Clasificación: sesgo y varianza

Matías Carrasco

2 de octubre de 2023

Índice

1. El problema de clasificación

1.1. El clasificador de Bayes

Consideremos el siguiente setting:

- Espacio de atributos $\mathcal{X} \subset \mathbb{R}^D$.
- Espacio de etiquetas $\mathcal{Y} = \{1, \ldots, m\}$ un conjunto finito.
- Espacio de observaciones $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$.
- Distribución desconocida $\mathcal{D}$ en $\mathcal{Z}$ (i.e. una distribución conjunta $p(\boldsymbol{x}, y)$).
- Conjunto de datos $S = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^N \sim \mathcal{D}^N$.
- Función de pérdida: $\text{Loss}(y, y') = \mathbb{1}_{y \neq y'}$.

El último punto sobre la función de pérdida requiere una aclaración. La función de pérdida $\text{Loss}(y, y') = \mathbb{1}_{y \neq y'}$ se llama **0-1 loss**, y equivale a penalizar los errores

de clasificación. Veremos que en casi todos los casos se suele usar una función de pérdida subrogada a la 0-1 loss, como será el caso por ejemplo con máxima verosimilitud.

Así como en un problema de regresión tenemos la *función de regresión* que representa la predicción óptima, en clasificación también tenemos una predicción óptima.

Definición (Clasificador de Bayes)**.** Fijado un $\boldsymbol{x} \in \mathcal{X}$, la predicción **óptima** en $\boldsymbol{x}$ es

$$c(\boldsymbol{x}) = \arg\min_{\widehat{y} \in \mathcal{Y}} \left\{ \mathbb{E}\left[\text{Loss}(\widehat{y}, y) \mid \boldsymbol{x}\right] \right\}$$

Como la pérdida es la 0-1 loss, se ve inmediatamente que

$$c(\boldsymbol{x}) = \arg\max_{\widehat{y} \in \mathcal{Y}} \left\{ \text{Prob}\left[\widehat{y} \mid \boldsymbol{x}\right] \right\}$$

La función $c : \mathcal{X} \to \mathcal{Y}$ se llama **clasificador de Bayes**.

Definición (Error de Bayes)**.** Fijamos un atributo $\boldsymbol{x} \in \mathcal{X}$. El error esperado incurrido por el clasificador óptimo en $\boldsymbol{x}$:

$$\text{N}(\boldsymbol{x}) = \mathbb{E}\left[\text{Loss}(c(\boldsymbol{x}), y) \mid \boldsymbol{x}\right] = \text{Prob}\left[c(\boldsymbol{x}) \neq y \mid \boldsymbol{x}\right]$$

se llama **error de Bayes**. También se lo suele llamar *ruido* o *error irreducible*.

1.2. Ejemplo: clasificación en un tablero

Consideremos como ejemplo el siguiente problema de predicción del color de un casillero en un tablero:

- Atributos: $\mathcal{X}$, $\boldsymbol{x} = (i, j)$ la posición en el tablero.
- Etiquetas: $\mathcal{Y}$, $y =$ rojo(0)/azul(1).
- Datos: $S = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{N}$.
- Distribución: $\mathcal{D}$ está dada por
  - $\boldsymbol{x}$ es uniforme en el tablero;
  - La curva verde (ver Fig. 1) divide el tablero en dos regiones, superior e inferior, y

$$\text{Prob}(\text{azul}|\boldsymbol{x}) = \begin{cases} 3/4 & \text{si } \boldsymbol{x} \text{ superior} \\ 1/4 & \text{si } \boldsymbol{x} \text{ inferior} \end{cases}$$

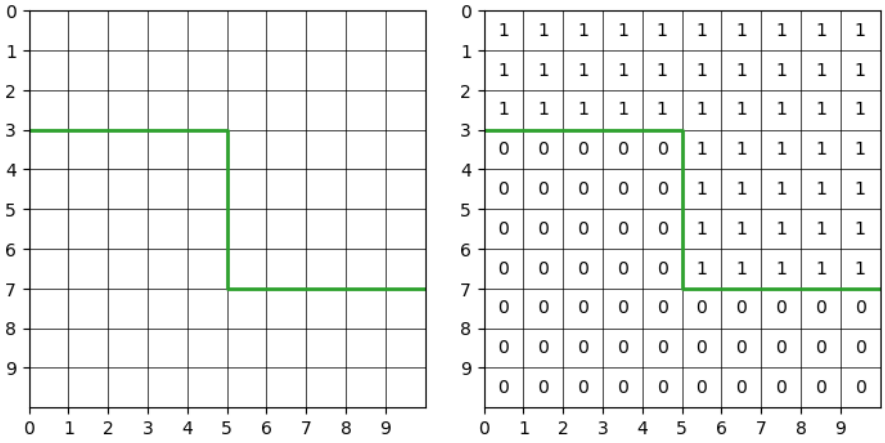

Figura 1: Ejemplo - clasificador de Bayes en el tablero.

