Proyecto Profesional

Maestría en Ciencia de Datos

Emmanuel Alcalá

Departamento de Matemáticas y Física

Definición del problema

> "Si me dieran una hora para salvar el planeta, gastaría 59 minutos definiendo el problema y un minuto resolviéndolo"
>
> -- *Albert Einstein*

El propósito de esta sesión es:

- No saltarse a analizar los datos sin *antes* haber identificado el problema. No es correcto tener una respuesta y luego buscar la pregunta.
  - fishin expedition: proyecto que nunca se enmarcó correctamente y luego se tortura a los datos para encontrar relaciones *inesperadas*
- No tomar proyectos que excedan las capacidades (por ejemplo, que no puedan terminarse en un tiempo razonable).
- Asegurarse de que los datos (la evidencia) que tenemos permitan responder o resolver el problema.
- Ninguna cantidad de competencia técnica o rigor estadístico puede compensar haber resuelto un problema inútil.

Cómo definir un problema

La ciencia de datos es tan científica como otras ciencias.

- Planteamiento: descripción concisa de un tema o condición para mejorar. Identifica una brecha entre *estado actual* y *estado deseado*.
  - Identificar y explicar. Describir el contexto actual, en dónde ocurre el problema, qué impacto tiene, y cuál podría ser una mejora. -

Ejemplo

Supongamos que queremos predecir si alguien tiene cáncer a partir de imágenes de resonancia magnética[1].

- ¿Cuál es el problema?
  - No es que los médicos entrenados en imagenología no sepan cómo se ve un cáncer con RM.
  - Tampoco queremos mejorar el proceso de toma de imágenes -aunque puede ser un problema.
  - El problema: ¿puede mejorarse la tasa de predicciones correctas que hace un médico usando algoritmos de clasificación?
  - Dado un conjunto de pacientes tienen cáncer, ¿qué proporción de casos identifican correctamente? ¿Podemos mejorar esa tasa?

[1] Bishop, C. (2006). *Pattern Recognition and Machine Learning*. Ch. 1

Ya identificamos algunas cosas:

- El estado de cosas actual es $\hat{\pi}_{\text{médico}}$: proporción de casos correctamente identificados.
- Un estado deseado es $\hat{\pi}_{\text{algoritmo}} \geq \hat{\pi}_{\text{médico}}$: minimizar la probabilidad de asignar un paciente a la clase equivocada (minimizar la probabilidad de cometer un error). El siguiente paso es hacer una descripción *concisa y clara* del problema.
- ¿Cómo evaluaremos nuestra solución en términos que sean comparables a nuestro *benchmark*?
- ¿Cómo saber cuándo detenernos, cuándo es suficientemente buena una solución?
- En este ejemplo, tenemos que plantear el problema en términos de teoría de la decisión.

Descripción del problema

Sean $\mathbf{x}$ un vector de intensidad de pixeles de una imagen, y C_k tal que:

$$\text{Presencia de cáncer es la clase } \mathcal{C}_1$$
$$\text{Ausencia de cáncer es la clase } \mathcal{C}_2$$

Escogemos una variable binaria $\mathbf{t}$ tal que

$$t = 0 \text{ si la clase es } \mathcal{C}_1$$
$$t = 1 \text{ si la clase es } \mathcal{C}_2$$

El problema consiste en dos pasos:

- Inferencia: determinar la distribución conjunta $p(\mathbf{x}, \mathcal{C}_k)$ o $p(\mathbf{x}, t)$ a partir de datos de entrenamiento.
- Decisión: una vez estimamos $p(\mathbf{x}, \mathcal{C}_k)$ debemos *decidir* algo. Dado que tenemos datos $\mathbf{x}$ queremos saber la probabilidad de $\mathcal{C}_k$ condicional a los datos obtenidos.

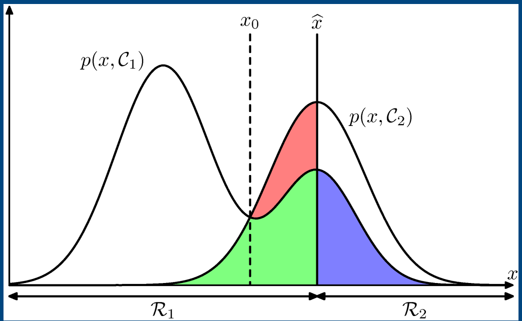
$$p(\mathcal{C}_k|\mathbf{x}) = \frac{p(\mathbf{x}|\mathcal{C}_k)p(\mathcal{C}_k)}{p(\mathbf{x})}$$

- $p(\mathcal{C}_1)$ es la probabilidad de que un paciente tenga cáncer antes de que tenga lugar la medición.
- $p(\mathcal{C}_1|\mathbf{x})$ es la probabilidad posterior *después* medición.

