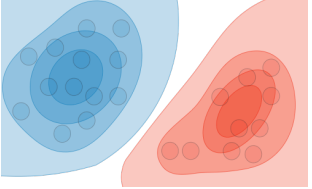
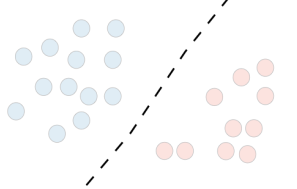


نموذج توليدي (Generative)	نموذج تمييزي (Discriminative)	
تقدير $P(x y)$ ثم استنتاج $P(y x)$	التقدير المباشر لـ $P(y x)$	الهدف
التوزيع الاحتمالي للبيانات	حدود القرار	ماذا يتعلم
		توضيح
GDA, بايز البسيط (Naive Bayes)	الانحدار (Regression)، آلة المتجهات الداعمة (SVM)	أمثلة

مرجع سريع تعلم آلي

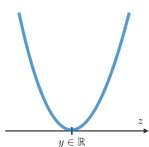
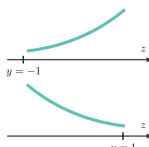
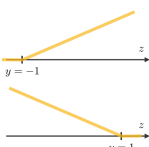
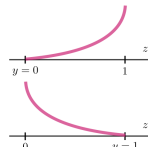
افشين عميدى و شروين عميدى

١٤ ربيع الثاني، ١٤٤١

٢.١ الرموز ومفاهيم أساسية

□ **الفرضية (Hypothesis)** - الفرضية، ويرمز لها بـ h_θ ، هي النموذج الذي نختاره. إذا كان لدينا المدخل $x^{(i)}$ ، فإن المخرج الذي سيتوقعه النموذج هو $h_\theta(x^{(i)})$.

□ **دالة الخسارة (Loss function)** - دالة الخسارة هي الدالة $L: (z, y) \in \mathbb{R} \times Y \mapsto L(z, y) \in \mathbb{R}$ التي تأخذ كمدخلات القيمة المتوقعة z والقيمة الحقيقية y وتعطينا الاختلاف بينهما. الجدول التالي يحتوي على بعض دوال الخسارة الشائعة:

خطأ أصغر تربيع (Least squared error)	خسارة لوجستية (Logistic loss)	خسارة مفصلية (Hinge loss)	الانتروبيا التقاطعية (Cross-entropy)
$\frac{1}{2}(y - z)^2$	$\log(1 + \exp(-yz))$	$\max(0, 1 - yz)$	$-\left[y \log(z) + (1 - y) \log(1 - z)\right]$
			
الانحدار الخطي (Linear regression)	الانحدار اللوجستي (Logistic regression)	آلة المتجهات الداعمة (SVM)	الشبكات العصبية (Neural Network)

□ **دالة التكلفة (Cost function)** - دالة التكلفة J تستخدم عادة لتقييم أداء نموذج ما، ويتم تعريفها مع دالة الخسارة L كالتالي:

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

١ تعلم مراقب

تمت الترجمة بواسطة فارس القنيعير. تمت المراجعة بواسطة زيد البافعي.

١.١ مقدمة للتعلم المُوقَّع

إذا كان لدينا مجموعة من نقاط البيانات $\{x^{(1)}, \dots, x^{(m)}\}$ مرتبطة بمجموعة مخرجات $\{y^{(1)}, \dots, y^{(m)}\}$ ، نريد أن نبني مُصنَّف يتعلم كيف يتوقع y من x .

□ **نوع التوقع** - أنواع نماذج التوقع المختلفة موضحة في الجدول التالي:

المُخرَج	الانحدار (Regression)	التصنيف (Classification)
مستمر	انحدار خطي (Linear regression)	صنف
أمثلة	انحدار خطي (Linear regression)	انحدار لوجستي (Logistic regression)، آلة المتجهات الداعمة (SVM)، بايز البسيط (Naive Bayes)

□ **نوع النموذج** - أنواع النماذج المختلفة موضحة في الجدول التالي:

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

ملاحظة: قانون التحديث هذا يعتبر حالة خاصة من النزول الاشتقاقي (Gradient descent).

□ الانحدار الموزون محلياً (LWR) – الانحدار الموزون محلياً (Locally Weighted Regression)، ويعرف بـ LWR، هو نوع من الانحدار الخطي يزن كل عينة تدريب أثناء حساب دالة التكلفة باستخدام $w^{(i)}(x)$ ، التي يمكن تعريفها باستخدام المُدخل $\tau \in \mathbb{R}$ (parameter) كالتالي:

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

٢.٢.١ التصنيف والانحدار اللوجستي

□ دالة سيجمويد (Sigmoid) – دالة سيجمويد g ، وتعرف كذلك بالدالة اللوجستية، تعرّف كالتالي:

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

□ الانحدار اللوجستي (Logistic regression) – نفترض هنا أن $y|x; \theta \sim \text{Bernoulli}(\phi)$. فيكون لدينا:

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

ملاحظة: ليس هناك حل رياضي مغلق للانحدار اللوجستي.

□ انحدار سوفت ماكس (Softmax) – ويطلق عليه الانحدار اللوجستي متعدد الأصناف (multiclass logistic regression)، يستخدم لتعميم الانحدار اللوجستي إذا كان لدينا أكثر من صنفين. في العرف يتم تعيين $\theta_K = 0$ بحيث تجعل مُدخل بيرنولي (Bernoulli) ϕ_i لكل فئة i يساوي:

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

٢.٢.١ النماذج الخطية العامة (Generalized Linear Models - GLM)

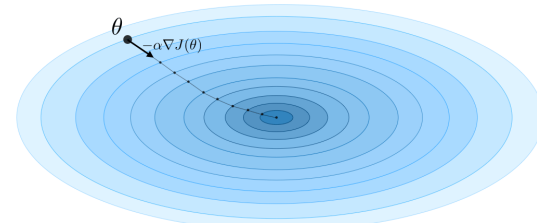
□ العائلة الأسية (Exponential family) – يطلق على صنف من التوزيعات (distributions) بأنها تنتمي إلى العائلة الأسية إذا كان يمكن كتابتها بواسطة مُدخل قانوني (canonical parameter) η ، إحصاء كافٍ (sufficient statistic) $T(y)$ ، ودالة تجزئة لوغاريتمية $a(\eta)$ ، كالتالي:

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

ملاحظة: كثيراً ما سيكون $T(y) = y$. كذلك فإن $\exp(-a(\eta))$ يمكن أن تفسر كمدخل تسوية (normalization) للتأكد من أن الاحتمالات يكون حاصل جمعها يساوي واحد.

□ النزول الاشتقاقي (Gradient descent) – لنعرّف معدل التعلم $\alpha \in \mathbb{R}$ ، يمكن تعريف القانون الذي يتم تحديث خوارزمية النزول الاشتقاقي من خلاله باستخدام معدل التعلم ودالة التكلفة J كالتالي:

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



ملاحظة: في النزول الاشتقاقي العشوائي (Stochastic gradient descent (SGD)) يتم تحديث المُعاملات (parameters) بناءً على كل عينة تدريب على حدة، بينما في النزول الاشتقاقي الحُرْمي (batch gradient descent) يتم تحديثها باستخدام حُرْم من عينات التدريب.

