مرجع سريع للتعلّم المُوَجَّه

افشين عميدى و شروين عميدى

١٤ ربيع الثاني، ١٤٤١

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

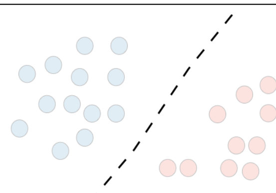
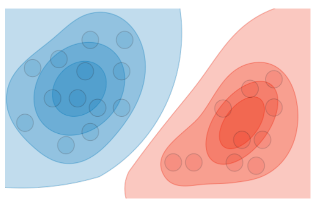
مقدمة للتعلّم المُوَجَّه

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

❒ **نوع التوقّع** – أنواع نماذج التوقّع المختلفة موضحة في الجدول التالي:

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

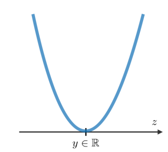
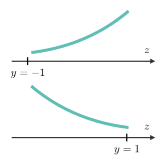
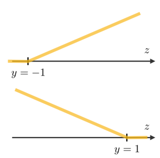
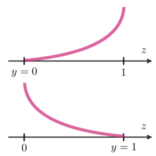
❒ **نوع النموذج** – أنواع النماذج المختلفة موضحة في الجدول التالي:

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

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 $L:(z,y)\in\mathbb{R}\times Y\longmapsto L(z,y)\in\mathbb{R}$ التي تأخذ كمدخلات القيمة المتوقعة z والقيمة الحقيقية y وتعطينا الاختلاف بينهما. الجدول التالي يحتوي على بعض دوال الخسارة الشائعة:

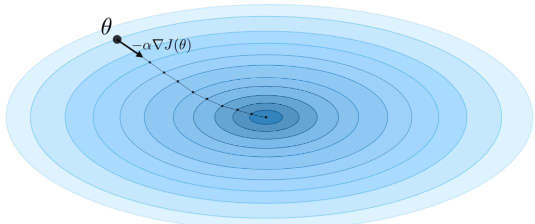
خطأ أصغر تربيع (Least squared error)	خسارة لوجستية (Logistic loss)	خسارة مفصلية (Hinge loss)	الانتروبيا التقاطعية (Cross-entropy)
$\frac{1}{2}(y-z)^2$	$\log(1+\exp(-yz))$	$\max(0,1-yz)$	$-\Big[y\log(z)+(1-y)\log(1-z)\Big]$
			
الانحدار الخطّي (Linear regression)	الانحدار اللوجستي (Logistic regression)	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:

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

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

$$\boxed{\theta\longleftarrow\theta-\alpha\nabla J(\theta)}$$



ملاحظة: في النزول الاشتقاقي العشوائي (*Stochastic gradient descent (SGD)*) يتم تحديث المُعاملات (*parameters*) بناءاً على كل عينة تدريب على حدة، بينما في النزول الاشتقاقي الحُزَمي (*batch gradient descent*) يتم تحديثها باستخدام حُزَم 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 اللوغاريثمية (log-likelihood) $\ell(\theta)=\log(L(\theta))$ حيث أنها أسهل في التحسين (optimize). فيكون لدينا:

$$\boxed{\theta^{\textrm{opt}}=\underset{\theta}{\textrm{arg max }}L(\theta)}$$

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

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

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

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

الانحدار الخطّي (Linear regression)

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

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

$$\boxed{\theta=(X^TX)^{-1}X^Ty}$$

❒ **خوارزمية أصغر معدل تربيع LMS** – إذا كان لدينا معدل التعلّم 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:

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

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 يَزِن كل عينة تدريب أثناء حساب دالة التكلفة باستخدام $w^{(i)}(x)$، التي يمكن تعريفها باستخدام المُدخل (parameter) $\tau\in\mathbb{R}$ كالتالي:

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

التصنيف والانحدار اللوجستي

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

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

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

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

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 بيرنوللي (Bernoulli) ϕ_i لكل فئة i يساوي:

$$\boxed{\phi_i=\frac{\exp(\theta_i^Tx)}{\displaystyle\sum_{j=1}^K\exp(\theta_j^Tx)}}$$

النماذج الخطية العامة (Generalized Linear Models - GLM)

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

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

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

تم تلخيص أكثر التوزيعات الأسيّة استخداماً في الجدول التالي:

