangleq \sigma_{XY}^2 = E[(X - \mu_X)(Y - \mu_Y)] = E[XY] - \mu_X \mu_Y$$

□ **همبستگی** – با نمایش انحراف معیار  $X$  و  $Y$  به صورت  $\sigma_X$  و  $\sigma_Y$ ، همبستگی مابین دو متغیر تصادفی  $X$  و  $Y$  که با  $\rho_{XY}$  نمایش داده می‌شود به صورت زیر تعریف می‌شود:

$$\rho_{XY} = \frac{\sigma_{XY}^2}{\sigma_X \sigma_Y}$$

نکته‌ی: برای هر دو متغیر تصادفی دلخواه  $X$  و  $Y$ ، داریم  $\rho_{XY} \in [-1, 1]$ . اگر  $X$  و  $Y$  مستقل باشند، داریم  $\rho_{XY} = 0$ .

□ **توزیع‌های احتمالی اصلی** – توزیع‌های زیر توزیع‌های احتمالی اصلی هستند که بهتر است به خاطر بسپارید:

نوع	توزیع	PDF	$\psi(\omega)$	$E[X]$	$\text{Var}(X)$
(D)	$X \sim \mathcal{B}(n, p)$ Binomial	$P(X = x) = \binom{n}{x} p^x q^{n-x}$ $x \in \llbracket 0, n \rrbracket$	$(pe^{i\omega} + q)^n$	$np$	$npq$
	$X \sim \text{Po}(\mu)$ Poisson	$P(X = x) = \frac{\mu^x}{x!} e^{-\mu}$ $x \in \mathbb{N}$	$e^{\mu(e^{i\omega} - 1)}$	$\mu$	$\mu$
(C)	$X \sim \mathcal{U}(a, b)$ Uniform	$f(x) = \frac{1}{b-a}$ $x \in [a, b]$	$\frac{e^{i\omega b} - e^{i\omega a}}{(b-a)i\omega}$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
	$X \sim \mathcal{N}(\mu, \sigma)$ Gaussian	$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$ $x \in \mathbb{R}$	$e^{i\omega\mu - \frac{1}{2}\omega^2\sigma^2}$	$\mu$	$\sigma^2$
	$X \sim \text{Exp}(\lambda)$ Exponential	$f(x) = \lambda e^{-\lambda x}$ $x \in \mathbb{R}_+$	$\frac{1}{1 - i\omega/\lambda}$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$

## ۵/۵ تخمین پارامتر

□ **نمونه‌ی تصادفی** – یک نمونه‌ی تصادفی مجموعه‌ای از  $n$  متغیر تصادفی  $X_1, \dots, X_n$  است که از هم مستقل هستند و توزیع یکسانی با  $X$  دارند.

□ **تخمین‌گر** – یک تخمین‌گر  $\hat{\theta}$  تابعی از داده‌ها است که برای به‌دست‌آوردن مقدار نامشخص یک پارامتر در یک مدل  $\theta$  آماری به کار می‌رود.

□ **پیش‌قدر** – پیش‌قدر یک تخمین‌گر  $\hat{\theta}$  به عنوان اختلاف بین امید ریاضی توزیع  $\hat{\theta}$  و مقدار واقعی تعریف می‌شود. یعنی:

$$\text{Bias}(\hat{\theta}) = E[\hat{\theta}] - \theta$$

نکته: یک تخمین‌گر بدون پیش‌قدر است اگر داشته باشیم  $E[\hat{\theta}] = \theta$ .

## ۶ جبر خطی و حسابان

• ضرب خارجی : برای هر  $x \in \mathbb{R}^m$  و  $y \in \mathbb{R}^n$  داریم :

$$xy^T = \begin{pmatrix} x_1 y_1 & \cdots & x_1 y_n \\ \vdots & & \vdots \\ x_m y_1 & \cdots & x_m y_n \end{pmatrix} \in \mathbb{R}^{m \times n}$$

ترجمه به فارسی توسط عرفان نوری. بازبینی توسط محمد کریمی.