Es fácil ver que el clasificador de Bayes es

$$c^*(\boldsymbol{x}) = \begin{cases} 1 & \text{si } \boldsymbol{x} \text{ superior} \\ 0 & \text{si } \boldsymbol{x} \text{ inferior} \end{cases}$$

y el error de Bayes es

$$\text{Prob}_{(\boldsymbol{x},y)\sim\mathcal{D}}\left\{y \neq c^*(\boldsymbol{x})\right\} = 1/4$$

2. K vecinos más cercanos (KNN)

Es un clasificador clásico y simple. El principio básico detrás de su funcionamiento es que instancias con atributos cercanos tienen etiquetas similares. Para ello es necesario elegir una forma de medir la distancia o cercanía entre puntos en el espacio de atributos $\mathcal{X}$.

Las distancias mas comunes son:

L_2 **o Euclidia:** $d_2(\boldsymbol{x}, \boldsymbol{x}') = \left(\sum_{j=1}^{D} |x_j - x'_j|^2\right)^{1/2}$

L_1 **o Manhattan:** $d_1(\boldsymbol{x}, \boldsymbol{x}') = \sum_{j=1}^{D} |x_j - x'_j|$

L_p **o Minkowski:** $d_p(\boldsymbol{x}, \boldsymbol{x}') = \left(\sum_{j=1}^{D} |x_j - x'_j|^p\right)^{1/p}$

Como todo algoritmo de Machine Learning su input son los datos de entrenamiento S y su salida es un modelo (en este caso un clasificador) $c(\boldsymbol{x})$.

El modelo resultante $c(\boldsymbol{x})$ se calcula de la siguiente manera:

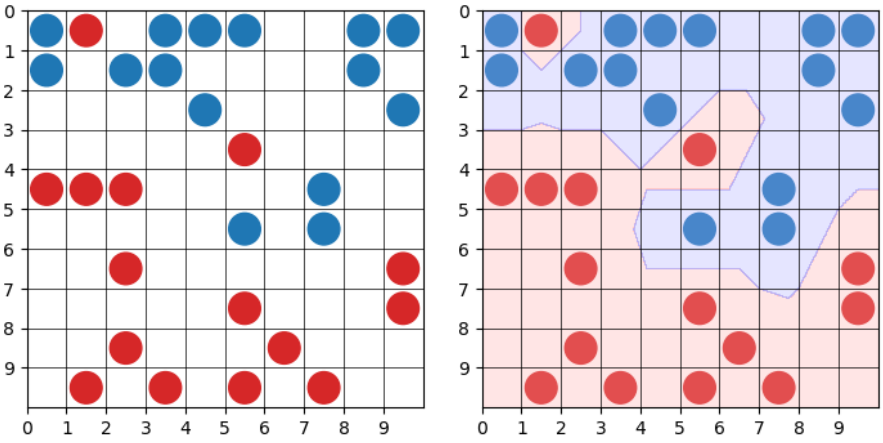

Figura 2: A la izquierda una muestra de datos S de ejemplo en el problema de predicción en el tablero. A la derecha el clasificador de K vecinos más cercanos con $K = 1$.

1. Calcula la distancia entre $\boldsymbol{x}$ y todos los puntos en S.
2. Ordena las distancias de menor a mayor.
3. Selecciona los K puntos de S más cercanos a x: $V_K(\boldsymbol{x})$.
4. Voto mayoritario: $c(\boldsymbol{x}) = \arg\max_c \text{Prob}_K(y = c \mid \boldsymbol{x})$ donde

$$\text{Prob}_K(y = c \mid \boldsymbol{x}) = \frac{1}{K} \#\left\{\boldsymbol{x}_i \in V_K(\boldsymbol{x}) : y_i = c\right\}$$

Sus hiperparámetros más comunes son: el valor de K y la distancia.

3. Sesgo y varianza en clasificación

Para ilustrar los conceptos de sesgo y varianza en clasificación usaremos el ejemplo del tablero con el algoritmo de K vecinos más cercanos. En la Fig. 2 tenemos una muestra de datos S de ejemplo con $N = 30$ observaciones. A su vez se muestra el clasificador KNN para $K = 1$.