□ الأرجحية (Likelihood) – تستخدم أرجحية النموذج $L(\theta)$ ، حيث أن θ هي المُدخلات، للبحث عن المُدخلات θ الأحسن عن طريق تعظيم (maximizing) الأرجحية. عملياً يتم استخدام الأرجحية اللوغاريتمية (log-likelihood) $\ell(\theta) = \log(L(\theta))$ حيث أنها أسهل في التحسين (optimize). فيكون لدينا:

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

□ خوارزمية نيوتن (Newton's algorithm) – خوارزمية نيوتن هي طريقة حسابية للعثور على θ بحيث يكون $\ell'(\theta) = 0$. قاعدة التحديث للخوارزمية كالتالي:

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

ملاحظة: هناك خوارزمية أعم وهي متعددة الأبعاد (multidimensional)، يطلق عليها خوارزمية نيوتن-رافسون (Newton-Raphson)، ويتم تحديثها عبر القانون التالي:

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

١.٢.١ الانحدار الخطي (Linear regression)

هنا نفترض أن $y|x; \theta \sim \mathcal{N}(\mu, \sigma^2)$

□ المعادلة الطبيعية (Normal) – إذا كان لدينا المصفوفة X ، القيمة θ التي تقلل من دالة التكلفة يمكن حلها رياضياً بشكل مغلق (closed-form) عن طريق:

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

□ خوارزمية أصغر معدل تربيع LMS – إذا كان لدينا معدل التعلم α ، فإن قانون التحديث لخوارزمية أصغر معدل تربيع (Least Mean Squares (LMS)) لمجموعة بيانات من m عينة، ويطلق عليه قانون تعلم ويدرو-هوف (Widrow-Hoff)، كالتالي:

تم تلخيص أكثر التوزيعات الأسية استخداماً في الجدول التالي:

التوزيع	η	$T(y)$	$a(\eta)$	$b(y)$
برنولي (Bernoulli)	$\log\left(\frac{\phi}{1-\phi}\right)$	y	$\log(1 + \exp(\eta))$	1
جاوسي (Gaussian)	μ	y	$\frac{\eta^2}{2}$	$\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{y^2}{2}\right)$
بواسون (Poisson)	$\log(\lambda)$	y	e^η	$\frac{1}{y!}$
هندسي (Geometric)	$\log(1 - \phi)$	y	$\log\left(\frac{e^\eta}{1 - e^\eta}\right)$	1

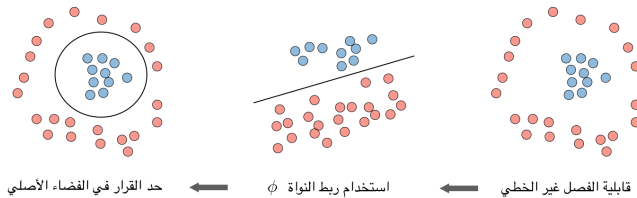
❑ الخسارة المفصليّة (Hinge loss) - تستخدم الخسارة المفصليّة في حل SVM ويعرف على النحو التالي:

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

❑ النواة (Kernel) - إذا كان لدينا دالة ربط الخصائص (features) ϕ ، يمكننا تعريف النواة K كالتالي:

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

عملياً، يمكن أن نُعرّف الدالة K عن طريق المعادلة $K(x, z) = \exp\left(-\frac{\|x - z\|^2}{2\sigma^2}\right)$ (Gaussian kernel)، وهي تستخدم بكثرة. ويطلق عليها النواة الجاوسية.



ملاحظة: نقول أننا نستخدم "حيلة النواة" (kernel trick) لحساب دالة التكلفة عند استخدام النواة لأننا في الحقيقة لا نحتاج أن نعرف التحويل الصريح ϕ ، الذي يكون في الغالب شديد التعقيد. ولكن، نحتاج أن نحسب القيم $K(x, z)$.

❑ اللاغرانجي (Lagrangian) - يتم تعريف اللاغرانجي $\mathcal{L}(w, b)$ على النحو التالي:

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

ملاحظة: المعاملات (coefficients) β_i يطلق عليها مضروبات لاغرانج (Lagrange multipliers).

٤.١ التعلم التوليدي (Generative Learning)

النموذج التوليدي في البداية يحاول أن يتعلم كيف تم توليد البيانات عن طريق تقدير $P(x|y)$ ، التي يمكن حينها استخدامها لتقدير $P(y|x)$ باستخدام قانون بايز (Bayes' rule).

١.٤.١ تحليل التمايز الجاوسي (Gaussian Discriminant Analysis)

❑ الإطار - تحليل التمايز الجاوسي يفترض أن y و $x|y=0$ و $x|y=1$ بحيث يكونوا كالتالي:

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

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

❑ التقدير - الجدول التالي يلخص التقديرات التي يمكننا التوصل لها عند تعظيم الأرجحية (likelihood):

❑ افتراضات GLMs - تهدف النماذج الخطية العامة (GLM) إلى توقع المتغير العشوائي y كدالة لـ $x \in \mathbb{R}^{n+1}$ وتستند إلى ثلاثة افتراضات:

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

ملاحظة: أصغر تربيع (least squares) الاعتيادي و الانحدار اللوجستي يعتبران من الحالات الخاصة للنماذج الخطية العامة.

٣.١ آلة المتجهات الداعمة (Support Vector Machines)

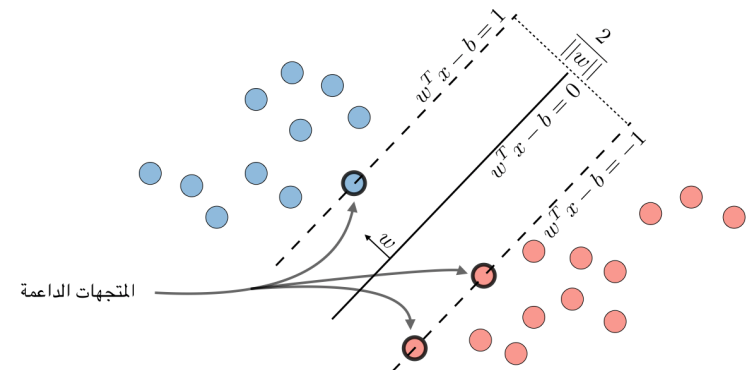
تهدف آلة المتجهات الداعمة (SVM) إلى العثور على الخط الذي يعظم أصغر مسافة إليه:

❑ مُصنّف الهامش الأحسن (Optimal margin classifier) - يُعرّف مُصنّف الهامش الأحسن h كالتالي:

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

حيث $(w, b) \in \mathbb{R}^n \times \mathbb{R}$ هو الحل لمشكلة التحسين (optimization) التالية:

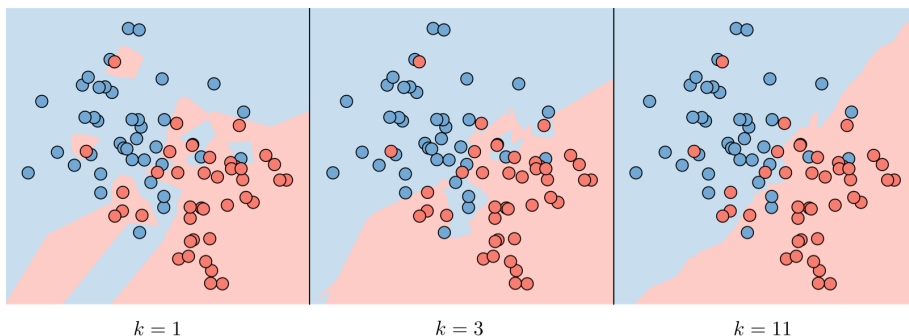
$$\min \frac{1}{2} \|w\|^2 \quad \text{بحيث أن} \quad y^{(i)} (w^T x^{(i)} - b) \geq 1$$



٦.١ طرق أخرى غير بارامترية (non-parametric)

□ **خوارزمية أقرب الجيران (k -nearest neighbors)** - تعتبر خوارزمية أقرب الجيران، وتعرف بـ k -NN، طريقة غير بارامترية، حيث يتم تحديد نتيجة عينة من البيانات من خلال عدد k من البيانات المجاورة في مجموعة التدريب. ويمكن استخدامها للتصنيف والانحدار.

