INTELIGENCIA ARTIFICIAL Y BIG DATA

Sistemas de Aprendizaje Automático

Tema 3 Porqué no usar k=1

Aprendizaje Supervisado
Clasificación

Aunque k=1 pueda generar la mejor tasa de aciertos en ciertos casos, en general **no se recomienda tomarlo como valor de configuración**.

Elegir k=1 implica riesgos significativos de sobreajuste y sensibilidad al ruido, lo que puede afectar el desempeño del modelo en datos reales.

Es importante considerar el contexto y la prioridad del proyecto.

Se debe valorar precisión vs. robustez y la forma en la que trabaja k-NN

Aprendizaje Supervisado
Clasificación

¿Cuándo ¡podría! considerarse utilizar K=1?:

En problemas con separación de clases muy clara:

- Si las clases están completamente separadas en el espacio de características y no hay ruido, K=1 puede ser una opción viable.

Con datos perfectamente limpios:

- En contextos controlados donde se sabe que no hay ruido ni outliers, K=1 puede funcionar bien.

Cuando la precisión es la única métrica importante:

- Si el objetivo es maximizar la precisión en el conjunto de datos actual y no se espera usar el modelo en otros escenarios.

Aprendizaje Supervisado
Clasificación

¿Cuándo NO es recomendable usar K=1?:

Cuando el conjunto de datos tiene ruido o outliers:

- K=1 es extremadamente sensible a datos atípicos, ya que solo considera el punto más cercano. Esto puede generar errores graves en datos no vistos.

Cuando la generalización es crítica:

- Aunque K=1 maximice la precisión en validación cruzada, puede fallar en generalizar bien a nuevos datos, especialmente si el conjunto de datos de entrenamiento no representa perfectamente el problema real.

Cuando hay un tamaño de muestra pequeño:

- Con pocos datos, el riesgo de que K=1 capte patrones específicos (en lugar de generales) es alto.

Aprendizaje Supervisado
Clasificación

¿Por qué seleccionar un K mayor aunque baje la tasa de aciertos?:

Mayor robustez:

- Valores más altos de K suavizan las predicciones al considerar múltiples vecinos, lo que reduce la influencia del ruido y los datos atípicos.

Mejor generalización:

- Aunque la precisión pueda bajar ligeramente, un K mayor tiende a capturar patrones más generales, aumentando la confianza en datos no vistos.

Evitar sobreajuste:

- K=1 tiende a ajustarse demasiado a los datos de entrenamiento. Un K mayor equilibra la flexibilidad del modelo con su capacidad de generalizar.

Compromiso razonable:

- Usar un valor como K=3 o K=5 sigue siendo lo suficientemente bajo como para capturar patrones locales, pero lo suficientemente alto para reducir el impacto del ruido.

