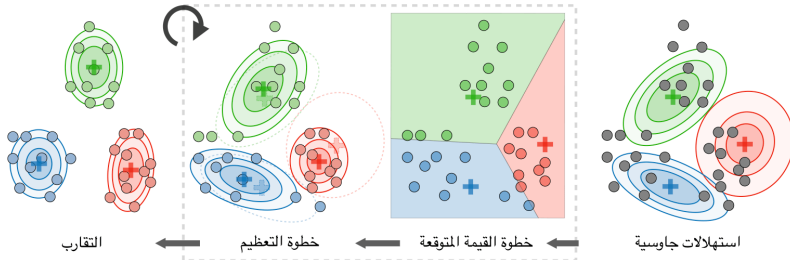


$$\theta_i = \operatorname{argmax}_{\theta} \sum_i \int_{z^{(i)}} Q_i(z^{(i)}) \log \left(\frac{P(x^{(i)}, z^{(i)}; \theta)}{Q_i(z^{(i)})} \right) dz^{(i)}$$



مرجع سريع للتعلم غير المُؤجَّه

افشين عميدى و شروين عميدى

١٤ ربيع الثاني، ١٤٤١

تمت الترجمة بواسطة رضوان لغوينسات. تمت المراجعة بواسطة فارس القنيعير.

مقدمة للتعلم غير المُؤجَّه

□ الهدف من التعلم غير المُؤجَّه هو إيجاد الأنماط الخفية في البيانات غير المُعلَّمة $\{x^{(1)}, \dots, x^{(m)}\}$ الحافز -

□ متباينة جينسن - لتكن f دالة محدبة و X متغير عشوائي. لدينا المتباينة التالية:

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

تعظيم القيمة المتوقعة (Expectation-Maximization)

□ المتغيرات الكامنة - المتغيرات الكامنة هي متغيرات مخفية غير معبينة تزيد من صعوبة مشاكل التقدير، غالباً ما ترمز بالحرف z . في مايلي الإعدادات الشائعة التي تحتوي على متغيرات كامنة:

الإعداد	المتغير الكامن z	$x z$	ملاحظات
خليط من k توزيع جاسي	Multinomial(ϕ)	$\mathcal{N}(\mu_j, \Sigma_j)$	$\mu_j \in \mathbb{R}^n, \phi \in \mathbb{R}^k$
تحليل عاملي	$\mathcal{N}(0, I)$	$\mathcal{N}(\mu + \Lambda z, \psi)$	$\mu_j \in \mathbb{R}^n$

□ خوارزمية - تعظيم القيمة المتوقعة (Expectation-Maximization) هي عبارة عن طريقة فعالة لتقدير المُدخل θ عبر تقدير تقدير الأرجحية الأعلى (maximum likelihood estimation)، ويتم ذلك بشكل تكراري حيث يتم إيجاد حد أدنى للأرجحية ثم يتم تحسين (optimizing) ذلك الحد الأدنى كما يلي:

• الخطوة E: حساب الاحتمال البعدي $Q_i(z^{(i)})$ بأن تصدر كل نقطة $x^{(i)}$ من مجموعة (cluster) $z^{(i)}$ كما يلي:

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

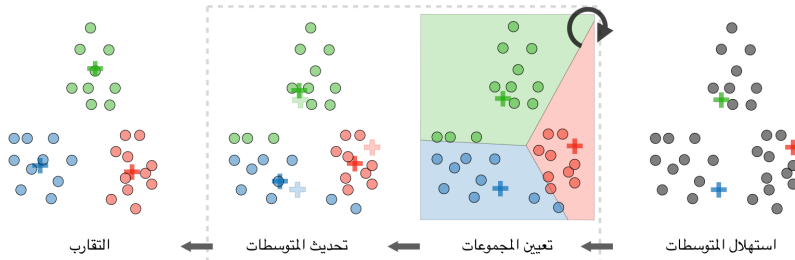
• الخطوة M: يتم استعمال الاحتمالات البعدية $Q_i(z^{(i)})$ كأوزان خاصة لكل مجموعة (cluster) على النقط $x^{(i)}$ ، لكي يتم تقدير نموذج لكل مجموعة بشكل منفصل، و ذلك كما يلي:

التجميع بالمتوسطات k (k-mean clustering)

نرمز لمجموعة النقط i ب $c^{(i)}$ ، ونرمز ب μ_j مركز المجموعات j .

□ خوارزمية - بعد الاستهلال العشوائي للنقاط المركزية (centroids) للمجموعات $\mu_1, \mu_2, \dots, \mu_k \in \mathbb{R}^n$ ، التجميع بالمتوسطات k تكرر الخطوة التالية حتى التقارب:

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



□ دالة التحريف (distortion function) - لكي نتأكد من أن الخوارزمية تقارب، ننظر إلى دالة التحريف المعرفة كما يلي:

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

التجميع الهرمي

□ **خوارزمية** - هي عبارة عن خوارزمية تجميع تعتمد على طريقة تجميع هرمية تبني مجموعات متداخلة بشكل متتال.

□ **الأنواع** - هنالك عدة أنواع من خوارزميات التجميع الهرمي التي ترمي إلى تحسين دوال هدف (objective functions) مختلفة، هذه الأنواع ملخصة في الجدول التالي:

ربط واُزد (ward linkage)	الربط المتوسط	الربط الكامل
تصغير المسافة داخل المجموعة	تصغير متوسط المسافة بين أزواج المجموعات	تصغير المسافة العظمى بين أزواج المجموعات

مقاييس تقدير المجموعات

في التعلّم غير المُوجّه من الصعب غالباً تقدير أداء نموذج ما، لأن القيم الحقيقية تكون غير متوفرة كما هو الحال في التعلّم المُوجّه.