Al cambiar el conjunto de datos S cambia el clasificador, y por lo tanto cambia la predicción $c_S(\boldsymbol{x})$ en un $\boldsymbol{x}$ dado.

Definición (Predicción más frecuente)**.** Fijemos un atributo $\boldsymbol{x} \in \mathcal{X}$. Llamamos $\bar{c}(\boldsymbol{x})$ la **predicción más frecuente** en $\boldsymbol{x}$ al variar el conjunto de datos S:

$$\bar{c}(x) = \arg\max_{\widehat{y} \in \mathcal{Y}} \left\{\text{Prob}_{S \sim D^N}\left[c_S(\boldsymbol{x}) = \widehat{y}\right]\right\}$$

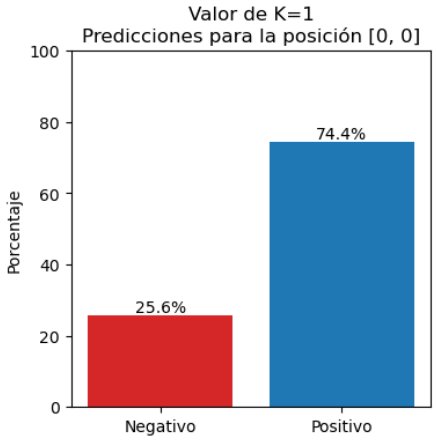

Figura 3: Simulación de 5mil conjuntos de datos S y cálculo de la predicción (con $K = 1$) en el casillero $[0, 0]$.

En el ejemplo del tablero, es muy difícil calcular la predicción más frecuente explícitamente. Sin embargo, podemos hacer una simulación. En la Fig. 3 se muestra una simulación de la distribución de predicciones para el casillero $[0, 0]$ usando 5mil datasets S aleatorios:

- Fijamos el casillero $\boldsymbol{x} = [0, 0]$.
- Generamos $M = 5k$ muestras $\{S_j\}_{j=1}^{M}$ todas ellas de tamaño $N = 30$.
- Para cada muestra S_j ajustamos KNN c_j con $K = 1$.
- Calculamos las predicciones $\{c_j(\boldsymbol{x})\}_{j=1}^{M}$.

Vemos así que en este caso $\bar{c}([0, 0]) = 1$.

En esta simulación, el error en $\boldsymbol{x} = [0, 0]$ se aproxima de la siguiente manera

$$\text{Prob}_{S\sim D^N, y}\{y \neq c_S(\boldsymbol{x}) \mid \boldsymbol{x} = [0, 0]\} \approx \frac{1}{M}\sum_{j=1}^{M} \text{Prob}_y\left(y \neq c_j(\boldsymbol{x}) \mid \boldsymbol{x} = [0, 0]\right)$$

Descomponiendo en $c_j(\boldsymbol{x}) = 0$ o 1 tenemos:

$$0{,}256 \times \underbrace{\text{Prob}_y\left(y \neq 0 \mid \boldsymbol{x} = [0, 0]\right)}_{0{,}75} + 0{,}744 \times \underbrace{\text{Prob}_y\left(y \neq 1 \mid \boldsymbol{x} = [0, 0]\right)}_{0{,}25} = 0{,}378$$

A partir de la predicción más frecuente y de la predicción óptima podemos definir el sesgo.

Definición. Fijamos un atributo $\boldsymbol{x} \in \mathcal{X}$. El **sesgo** en $\boldsymbol{x}$ se define como

$$\mathrm{B}(\boldsymbol{x}) = \mathrm{Loss}\left[c^*(\boldsymbol{x}), \bar{c}(\boldsymbol{x})\right] = \begin{cases} 1 & \text{si } c^*(\boldsymbol{x}) \neq \bar{c}(\boldsymbol{x}) \\ 0 & \text{si } c^*(\boldsymbol{x}) = \bar{c}(\boldsymbol{x}) \end{cases}$$

En el ejemplo del tablero de la Fig. 3, como tenemos $c^*([0,0]) = 1$ y $\bar{c}([0,0]) = 1$ el sesgo es $B([0,0]) = 0$.

Por otro lado tenemos la varianza.