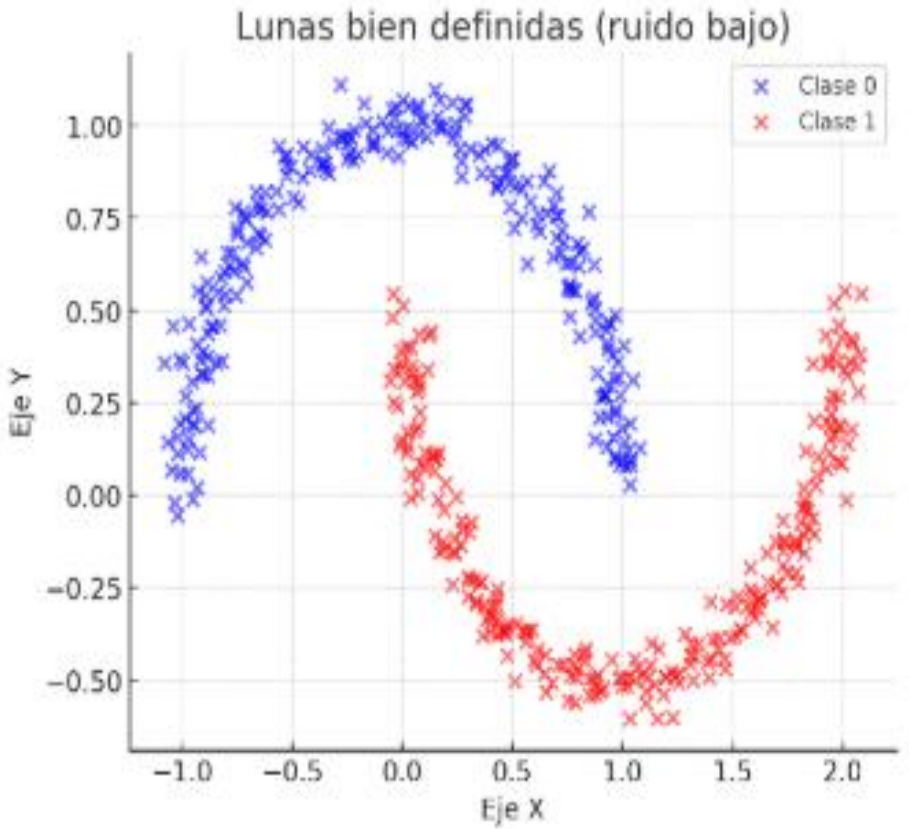
Aprendizaje Supervisado
Clasificación

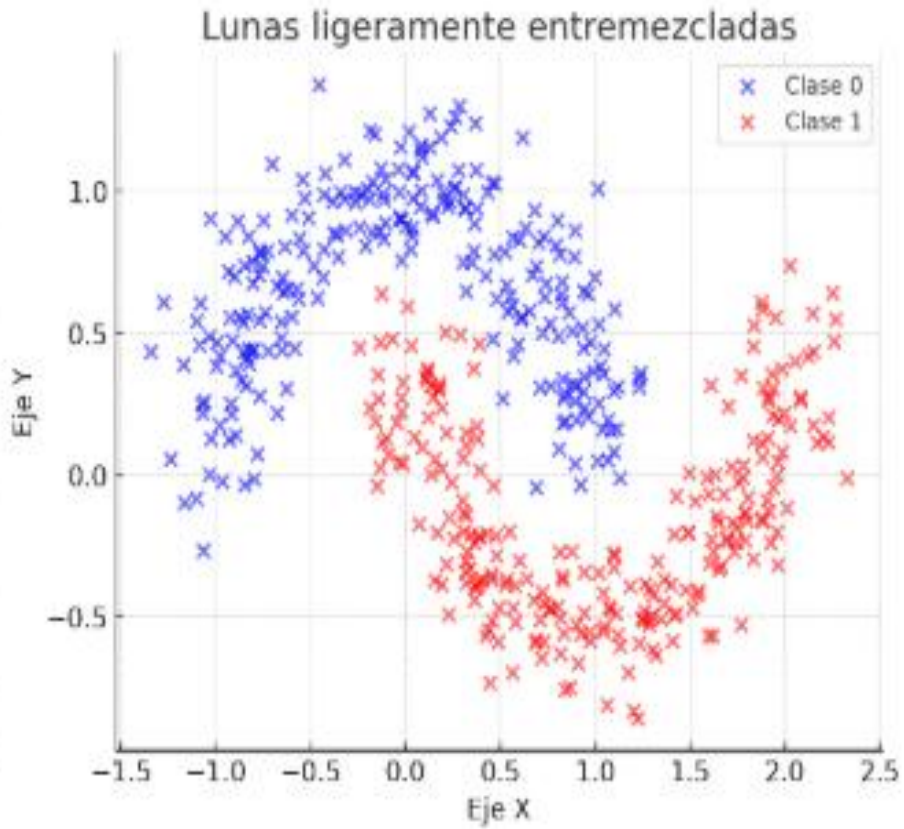
Recomendación general:

- **Se debe elegir un valor de K mayor que 1**, como K=3, K=5 o superior, en la mayoría de los casos. Aunque esto pueda reducir ligeramente la precisión en validación, suele producir un modelo más robusto y confiable para datos nuevos.
- Si se decide usar k=1, se debe de justificar con base en la naturaleza de los datos (por ejemplo, ausencia de ruido o clases bien separadas).