## ۶/۱ نمادها

□ **بردار** –  $x \in \mathbb{R}^n$  یک بردار با  $n$  درایه است، که  $x_i \in \mathbb{R}$  درایه  $i$  ام می باشد :

$$x = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \in \mathbb{R}^n$$

□ **ماتریس** –  $A \in \mathbb{R}^{m \times n}$  یک بردار با  $m$  سطر و  $n$  ستون است، که در آن  $A_{i,j} \in \mathbb{R}$  درایه های است که در سطر  $i$  ام و ستون  $j$  ام قرار دارد :

$$A = \begin{pmatrix} A_{1,1} & \cdots & A_{1,n} \\ \vdots & & \vdots \\ A_{m,1} & \cdots & A_{m,n} \end{pmatrix} \in \mathbb{R}^{m \times n}$$

نکته : بردار  $x$  که در بالا تعریف شد را می توان به صورت یک ماتریس  $1 \times n$  در نظر گرفت که به طور خاص به آن بردار ستونی گویند.□ **ماتریس همانی** – ماتریس همانی  $I \in \mathbb{R}^{n \times n}$  یک ماتریس مربعی است که درایه های قطری آن همه مقدار ۱ و بقیه درایه ها مقدار ۰ دارند :

$$I = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & \ddots & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & 1 \end{pmatrix}$$

نکته : برای همه ی ماتریس های  $A \in \mathbb{R}^{n \times n}$  داریم  $A \times I = I \times A = A$ .□ **ماتریس قطری** – ماتریس  $D \in \mathbb{R}^{n \times n}$  یک ماتریس مربعی است که درایه های قطری آن مقادیر غیرصفر دارند و بقیه درایه ها صفر هستند :

$$D = \begin{pmatrix} d_1 & 0 & \cdots & 0 \\ 0 & \ddots & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & d_n \end{pmatrix}$$

نکته :  $D$  همچنین به صورت  $diag(d_1, \dots, d_n)$  هم نمایش داده می شود.

## ۶/۲ عملیات ماتریسی

## ۶/۲/۱ ضرب

□ **بردار با بردار** – دو نوع عملیات ضرب بردار با بردار وجود دارد :• ضرب داخلی : برای هر  $x, y \in \mathbb{R}^n$  داریم :

$$x^T y = \sum_{i=1}^n x_i y_i \in \mathbb{R}$$

□ **ماتریس با بردار** – ضرب ماتریس  $A \in \mathbb{R}^{m \times n}$  و بردار  $x \in \mathbb{R}^n$  با اندازه ی  $m$  است به طوری که :

$$Ax = \begin{pmatrix} a_{r,1}^T x \\ \vdots \\ a_{r,m}^T x \end{pmatrix} = \sum_{i=1}^n a_{c,i} x_i \in \mathbb{R}^m$$

که  $a_{r,i}^T$  بردارهای سطری و  $a_{c,j}$  بردارهای ستونی  $A$ ، و  $x_i$  درایه های  $x$  هستند.□ **ماتریس با ماتریس** – ضرب ماتریس های  $A \in \mathbb{R}^{m \times n}$  و  $B \in \mathbb{R}^{n \times p}$  ماتریسی با اندازه ی  $n \times p$  است که :

$$AB = \begin{pmatrix} a_{r,1}^T b_{c,1} & \cdots & a_{r,1}^T b_{c,p} \\ \vdots & & \vdots \\ a_{r,m}^T b_{c,1} & \cdots & a_{r,m}^T b_{c,p} \end{pmatrix} = \sum_{i=1}^n a_{c,i} b_{r,i}^T \in \mathbb{R}^{n \times p}$$

که  $a_{r,i}^T, b_{r,i}^T$  بردارهای سطری و  $a_{c,j}$  و  $b_{c,j}$  بردارهای ستونی  $A$  و  $B$  هستند.