Definición. Fijamos un atributo $\boldsymbol{x}$ cualquiera. La **varianza** en $\boldsymbol{x}$ la definimos como

$$\mathrm{V}(\boldsymbol{x}) = \mathbb{E}_{S\sim D^N}\left\{ \mathrm{Loss}\left[\bar{c}(\boldsymbol{x}), c_S(\boldsymbol{x})\right] \right\} = \mathrm{Prob}_{S\sim D^N}\left\{ c_S(\boldsymbol{x}) \neq \bar{c}(\boldsymbol{x}) \right\}$$

En el ejemplo con $\boldsymbol{x} = [0,0]$ tenemos $\bar{c}([0,0]) = 1$ por lo que

$$\mathrm{V}([0,0]) = \mathrm{Prob}_{S\sim D^N}\left[c_S([0,0]) \neq 1\right] = 0{,}256$$

4. Descomposición del error

La descomposición del error, al igual que en regresión, toma la forma

$$\text{Error esperado} = \text{FUNCIÓN}\Big(\text{Sesgo}, \text{Varianza}, \text{Error Irreducible}\Big)$$

En el caso de clasificación, y para la 0-1 loss, la descomposición es un poco más compleja.

Fijemos un atributo $\boldsymbol{x}$. En clasificación **binaria** con la 0-1 loss la descomposición es

$$\underbrace{\mathrm{Prob}\left\{y \neq c_S(\boldsymbol{x}) \mid \boldsymbol{x}\right\}}_{\text{Error}} = \underbrace{\mathrm{B}(\boldsymbol{x})}_{\text{Sesgo}} + \big[\mathrm{Factor}_1\big] \times \underbrace{\mathrm{V}(\boldsymbol{x})}_{\text{Varianza}} + \big[\mathrm{Factor}_2\big] \times \underbrace{\mathrm{N}(\boldsymbol{x})}_{\text{Error Irreducible}}$$

- El factor de la varianza es

$$\mathrm{Factor}_1 = \begin{cases} 1 & \text{si } \mathrm{B}(\boldsymbol{x}) = 0 \\ -1 & \text{si } \mathrm{B}(\boldsymbol{x}) = 1 \end{cases}$$

- El factor del error irreducible es

$$\mathrm{Factor}_2 = 2\,\mathrm{Prob}\left\{c_S(\boldsymbol{x}) = c^*(\boldsymbol{x})\right\} - 1$$

Veamos cómo queda la descomposición en el ejemplo:

- Fijamos el casillero $\boldsymbol{x} = [0, 0]$.
- Tenemos:
  - Error esperado: $\text{Prob}\{y \neq c_S(\boldsymbol{x}) \mid \boldsymbol{x} = [0, 0]\} = 0{,}378$
  - Sesgo: $\text{B}([0, 0]) = 0$
  - Varianza: $\text{V}([0, 0]) = 0{,}256$
  - Factor_1: 1
  - Error irreducible: $\text{N}([0, 0]) = 0{,}25$
  - Factor_2:

$$2\,\text{Prob}\{c_S(\boldsymbol{x}) = c^*(\boldsymbol{x})\} - 1 \approx 2 \times 0{,}744 - 1 = 0{,}488$$

- La descomposición queda:

$$\underbrace{0{,}378}_{\text{Error}} = \underbrace{0}_{\text{Sesgo}} + [1] \times \underbrace{0{,}256}_{\text{Varianza}} + [0{,}488] \times \underbrace{0{,}25}_{\text{Error Irreducible}}$$

La descomposición de error para todos los casilleros la podemos ver (a partir de una simulación) en la Fig. 4. Vemos que la región cercana a la frontera de decisión del clasificador de Bayes (el óptimo) es en donde el error es mayor.

Veamos la demostración de la descomposición. Recordar que estamos tratando el caso de clasificación **binaria**.

Fijemos un atributo $\boldsymbol{x} \in \mathcal{X}$. Empezamos mostrando que

$$\text{Prob}\,[y \neq c_S(\boldsymbol{x}) \mid S, \boldsymbol{x}] = \text{Loss}\,[c_S(\boldsymbol{x}), c^*(\boldsymbol{x})] + [\text{Factor}_0]\,\text{N}(\boldsymbol{x})$$

en donde

$$\text{Factor}_0 = \begin{cases} 1 & \text{si } c_S(\boldsymbol{x}) = c^*(\boldsymbol{x}) \\ -1 & \text{si } c_S(\boldsymbol{x}) \neq c^*(\boldsymbol{x}). \end{cases}$$

Recordar que $\text{N}(\boldsymbol{x}) = \text{Prob}\,[y \neq c^*(\boldsymbol{x}) \mid \boldsymbol{x}]$.

