Introducción a la Ciencia de Datos

Maestría en Probabilidad y Estadística

Dr. Marco Antonio Aquino López

Centro de Investigación en Matemáticas

Agosto–Diciembre 2025

Motivación: más allá de clasificadores lineales

- Los clasificadores lineales dividen el espacio con una **frontera recta** (una recta en 2D, un hiperplano en dimensiones mayores).
- Sin embargo, muchos conjuntos de datos no se separan bien con fronteras lineales.
- Queremos un método **no paramétrico**, que no suponga una forma fija para la frontera.
- Idea central: la etiqueta de un punto nuevo se puede deducir a partir de las etiquetas de sus **vecinos más cercanos**.
- Esto nos conduce al clasificador de **k-vecinos más cercanos (k-NN)**.

¿Qué significa “cercanía”?

- Para aplicar k-NN, debemos responder: **¿qué tan cerca están dos puntos?**
- La noción de distancia depende del tipo de variables:
  - Numéricas: diferencias en coordenadas.
  - Categóricas: coincidencia o no en las categorías.
- Distintas métricas generan diferentes “vecinos más cercanos”.

Métricas comunes en espacios numéricos

Sea $x = (x_1, \ldots, x_d)$ y $z = (z_1, \ldots, z_d)$ dos observaciones.

- **Distancia de Minkowski** (parámetro $p \geq 1$):

$$d_p(x, z) = \left(\sum_{j=1}^{d} |x_j - z_j|^p \right)^{1/p}.$$

  - $p = 2$: **Euclidiana**.
  - $p = 1$: **Manhattan**.
  - $p \to \infty$: **Chebyshev**.

- La elección de p cambia la forma de las regiones de cercanía.

Métricas para variables categóricas

- Si los datos no son numéricos, necesitamos otra definición de "cercanía".
- La más común es la **distancia de Hamming**:

$$d_H(x, z) = \#\{j : x_j \neq z_j\}.$$

- Cuenta el número de coordenadas (atributos) en las que difieren dos observaciones.
- Ejemplo: "color = rojo/azul/verde" o "sí/no" en encuestas.

k-Vecinos más cercanos: definición

El clasificador de k-vecinos más cercanos se basa en la idea de que **los puntos cercanos se parecen**.

Definición: Dado un punto del espacio x, encontramos los k puntos de entrenamiento $X^{(1)}, \ldots, X^{(k)}$ más cercanos a x. La clasificación $Y(x)$ se asigna como la clase más frecuente entre los Y_i correspondientes.

- Los empates se resuelven al azar.
- Típicamente se usa la distancia euclidiana.
- k pequeño: clasificador muy local.
- k grande: más robusto a errores en los datos.

Interpretación: la decisión se toma por *votación local*, usando la noción de cercanía definida previamente.

ver animación.

Clasificador k-NN

Formalmente, el clasificador de k-vecinos más cercanos se define como

$$g_n(x) = \begin{cases} 1 & \text{si } \sum_{i=1}^{n} w_{n,i} 1_{\{Y_i=1\}} > \sum_{i=1}^{n} w_{n,i} 1_{\{Y_i=0\}}, \\ 0 & \text{en otro caso,} \end{cases}$$

donde los pesos se definen por

$$w_{n,i} = \begin{cases} \frac{1}{k}, & \text{si } X_i \text{ es uno de los } k \text{ vecinos más cercanos de } x, \\ 0, & \text{en otro caso.} \end{cases}$$

Interpretación del clasificador k-NN

- La decisión se toma por *votación mayoritaria* entre los k vecinos más cercanos.
- Cada vecino contribuye con el mismo peso $1/k$.
- En caso de empate, se puede decidir al azar o con reglas adicionales (por ejemplo, elegir la clase más frecuente en todo el conjunto).

Intuición:

a medida que disponemos de más datos y el número de vecinos k se ajusta de forma adecuada, las regiones de votación local reflejan cada vez mejor la estructura real de las probabilidades de clase en el espacio de entrada. Es decir, el clasificador va capturando cómo se distribuyen las clases alrededor de cada punto.

Lema sobre el vecino más cercano

Sea $X^{(k)}(x)$ el k-ésimo vecino más cercano a un punto x.
Lema: Si $X \sim \mu$, con x en el soporte de μ y $k/n \to 0$, entonces

$$\|X^{(k)}(x) - x\| \longrightarrow 0 \quad \text{en probabilidad.}$$

Intuición:

- Cuando el número de datos n crece, siempre hay puntos de entrenamiento cada vez más cerca de x.
- Por lo tanto, los "vecinos más cercanos" se aproximan realmente al punto x.

Demostración del lema

Sea $S_{x,\varepsilon}$ la bola centrada en x con radio $\varepsilon > 0$.
Queremos mostrar que

$$\Pr\left(\|X^{(k)}(x) - x\| > \varepsilon\right) \longrightarrow 0.$$

Observación clave:

$$\|X^{(k)}(x) - x\| > \varepsilon \iff \frac{1}{n}\sum_{i=1}^{n} 1_{\{X_i \in S_{x,\varepsilon}\}} < \frac{k}{n}.$$

- El lado izquierdo dice: "ninguno de los k vecinos está dentro de la bola de radio ε".
- Esto ocurre si la proporción de puntos de la muestra dentro de $S_{x,\varepsilon}$ es menor que k/n.