## ۶/۲/۲ دیگر عملیات

□ **ترانزاده** – ترانزاده ی ماتریس  $A \in \mathbb{R}^{m \times n}$  که با  $A^T$  نمایش داده می شود، ماتریسی است که مکان درایه های آن نسبت به قطر ماتریس برعکس شده اند :

$$\forall i, j, \quad A_{i,j}^T = A_{j,i}$$

نکته : برای ماتریس های  $A$  و  $B$ ، داریم  $(AB)^T = B^T A^T$ .□ **معکوس** – معکوس یک ماتریس مربعی معکوس پذیر  $A$  که با  $A^{-1}$  نمایش داده می شود، تنها ماتریسی است که :

$$AA^{-1} = A^{-1}A = I$$

نکته : همه ی ماتریس های مربعی معکوس پذیر نیستند. همچنین، برای ماتریس های مربعی معکوس پذیر  $A$  و  $B$ ، داریم  $(AB)^{-1} = B^{-1}A^{-1}$ .□ **اثر** – اثر ماتریس مربعی  $A$  که با  $\text{tr}(A)$  نمایش داده می شود، مجموع همه ی درایه های قطری ماتریس است.

$$\text{tr}(A) = \sum_{i=1}^n A_{i,i}$$

نکته : برای ماتریس های  $A$  و  $B$  داریم  $\text{tr}(A) = \text{tr}(A^T)$  و  $\text{tr}(AB) = \text{tr}(BA)$

❑ **ماتریس مثبت نیمه‌معین** – ماتریس  $A \in \mathbb{R}^{n \times n}$  یک ماتریس مثبت نیمه‌معین است که با  $A \succeq 0$  نمایش داده می‌شود اگر داشته باشیم:

$$A = A^T \quad \text{و} \quad \forall x \in \mathbb{R}^n, \quad x^T A x \geq 0$$

نکته: به طور مشابه، یک ماتریس  $A$  مثبت معین است ( $A \succ 0$ )، اگر یک ماتریس مثبت نیمه‌معین باشد که برای هر بردار غیرصفر  $x$  داشته باشیم  $x^T A x > 0$ .

❑ **مقدار ویژه، بردار ویژه** – برای یک ماتریس  $A \in \mathbb{R}^{n \times n}$ ، گوییم  $\lambda$  یک مقدار ویژه ماتریس  $A$  است اگر وجود داشته باشد بردار  $z \in \mathbb{R}^n \setminus \{0\}$ ، که یک بردار ویژه نام دارد، به طوری که:

$$Az = \lambda z$$

### ۶/۳/۲ قضیه

❑ **قضیه طیفی** – فرض کنید  $A \in \mathbb{R}^{n \times n}$  باشد. اگر  $A$  متقارن باشد، در این صورت  $A$  توسط یک ماتریس حقیقی متعامد  $U \in \mathbb{R}^{n \times n}$  قطری‌پذیر است. با نمایش  $\Lambda = \text{diag}(\lambda_1, \dots, \lambda_n)$  داریم:

$$\exists \Lambda \text{ diagonal, } A = U \Lambda U^T$$

❑ **تجزیه مقدار منفرد** – برای یک ماتریس  $A$  با ابعاد  $m \times n$ ، تجزیه مقدار منفرد یک تکنیک تقسیم‌بندی است که تضمین می‌کند یک ماتریس یکانی  $U \in \mathbb{R}^{m \times m}$ ، یک ماتریس قطری  $\Sigma \in \mathbb{R}^{m \times n}$ ، و یک ماتریس یکانی  $V \in \mathbb{R}^{n \times n}$  وجود دارند، به طوری که:

$$A = U \Sigma V^T$$

### ۶/۴ حسابان ماتریسی

❑ **گرادیان** – فرض کنید  $f: \mathbb{R}^{m \times n} \rightarrow \mathbb{R}$  یک تابع و  $A \in \mathbb{R}^{m \times n}$  یک ماتریس باشد. گرادیان  $f$  نسبت به  $A$  یک ماتریس با ابعاد  $m \times n$  است و با  $\nabla_A f(A)$  نمایش داده می‌شود، به طوری که:

$$\left( \nabla_A f(A) \right)_{i,j} = \frac{\partial f(A)}{\partial A_{i,j}}$$

نکته: گرادیان  $f$  تنها زمانی تعریف شده است که  $f$  تابعی باشد که یک عدد اسکالر خروجی دهد.