التوزيع	η	$T(y)$	$a(\eta)$	$b(y)$
بِرنوللي (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}	$\displaystyle\frac{1}{y!}$
هندسي (Geometric)	$\log(1-\phi)$	y	$\log\left(\frac{e^\eta}{1-e^\eta}\right)$	1

❒ **افتراضات GLMs** – تهدف النماذج الخطيّة العامة (GLM) إلى توقع المتغير العشوائي y كدالة لـ $x \in \mathbb{R}^{n+1}$، وتستند إلى ثلاثة افتراضات:

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

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

آلة المتجهات الداعمة (Support Vector Machines)

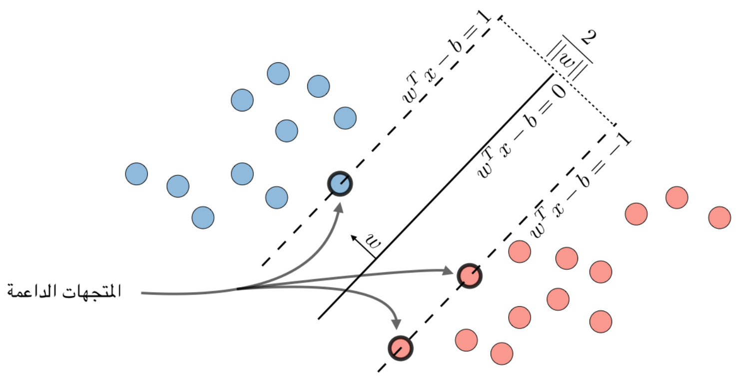
تهدف آلة المتجهات الداعمة (SVM) إلى العثور على الخط الذي يعظم أصغر مسافة إليه:

❒ **مُصنِّف الهامش الأحسن (Optimal margin classifier)** – يعرِّف مُصنِّف الهامش الأحسن h كالتالي:

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

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

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



ملاحظة: يتم تعريف الخط بهذه المعادلة $\boxed{w^T x - b = 0}$.

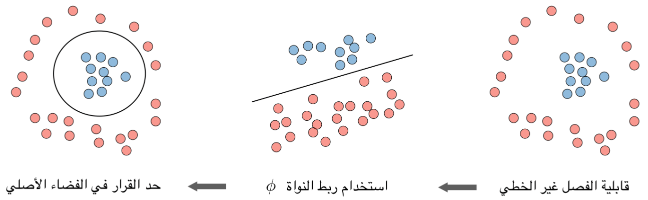
❒ **الخسارة المفصلية (Hinge loss)** – تستخدم الخسارة المفصلية في حل SVM ويعرف على النحو التالي:

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

❒ **النواة (Kernel)** – إذا كان لدينا دالة ربط الخصائص (features) ϕ، يمكننا تعريف النواة K كالتالي:

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

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



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 فقط أن نحسب القيم $K(x,z)$.

❒ **اللّاغرانجي (Lagrangian)** – يتم تعريف اللّاغرانجي $\mathcal{L}(w,b)$ على النحو التالي:

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

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

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).

تحليل التمايز الجاوسي (Gaussian Discriminant Analysis)

❒ **الإطار** – تحليل التمايز الجاوسي يفترض أن y و $x|y = 0$ و $x|y = 1$ بحيث يكونوا كالتالي:

$$\boxed{y \sim \textrm{Bernoulli}(\phi)}$$

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

❒ **التقدير** – الجدول التالي يلخص التقديرات التي يمكننا التوصل لها عند تعظيم الأرجحية (likelihood):

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

بايز البسيط (Naive Bayes)

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

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

❐ الحل – تعظيم الأرجحية اللوغاريثمية (log-likelihood) يعطينا الحلول التالية إذا كان $k \in \{0,1\}, l \in [\![1,L]\!]$:

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

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

الطرق الشجرية (tree-based) والتجميعية (ensemble)

هذه الطرق يمكن استخدامها لكلٍ 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 الخوارزمية المشهورة.