Demostración del lema (cont.)

Por la ley de los grandes números,

$$\frac{1}{n}\sum_{i=1}^{n} 1_{\{X_i \in S_{x,\varepsilon}\}} \longrightarrow \mu(S_{x,\varepsilon}),$$

donde μ es la distribución de X.

- Como x está en el soporte de μ, se cumple que $\mu(S_{x,\varepsilon}) > 0$ para todo $\varepsilon > 0$.
- Además, si $k/n \to 0$, entonces $\frac{k}{n} \to 0$.
- Para n suficientemente grande se tendrá

$$\mu(S_{x,\varepsilon}) > \frac{k}{n}.$$

Por lo tanto, la probabilidad de que $\|X^{(k)}(x) - x\| > \varepsilon$ tiende a 0.

$$\|X^{(k)}(x) - x\| \longrightarrow 0 \quad \text{en probabilidad.}$$

Celdas de Voronoi

- Dado un conjunto de puntos de entrenamiento $X_1, \ldots, X_n$, podemos dividir todo el espacio en regiones.
- Cada región contiene todos los puntos del espacio más cercanos a un X_i que a cualquier otro.
- Estas regiones se llaman **celdas de Voronoi**.
- En el clasificador 1-NN, la clase de cualquier punto x se determina por la clase del centro de la celda de Voronoi en la que x cae.

ver animación

1-NN: el vecino más cercano

- En el clasificador 1-NN, cada punto de prueba x se etiqueta según la clase de su vecino más cercano $X^{(1)}(x)$.
- Geométricamente, esto induce una partición del espacio en **celdas de Voronoi**.
- Pregunta clave: ¿qué tan bueno es 1-NN en términos de error de clasificación?

Notación para analizar el error

- Definimos $\eta(x) = \Pr(Y = 1 \mid X = x)$.
- El clasificador de Bayes asigna la clase mayoritaria:

$$g^*(x) = \mathbf{1}\{\eta(x) > 1/2\}.$$

- Su error esperado es

$$L^* = \mathbb{E}\big[\min\{\eta(X), 1 - \eta(X)\}\big].$$

- Queremos comparar L_{1NN} (error de 1-NN) contra L^*.

Cálculo del error en 1-NN

Condicionemos en un punto fijo $X = x$:

- La etiqueta real es $Y \sim \text{Bernoulli}(\eta(x))$, donde $\eta(x) = \Pr(Y = 1 \mid X = x)$.
- La etiqueta del vecino más cercano se comporta como $Y^{(1)}(x) \sim \text{Bernoulli}(\eta(x))$, casi independiente de Y.

¿Cuándo se comete error? El clasificador falla cuando Y y $Y^{(1)}(x)$ no coinciden.

$$\Pr(Y^{(1)}(x) \neq Y \mid X = x) = \Pr(Y = 1, Y^{(1)} = 0) + \Pr(Y = 0, Y^{(1)} = 1).$$

Cálculo de cada caso:

$$\Pr(Y = 1, Y^{(1)} = 0) = \eta(x)\,[1 - \eta(x)], \quad \Pr(Y = 0, Y^{(1)} = 1) = [1 - \eta(x)]\,\eta(x).$$

Sumando:

$$\Pr(Y^{(1)}(x) \neq Y \mid X = x) = 2\,\eta(x)\,[1 - \eta(x)].$$

Finalmente, promediando sobre la distribución de X:

$$L_{1NN} = \mathbb{E}[\,2\,\eta(X)\,(1 - \eta(X))\,]\,.$$

Conclusión importante

- Comparando con L^*, se obtiene la famosa cota:

$$L^* \leq L_{1NN} \leq 2L^*.$$

- Interpretación:
  - 1-NN nunca es peor que el doble del error de Bayes.
  - Si las clases están bien separadas ($\eta(x) \approx 0$ o 1), entonces $L_{1NN} \approx L^*$.

Caso general: k impar

Para $k > 1$, el clasificador k-NN decide por votación mayoritaria entre los k vecinos más cercanos.

$$g_{kNN}(x) = \begin{cases} 1 & \text{si al menos la mitad de los } k \text{ vecinos tienen } Y = 1, \\ 0 & \text{en otro caso.} \end{cases}$$

El error en un punto x se calcula considerando todas las posibles combinaciones de etiquetas de los k vecinos.

Probabilidad de error en x

Sea $\eta(x) = \Pr(Y = 1 \mid X = x)$.
Cada vecino cercano se comporta como una variable Bernoulli($\eta(x)$).
Error en x:

$$L_{kNN}(x) = \Pr\big(\text{la votación mayoritaria difiere de la clase real} \mid X = x\big).$$

Esto equivale a sumar probabilidades de que la mayoría de los k vecinos den una etiqueta distinta a la más probable.