❑ **هسیان** – فرض کنید  $f: \mathbb{R}^n \rightarrow \mathbb{R}$  یک تابع و  $x \in \mathbb{R}^n$  یک بردار باشد. هسیان  $f$  نسبت به  $x$  یک ماتریس متقارن با ابعاد  $n \times n$  است و با  $\nabla_x^2 f(x)$  نمایش داده می‌شود، به طوری که:

$$\left( \nabla_x^2 f(x) \right)_{i,j} = \frac{\partial^2 f(x)}{\partial x_i \partial x_j}$$

نکته: هسیان تابع  $f$  تنها زمانی تعریف شده است که  $f$  تابعی با خروجی اسکالر باشد.

❑ **عملیات گرادیانی** – برای ماتریس‌های  $A$ ،  $B$  و  $C$ ، ویژگی‌های زیر را به خاطر داشته باشید:

$$\nabla_A \text{tr}(AB) = B^T$$

$$\nabla_A \text{tr} f(A) = (\nabla_A f(A))^T$$

$$\nabla_A \text{tr}(ABA^T C) = CAB + C^T A B^T$$

$$\nabla_A |A| = |A| (A^{-1})^T$$

❑ **دترمینان** – دترمینان یک ماتریس مربعی  $A \in \mathbb{R}^{n \times n}$  که با  $|A|$  یا  $\det(A)$  نمایش داده می‌شود، به صورت یک عبارت بازگشتی بر روی  $A_{i,j}$ ، که ماتریس  $A$  بدون سطر  $i$ –ام و ستون  $j$ –ام است، به صورت زیر تعریف می‌شود:

$$\det(A) = |A| = \sum_{j=1}^n (-1)^{i+j} A_{i,j} |A_{i,j}|$$

نکته:  $A$  معکوس‌پذیر است اگر و فقط اگر  $|A| \neq 0$ . همچنین  $|AB| = |A||B|$  و  $|A^T| = |A|$ .

### ۶/۳ ویژگی‌های ماتریس‌ها

#### ۶/۳/۱ تعاریف

❑ **تجزیه متقارن** – یک ماتریس دلخواه  $A$  را می‌توان با استفاده از اجزای متقارن و غیرمتقارن آن به صورت زیر نشان داد:

$$A = \underbrace{\frac{A + A^T}{2}}_{\text{Symmetric}} + \underbrace{\frac{A - A^T}{2}}_{\text{Antisymmetric}}$$

❑ **نرم** – نرم تابع  $N: V \rightarrow [0, +\infty]$  است که  $V$  یک فضای برداری است، و به گونه‌ای است که برای هر  $x, y \in V$  داریم:

$$N(x + y) \leq N(x) + N(y)$$

$$N(ax) = |a| N(x) \text{ برای عدد اسکالر } a$$

$$\text{اگر } N(x) = 0 \text{ باشد در این صورت } x = 0$$

برای  $x \in V$ ، نرم‌هایی که بیشتر استفاده می‌شوند در جدول زیر آمده‌اند:

نرم	نماد	تعریف	کاربرد
Manhattan, $L^1$	$\ x\ _1$	$\sum_{i=1}^n  x_i $	LASSO
Euclidean, $L^2$	$\ x\ _2$	$\sqrt{\sum_{i=1}^n x_i^2}$	Ridge
$p$ -norm, $L^p$	$\ x\ _p$	$\left( \sum_{i=1}^n x_i^p \right)^{\frac{1}{p}}$	Hölder inequality
Infinity, $L^\infty$	$\ x\ _\infty$	$\max_i  x_i $	Uniform convergence

❑ **وابستگی خطی** – مجموعه‌ای از بردارها وابستگی خطی دارند اگر یکی از بردارهای مجموعه را بتوان به صورت ترکیب خطی دیگر بردارها تعریف کرد.

نکته: اگر نتوان هیچ برداری را به این شکل تعریف کرد، در این صورت بردارها استقلال خطی دارند.

❑ **رتبه ماتریس** – رتبه یک ماتریس  $A$  که با  $\text{rank}(A)$  نمایش داده می‌شود، تعداد ابعاد فضایی است که توسط ستون‌های آن ایجاد می‌شود. این مقدار برابر است با حداکثر تعداد ستون‌های  $A$  که استقلال خطی داشته باشند.