Clase 4: Más Allá de la Linealidad

Matías Leoni

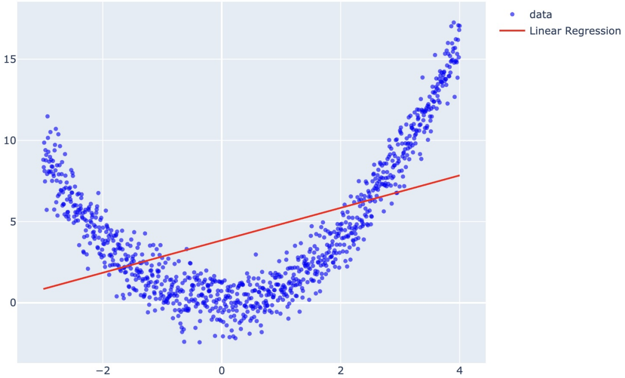
Curso de Aprendizaje Automático I
Maestría en IA - Universidad de San Andrés

8 de Julio de 2025

Agenda de la Clase (Revisada)

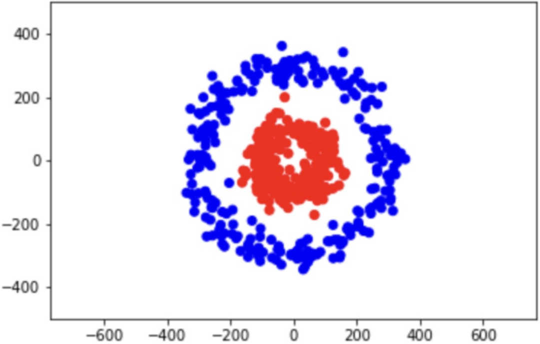
- **Parte 1:** Limitaciones de los modelos lineales.
- **Parte 2:** Estrategias para capturar no-linealidad.
  - Regresión Polinómica.
  - Otras familias de funciones: Splines, GAMs.
- **Parte 3:** Máquinas de Vectores de Soporte (SVM).
  - El clasificador de margen máximo.
  - El truco del Kernel y el Teorema de Mercer.
- **Parte 4:** Comparación de SVM vs. Regresión Logística.

¿Qué pasa cuando la realidad no es una línea recta?



El Supuesto Fundamental: Linealidad

- Los modelos lineales asumen que el límite de decisión (clasificación) o la tendencia (regresión) es una línea, un plano o un hiperplano.
- **Limitación principal:** Si la relación real entre las características y el resultado es no-lineal, el modelo tendrá un alto sesgo (bias) y un bajo poder predictivo.
- No puede capturar patrones complejos como curvas, círculos o interacciones intrincadas entre variables.

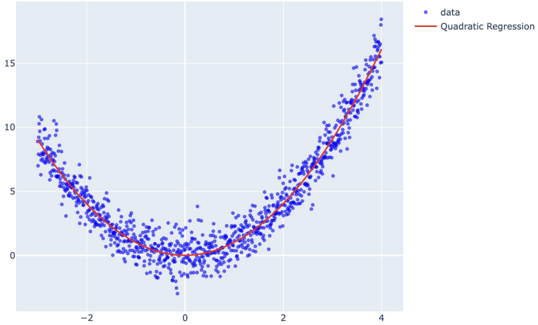


Si el modelo es simple, ¡compliquemos los datos!

- La idea es crear nuevas características (o usar funciones más complejas) para permitir que el andamiaje lineal capture patrones no lineales.
- Es como darle al modelo "nuevos ángulos" para ver el problema.
- $\implies$ **Feature Engineering**: Transformar los datos para que el modelo pueda aprender mejor.

Estrategia 1: Regresión Polinómica

- Añadimos potencias de las características originales como nuevas características.
- En lugar de: $y \approx \beta_0 + \beta_1 X$
- Ajustamos un modelo lineal a las características transformadas: $y \approx \beta_0 + \beta_1 X + \beta_2 X^2 + \cdots + \beta_d X^d$
- **Clave:** El modelo sigue siendo *lineal en los coeficientes* β_j.



