

Aprendizaje Automático – Guía de Ejercicios *

Departamento de Computación – FCEyN
Universidad de Buenos Aires

Primer cuatrimestre 2024
Versión: 20 de marzo de 2024

1. Repaso probabilidad y estadística

Para resolver los siguientes ejercicios recomendamos leer el **Capítulo 6** del libro Mathematics for Machine Learning de Deisenroth, Faisal y Soon Ong. <https://mml-book.github.io/book/mml-book.pdf>

Ejercicio 1.1. Explique por qué los siguientes eventos son independientes de a pares pero no independientes entre todos. Dadas 2 monedas,

- (a) la primera moneda es cara;
- (b) la segunda moneda cara;
- (c) las dos monedas son iguales.

Ejercicio 1.2. Demostrar el teorema de Probabilidad Total: dados una partición $\{A_i\}_{i=1}^n$ del espacio muestral tal que $P(A_i) > 0$ para todo i , y un evento B :

$$P(B) = \sum_{i=1}^n P(B | A_i) \cdot P(A_i)$$

Ejercicio 1.3.

- (a) Sugerencia, mirar este video: [https://www.youtube.com/watch?v=HZGCoVF3YvM&t=57s\(3blue1brown Bayes\)](https://www.youtube.com/watch?v=HZGCoVF3YvM&t=57s(3blue1brown Bayes))
- (b) Demostrar el teorema de Bayes: dados dos eventos A y B tal que $P(B) > 0$,

$$P(A | B) = \frac{P(B | A) \cdot P(A)}{P(B)}$$

- (c) Un local vende dos marcas de televisores, A y B. El 40 % de los televisores vendidos son de la marca A y un 30 % de ellos tienen un defecto. Por otro lado, el 20 % de los televisores vendidos son de la marca B y el 10 % tienen un defecto. Si un televisor tiene un defecto, ¿cuál es la probabilidad de que sea de la marca A?
- (d) Supongamos que tienes dos máquinas, A y B, que producen tornillos. La longitud de los tornillos producidos por cada máquina sigue una distribución normal. Se sabe que la máquina A produce tornillos con una longitud media de 10 cm y una desviación estándar de 0.5 cm, mientras que la máquina B produce tornillos con una longitud media de 10.5 cm y una desviación estándar de 0.7 cm.

Ahora, supongamos que se selecciona un tornillo al azar de la producción combinada de ambas máquinas y se encuentra que tiene una longitud de 10.2 cm. ¿Cuál es la probabilidad de que este tornillo provenga de la máquina B?

Ejercicio 1.4.

- (a) Explicar con tus palabras qué es la media y qué es el desvío estándar.
- (b) En una fábrica de producción de caramelos, se mide la longitud de los caramelos producidos. Si la longitud media es de 5cm y un desvío de $0,05cm^2$. Siendo que tener un caramelo de más de 5.05 cm o menos de 4.95 cm se considera defectuoso, ¿qué significa esto en términos de la calidad de los caramelos producidos por esa fabrica?. ¿Hicieron alguna suposición sobre la distribución?

* Algunos ejercicios fueron adaptados de los libros “Machine Learning”, de Tom Mitchell (McGraw-Hill, 1997); “Pattern Recognition and Machine Learning”, de Christopher Bishop (Springer, 2006); y “An Introduction to Statistical Learning”, de James, Witten, Hastie & Tibshirani (Springer, 2015).

Ejercicio 1.5.

Has realizado un experimento en el que lanzaste una moneda 10 veces y la secuencia observada de resultados fue HHHHTHTTHHT.

