

Maldición de la dimensión & Aprendizaje de máquina

Erik Amézquita¹

`erik.amezquita@cimat.mx`

¹Departamento de Matemáticas, UG

Extensión del Conocimiento
18 de mayo 2018

Aprendizaje de Máquina (Machine Learning)

Ciencia de datos, Redes neuronales, Deep learning, Big data, Minería de datos, inteligencia artificial, Análisis topológico de datos,...



what society thinks I do



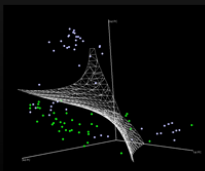
what my friends think I do



what my parents think I do

$$\begin{aligned} L_y &= \frac{1}{2} \|w\|^2 = \frac{1}{2} \sum_i a_i y_i (x_i \cdot w + b) + \sum_i a_i \\ a_i &\geq 0, \forall i \\ w &= \sum_i a_i y_i x_i, \sum_i a_i y_i = 0 \\ \nabla g(\theta_t) &= \frac{1}{n} \sum_{i=1}^n \nabla \ell(x_i, y_i; \theta_t) + \nabla r(\theta_t) \\ \theta_{t+1} &= \theta_t - \eta_t \nabla \ell(x_i(t), y_i(t); \theta_t) - \eta_t \cdot \nabla r(\theta_t) \\ \mathbb{E}_{i(t)} [\ell(x_i(t), y_i(t); \theta_t)] &= \frac{1}{n} \sum_i \ell(x_i, y_i; \theta_t) \end{aligned}$$

what other programmers think I do



what I think I do

```
>>> from scipy import SVM
```

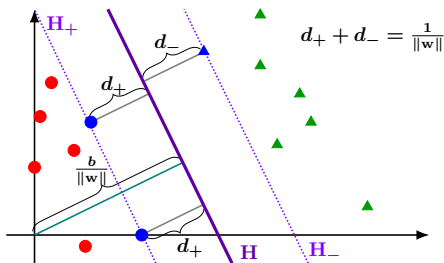
what I really do



Pero hay muchas cosas que pueden salir mal...

Máquinas de Soporte Vectorial (SVM)

Queremos dividir el plano en dos de modo que cada lado corresponde a un grupo distinto.



Nuestros objetos los transformamos de algún modo en puntos de alguna dimensión.

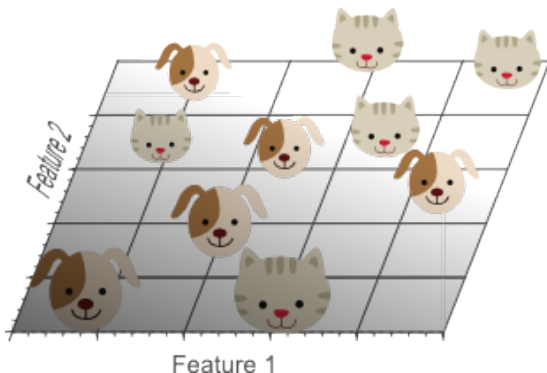
Clasificar perros y gatos en 1D

- 1 número por imagen
- Veo los datos en la recta real
- Está difícil partir la línea en 2 pedazos, con perros a la izquierda y gatos a la derecha.



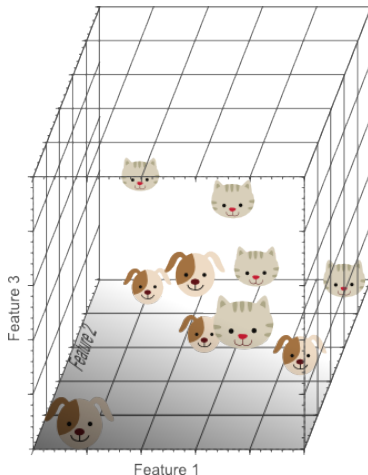
Clasificar perros y gatos en 2D

- 2 números por imagen
- Veo los objetos en el plano
- Sigue difícil partir el plano en 2 pedazos, con perros a la izquierda y gatos a la derecha.