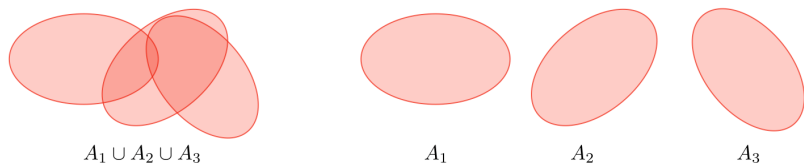
ملاحظة: كلما زاد الفدخل k ، كلما زاد الانحياز ($bias$)، وكلما نقص k ، زاد التباين ($variance$).



٧.١ نظرية التعلم

□ **حد الاتحاد (Union bound)** - لنجعل A_1, \dots, A_k تمثل k حدث. فيكون لدينا:

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



□ **مراجعة هوفدينج (Hoeffding)** - لنجعل Z_1, \dots, Z_m تمثل m متغير مستقلة وموزعة بشكل مماثل (iid) مأخوذة من توزيع برنولي (Bernoulli distribution) ذا مدخل ϕ . لنجعل $\hat{\phi}$ متوسط العينة (sample mean) و $\gamma > 0$ ثابت. فيكون لدينا:

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

ملاحظة: هذه المتراجحة تعرف كذلك بحد تشرنوف (Chernoff bound).

□ **خطأ التدريب** - ليكن لدينا المُصنّف h ، يمكن تعريف خطأ التدريب $\hat{\epsilon}(h)$ ، ويعرف كذلك بالخطر التجريبي أو الخطأ التجريبي، كالتالي:

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

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

٢.٤.١ بايز البسيط (Naive Bayes)

□ **الافتراض** - يفترض نموذج بايز البسيط أن جميع الخصائص لكل عينة بيانات مستقلة (independent):

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

□ **الحل** - تعظيم الأرجحية اللوغاريتمية (log-likelihood) يعطينا الحلول التالية إذا كان $k \in \{0, 1\}, l \in [1, L]$:

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

ملاحظة: بايز البسيط يستخدم بشكل واسع لتصنيف النصوص واكتشاف البريد الإلكتروني المزعج.

٥.١ الطرق الشجرية (tree-based) والتجميعية (ensemble)

هذه الطرق يمكن استخدامها لكل من مشاكل الانحدار (regression) والتصنيف (classification).

□ **التصنيف والانحدار الشجري (CART)** - والاسم الشائع له أشجار القرار (decision trees)، يمكن أن يمثل كأشجار ثنائية (binary trees). من المزايا لهذه الطريقة إمكانية تفسيرها بسهولة.

□ **الغابة العشوائية (Random forest)** - هي أحد الطرق الشجرية التي تستخدم عدداً كبيراً من أشجار القرار مبنية باستخدام مجموعة عشوائية من الخصائص. بخلاف شجرة القرار البسيطة لا يمكن تفسير النموذج بسهولة، ولكن أدائها العالي جعلها أحد الخوارزميات المشهورة.

ملاحظة: أشجار القرار نوع من الخوارزميات التجميعية (ensemble).

□ **التعزيز (Boosting)** - فكرة خوارزميات التعزيز هي دمج عدة خوارزميات تعلم ضعيفة لتكوين نموذج قوي. الطرق الأساسية ملخصة في الجدول التالي:

التعزيز التكيّفي (Adaptive boosting)	التعزيز الاشتقاقي (Gradient boosting)
- يتم التركيز على مواطن الخطأ لتحسين النتيجة في الخطوة التالية. - "Adaboost"	- يتم تدريب خوارزميات التعلم الضعيفة على الأخطاء المتبقية.

□ تقريباً صحيح احتمالياً ((Probably Approximately Correct (PAC)) - هو إطار يتم من خلاله إثبات ٢ تعلم غير مراقب العديد من نظريات التعلم، ويحتوي على الافتراضات التالية:

- مجموعات التدريب والاختبار يتبعان نفس التوزيع.
- عينات التدريب تؤخذ بشكل مستقل.

١.٢ مقدمة للتعلم غير المُوجّه

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

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

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

١.١.٢ تعظيم القيمة المتوقعة (Expectation-Maximization)

□ المتغيرات الكامنة - المتغيرات الكامنة هي متغيرات مخفية غير معانية 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 جاوسي	Multinomial(ϕ)	$\mathcal{N}(\mu_j, \Sigma_j)$	$\mu_j \in \mathbb{R}^n, \phi \in \mathbb{R}^k$
تحليل عاملي	$\mathcal{N}(0, I)$	$\mathcal{N}(\mu + \Lambda z, \psi)$	$\mu_j \in \mathbb{R}^n$

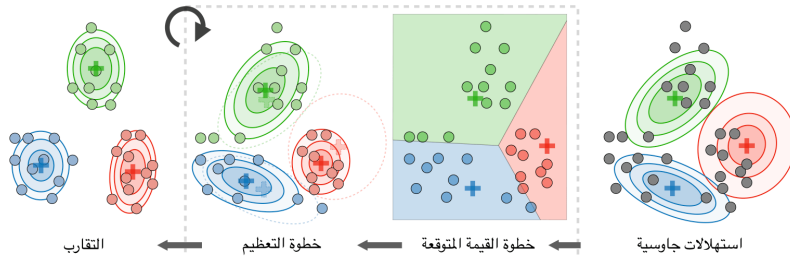
□ خوارزمية - تعظيم القيمة المتوقعة (Expectation-Maximization) هي عبارة عن طريقة فعالة لتقدير المُدخل θ عبر تقدير تقدير الأرجحية الأعلى (maximum likelihood estimation)، ويتم ذلك بشكل تكراري حيث يتم إيجاد حد أدنى للأرجحية ثم يتم تحسين (optimizing) ذلك الحد الأدنى كما يلي:

• 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، وذلك كما يلي:

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



□ مجموعة تكسيرية (Shattering Set) - إذا كان لدينا المجموعة $S = \{x^{(1)}, \dots, x^{(d)}\}$ ومجموعة مُصنّفات \mathcal{H} نقول أن \mathcal{H} تكسر S (H shatters S) إذا كان لكل مجموعة علامات (labels) $\{y^{(1)}, \dots, y^{(d)}\}$ لدينا:

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

□ مبرهنة الحد الأعلى (Upper bound theorem) - لنجعل \mathcal{H} فئة فرضية محدودة (finite hypothesis class) بحيث $|\mathcal{H}| = k$ ، و δ وحجم العينة m ثابتين. حينها سيكون لدينا، مع احتمال على الأقل $1 - \delta$ ، التالي:

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

□ بُعد فابنيك - تشرفونيكس (Vapnik-Chervonenkis - VC) لفئة فرضية غير محدودة (infinite hypothesis class \mathcal{H} ، ويرمز له بـ $VC(\mathcal{H})$ ، هو حجم أكبر مجموعة (set) التي تم تكسيرها بواسطة \mathcal{H} (shattered by \mathcal{H}).

ملاحظة: بُعد فابنيك [تشرفونيكس VC] \mathcal{H} مجموعة التصنيفات الخطية في بُعدين {يساوي ٣}.



□ مبرهنة فابنيك (Vapnik theorem) - ليكن لدينا \mathcal{H} ، مع $d = VC(\mathcal{H})$ وعدد عينات التدريب m سيكون لدينا، مع احتمال على الأقل $1 - \delta$ ، التالي:

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

□ **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 التششت 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 معرف كالتالي:

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

٥.١.٢ 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:

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

ملحوظة: المتجه الذاتي المرتبط بأكبر 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.