Se pueden tener varios objetivos:

- Minimizar las asignaciones de $\mathbf{x}$ a la clase incorrecta.
    - Partir $\mathbf{x}$ en dos regiones de decisión $\mathcal{R}_k$, tal que los puntos en $\mathcal{R}_k$ son asignados a la clase $\mathcal{C}_k$.
    - Un error ocurre cuando un valor de x que pertenece a $\mathcal{C}_1$ es asignado a $\mathcal{C}_2$ o viceversa. La probabilidad de un error es:

$$p(\text{error}) = p(\mathbf{x} \in \mathcal{R}_1, \mathcal{C}_2) + p(\mathbf{x} \in \mathcal{R}_2, \mathcal{C}_1) = \underbrace{\int_{\mathcal{R}_1} p(\mathbf{x}, \mathcal{C}_2)\mathrm{d}\mathbf{x}}_{\text{Error en la región 1}} + \underbrace{\int_{\mathcal{R}_2} p(\mathbf{x}, \mathcal{C}_1)\mathrm{d}\mathbf{x}}_{\text{Error en la región 2}}$$



$\hat{x}$ es la regla de decisión. Conforme $\hat{x} \to x_0$, la zona roja desaparece, pero la azul crece.

La suma de las áreas verde y azul es constante.

Se pueden tener varios objetivos:

- Minimizar las asignaciones de $\mathbf{x}$ a la clase incorrecta.
- Optimizar otra variable. Hay dos tipos de asignaciones incorrectas:
  - Que tenga cáncer pero se clasifique como $\mathcal{C}_2$.
  - Que no tenga cáncer pero se clasifique como $\mathcal{C}_1$.
  - Cuando se minimiza la mala clasificación, se puede reducir solo el segundo error (área roja), pero no el primero (área azul).
  - Evidentemente, el primer error *es más costoso* que el segundo: las consecuencias de tener cáncer y no ser diagnosticado son peores que las de no tenerlo y ser diagnosticado.
  - Es mejor minimizar los errores del primer tipo.

- Función de costo:
  - Podemos asignar diferentes pesos a cada tipo de diagnóstico, de tal manera que refleje el hecho de que un tipo de error es más costoso.

$$\begin{array}{c c} & \begin{array}{cc} \text{cancer} & \text{normal} \end{array} \\ \begin{array}{c} \text{cancer} \\ \text{normal} \end{array} & \begin{bmatrix} 0 & 1000 \\ 1 & 0 \end{bmatrix} \end{array}$$

  - El costo de ser diagnosticado como normal, si se tiene cáncer, es de 1000.
  - El costo de ser diagnosticado con cáncer pero ser normal es de 1 (i.e., sí hay un costo en este tipo de error).
  - Ser diagnosticado correctamente tiene un costo de 0.

- Suponer que para un valor nuevo de $\mathbf{x}$, la clase verdadera es $\mathcal{C}_k$ y la asignamos a $\mathcal{C}_k$, en donde j puede o puede no ser igual a k.
- Incurrimos en costo L_{kj} tomado de la matriz de costos.
- El propósito ahora es reducir el costo esperado, que es una suma ponderada del costo L_{kj} con la probabilidad $p(\mathbf{x}, \mathcal{C}_k)$

$$\mathbf{E}[L] = \sum_j \sum_k \int_{\mathcal{R}_j} L_{kj} p(\mathbf{x}, \mathcal{C}_k) \, \mathrm{d}\mathbf{x}$$

- Si un $\mathbf{x}$ es asignado a $\mathcal{R}_i$, consideramos minimizar $\sum_k L_{kj} p(\mathbf{x}, \mathcal{C}_k)$, que es equivalente a minimizar $\sum_k L_{kj} p(\mathcal{C}_k | \mathbf{x})$.
- Por ejemplo, si un paciente con dx $\mathbf{x}$ tiene 0.1 de probabilidad de tener cáncer, $p(\mathcal{C}_{\text{cáncer}} | \mathbf{x}) = 0.1$ y $p(\mathcal{C}_{\text{normal}} | \mathbf{x}) = 0.9$, el costo esperado es

$$\begin{aligned} 0 \times 0.1 + 1 \times 0.9 &= 0.9, \text{ si se clasifica como cáncer} \\ 1000 \times 0.1 + 0 \times 0.9 &= 100, \text{ si se clasifica como normal} \end{aligned}$$

Take-home message

- Resolver un problema implica proponer una mejora, pero debemos definir en qué consiste una mejora (e.g., minimizar un error por sí mismo no es una mejora, debemos tener el cuenta el costo de minimizar ciertos errores).
- Como se dijo antes, la ciencia de datos debe ser tan científica como otras ciencias.
- Uno de los problemas actuales de varias disciplinas científicas es el de la reproducibilidad: resultados no se pueden reproducir con nuevos experimentos.
- ¿Podría suceder esto en CD? ¿Qué consecuencias tendría?
- Si la CD tiene el propósito de adquirir valor de los datos, ¿qué valor tiene información que no se puede reproducir?

Si el procedimiento que seguimos resuelve en negativo una pregunta, no debemos desalentarnos, ni cambiar constantemente la pregunta hasta obtener un resultado que nos satisfaga (**sesgo de confirmación**).
Una respuesta negativa también tiene valor.
Recordar