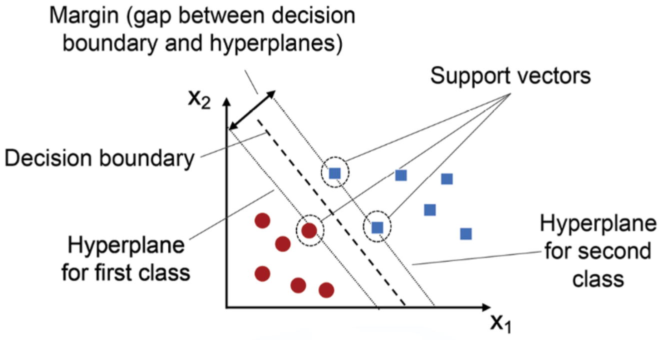
Estrategia 2: Otras Familias de Funciones

- La regresión polinómica es solo una opción. Existen métodos más avanzados y flexibles:
- **Funciones Escalonadas (Step Functions):**
  - Dividen el rango de una variable en K regiones y ajustan una constante en cada una. Útil para capturar saltos.
- **Splines:**
  - Ajustan polinomios de bajo grado (ej. cúbicos) en diferentes regiones de los datos y los "unen" suavemente en puntos llamados "nudos". Muy potentes y flexibles.
- **Modelos Aditivos Generalizados (GAMs):**
  - Permiten modelar la relación de y con cada variable X_j usando una función suave y no-lineal $f_j(X_j)$, y luego sumarlas.

$$y = \beta_0 + \sum_{j=1}^{p} f_j(X_j) + \epsilon$$

Encontrando la “calle” más ancha entre las clases.

- SVM es un clasificador que formaliza esta idea.
- Comencemos con el caso más simple: datos perfectamente separables por una línea.



El Clasificador de Margen Máximo

- Un hiperplano se define como: $b + w_1X_1 + \cdots + w_pX_p = 0$.
- Para clasificar, usamos el signo: $y_i \in \{-1, 1\}$.
- El problema de optimización consiste en **maximizar el margen**.

Problema de Optimización (Hard Margin)

$$\begin{aligned} \underset{\vec{w},b}{\text{minimizar}} \quad & |\vec{w}|^2 \\ \text{sujeto a} \quad & y_i(\vec{w}.\vec{x}_i + b) \geq 1, \quad \forall i \end{aligned}$$

- Los puntos que cumplen la igualdad $y_i(\dots) = 1$ son los **Vectores de Soporte**.

Clasificador de Vector Soporte (Margen Blando)

- ¿Qué pasa si los datos no son perfectamente separables? Permitimos algunas "violaciones" del margen.
- Introducimos **variables de holgura** ($\zeta_i \geq 0$) y un **coste de regularización** C.

Problema de Optimización (Soft Margin)

$$\begin{aligned} \underset{\vec{w},b}{\text{minimizar}} \quad & |\vec{w}|^2 + C\sum_{i=1}^{n}\zeta_i \\ \text{sujeto a} \quad & y_i(\vec{w}.\vec{x}_i + b) \geq 1 - \zeta_i, \quad \text{and} \quad \zeta_i \geq 0 \quad \forall i \end{aligned}$$

- C es un hiperparámetro que controla el trade-off sesgo-varianza.
- Si $C \to 0$, se permiten muchas violaciones para obtener una calle más ancha.
- Si $C \to \infty$, se recupera el problema **hard**.

¿Cómo pasamos a fronteras no lineales?

- El problema de optimización de SVM solo depende de los **productos punto** $\langle x_i, x_j \rangle$ entre observaciones.
- **Idea clave:** Sustituir el producto punto $\langle x_i, x_j \rangle$ por una función **Kernel** $K(x_i, x_j)$ que mida la similaridad en un espacio de mayor dimensión.
- Esto nos permite obtener una frontera no lineal en el espacio original, pero resolviendo el mismo problema de optimización lineal en el espacio transformado. ¡Sin calcular la transformación!

