(https://stanford.edu/%7Eshervine/teaching/cs-229.html)

راهنمای کوتاه یادگیری بدون نظارت

مبانی یادگیری بدون نظارت

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

نابرابری ینسن - فرض کنید f تابعی محدب و X یک متغیر تصادفی باشد. در این صورت نابرابری زیر را داریم:

$$\boxed{E[f(X)] \geqslant f(E[X])}$$

خوشه‌بندی

بیشینه‌سازی امید ریاضی

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

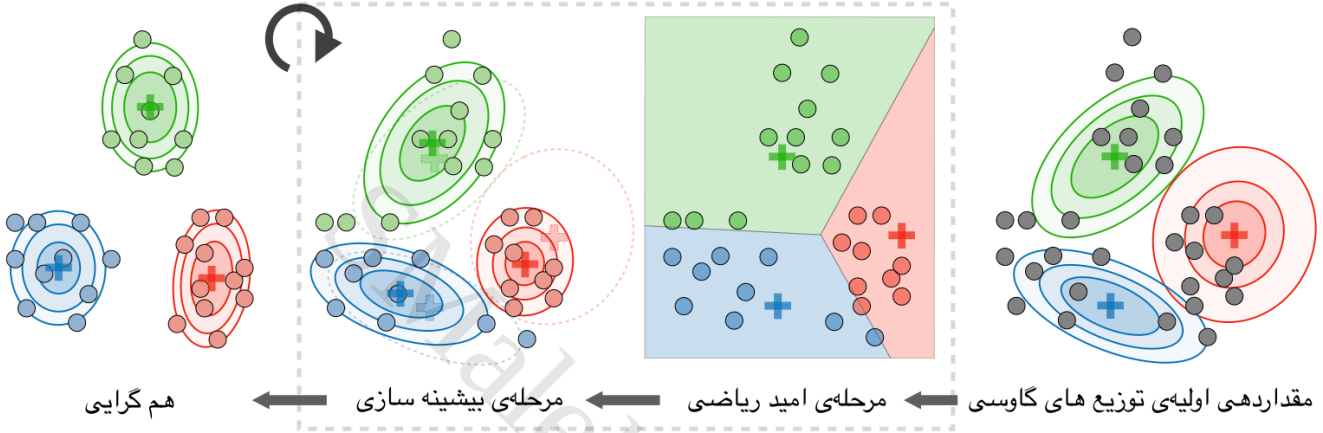
الگوریتم - الگوریتم بیشینه‌سازی امید ریاضی روشی بهینه برای تخمین پارامتر θ از طریق تخمین درستی بشینه در اختیار قرار می‌دهد. این کار با تکرار مرحله‌ی به دست آوردن یک کران پایین برای درستی (مرحله‌ی امید ریاضی) و همچنین بهینه‌سازی آن کران پایین (مرحله‌ی بیشینه‌سازی) طبق توضیح زیر انجام می‌شود:

• مرحله‌ی 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:

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

• مرحله‌ی بیشینه‌سازی: 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 می‌شوند:

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



خوشه‌بندی k-میانگین

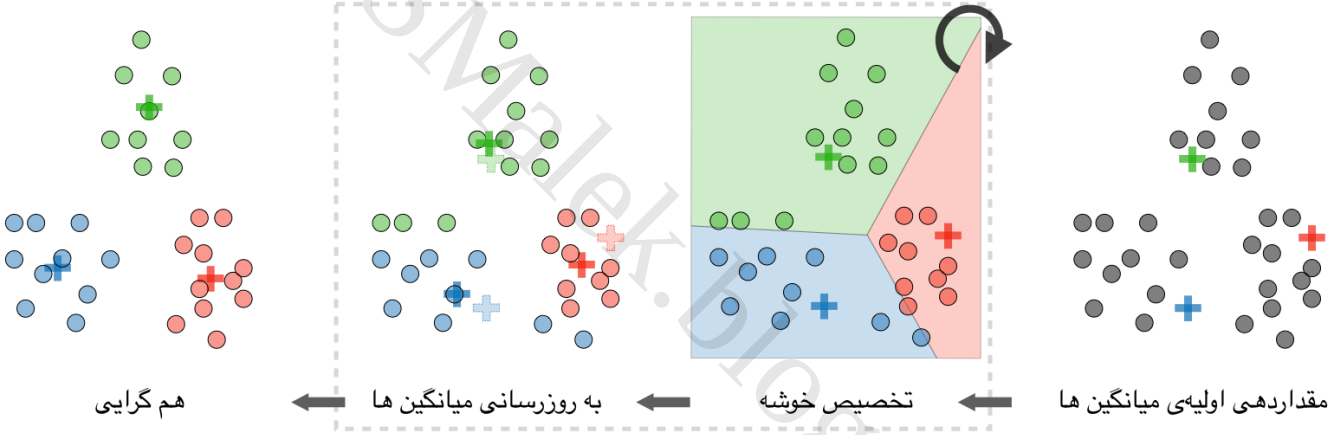
توجه کنید که $c^{(i)}$ خوشه‌ی نمونه داده‌ی i و μ_j مرکز خوشه‌ی j است.

