

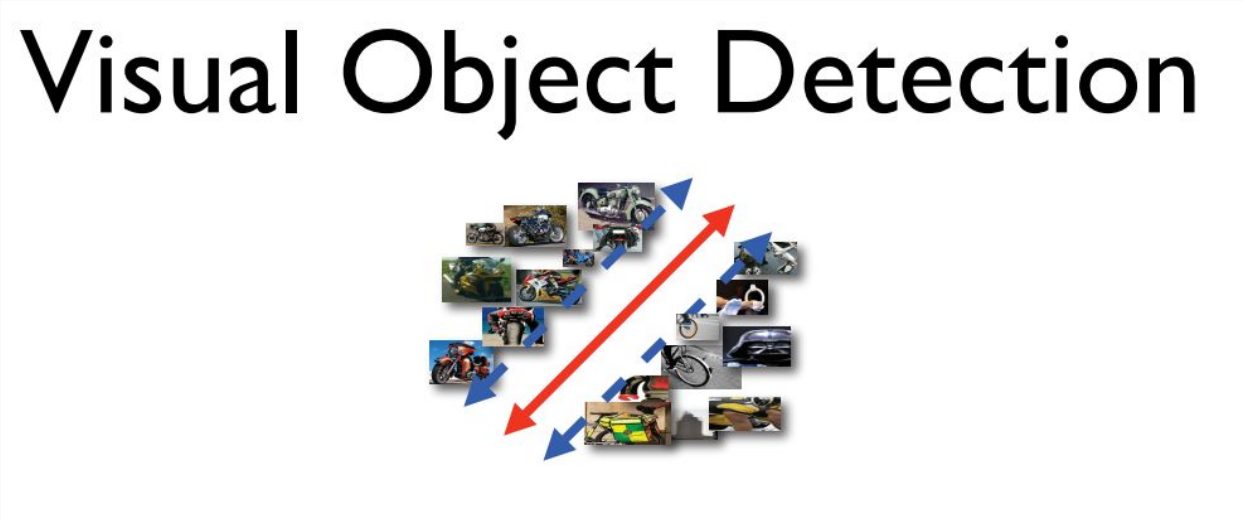
Ayudantía 11

Exemplar SVM y Reinforcement Learning (RL)

Por Benjamín Santander y Nicolás San Martín

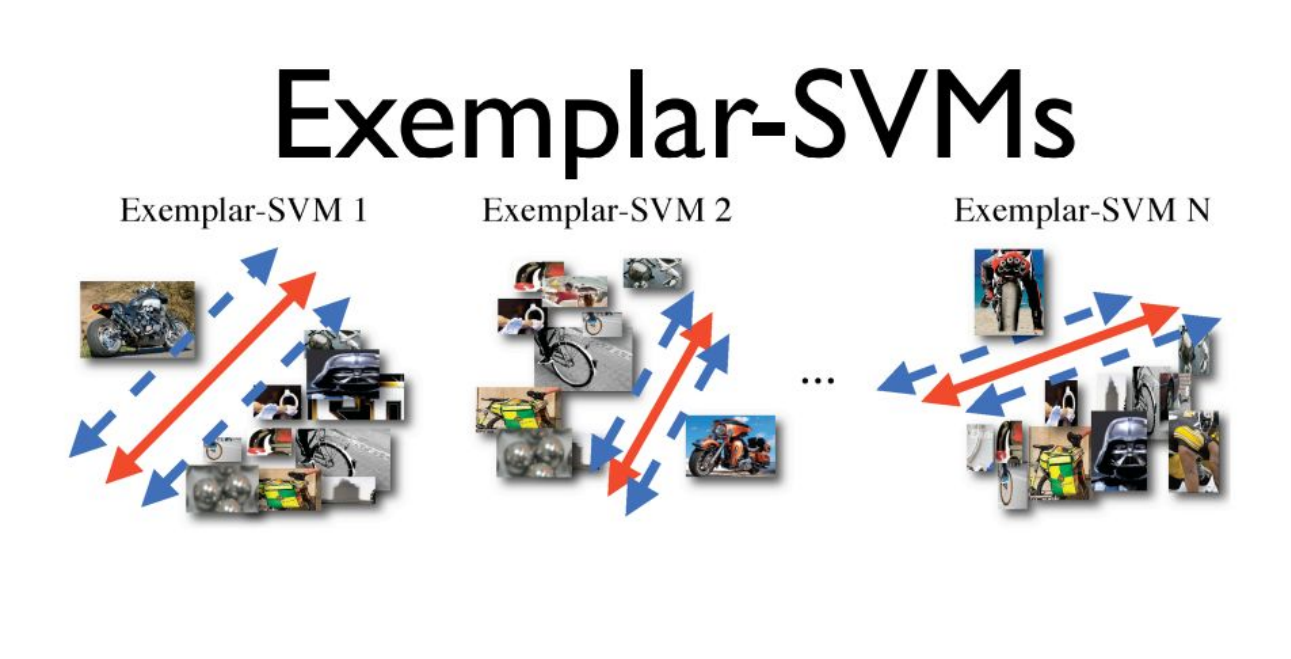
SVM Tradicional

Realizamos separaciones por conjuntos de datos para cada categoría



Exemplar SVM

Tomamos un ejemplo de esa categoría y lo enfrentamos a todos los demás valores



Como distintos modelos clasifican un objeto.