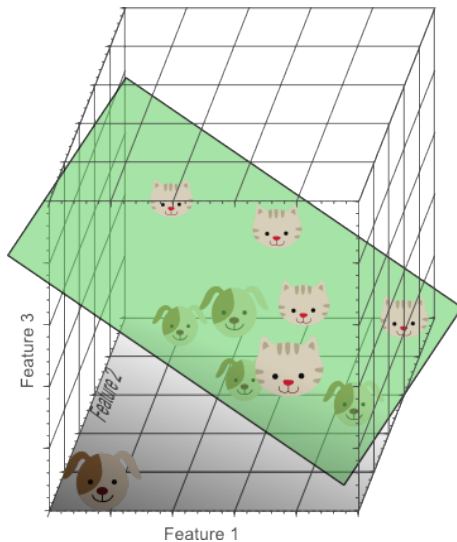


Clasificar perros y gatos en 3D

- 3 números por imagen
- Vemos los objetos en el espacio
- Ahora sí se puede dividir el espacio en 2 pedazos bonitos



¡Clasificar perros y gatos en 3D: Sí se puede!



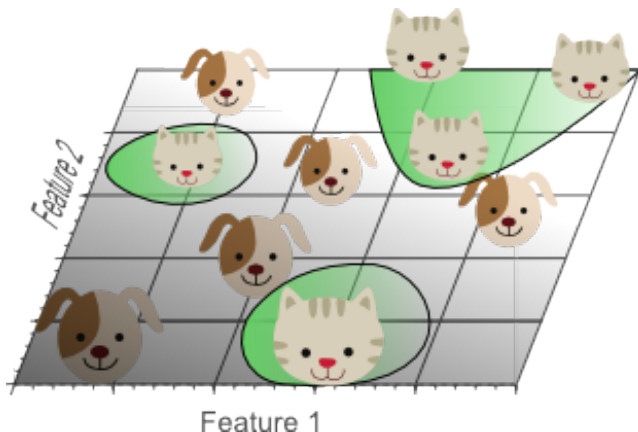
¿Seguimos aumentando dimensiones hasta clasificar perfecto? NO

A mayor dimensión, las cosas empiezan a bailar la macarena: es la **Maldición de la Dimensión**

- Concentración de medida
- Datos muy muy raros (dispersos)
- La geometría juega trucos raros
- Nuestra intuición deja de funcionar

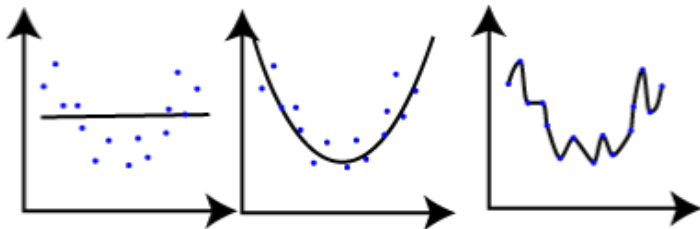
Maldición 1: Sobreajuste

- Esto sólo sirve para nuestro conjunto de datos específico.
- Seguro fracasa con datos nuevos



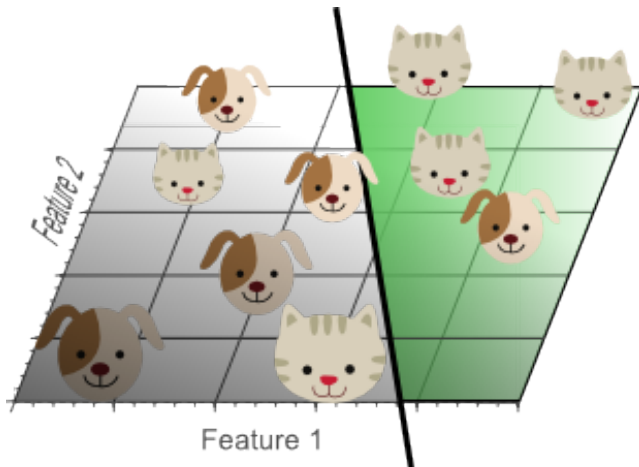
Sobreajuste \equiv complicar las cosas

- Debemos tener en cuenta que siempre ocurren errores
- Se debe buscar la imagen más simple que de una idea general de los datos
- Tira y encoge entre simplicidad y ajustar bien