1. Suponiendo que la moneda es justa (es decir, la probabilidad de que salga cara ($P(H)$) = 0,5 y la probabilidad de que salga cruz ($P(T)$) = 0,5), calcule la probabilidad de observar la secuencia dada.
2. Es más probable que la moneda esté sesgada hacia la cara en 70 % ó que la moneda este balanceada dados los datos observados. (HHHTHTTHHT).
3. Imaginen ahora que queremos estimar la carga de la moneda lo mejor posible dados los datos de la tirada. Plantear los pasos a seguir para encontrar este valor. Tip: están calculando máxima verosimilitud.
4. Calcularla :)
5. Sugerencia, ver: <https://www.youtube.com/watch?v=Dn6b9fCIUpM&t=195s> (statquest)

2. Introducción

Ejercicio 2.1. Revisar y completar el notebook `notebook_1_herramientas.ipynb`.

Ejercicio 2.2. Describir para los siguientes problemas si se trata de aprendizaje supervisado o aprendizaje no supervisado. Especificar qué medida de performance y de un ejemplo de una base de datos que permita encarar el problema.

- (a) detección de discurso de odio en tweets;
- (b) recomendación de películas;
- (c) diagnóstico de tumores por imágenes;
- (d) autocompletar textos;
- (e) segmentación comercial de clientes;
- (f) autenticación biométrica (ej: huellas dactilares);
- (g) detección de fraude en tarjetas de crédito.

Ejercicio 2.3. Determinar para los siguientes problemas de aprendizaje supervisado si se trata de problemas de clasificación o de regresión. Para cada caso, indique un ejemplo de instancia (el valor de sus atributos) junto a una etiqueta posible especificando el tipo de cada valor.

- (a) Dado un tweet, determinar si habla en contra o a favor de un candidato presidencial.
- (b) Predecir cuánto gastará una empresa en luz el próximo semestre.
- (c) Dado un tweet, predecir la probabilidad de que hable en contra o a favor de un candidato.
- (d) Predecir a qué distancia de la facultad vive una persona.
- (e) Predecir si se gastará más o menos que \$50.000 por mes de luz el próximo semestre.
- (f) Predecir la probabilidad de que se gaste más o menos que \$50.000 por mes de luz el próximo semestre.
- (g) Predecir la nota que tendrá un alumno en un examen cuya nota puede ser 0, 1, 2, ..., 10
- (h) Predecir la nota que tendrá un alumno en un examen cuya nota puede ser "A", "R" o "I".
- (i) Predecir dónde vive una persona.
- (j) Predecir la próxima palabra a autocompletar dadas las oraciones anteriores.
- (k) Predecir el valor que tomará el dolar en los próximos diez días.

Ejercicio 2.4. Sea un problema de clasificación en el cual cada instancia tiene 2 atributos numéricos (coordenadas x e y) y pertenece a una de dos clases posibles (blanco o negro).

Se tienen tres tipos de hipótesis ilustrados en la Figura que representan (a) rectas, (b) líneas verticales (hasta 30 líneas), (c) elipses (tantas como se quiera).

Para cada uno de ellos, se pide:

- Describir el espacio de hipótesis H ;
- Identificar los parámetros de la hipótesis (el conjunto de valores que permiten describir una hipótesis en concreto, θ).¹

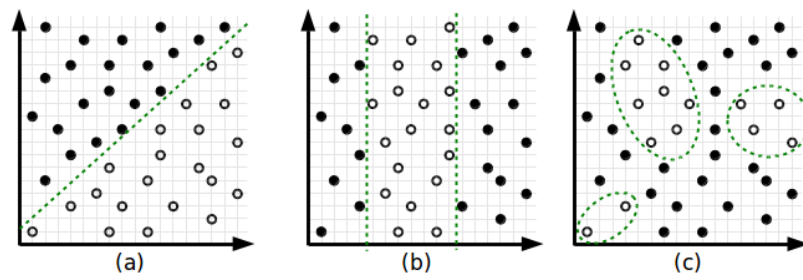


Figura 1: Tipos de Hipótesis

Ejercicio 2.5. Completar el notebook `notebook_2_titanic.ipynb`.

¹“Machine Learning”, de Tom Mitchell (McGraw-Hill, 1997);