El Fundamento Matemático: Teorema de Mercer

Pregunta Fundamental

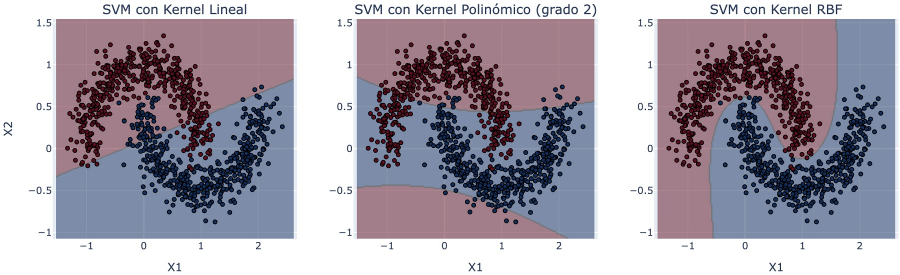
¿Puede cualquier función de similaridad $K(x_i, x_j)$ ser usada como un kernel válido? **La respuesta es NO.**

- **Condición de Mercer:** Una función K es un kernel admisible si y solo si la matriz del kernel (o Matriz de Gram), G, donde $G_{ij} = K(x_i, x_j)$, es **semidefinida positiva** para cualquier conjunto de puntos $\{x_1, \ldots, x_n\}$.
- **¿Por qué es importante? (Para mostrar en el pizarrón)**
  - Este teorema garantiza que si la condición se cumple, *existe* una transformación a un espacio de mayor dimensión $\phi(x)$ tal que $K(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle$.
  - Conecta la condición matemática (matriz semidefinida positiva) con la intuición geométrica (producto punto en un espacio transformado).

Kernels más Comunes

- **1. Kernel Lineal:**
  - Caso base, no hay transformación. Recuperamos el Clasificador de Vector Soporte.
  - $K(x_i, x_j) = x_i^T x_j$
- **2. Kernel Polinómico:**
  - Crea fronteras polinómicas de grado d.
  - $K(x_i, x_j) = (\gamma x_i^T x_j + r)^d$
- **3. Kernel de Función de Base Radial (RBF):**
  - El más popular y flexible. Puede crear fronteras muy complejas. La similaridad decae con la distancia euclidiana.
  - $K(x_i, x_j) = \exp(-\gamma ||x_i - x_j||^2)$

Visualizando el Poder de los Kernels



Mismo algoritmo (SVM), diferentes kernels, resultados drásticamente distintos.

SVM vs. Regresión Logística

- Ambos son excelentes algoritmos de clasificación.
- La elección depende de la naturaleza del problema y los datos.
- ¿Cuándo elegir uno sobre el otro?

Tabla Comparativa: SVM vs. Regresión Logística

Característica	**Regresión Logística**	**SVM**
Frontera de Decisión	Lineal.	Lineal o No-lineal (vía kernels).
Salida del Modelo	Produce probabilidades directas ($P(Y=1\|X)$).	Produce "scores" (distancia al hiperplano).
Sensibilidad	Sensible a todos los puntos.	Solo depende de los vectores de soporte. Menos sensible a outliers.
Rendimiento	Ideal si la frontera es lineal. Rápido y eficiente.	Potente para fronteras complejas y no lineales.
Interpretabilidad	Alta (coeficientes β interpretables).	Baja (especialmente con kernels no lineales, es "caja negra").
Hiperparámetros	Regularización (ej. C).	Regularización (C) y parámetros del kernel (ej. γ, d).

Guía Rápida: ¿Cuándo usar cuál?

Use Regresión Logística cuando...

- Necesita probabilidades como salida.
- La interpretabilidad del modelo es crucial.
- Cree que la frontera de decisión es (aproximadamente) lineal.
- Tiene un dataset muy grande y necesita velocidad.

Use SVM cuando...

- La máxima precisión predictiva es el objetivo.
- Sospecha que la frontera de decisión es compleja y no-lineal.
- Tiene un número alto de características (features).
- El dataset no es excesivamente grande (por coste computacional).

Conclusiones Clave

- Los modelos lineales son una base fantástica, pero suponen linealidad.
- Podemos capturar no-linealidad **explícitamente** (polinomios, splines) o **implícitamente**.
- Las **SVM con kernels** lo hacen de forma implícita y elegante, encontrando fronteras complejas al maximizar el margen en un espacio de alta dimensión.
- La elección del modelo (Regresión Logística vs. SVM) implica un trade-off fundamental entre **interpretabilidad y poder predictivo**.

Próximos Pasos

- **Taller Práctico:**
  - Aplicaremos SVM a un dataset real.
  - Visualizaremos el efecto de los diferentes kernels (lineal, RBF).
  - Compararemos su rendimiento con el de la Regresión Logística.
  - Pondremos a prueba el trade-off del parámetro C.
- **Próxima Clase Teórica:**
  - Árboles de Decisión. ¡Otra forma poderosa de abordar la no-linealidad!