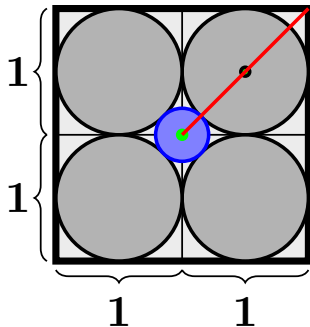
A veces no se conseguirá el ajuste perfecto

- Nuestros datos solo representan un pedazo pequeño de la realidad.
- Es mejor tener cierta flexibilidad para datos nuevos.
- modelo que use menos dimensiones es mejor para evitar caer en la maldición



La geometría de altas dimensiones está bien rara

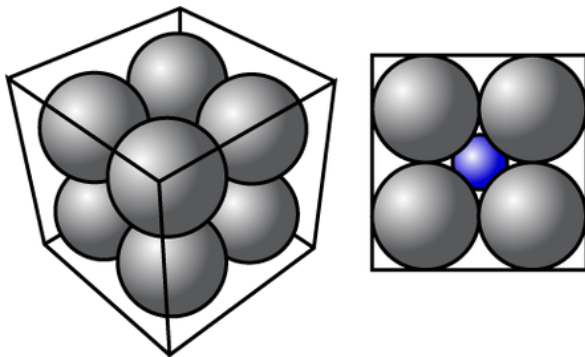
- Pensemos en un cuadrado de lado 2.
- Tenemos 4 círculos tangentes de diámetro 1 c/u.
- ¿Cuánto mide el radio r_2 del círculo tangente azul?



$$r_2 = \frac{\sqrt{2} - 1}{2} \approx 0.21$$

Ahora lo mismo en 3D

- Tenemos 8 esferas tangentes de diámetro 1 c/u.
- ¿Cuánto mide el radio de la esfera azul tangente a las otras 8?



$$r_3 = \frac{\sqrt{3} - 1}{2} \approx 0.37$$

Maldición 2: Nuestra intuición truena

En 9D con $2^9 = 512$ esferas vemos que la esfera azul toca el borde del cubo

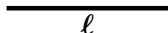
$$r_9 = \frac{\sqrt{9} - 1}{2} = 1$$

En 10D y para arriba la esfera azul atraviesa el borde del cubo


$$r_{10} = \frac{\sqrt{10} - 1}{2} \approx 1.08$$

Volumen en general

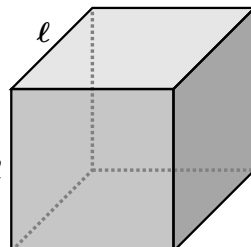
Longitud \equiv área \equiv volumen (no lo usen en su tarea)



$\text{Vol}(\text{---}_\ell) = \ell$



$\text{Vol}(\square_\ell) = \ell \times \ell = \ell^2$



$\text{Vol}(\text{cube}_\ell) = \ell \times \ell \times \ell = \ell^3$

Volumen del hipercubo

Para el hipercubo de d dimensiones y lado ℓ , su volumen es

$$\text{Vol} \left(\text{cube}_{\ell}^d \right) = \overbrace{\ell \times \ell \times \cdots \times \ell}^{d \text{ veces}} = \ell^d.$$

Si se tiene que el hipercubo es unitario, $\ell = 1$, entonces

$$\text{Vol} \left(\text{cube}_1^d \right) = 1.$$

¿Cómo luce el hipercubo unitario de 100 dimensiones?