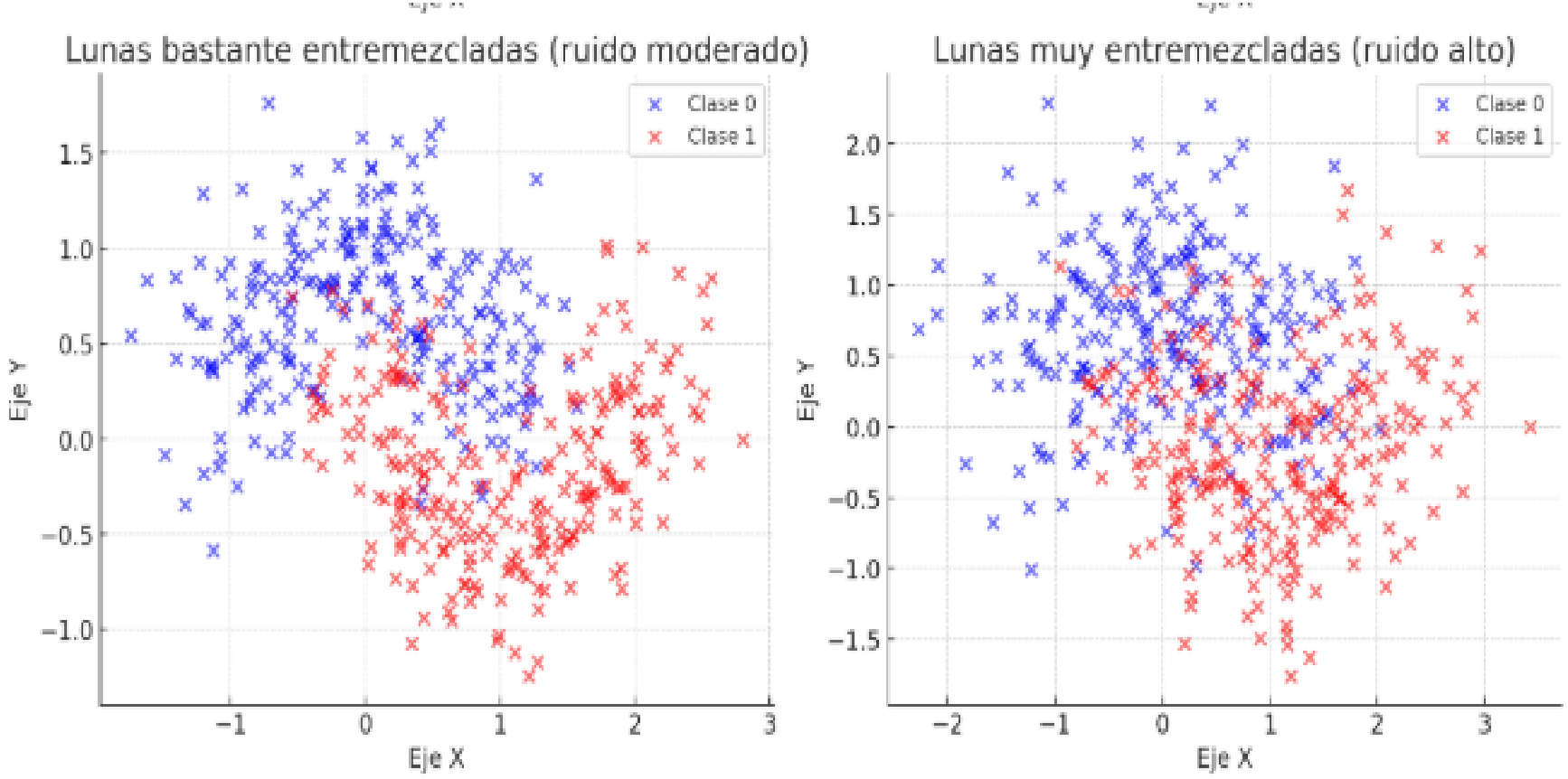
En resumen, el valor de k debe equilibrar precisión y generalización. Sacrificar un poco de precisión a favor de un modelo más robusto suele ser mejor práctica en la mayoría de los escenarios.

Aprendizaje Supervisado
Clasificación





Aprendizaje Supervisado
Clasificación



Aprendizaje Supervisado
Clasificación

¿Cómo debemos interpretar los resultados del conjunto de entrenamiento-prueba en el contexto del comportamiento general del modelo en producción?. Aunque los datos de prueba son efectivamente "nuevos" para el modelo durante la validación, existen razones sólidas para ser cautelosos al elegir k=1, incluso si tiene un rendimiento casi perfecto en ese conjunto.

1.El conjunto de prueba no representa toda la población posible:

1. Aunque los datos de prueba no han sido vistos por el modelo, siguen siendo parte del mismo conjunto de datos que los datos de entrenamiento. Esto significa que comparten la misma distribución y características.
2. En producción, los datos reales podrían contener variaciones o distribuciones diferentes a las observadas en el conjunto de prueba. Un modelo con k=1, que toma decisiones basadas en un único punto, no es capaz de adaptarse a estas variaciones.

2.Ruido y errores en datos futuros:

1. En un entorno de producción, los datos pueden incluir ruido, valores atípicos o errores.
2. Con k>1, las decisiones se suavizan porque se basan en un consenso de vecinos, lo que ayuda a mitigar los efectos de anomalías individuales.

Aprendizaje Supervisado
Clasificación

3.Tendencia al sobreajuste en conjuntos pequeños:

Un modelo con k=1 puede dar buenos resultados en un conjunto de prueba pequeño, pero esto no significa que sea generalizable. Esto ocurre porque el modelo puede aprovechar patrones muy específicos del conjunto de prueba que no son relevantes para nuevos datos.

En cambio, k>1 tiende a capturar tendencias generales en los datos, sacrificando precisión local en favor de una mayor robustez global.

4.Robustez frente a distribuciones cambiantes:

Los datos futuros podrían no seguir exactamente la misma distribución que los datos de prueba. Al depender de un solo vecino, k=1 es mucho más sensible a cualquier cambio en la distribución subyacente, lo que lo hace menos confiable en escenarios dinámicos.