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

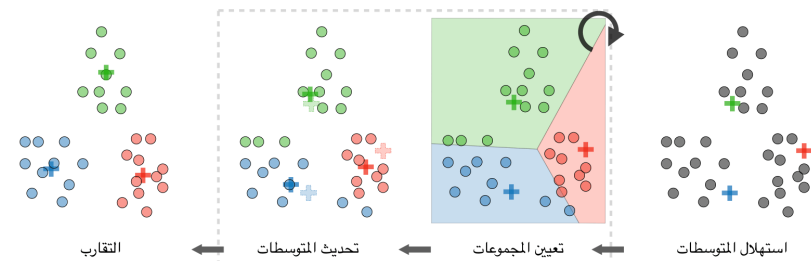
• الخطوة ٢: حساب $\Sigma = \frac{1}{m} \sum_{i=1}^m x^{(i)} x^{(i)T} \in \mathbb{R}^{n \times n}$ ، وهي متناظرة وذات قيم ذاتية حقيقية.

٢.١.٢ 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:

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



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

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

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

٤.١.٢ 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.

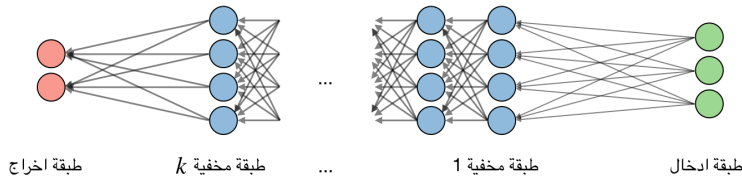
٣ تعلم متعمق

تمت الترجمة بواسطة امجد الخطابي. تمت المراجعة بواسطة زيد اليافعي.

١.٣ الشبكة العصبونية الاصطناعية (Neural Networks)

الشبكة العصبونية الاصطناعية هي عبارة عن نوع من النماذج يبني من عدة طبقات ، اكثر هذه الانواع استخداما هي الشبكات الالتفافية و الشبكات العصبونية المتكررة

□ البنية - المصطلحات حول بنية الشبكة العصبونية موضح في الشكل ادناه



عبر تدوين i كالتبقة رقم i و j للدلالة على رقم الوحدة الخفية في تلك الطبقة ، نحصل على □

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

حيث نعرف w, b, z كالوزن ، و معامل التعديل ، و الناتج حسب الترتيب.

□ دالة التفعيل (Activation function) - دالة التفعيل تستخدم في نهاية الوحدة الخفية لتضمن المكونات الغير خطية للنموذج. هنا بعض دوال التفعيل الشائعة

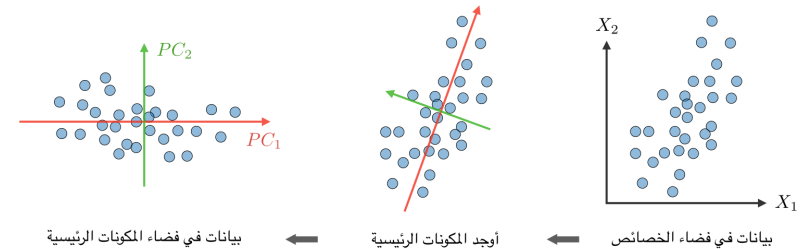
Leaky ReLU	ReLU	Tanh	Sigmoid
$g(z) = \max(\epsilon z, z)$ $\epsilon \ll 1$	$g(z) = \max(0, z)$	$g(z) = \frac{e^z - e^{-z}}{e^z + e^{-z}}$	$g(z) = \frac{1}{1 + e^{-z}}$

□ دالة الانتروبيا التقاطعية للخسارة (Cross-entropy loss) - في سياق الشبكات العصبونية، دالة الانتروبيا $L(z, y)$ تستخدم و تعرف كالآتي □

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

□ معدل التعلم (Learning rate) - معدل التعلم، يرمز ، و هو مؤشر في اي تجاة يتم تحديث الازان. يمكن تثبيت هذا المعامل او تحديثه بشكل تأقلمي . حاليا اكثر النسب شيوعا تدعى Adam ، وهي طريقة تجعل هذه النسبة سرعة التعلم بشكل تأقلمي α او η ب ،

- 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 $\text{span}_{\mathbb{R}}(u_1, \dots, u_k)$.



٦.١.٢ تحليل المكونات المستقلة

هي طريقة تهدف إلى إيجاد المصادر التوليدية الكامنة.

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

$$x = As$$

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 ل $x = As = W^{-1}s$ كالآتي:

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

• لكن $\{x^{(i)}, i \in [1, m]\}$ بيانات التمرن و g دالة سيجمويد، اكتب الأرجحية اللوغاريتمية (log likelihood) كالآتي:

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

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

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

□ **LSTM** - ذاكرة طويلة قصيرة الذاكرة (long short-term memory) هي نوع من نموذج ال RNN تستخدم لتجنب مشكلة اختفاء الانحدار عبر اضافة بوابات النسيان.

٤.٣ التعلم و التحكم المعزز (Reinforcement Learning)

الهدف من التعلم المعزز للتعلم الذكي هو التعلم لكيفية التأقلم في اي بيئة.

□ **عملية ماركوف لاتخاذ القرار** - عملية ماركوف لاتخاذ القرار هي سلسلة خماسية $(S, A, \{P_{sa}\}, \gamma, R)$ حيث

• S هي مجموعة من حالات البيئة

• A هي مجموعة من حالات الاجراءات

• $\{P_{sa}\}$ هو حالة احتمال الانتقال من الحالة $s \in S$ و $a \in A$

• $\gamma \in [0, 1]$ هي عامل الخصم

• $R : S \times A \rightarrow \mathbb{R}$ و $R : S \rightarrow \mathbb{R}$ هي دالة المكافأة والتي تعمل الخوارزمية على جعلها اعلى قيمة

□ **دالة القواعد** - دالة القواعد $\pi : S \rightarrow A$ هي التي تقوم بترجمة الحالات الى اجراءات.

ملاحظة: نقول ان النموذج ينفذ القاعدة المعينه π للحالة المعطاة s ان نتخذ الاجراء $a = \pi(s)$.

□ **دالة القاعدة** - لاي قاعدة معطاة π و حالة s ، نقوم بتعريف دالة القيمة V^π كما يلي:

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

□ **معادلة بيلمان** - معادلات بيلمان المثلى تشخص دالة القيمة دالة القيمة V^{π^*} : للقاعدة المثلى

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

π^* للحالة المعطاة s تعطى كالتالي: ملاحظة: نلاحظ ان القاعدة المثلى

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

□ **خوارزمية تكرار القيمة (Value iteration algorithm)** - خوارزمية تكرار القيمة تكون في خطوتين:

١. نقوم بوضع قيمة اولية □

$$V_0(s) = 0$$

٢. نقوم بتكرير القيمة حسب القيم السابقة □

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

□ **التغذية الخلفية (Backpropagation)** - التغذية الخلفية هي طريقة لتحديث الاوزان في الشبكة العصبونية عبر اعتبار القيم الحقيقية للنتاج مع القيمة المطلوبة للخرج. المشتقة بالنسبة للوزن w يتم حسابها باستخدام قاعدة التسلسل و تكون عبر الشكل الاتي □

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

كنتيجة ، الوزن سيتم تحديثه كالتالي:

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

□ **تحديث الاوزان** - في الشبكات العصبونية ، يتم تحديث الاوزان كما يلي:

• الخطوة ١: خذ حزمة من بيانات التدريب

• الخطوة ٢: قم بعملية التغذية الامامية لحساب الخسارة الناتجة

• الخطوة ٣: قم بتغذية خلفية للخسارة للحصول على دالة الانحدار

• الخطوة ٤: استخدم قيم الانحدار لتحديث اوزان الشبكة

□ **الاسقاط (Dropout)** - الاسقاط هي طريقة الغرض منها منع التكيف الزائد للنموذج في بيانات التدريب عبر اسقاط بعض الواحدات في الشبكة العصبونية، العصبونات يتم اما اسقاطها باحتمالية p او الحفاظ عليها باحتمالية $1 - p$.