¿Cómo cubrir el 20 % de nuestros hipercubos unitarios?

- En dimensión d queremos hallar un lado ℓ tal que

$$(\ell)^d = 0.2$$

- Por ejemplo,

$$(0.20)^1 = 0.2 \quad (0.45)^2 = 0.2$$

$$(0.58)^3 = 0.2 \quad (0.98)^{100} = 0.2$$

- En 100D hay 5 cubos diferentes de lado $\ell = 0.98$ dentro de un cubo de lado 1.

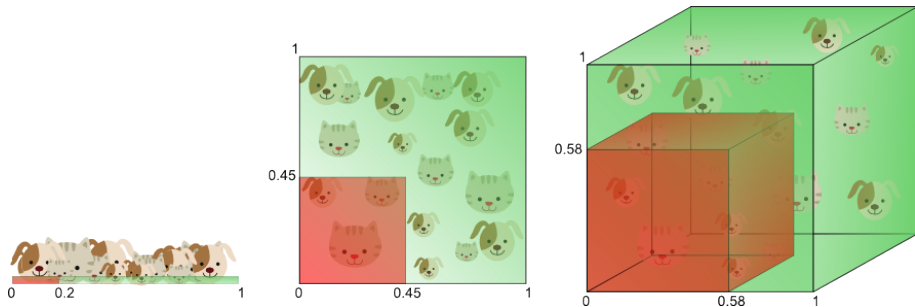
Maldición 3: Datos muy dispersos

1D: Necesitamos 20 puntos para cubrir el 20 %.

2D: Necesitamos $45^2 = 2025$ puntos para cubrir el 20 %.

3D: Necesitamos $58^3 \approx 200,000$ para cubrir el 20 %.

100D: ¡Necesitamos 98^{100} puntos para cubrir el 20 %!



¿Y el volumen de las hiperesferas?

- Pensemos en esferas de radio R

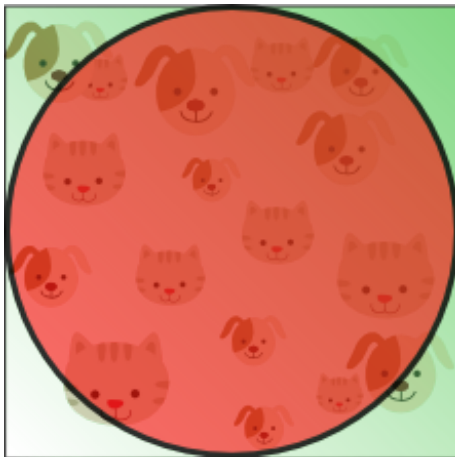
$$\text{Vol}(\bigcirc_R) = \pi R^2, \quad \text{Vol}(\bigodot_R) = \frac{4}{3}\pi R^3$$

- En general, las esferas desaparecen

$$\text{Vol}\left(\bigodot_R^d\right) = \frac{2\pi^{d/2}}{\Gamma(d/2)} R^d \xrightarrow{d \rightarrow \infty} 0$$

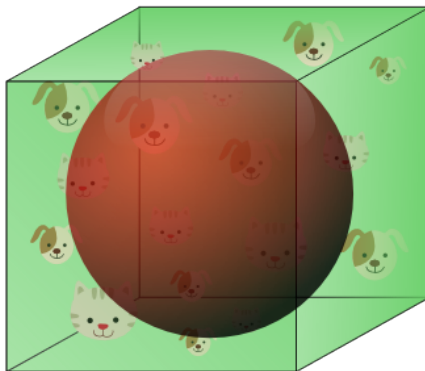
De vuelta a perros y gatos

- Pensemos en el círculo inscrito en el cuadrado unitario.
- Las 4 esquinas son feas y queremos estar en el círculo.