Expresión general del error

La probabilidad de error se puede escribir como

$$L_{kNN}(x) = \sum_{j=0}^{k} \binom{k}{j} \eta(x)^j \left(1-\eta(x)\right)^{k-j} \Big(\eta(x)\, \mathbf{1}_{\{j<k/2\}} + (1-\eta(x))\, \mathbf{1}_{\{j>k/2\}} \Big).$$

- El término $\binom{k}{j}\eta^j(1-\eta)^{k-j}$ es la probabilidad de que j vecinos tengan etiqueta 1.
- La parte final indica cuándo la votación produce error: si la mayoría vota en contra de la clase verdadera.

Cadena de cotas

Un resultado clásico establece que:

$$L^* \leq \cdots \leq L_{(2k+1)NN} \leq L_{(2k-1)NN} \leq \cdots \leq L_{3NN} \leq L_{1NN} \leq 2L^*.$$

- Al aumentar k (manteniéndolo impar), el error esperado disminuye.
- En el límite, si $k \to \infty$ y $k/n \to 0$, el clasificador es **consistente**, es decir, $L_{kNN} \to L^*$.

Variante: k-NN ponderado

- En el k-NN clásico, cada vecino contribuye con el mismo peso $1/k$.
- Una extensión es dar más importancia a los **vecinos más cercanos**.
- Definición general:

$$g_n(x) = \begin{cases} 1 & \text{si } \sum_{i=1}^{n} w_{n,i} 1_{\{Y_i=1\}} > \sum_{i=1}^{n} w_{n,i} 1_{\{Y_i=0\}}, \\ 0 & \text{en otro caso,} \end{cases}$$

donde los pesos $w_{n,i}$ dependen de la distancia entre x y X_i.

Esquemas de ponderación

- **Uniforme (clásico):** $w_{n,i} = 1/k$ si X_i está entre los k más cercanos.
- **Inverso a la distancia:**
$$w_{n,i} \propto \frac{1}{d(x, X_i)}.$$
- **Inverso al cuadrado de la distancia:**
$$w_{n,i} \propto \frac{1}{d(x, X_i)^2}.$$
- **Kernel local:** $w_{n,i} = K\left(\frac{d(x,X_i)}{h}\right)$ con K decreciente (ejemplo: gaussiano).

Idea: vecinos más cercanos aportan más evidencia, vecinos lejanos menos.

Bosquejo del algoritmo k-NN

Entrada:

- Conjunto de entrenamiento $\{(X_1, Y_1), \ldots, (X_n, Y_n)\}$.
- Un punto nuevo x a clasificar.
- Número de vecinos k.

Algoritmo:

1. Calcular la distancia entre x y cada punto X_i en el conjunto de entrenamiento.
2. Ordenar los puntos de entrenamiento de acuerdo con su distancia a x.
3. Seleccionar los k puntos más cercanos: $X^{(1)}, \ldots, X^{(k)}$.
4. Recoger las etiquetas correspondientes: $Y^{(1)}, \ldots, Y^{(k)}$.
5. Asignar a x la clase más frecuente entre esos k vecinos.

Salida: clase predicha para x.

Maldición de la dimensionalidad

- Cuando la dimensión d aumenta, la noción de "cercanía" pierde fuerza.
- Las distancias entre puntos tienden a ser similares: el vecino más cercano no está mucho más cerca que el más lejano.
- Esto implica que se necesitan **muchos más datos** para cubrir adecuadamente el espacio.
- Consecuencia: k-NN se vuelve poco fiable en alta dimensión sin una reducción de dimensionalidad o selección de variables.

Ventajas y desventajas de k-NN

Ventajas:

- Sencillo de entender e implementar.
- No paramétrico: no asume forma de la frontera de decisión.
- Se adapta bien a fronteras no lineales.

Desventajas:

- Costoso en predicción: requiere calcular distancias a todos los puntos de entrenamiento.
- Sensible a la escala y al ruido (importancia del preprocesamiento).
- Mal desempeño en alta dimensión (maldición de la dimensionalidad).

Extensiones de k-NN

- **k-NN ponderado:** vecinos más cercanos tienen más peso.
- **Métodos aproximados:** KD-trees, Ball-trees para acelerar la búsqueda.
- **Metric learning:** aprender la métrica de distancia más adecuada para el problema.
- **Reducción de dimensionalidad:** PCA para mejorar desempeño en alta dimensión.

Referencias

- Stone, C. (1977). Consistency of nearest neighbor classifiers. *Annals of Statistics*.
- Devroye, L., Györfi, L., Lugosi, G. (1996). *A Probabilistic Theory of Pattern Recognition*. Springer.
- Hastie, Tibshirani, Friedman (2009). *The Elements of Statistical Learning*. Springer.
- Bishop, C. (2006). *Pattern Recognition and Machine Learning*. Springer.
- Cover, T., Hart, P. (1967). *Nearest Neighbor Pattern Classification*. IEEE Transactions on Information Theory.
- Duda, R., Hart, P., Stork, D. (2001). *Pattern Classification*. Wiley-Interscience.