٢.٣ الشبكات العصبونية الالتفافية (CNN)

□ **احتياج الطبقة الالتفافية** - عبر رمز W لحجم المدخل ، F حجم العصبونات للطبقة الالتفافية ، P عدد الحشوات الصفرية ، فإن N عدد العصبونات لكل حجم معطى يحسب عبر الاتي:

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

□ **تنظيم الحزمة (Batch normalization)** - هي خطوه من قيم التحسين الخاصة γ, β والتي تعدل الحزمة $\{x_i\}$. لنجعل μ_B, σ_B^2 المتوسط و الانحراف للحزمة المعنية و نريد تصحيح هذه الحزمة، يتم ذلك كالتالي:

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

في الغالب تتم بعد الطبقة الالتفافية أو المتصلة كليا و قبل طبقة التغيرات الغير خطية و تهدف للسماح للسرعات التعليم العالية للتقليل من الاعتمادية القوية للقيم الاولى.

٣.٣ الشبكات العصبونية التكرارية (RNN)

□ **انواع البوابات** - هنا الانواع المختلفة التي ممكن مواجهتها في الشبكة العصبونية الاعتيادية:

بوابة ادخال	بوابة نسيان	بوابة اخراج	بوابة منفذ
كتابة ام عدم كتابة الى الخلية؟	مسح ام عدم مسح الخلية؟	مدى الافصاح عن الخلية ؟	كمية الكتابة الى الخلية ؟

□ **تقدير الامكانية القصوى** - تقديرات الامكانية القصوى □ تقدير الاحتمال الأرجح □ احتماليات انتقال الحالة تكون ٤ لنصائح وحيل تعلم كما يلي □

تمت الترجمة بواسطة فارس القنيعير. تمت المراجعة بواسطة زيد اليافعي.

$$P_{sa}(s') = \frac{\text{اوقات تنفيذ الاجراء } a \text{ في الحالة } s \text{ وانتقلت الى } s'}{\text{اوقات تنفيذ الاجراء } a \text{ في الحالة } s}$$

١.٤ مقاييس التصنيف

في سياق التصنيف الثنائي، هذه المقاييس (metrics) المهمة التي يجدر مراقبتها من أجل تقييم أداء النموذج.

□ **مصفوفة الدقة (confusion matrix)** - تستخدم مصفوفة الدقة لأخذ تصور شامل عند تقييم أداء النموذج. وهي تعرّف كالتالي:

التصنيف المتوقع			
	-	+	
التصنيف الفعلي	FN False Negatives Type II error	TP True Positives	+
	TN True Negatives	FP False Positives Type I error	-

□ **المقاييس الأساسية** - المقاييس التالية تستخدم في العادة لتقييم أداء نماذج التصنيف:

المقياس	المعادلة	التفسير
الضبط (accuracy)	$\frac{TP + TN}{TP + TN + FP + FN}$	الأداء العام للنموذج
Precision	$\frac{TP}{TP + FP}$	دقة التوقعات الإيجابية (positive)
Recall Sensitivity	$\frac{TP}{TP + FN}$	تغطية عينات التوقعات الإيجابية الفعلية
Specificity	$\frac{TN}{TN + FP}$	تغطية عينات التوقعات السلبية الفعلية
درجة F1	$\frac{2TP}{2TP + FP + FN}$	مقياس هجين مفيد للأصناف غير المتوازنة (unbalanced)

□ **منحنى دقة الأداء (ROC)** - منحنى دقة الأداء، ويطلق عليه ROC، هو رسمه لمعدل التصنيفات الإيجابية الصحيحة (TPR) مقابل معدل التصنيفات الإيجابية الخاطئة (FPR) باستخدام قيم حد (threshold) متغيرة. هذه المقاييس ملخصة في الجدول التالي:

٣.٤ اختيار النموذج

□ مفردات - عند اختيار النموذج، نفرق بين ٣ أجزاء من البيانات التي لدينا كالتالي:

مجموعة اختبار	مجموعة تحقق	مجموعة تدريب
- النموذج يعطي التوقعات - بيانات لم يسبق رؤيتها من قبل	- يتم تقييم النموذج - غالباً 20% من مجموعة البيانات - يطلق عليها كذلك المجموعة المُجَبَّة أو مجموعة التطوير	- يتم تدريب النموذج - غالباً 80% من مجموعة البيانات

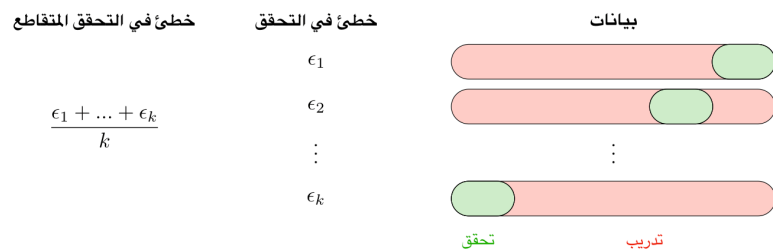
بمجرد اختيار النموذج، يتم تدريبه على مجموعة البيانات بالكامل ثم يتم اختباره على مجموعة اختبار لم يسبق رؤيتها من قبل. كما هو موضح في الشكل التالي:



□ التحقق المتقاطع (Cross-validation) - التحقق المتقاطع، وكذلك يختصر بـ CV، هو طريقة تستخدم لاختيار نموذج بحيث لا يعتمد بشكل كبير على مجموعة بيانات التدريب المبدئية. أنواع التحقق المتقاطع المختلفة ملخصة في الجدول التالي:

Leave-p-out	k-fold
- التدريب على $n - p$ عينة والتقييم باستخدام الـ p عينات المتبقية - الحالة $p = 1$ يطلق عليها الإبقاء على واحد (leave-one-out)	- التدريب على $k - 1$ جزء والتقييم باستخدام الجزء الباقي - بشكل عام $k = 5$ أو 10

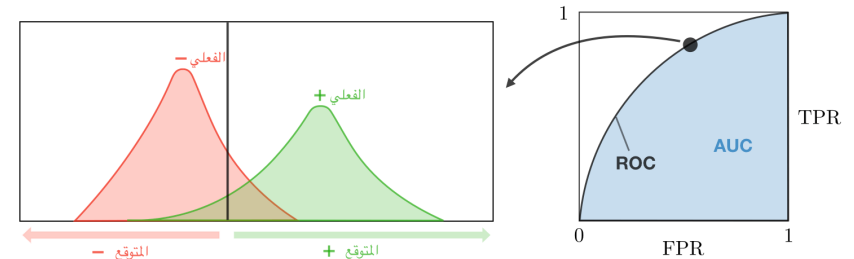
الطريقة الأكثر استخداماً يطلق عليها التحقق المتقاطع س جزءاً (k-fold)، ويتم فيها تقسيم البيانات إلى k جزء، بحيث يتم تدريب النموذج باستخدام $k - 1$ والتحقق باستخدام الجزء المتبقي، ويتم تكرار ذلك k مرة. يتم بعد ذلك حساب معدل الأخطاء في الأجزاء k ويسمى خطأ التحقق المتقاطع.