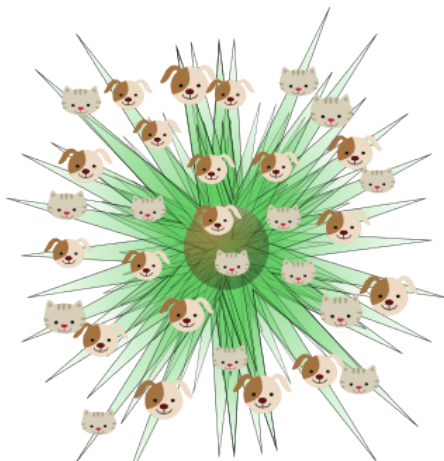
En 3D hay más espacio en las esquinas

- Pensemos en la esfera inscrita en el cubo unitario
- Tenemos ahora 8 esquinas feas.
- Cada vez habrá más volumen en las esquinas.



Maldición 4: Todo se parece a todo

- En 8D hay $2^8 = 256$ esquinas diferentes que concentran el 98 % del volumen total.
- El cubo en 8D luce más bien como un erizo.
- Todos los datos están en esquinas y es difícil distinguirlos.



¿Cómo se curan las maldiciones?

- No hay una respuesta única y mágica.
- La mayoría de veces las curas, si existen, son artesanales.
- Hay enfoques estándar que pueden funcionar, pero debe de tenerse cuidado.
- *A veces la cura resulta peor que la enfermedad.*



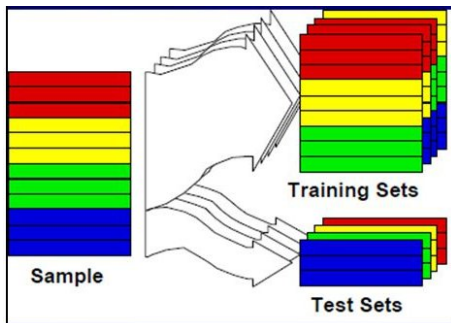
Pócima 1: Conseguir datos (casi) infinitos

- Una maldición es que a mayor dimensión, los datos son muchísimo más malos.
- Entre más datos tengamos, es más fácil ver ciertas tendencias
- El **problema** es que casi siempre es imposible.
- De hecho, la necesidad de datos crece exponencialmente



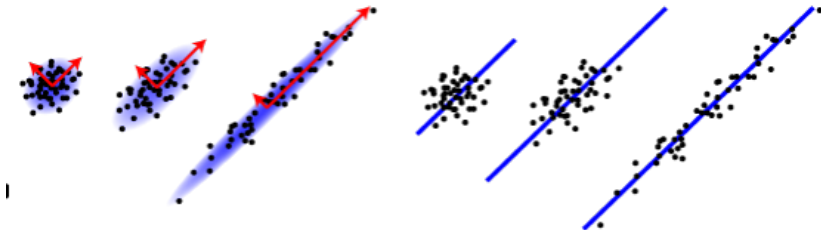
Póxima 2: Validación cruzada (Cross-validation)

- Partimos nuestros datos en dos: entrenamiento y prueba
- Tomamos el subconjunto de entrenamiento y definimos un patrón.
- Con la prueba verificamos que entrenamos bien.
- Volvemos a partir y repetimos muchas veces.
- Si tenemos pocos o malos datos, sólo nos engañamos a nosotros mismos.



Póxima 3: Análisis de Componentes Principales (PCA)

- Quizá no necesitamos tantas dimensiones.
- podemos deducir a través de unas dimensiones el resto.
- Reducimos dimensiones en función de mayor varianza.
- ¿Cómo sabemos que estamos midiendo las variables correctas en primer lugar?



Referencias



Vincent Spruyt *The Curse of Dimensionality* 2014. <http://www.visiondummy.com/2014/04/curse-dimensionality-affect-classification/>



Jesse Johnson *The curse of dimensionality* 2013 <https://shapeofdata.wordpress.com/2013/04/02/the-curse-of-dimensionality/>