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

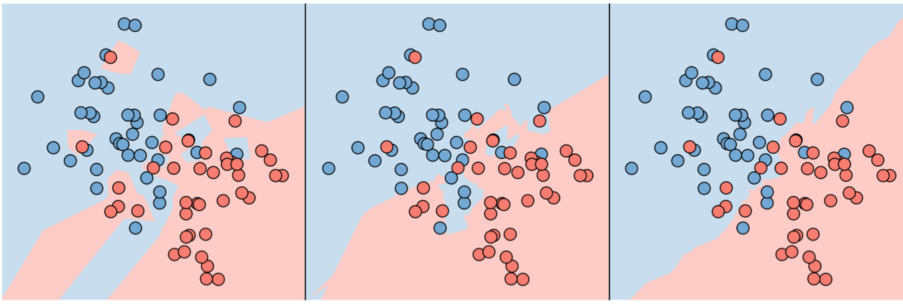
❐ **التعزيز (Boosting)** – فكرة خوارزميات التعزيز هي دمج عدة خوارزميات تعلم ضعيفة لتكوين نموذج قوي. الطرق الأساسية ملخصة في الجدول التالي:

التعزيز التَكَيُّفي (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.

طرق أخرى غير بارامترية (non-parametric)

❐ **خوارزمية أقرب الجيران (*k*-nearest neighbors)** – تعتبر خوارزمية أقرب الجيران، وتعرف بـ kNN، 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 المُدخل k، كلما زاد الانحياز *(bias)*، وكلما نقص k، زاد التباين *(variance)*.

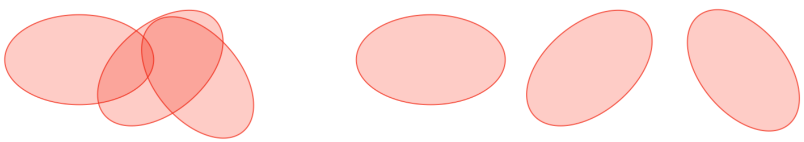


$k=1$ $k=3$ $k=11$

نظرية التعلُّم

❐ **حد الاتحاد (Union bound)** – لنجعل $A_1,...,A_k$ تمثل k حدث. فيكون لدينا:

$$\boxed{P(A_1\cup...\cup A_k)\leqslant P(A_1)+...+P(A_k)}$$



$A_1\cup A_2\cup A_3$ A_1 A_2 A_3

❐ **متراجحة هوفدينج (Hoeffding)** – لنجعل $Z_1,..,Z_m$ تمثل m متغير مستقلة وموزعة بشكل مماثل (iid) مأخوذة من توزيع برنولي (Bernoulli distribution) ذا مُدخل ϕ. لنجعل $\widehat{\phi}$ متوسط العينة (sample mean) و $\gamma>0$ ثابت. فيكون لدينا:

$$\boxed{P(|\phi-\widehat{\phi}|>\gamma)\leqslant 2\exp(-2\gamma^2m)}$$

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

❐ خطأ التدريب – ليكن لدينا المُصنِّف h، يمكن تعريف خطأ التدريب $\widehat{\epsilon}(h)$، ويعرف كذلك بالخطر التجريبي أو الخطأ التجريبي، كالتالي:

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

❐ **تقريباً صحيح احتمالياً (Probably Approximately Correct (PAC))** – هو إطار يتم من خلاله إثبات العديد من نظريات التعلم، ويحتوي على الافتراضات التالية:

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

❐ **مجموعة تكسيرية (Shattering Set)** – إذا كان لدينا المجموعة $S=\{x^{(1)},...,x^{(d)}\}$، ومجموعة مُصنِّفات $\mathcal{H}$، نقول أن $\mathcal{H}$ تكسر S (H shatters S) إذا كان لكل مجموعة علامات (labels) $\{y^{(1)}, ..., y^{(d)}\}$ لدينا:

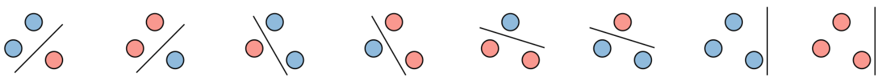
$$\boxed{\exists h\in\mathcal{H}, \quad \forall i\in[\![1,d]\!], \quad h(x^{(i)})=y^{(i)}}$$

❐ **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:

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

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

ملاحظة: بُعْد فابنيك تشرفونيكس VC لـ $\mathcal{H} =$ {مجموعة التصنيفات الخطية في بُعدين} يساوي ۳.



❐ **مبرهنة فابنيك (Vapnik theorem)** – ليكن لدينا $\mathcal{H}$، مع $\textrm{VC}(\mathcal{H})=d$ وعدد عيّنات التدريب m. سيكون لدينا، مع احتمال على الأقل $1-\delta$، التالي:

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