□ ضبط (Regularization) - عملية الضبط تهدف إلى تفادي فرط التخصيص (overfit) للنموذج، وهو بذلك يتعامل مع مشاكل التباين العالي. الجدول التالي يلخص أنواع وطرق الضبط الأكثر استخداماً:

المقياس	المعادلة	مرادف
True Positive Rate TPR	$\frac{TP}{TP + FN}$	Recall, sensitivity
False Positive Rate FPR	$\frac{FP}{TN + FP}$	1-specificity

□ المساحة تحت منحنى دقة الأداء المساحة تحت المنحنى (AUC) - المساحة تحت منحنى دقة الأداء المساحة تحت المنحنى، ويطلق عليها AUC أو AUROC، هي المساحة تحت ROC كما هو موضح في الرسم التالية:



٢.٤ مقاييس الانحدار

□ المقاييس الأساسية - إذا كان لدينا نموذج الانحدار f ، فإن المقاييس التالية غالباً ما تستخدم لتقييم أداء النموذج:

المجموع الكلي للمربعات	مجموع المربعات المُفسَّر	مجموع المربعات المتبقي
$SS_{tot} = \sum_{i=1}^m (y_i - \bar{y})^2$	$SS_{reg} = \sum_{i=1}^m (f(x_i) - \bar{y})^2$	$SS_{res} = \sum_{i=1}^m (y_i - f(x_i))^2$

□ مُعامل التحديد (Coefficient of determination) - مُعامل التحديد، وغالباً يرمز له بـ R^2 أو r^2 ، يعطي قياس لمدى مطابقة النموذج للنتائج الملحوظة، ويعرف كما يلي:

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

□ المقاييس الرئيسية - المقاييس التالية تستخدم غالباً لتقييم أداء نماذج الانحدار، وذلك بأن يتم الأخذ في الحسبان عدد المتغيرات n المستخدمة فيها:

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

حيث L هو الأرجحية، و $\hat{\sigma}^2$ تقدير التباين الخاص بكل نتيجة.

			توضيح التصنيف
			توضيح التعلم العميق
إجراء الضبط (regularization) - الحصول على المزيد من البيانات		- زيادة تعقيد النموذج - إضافة المزيد من الخصائص - تدريب لمدة أطول	العلاجات الممكنة

Elastic Net	Ridge	LASSO
المفاضلة بين اختيار المتغيرات والمعاملات الصغيرة	يجعل المعاملات أصغر	- يقلص المعاملات إلى ٠ - جيد لاختيار المتغيرات
$\dots + \lambda \left[(1 - \alpha) \ \theta\ _1 + \alpha \ \theta\ _2^2 \right]$ $\lambda \in \mathbb{R}, \quad \alpha \in [0, 1]$	$\dots + \lambda \ \theta\ _2^2$ $\lambda \in \mathbb{R}$	$\dots + \lambda \ \theta\ _1$ $\lambda \in \mathbb{R}$

٤.٤ التشخيصات

□ **تحليل الخطأ** - تحليل الخطأ هو تحليل السبب الرئيسي للفرق في الأداء بين النماذج الحالية والنماذج المثالية.

□ **تحليل استثنائي (Ablative analysis)** - التحليل الاستثنائي هو تحليل السبب الرئيسي للفرق في الأداء بين النماذج الحالية والنماذج المبدئية (baseline).

□ **الانحياز (Bias)** - الانحياز للنموذج هو الفرق بين التنبؤ المتوقع والنموذج الحقيقي الذي نحاول تنبؤه للبيانات المعطاة.

□ **التباين (Variance)** - تباين النموذج هو مقدار التغير في تنبؤ النموذج لنقاط البيانات المعطاة.

□ **موازنة الانحياز والتباين (Bias/variance tradeoff)** - كلما زادت بساطة النموذج، زاد الانحياز، وكلما زاد تعقيد النموذج، زاد التباين.

Overfitting	Just right	Underfitting	
- خطأ التدريب منخفض جداً - خطأ التدريب أقل بكثير من خطأ الاختبار - تباين عالي	- خطأ التدريب أقل بقليل من خطأ الاختبار	- خطأ التدريب عالي - خطأ التدريب قريب من خطأ الاختبار - انحياز عالي	الأعراض
			توضيح الانحدار

٥ بالطبع تنشيطية

١.٥ الاحتمالات والإحصاءات

تمت الترجمة بواسطة محمود أصلان. تمت المراجعة بواسطة فارس القنيعير.

٢.٥ مقدمة في الاحتمالات والتوافيق

□ **فضاء العينة** - يعرف فضاء العينة لتجربة ما بمجموعة كل النتائج الممكنة لهذه التجربة ويرمز لها بـ S .

□ **الحدث** - أي مجموعة جزئية E من فضاء العينة تعتبر حدثاً. أي، الحدث هو مجموعة من النتائج الممكنة للتجربة. إذا كانت نتيجة التجربة محتواة في E ، عندها نقول أن الحدث E وقع.

□ **مسلمات الاحتمالات** - لكل حدث E ، نرمز لإحتمال وقوعه بـ $P(E)$.

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

□ **التباديل** - التبديل هو عبارة عن عدد الاختيارات لـ r غرض من مجموعة مكونة من n غرض بترتيب محدد. عدد هكذا تراتيب يرمز له بـ $P(n, r)$ ، المعرف كالتالي:

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

□ **التوافيق** - التوفيق هو عدد الاختيارات لـ r غرض من مجموعة مكونة من n غرض بدون إعطاء الترتيب أية أهمية. عدد هكذا توافيق يرمز له بـ $C(n, r)$ ، المعرف كالتالي:

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

ملاحظة: لكل $0 \leq r \leq n$ ، يكون لدينا $P(n, r) \geq C(n, r)$

٣.٥ الاحتمال الشرطي

□ **قاعدة بايز** - إذا كانت لدينا الأحداث A و B بحيث $P(B) > 0$ ، يكون لدينا:

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

ملاحظة: لدينا $P(A \cap B) = P(A)P(B|A) = P(A|B)P(B)$

□ **القسم** - ليكن $\{A_i, i \in [1, n]\}$ بحيث لكل i لدينا $A_i \neq \emptyset$. نقول أن $\{A_i\}$ قسم إذا كان لدينا:

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

ملاحظة: لأي حدث B في فضاء العينة، لدينا $P(B) = \sum_{i=1}^n P(B|A_i)P(A_i)$

□ **النسخة الموسعة من قاعدة بايز** - ليكن $\{A_i, i \in [1, n]\}$ قسم من فضاء العينة. لدينا:

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

□ **الاستقلال** - يكون حدثين A و B مستقلين إذا وفقط إذا كان لدينا:

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

٤.٥ المتحولات العشوائية

□ **المتحول العشوائي** - المتحول العشوائي، ويرمز له عادة بـ X ، هو دالة تربط كل عنصر في فضاء العينة إلى خط الأعداد الحقيقية.

□ **دالة التوزيع التراكمي (CDF)** - تعرف دالة التوزيع التراكمي F ، والتي تكون غير متناقصة بشكل رتيب وتحقق $\lim_{x \rightarrow -\infty} F(x) = 0$ و $\lim_{x \rightarrow +\infty} F(x) = 1$ ، كالتالي:

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

ملاحظة: لدينا $P(a < X \leq b) = F(b) - F(a)$

□ **دالة الكثافة الإحتمالية (PDF)** - دالة الكثافة الاحتمالية f هي احتمال أن يأخذ X قيمة بين قيمتين متجاورتين من قيم المتحول العشوائي.

□ **علاقات تتضمن دالة الكثافة الاحتمالية ودالة التوزيع التراكمي** - هذه بعض الخصائص التي من المهم معرفتها في الحالتين المتقطعة (D) والمستمرة (C).