De hecho, si $c_S(\boldsymbol{x}) = c^*(\boldsymbol{x})$ entonces el primer término (del lado derecho de la igualdad) es cero, el factor es uno y la igualdad es clara. Si $c_S(\boldsymbol{x}) \neq c^*(\boldsymbol{x})$ la afirmación es

$$\text{Prob}\,[y \neq c_S(\boldsymbol{x}) \mid S, \boldsymbol{x}] = 1 - \text{Prob}\,[y \neq c^*(\boldsymbol{x}) \mid \boldsymbol{x}]$$

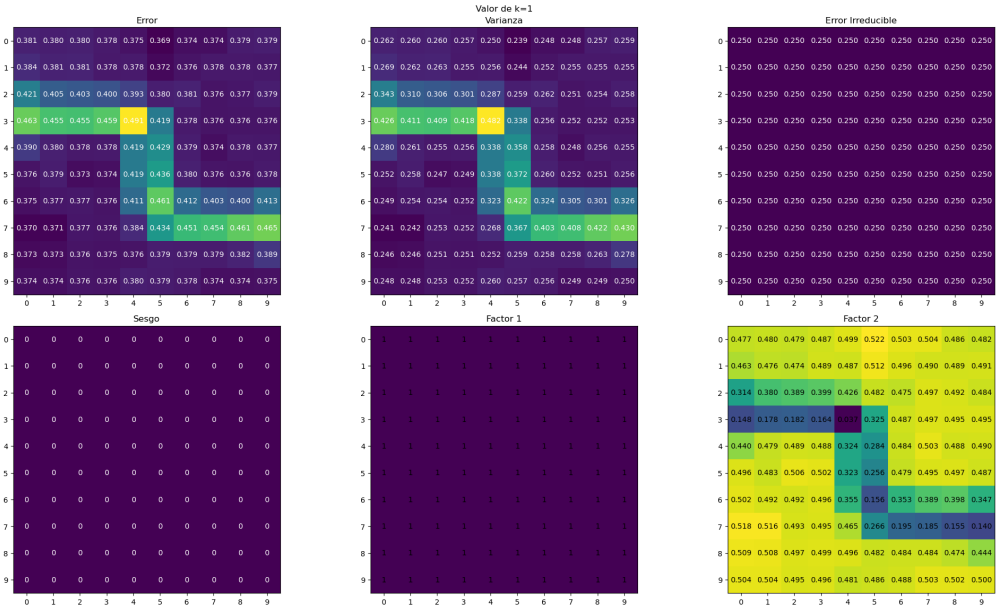

Figura 4: Descomposición del error simulada con $5k$ datasets en el ejemplo del tablero para el algoritmo KNN con $K = 1$.

lo cual es cierto por tratarse de un problema de clasificación binaria.

Veamos ahora que

$$\text{Prob}_{S\sim D^N}\left[c_S(\boldsymbol{x}) \neq c^*(\boldsymbol{x})\right] = \text{B}(\boldsymbol{x}) + \left[\text{Factor}_1\right]\text{V}(\boldsymbol{x})$$

Si $B(\boldsymbol{x}) = \text{Loss}\left[\bar{c}(\boldsymbol{x}), c^*(\boldsymbol{x})\right] = 0$, entonces $\bar{c}(\boldsymbol{x}) = c^*(\boldsymbol{x})$, y recordando que

$$\text{V}(\boldsymbol{x}) = \text{Prob}_{S\sim D^N}\left\{c_S(\boldsymbol{x}) \neq \bar{c}(\boldsymbol{x})\right\}$$

vemos que la igualdad es clara con $\text{Factor}_1 = 1$. Por el contrario, si $B(\boldsymbol{x}) = 1$, es decir $\bar{c}(\boldsymbol{x}) \neq c^*(\boldsymbol{x})$, la igualdad también es clara por ser un problema de clasificación binaria (complemento igual que antes).

Por último, resta observar que el valor promedio de Factor_0 al variar $S \sim D^N$ es igual a

$$1 \times \text{Prob}_{S\sim D^N}\left\{c_S(\boldsymbol{x}) = c^*(\boldsymbol{x})\right\} + (-1) \times \text{Prob}_{S\sim D^N}\left\{c_S(\boldsymbol{x}) \neq c^*(\boldsymbol{x})\right\}$$

y otra vez, usando que es un problema de clasificación binaria, y por lo tanto

$$\text{Prob}_{S\sim D^N}\left\{c_S(\boldsymbol{x}) \neq c^*(\boldsymbol{x})\right\} = 1 - \text{Prob}_{S\sim D^N}\left\{c_S(\boldsymbol{x}) = c^*(\boldsymbol{x})\right\}$$

vemos que

$$\begin{aligned}\mathbb{E}_{S\sim D^N}\left[\text{Factor}_0\right] &= 2 \times \text{Prob}_{S\sim D^N}\left\{c_S(\boldsymbol{x}) = c^*(\boldsymbol{x})\right\} - 1 \\ &= \text{Factor}_2\end{aligned}$$

Esto demuestra la descomposición.