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

$$c^{(i)} = \arg\min_j ||x^{(i)} - \mu_j||^2$$

و

$$\mu_j = \frac{\sum_{i=1}^{m} 1_{\{c^{(i)}=j\}} x^{(i)}}{\sum_{i=1}^{m} 1_{\{c^{(i)}=j\}}}$$



تابع اعوجاج - برای تشخیص اینکه الگوریتم به هم‌گرایی رسیده است، به تابع اعوجاج که به صورت زیر تعریف می‌شود رجوع می‌کنیم:

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

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 کامل	پیوند میانگین	پیوند بخشی
کمینه‌کردن حداکثر فاصله بین هر دو جفت خوشه	کمینه‌کردن فاصله‌ی میانگین بین هر دو جفت خوشه	کمینه‌کردن فاصله‌ی درونِ خوشه

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 - 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:

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

شاخص **Calinski-Harabasz** - 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:

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

شاخص Calinski-Harabasz $s(k)$ 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:

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

کاهش ابعاد

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 - برای ماتریس دلخواه $A \in \mathbb{R}^{n \times n}$، λ مقدار ویژه‌ی ماتریس A است اگر وجود داشته باشد بردار $z \in \mathbb{R}^n \backslash \{0\}$ که به آن بردار ویژه می‌گویند، به طوری که:

$$\boxed{Az = \lambda z}$$

قضیه‌ی 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 $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_n)$ داریم:

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

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 بیشینه کردن واریانس داده‌ها، به صورت زیر تصویر می‌کند:

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

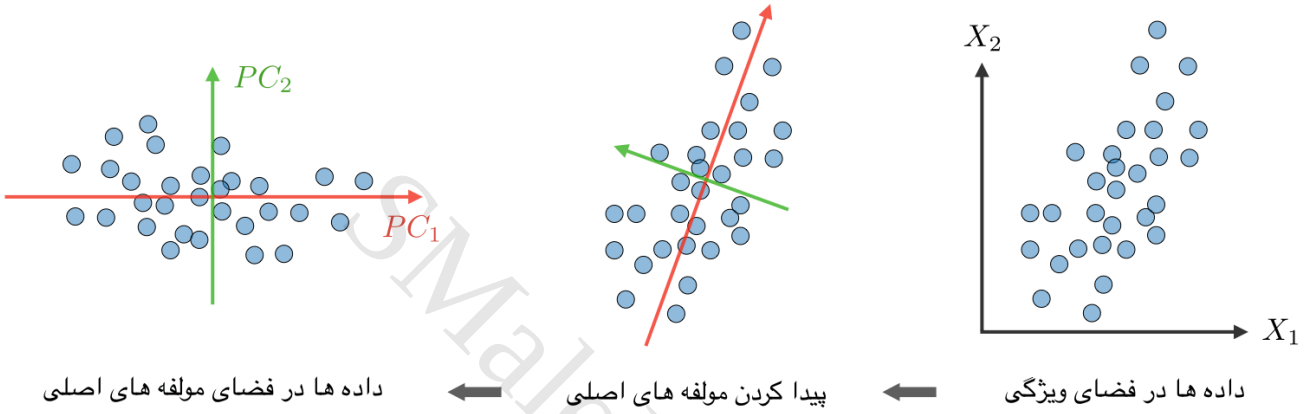
$$\boxed{x_j^{(i)} \leftarrow \frac{x_j^{(i)} - \mu_j}{\sigma_j}} \quad \text{و} \quad \boxed{\mu_j = \frac{1}{m}\sum_{i=1}^{m} x_j^{(i)}} \quad \text{و} \quad \boxed{\sigma_j^2 = \frac{1}{m}\sum_{i=1}^{m}(x_j^{(i)} - \mu_j)^2}$$

• مرحله‌ی 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.

• مرحله‌ی ۳: بردارهای $u_1, \ldots, u_k \in \mathbb{R}^n$ که k بردارهای ویژه‌ی اصلی متعامد Σ هستند محاسبه می‌شوند. این بردارهای ویژه متناظر با k مقدار ویژه با بزرگ‌ترین مقدار هستند.

• مرحله‌ی ۴: داده‌ها بر روی فضای $\text{span}_{\mathbb{R}}(u_1, \ldots, u_k)$ تصویر می‌شوند.

این رویه واریانس را در فضای k-بعدی به دست آمده بیشینه می‌کند.



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 $s = (s_1, \ldots, s_n)$ تولید شده است، که s_iها 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:

$$\boxed{x = As}$$

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 با 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 گرادیان تصادفی افزایشی برای هر نمونه از داده‌های آموزش $x^{(i)}$ به گونه‌ای است که برای به‌روزرسانی W داریم:

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