□ **معامل الظل (silhouette coefficient)** - إذا رمزنا a و b لمتوسط المسافة بين عينة وكل النقط المنتمية لنفس الصنف، و بين عينة وكل النقط المنتمية لأقرب مجموعة، المعامل الظلي s لعينة واحدة معرف كالتالي:

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

□ **مؤشر كالينسكي-هارباز (Calinski-Harabaz index)** - إذا رمزنا b لعدد المجموعات، فإن B_k و W_k مصفوفتي التشتت بين المجموعات وداخلها تعرف كالتالي:

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

مؤشر كالينسكي-هارباز $s(k)$ يشير إلى جودة نموذج تجميعي في تعريف مجموعاته، بحيث كلما كانت النتيجة أعلى كلما دل ذلك على أن المجموعات أكثر كثافة وأكثر انفصلاً فيما بينها. هذا المؤشر معرّف كالتالي:

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

تحليل المكون الرئيس

إنها طريقة لتقليص الأبعاد ترمي إلى إيجاد الاتجاهات المعظمة للتباين من أجل إسقاط البيانات عليها.

□ **قيمة ذاتية (eigenvalue)**، **متجه ذاتي (eigenvector)** - لتكن $A \in \mathbb{R}^{n \times n}$ مصفوفة، نقول أن λ قيمة ذاتية للمصفوفة A إذا وُجد متجه $z \in \mathbb{R}^n \setminus \{0\}$ يسمى متجهاً ذاتياً، بحيث:

$$Az = \lambda z$$

□ **مبرهنة الطيف (Spectral theorem)** - لتكن $A \in \mathbb{R}^{n \times n}$ إذا كانت A متناظرة فإنها يمكن أن تكون شبه قطرية عن طريق مصفوفة متعامدة حقيقية $U \in \mathbb{R}^{n \times n}$ إذا رمزنا $\Lambda = \text{diag}(\lambda_1, \dots, \lambda_n)$ ، لدينا:

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

ملحوظة: المتجه الذاتي المرتبط بأبزر قيمة ذاتية يسمى بالمتجه الذاتي الرئيسي (principal eigenvector) للمصفوفة A .

□ **خوارزمية** - تحليل المكون الرئيس (Principal Component Analysis (PCA)) طريقة لخفض الأبعاد تهدف إلى إسقاط البيانات على k بُعد بحيث يتم تعظيم التباين (variance)، خطواتها كالتالي:

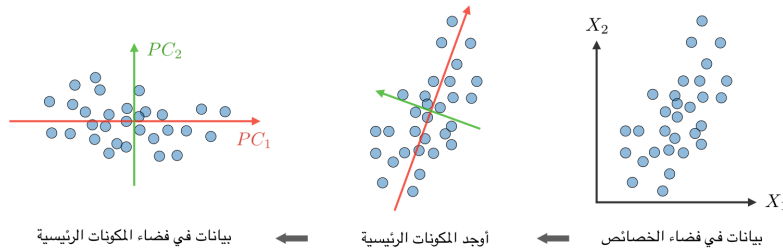
• **الخطوة ١:** تسوية البيانات بحيث تصبح ذات متوسط يساوي صفر وانحراف معياري يساوي واحد.

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

• **الخطوة ٢:** حساب $\Sigma = \frac{1}{m} \sum_{i=1}^m x^{(i)} x^{(i)T} \in \mathbb{R}^{n \times n}$ ، وهي متناظرة وذات قيم ذاتية حقيقية.

• **الخطوة ٣:** حساب $u_1, \dots, u_k \in \mathbb{R}^n$ المتجهات الذاتية الرئيسية المتعامدة ل Σ وعددها k ، بعبارة أخرى، k من المتجهات الذاتية المتعامدة ذات القيم الذاتية الأكبر.

• **الخطوة ٤:** إسقاط البيانات على $\text{span}_{\mathbb{R}}(u_1, \dots, u_k)$.



تحليل المكونات المستقلة

هي طريقة تهدف إلى إيجاد المصادر التوليدية الكامنة.

□ **افتراضات** - لنفترض أن بياناتنا x تم توليدها عن طريق المتجه المصدر $s = (s_1, \dots, s_n)$ إذا n بُعد، حيث s_i متغيرات عشوائية مستقلة، وذلك عبر مصفوفة خلط غير منفردة (mixing and non-singular) A كالتالي:

$$x = As$$

الهدف هو العثور على مصفوفة الفصل $W = A^{-1}$.

□ **خوارزمية تحليل المكونات المستقلة (ICA) لبيل وسجنوسكي (Bell and Sejnowski)** - هذه الخوارزمية تجد مصفوفة الفصل W عن طريق الخطوات التالية:

• اكتب الاحتمال $s = W^{-1}x$ كالتالي:

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

- لتكن $\{x^{(i)}, i \in \llbracket 1, m \rrbracket\}$ بيانات التمرن و g دالة سيجمويد، اكتب الأرجحية اللوغاريتمية (log likelihood) كالتالي:

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

هكذا، باستخدام الصعود الاشتقاقي العشوائي (stochastic gradient ascent)، لكل عينة تدريب $x^{(i)}$ نقوم بتحديث W كما يلي:

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