الحالة	دالة التوزيع التراكمي F	دالة الكثافة الاحتمالية f	خصائص دالة الكثافة الاحتمالية
(D)	$F(x) = \sum_{x_i \leq x} P(X = x_i)$	$f(x_j) = P(X = x_j)$	$\sum_j f(x_j) = 1$ و $0 \leq f(x_j) \leq 1$
(C)	$F(x) = \int_{-\infty}^x f(y)dy$	$f(x) = \frac{dF}{dx}$	$f(x) \geq 0$ و $\int_{-\infty}^{+\infty} f(x)dx = 1$

□ **التباين** - تباين متحول عشوائي، والذي يرمز له عادةً بـ $\text{Var}(X)$ أو σ^2 ، هو مقياس لانتشار دالة توزيع هذا المتحول. يحسب بالشكل التالي:

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

□ **الانحراف المعياري** - الانحراف المعياري لمتحول عشوائي، والذي يرمز له عادةً بـ σ ، هو مقياس لانتشار دالة توزيع هذا المتحول بما يتوافق مع وحدات قياس المتحول العشوائي. يحسب بالشكل التالي:

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

الحالة	الكثافة الهامشية	الدالة التراكمية
(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)	$f_X(x) = \int_{-\infty}^{+\infty} f_{XY}(x, y) dy$	$F_{XY}(x, y) = \int_{-\infty}^x \int_{-\infty}^y f_{XY}(x', y') dx' dy'$

□ **الاستقلال** - يقال عن متحولين عشوائيين X و Y أنهما مستقلين إذا كان لدينا:

$$\psi_{X+Y}(\omega) = \psi_X(\omega) \times \psi_Y(\omega)$$

□ **التغاير** - نعرف تغاير متحولين عشوائيين X و Y ، والذي نرمز له بـ σ_{XY}^2 أو بالرمز الأكثر شيوعاً $\text{Cov}(X, Y)$ ، كالتالي:

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

□ **الارتباط** - بأخذ σ_X, σ_Y كانحراف معياري لـ X و Y ، نعرف الارتباط بين المتحولات العشوائية X و Y ، والمرمز بـ ρ_{XY} ، كالتالي:

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

ملاحظة ١: لأي متحولات عشوائية X, Y لدينا $\rho_{XY} \in [-1, 1]$.

ملاحظة ٢: إذا كان X و Y مستقلين، فإن $\rho_{XY} = 0$.

□ **التوزيعات الأساسية** - فيما يلي التوزيعات الأساسية لأخذها بالاعتبار:

النوع	التوزيع	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
	$X \sim \text{Poisson}(\mu)$ Poisson	$P(X = x) = \frac{\mu^x}{x!} e^{-\mu}$ $x \in \mathbb{N}$	$e^{\mu(e^{i\omega} - 1)}$	μ	μ
(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	$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$ $x \in \mathbb{R}$	$e^{i\omega\mu - \frac{1}{2}\omega^2\sigma^2}$	μ	σ^2
	$X \sim \text{Exp}(\lambda)$ Exponential	$f(x) = \lambda e^{-\lambda x}$ $x \in \mathbb{R}_+$	$\frac{1}{1 - \frac{i\omega}{\lambda}}$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$

□ **التوقع وعزوم التوزيع** - فيما يلي المصطلحات المستخدمة للتعبير عن القيمة المتوقعة $E[X]$ ، الصيغة العامة للقيمة المتوقعة $E[g(X)]$ ، العزم رقم k $E[X^k]$ ودالة السمة $\psi(\omega)$ للحالات المتقطعة والمستمرة:

الحالة	$E[X]$	$E[g(X)]$	$E[X^k]$	$\psi(\omega)$
(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)	$\int_{-\infty}^{+\infty} x f(x) dx$	$\int_{-\infty}^{+\infty} g(x) f(x) dx$	$\int_{-\infty}^{+\infty} x^k f(x) dx$	$\int_{-\infty}^{+\infty} f(x) e^{i\omega x} dx$

□ **تحويل المتحولات العشوائية** - لتكن المتحولات العشوائية X و Y مرتبطة من خلال دالة ما. باعتبار f_X و f_Y دالتا التوزيع لـ X و Y على التوالي، يكون لدينا:

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

□ **قاعدة لايبنتز (Leibniz) للتكامل** - لتكن g دالة لـ x وربما لـ c ، ولتكن a و b حدود قد تعتمد على c . يكون لدينا:

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

□ **مراجعة تشبيشيف (Chebyshev)** - ليكن X متحولاً عشوائياً قيمته المتوقعة تساوي μ . إذا كان لدينا $k, \sigma > 0$ ، سنحصل على المراجعة التالية:

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

٥.٥ المتغيرات العشوائية الموزعة اشتراكياً

□ **الكثافة الشرطية** - الكثافة الشرطية لـ X بالنسبة لـ Y ، والتي يرمز لها عادةً بـ $f_{X|Y}$ ، تعرف بالشكل التالي:

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

□ **الاستقلال** - يقال عن متحولين عشوائيين X و Y أنهما مستقلين إذا كان لدينا:

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

□ **الكثافة الهامشية والتوزيع التراكمي** - من دالة الكثافة الاحتمالية المشتركة f_{XY} ، لدينا:

٦.٥ تقدير المُدخَل (Parameter)

□ **العينة العشوائية** - العينة العشوائية هي مجموعة من n متحول عشوائي X_1, \dots, X_n والتي تكون مستقلة وموزعة تطابقاً مع X .

□ **المُقَدَّر** - المُقَدَّر هو دالة للبيانات المستخدمة ويستخدم لاستنباط قيمة مُدخل غير معلوم ضمن نموذج إحصائي.

□ **الانحياز** - انحياز مُقَدَّر $\hat{\theta}$ هو الفرق بين القيمة المتوقعة لتوزيع $\hat{\theta}$ والقيمة الحقيقية، كالتالي:

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

ملاحظة: يقال عن مُقَدَّر أنه غير منحاز عندما يكون لدينا $E[\hat{\theta}] = \theta$.

□ **متوسط العينة** - يستخدم متوسط عينة عشوائية لتقدير المتوسط الحقيقي μ لتوزيع ما، عادةً ما يرمز له بـ \bar{X} ويعرف كالتالي:

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$$

ملاحظة: متوسط العينة غير منحاز، أي $E[\bar{X}] = \mu$.

□ **تباين العينة** - يستخدم تباين عينة عشوائية لتقدير التباين الحقيقي σ^2 لتوزيع ما، والذي يرمز له عادةً بـ s^2 أو $\hat{\sigma}^2$ ويعرّف بالشكل التالي:

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

ملاحظة: تباين العينة غير منحاز، أي $E[s^2] = \sigma^2$.

□ **مبرهنة النهاية المركزية** - ليكن لدينا عينة عشوائية X_1, \dots, X_n والتي تتبع لتوزيع معطى له متوسط μ وتباين σ^2 ، فيكون:

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

٧.٥ الجبر الخطي وحساب التفاضل والتكامل

تمت الترجمة بواسطة زيد اليافعي. تمت المراجعة بواسطة أمجد الخطابي و مازن مليباري.

٨.٥ الرموز العامة

□ **متجه (vector)** - نرمز لـ $x \in \mathbb{R}^k$ متجه يحتوي على n مدخلات، حيث $x_i \in \mathbb{R}$ يعتبر المدخل رقم i .

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

□ **مصفوفة (Matrix)** - نرمز لـ $A \in \mathbb{R}^{m \times n}$ مصفوفة تحتوي على m صفوف و n أعمدة، حيث $A_{i,j}$ يرمز للمدخل في الصف i و العمود j .

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

ملاحظة: المتجه x المعروف مسبقاً يمكن اعتباره مصفوفة من الشكل $n \times 1$ والذي يسمى بـ مصفوفة من عمود واحد.

□ **مصفوفة الوحدة (Identity)** - مصفوفة الوحدة $I \in \mathbb{R}^{n \times n}$ تعتبر مصفوفة مربعة تحتوي على المدخل ١ في قطر المصفوفة و ٠ في بقية المدخلات:

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

ملاحظة: جميع المصفوفات من الشكل $A \in \mathbb{R}^{n \times n}$ فإن $A \times I = I \times A = A$.

□ **مصفوفة قطرية (diagonal)** - المصفوفة القطرية هي مصفوفة من الشكل

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

ملاحظة: نرمز كذلك لـ D بـ $\text{diag}(d_1, \dots, d_n)$.