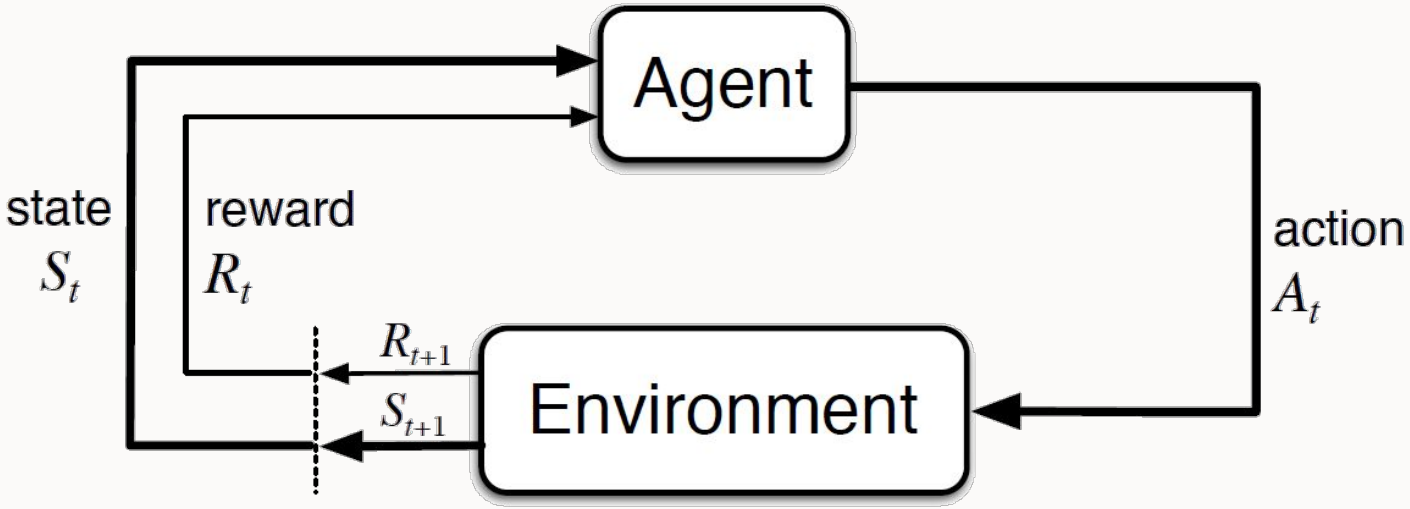


Ventajas y Desventajas

Aspecto	SVM Tradicional	Exemplar SVM
Objetivo	Clasificación general	Detección / búsqueda por ejemplo
Modelo	Un único clasificador	Un clasificador por ejemplo positivo
Generalización	Buena con datos balanceados	Limitada al ejemplo específico
Requiere muchos positivos	Sí	No (funciona con pocos o uno)
Costo computacional	Bajo a moderado	Alto (muchos modelos a entrenar)
Escalabilidad	Buena	Mala (no escala bien con muchos ejemplos)
Uso típico	Clasificación, reconocimiento de patrones	Recuperación visual, detección de objetos

¿Qué es?

Aprender a elegir acciones en un entorno tal que se maximice la **recompensa esperada**





¿Y el entorno?

Markov Decision Process (MDP)

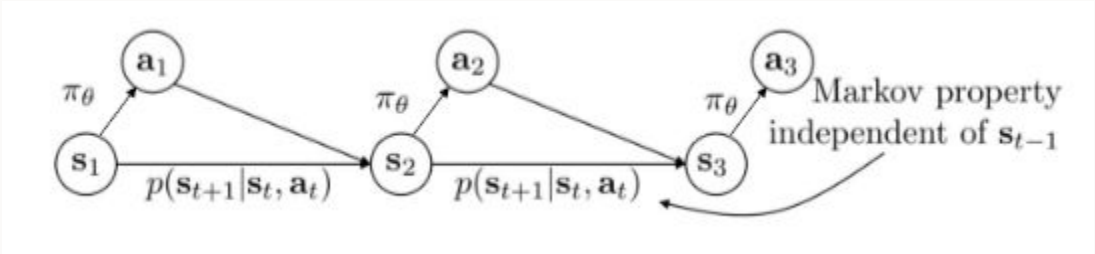
$$\mathcal{M} = \langle S, A, p, \gamma \rangle$$

- ***S*** → Conjunto de estados
- ***A*(s)** → Conjunto de acciones
- ***p*(s', r | s, a)** → Función de dinámica del entorno
- ***γ*** → Factor de descuento

Markov Decision Process (MDP)

Para aplicar reinforcement learning, realizamos un supuesto:

Propiedad Markoviana → **s'** y **r** dependen únicamente de **s** y **a**



Política de comportamiento (π)

El agente tiene como objetivo encontrar una política de comportamiento que le indique qué acción debe tomar en un instante determinado de un estado s_t

$$\pi_\theta(\mathbf{a}_t|\mathbf{s}_t)$$

Donde el objetivo es encontrar política una maximice la recompensa esperada

Value function

La *value function* es una estimación del retorno esperado que el agente puede obtener desde un cierto estado si sigue una política π

$$V^*(s) = \max_{\pi} \mathbb{E}\left[\sum_{t=0}^{H} \gamma^t R(s_t, a_t, s_{t+1}) \mid \pi, s_0 = s\right]$$

Función de valor del agente al iniciar en un estado *s* y seguir la política π

Tasa de descuento para recompensas más lejanas en el tiempo

Recompensa del agente en el instante de *t* al seguir la política π

Ejemplos solucionables con RL

Reinforcement Learning Lunar Lander



UArizona DataLab. (2024, 14 marzo). *Reinforcement Learning Lunar Lander* [Vídeo]. YouTube. https://www.youtube.com/watch?v=nQxmDH3mdCY

CartPole (Pole Balancing)



PilcoLearner. (2011, 26 mayo). *Cart-Pole swing-up* [Vídeo]. YouTube. https://www.youtube.com/watch?v=XiigTGKZfks