٩.٥ عمليات المصفوفات

□ **ضرب المتجهات** - توجد طريقتين لضرب متجه بمتجه:

• ضرب داخلي: (inner product) لـ $x, y \in \mathbb{R}^n$ نستنتج:

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

• ضرب خارجي: (outer product) لـ $x \in \mathbb{R}^n, y \in \mathbb{R}^n$ نستنتج:

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

□ **مصفوفة** - متجه: ضرب المصفوفة $A \in \mathbb{R}^{n \times m}$ والمتجه $x \in \mathbb{R}^n$ ينتجه متجه من الشكل $x \in \mathbb{R}^n$ حيث:

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

حيث $a_{r,i}^T$ يعتبر متجه الصفوف و $a_{c,j}$ يعتبر متجه الأعمدة لـ A كذلك x_i يرمز لعناصر x .

□ ضرب مصفوفة ومصفوفة - ضرب المصفوفة $A \in \mathbb{R}^{n \times m}$ و $A \in \mathbb{R}^{n \times p}$ ينتج عنه المصفوفة $A \in \mathbb{R}^{n \times p}$ • $N(x+y) \leq N(x) + N(y)$ حيث أن :

• لأي عدد a فإن $N(ax) = |a|N(x)$

• $N(x) = 0 \implies x = 0$

لأي $x \in V$ المعايير الأكثر إستخداماً ملخصة في الجدول التالي:

المعيار	الرمز	التعريف	مثال للإستخدام
L^1 Manhattan,	$\ x\ _1$	$\sum_{i=1}^n x_i $	LASSO regularization
L^2 Euclidean,	$\ x\ _2$	$\sqrt{\sum_{i=1}^n x_i^2}$	Ridge regularization
L^p -norm, p	$\ x\ _p$	$\left(\sum_{i=1}^n x_i^p\right)^{\frac{1}{p}}$	Hölder inequality
L^∞ Infinity,	$\ x\ _\infty$	$\max_i x_i $	Uniform convergence

□ الارتباط الخطي (Linear Dependence) - مجموعة المتجهات تعتبر تابعة خطياً إذا فقط إذا كل متجه يمكن كتابته بشكل خطي بإستخدام مجموعة من المتجهات الأخرى.

ملاحظة: إذا لم يتحقق هذا الشرط فإنها تسمى مستقلة خطياً.

□ رتبة المصفوفة (Rank) - رتبة المصفوفة A يرمز له ب $\text{rank}(A)$ وهو يصف حجم الفضاء المتجهي الذي نتج من أعمدة المصفوفة. يمكن وصفه كذلك بأقصى عدد من أعمدة المصفوفة A التي تمتلك خاصية أنها مستقلة خطياً.

□ مصفوفة شبه معرفة موجبة (Positive semi-definite) - المصفوفة $A \in \mathbb{R}^{n \times n}$ تعتبر مصفوفة شبه معرفة موجبة (PSD) ويرمز لها بالرمز $A \succeq 0$ إذا :

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

ملاحظة: المصفوفة A تعتبر مصفوفة معرفة موجبة إذا $A \succ 0$ وهي تعتبر مصفوفة (PSD) والتي تستوفي الشرط: لكل متجه غير الصفر x حيث $x^T A x > 0$.

□ القيم الذاتية (eigenvalue)، المتجه الذاتي (eigenvector) - إذا كان لدينا مصفوفة $A \in \mathbb{R}^{n \times n}$ ، القيمة λ تعتبر قيمة ذاتية للمصفوفة A إذا وجد متجه $z \in \mathbb{R}^n$ {0} يسمى متجه ذاتي حيث أن :

$$Az = \lambda z$$

□ النظرية الطيفية (spectral theorem) - نفرض $A \in \mathbb{R}^{n \times n}$ إذا كانت المصفوفة A متماثلة فإن A تعتبر مصفوفة قطرية بإستخدام مصفوفة متعامدة (orthogonal) $U \in \mathbb{R}^{n \times n}$ ويرمز لها بالرمز $\Lambda = \text{diag}(\lambda_1, \dots, \lambda_n)$ حيث أن:

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

حيث $a_{r,i}^T$ و $b_{r,i}^T$ يعتبر متجه الصفوف $a_{c,j}$ و $b_{c,j}$ متجه الأعمدة ل A و B على التوالي.

□ المنقول (Transpose) - منقول المصفوفة $A \in \mathbb{R}^{m \times n}$ يرمز له ب A^T حيث الصفوف يتم تبديلها مع الأعمدة :

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

ملاحظة: لأي مصفوفتين A و B ، نستنتج $(AB)^T = B^T A^T$.

□ المعكوس (Inverse) - معكوس أي مصفوفة A قابلة للعكس (Invertible) يرمز له ب A^{-1} ويعتبر المعكوس المصفوفة الوحيدة التي لديها الخاصية التالية :

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

ملاحظة: ليس جميع المصفوفات يمكن إيجاد معكوس لها. كذلك لأي مصفوفتين A و B نستنتج $(AB)^{-1} = B^{-1}A^{-1}$.

□ أثر المصفوفة (Trace) - أثر أي مصفوفة مربعة A يرمز له ب $\text{tr}(A)$ يعتبر مجموع العناصر التي في القطر:

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

ملاحظة: لأي مصفوفتين A و B لدينا $\text{tr}(AB) = \text{tr}(BA)$ و $\text{tr}(A^T) = \text{tr}(A)$.

□ المحدد (Determinant) - المحدد لأي مصفوفة مربعة من الشكل $A \in \mathbb{R}^{n \times n}$ يرمز له ب $|A|$ او $\det(A)$ يتم تعريفه بإستخدام $A_{i,j}$ والذي يعتبر المصفوفة A مع حذف الصف i والعمود j كالتالي :

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

ملاحظة: A يكون لديه معكوز إذا فقط إذا $|A| \neq 0$. كذلك $|AB| = |A||B|$ و $|A^T| = |A|$.

١٠.٥ خواص المصفوفات

□ التفكيك المتماثل (Symmetric Decomposition) - المصفوفة A يمكن التعبير عنها بإستخدام جزئين متماثل (Symmetric) وغير متماثل (Antisymmetric) كالتالي :

$$A = \underbrace{\frac{A + A^T}{2}}_{\text{متماثل}} + \underbrace{\frac{A - A^T}{2}}_{\text{غير متماثل}}$$

□ المعيار (Norm) - المعيار يعتبر دالة $N: V \rightarrow [0, +\infty]$ حيث V يعتبر فضاء متجه (Vector Space)، حيث أن لكل $x, y \in V$ لدينا :

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

□ مجزئ القيمة المفردة (singular value decomposition) - لأي مصفوفة A من الشكل $n \times m$ ، تفكيك القيمة المنفردة (SVD) يعتبر طريقة تحليل تضمن وجود $U \in \mathbb{R}^{m \times m}$ ، مصفوفة قطرية $\Sigma \in \mathbb{R}^{m \times n}$ و $V \in \mathbb{R}^{n \times n}$ حيث أن:

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

١١.٥ حساب المصفوفات

□ المشتقة في فضاءات عالية (gradient) - افترض $f: \mathbb{R}^{m \times n} \rightarrow \mathbb{R}$ تعتبر دالة و $f: \mathbb{R}^{m \times n} \rightarrow \mathbb{R}$ تعتبر مصفوفة $n \times m$ يرمز له $\nabla_A f(A)$ حيث أن:

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

ملاحظة: المشتقة العليا معرفة فقط إذا كانت الدالة f لديها مدى ضمن الأعداد الحقيقية.

□ هيشيان (Hessian) - افترض $f: \mathbb{R}^n \rightarrow \mathbb{R}$ تعتبر دالة و $x \in \mathbb{R}^n$ يعتبر متجه. الهيشيان ل f بالنسبة ل x تعتبر مصفوفة متماثلة من الشكل $n \times n$ يرمز لها بالرمز $\nabla_x^2 f(x)$ حيث أن:

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

ملاحظة: الهيشيان معرفة فقط إذا كانت الدالة f لديها مدى ضمن الأعداد الحقيقية.

□ الحساب في مشتقة الفضاءات العالية - لأي مصفوفات A, B, C فإن الخواص التالية مهمة:

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

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

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

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