Aprendizaje Automático – Guía de Ejercicios*

Departamento de Computación – FCEyN
Universidad de Buenos Aires

Segundo cuatrimestre 2018
Versión: 8 de septiembre de 2018

1. Herramientas

Ejercicio 1.1. Revisar y completar el notebook `notebook_1_herramientas.ipynb` disponible en la sección de Descargas.

2. Introducción

Ejercicio 2.1. ¿Cuál es la diferencia entre el aprendizaje supervisado, el aprendizaje no supervisado y el aprendizaje por refuerzos?

Ejercicio 2.2. ¿Cuál es la diferencia entre un problema de clasificación y uno de regresión?. Determinar para los siguientes problemas de aprendizaje supervisado si se trata de problemas de clasificación o de regresión.

I. Dado un tweet, determinar si habla en contra o a favor de un candidato presidencial.

II. Predecir cuánto gastará una empresa en luz el próximo semestre.

III. Predecir la nota que tendrá un alumno en un examen cuya nota puede ser $0, 1, 2, \ldots, 10$

IV. Predecir la nota que tendrá un alumno en un examen cuya nota puede ser "A", "R" o "I".

V. Predecir donde vive una persona.

VI. Predecir donde vivirá una persona dentro de 5 años.

VII. Predecir si se gastará más o menos que $50.000 por mes de luz el próximo semestre.

VIII. Predecir la probabilidad de que una persona haya comprado un bote el último año. **TIP**: Pensar con qué etiquetas se entrena al algoritmo.

¿Qué diferencia hay entre los items (v) y (vi)? ¿Qué relación hay entre el momento de toma de los atributos y la etiqueta?

Ejercicio 2.3. Describir la tarea, medida de performance y experiencia para (a) filtro de spam; (b) dictado de textos; (c) autenticación biométrica (ej: huellas dactilares); (d) detección de fraude en tarjetas de crédito.

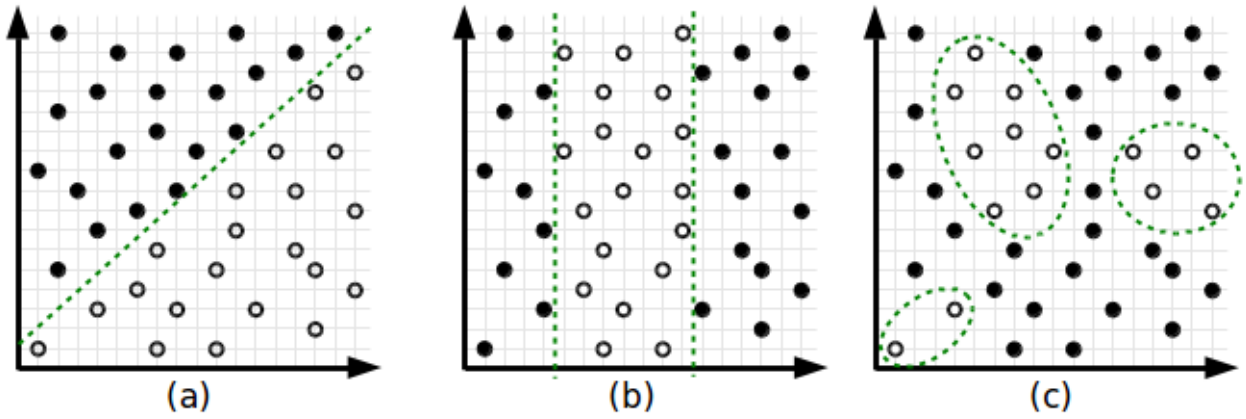
Ejercicio 2.4. Sea un problema de clasificación en el cual cada instancia tiene 2 atributos numéricos (coordenadas x e y) y pertenece a una de dos clases posibles (blanco o negro). Para cada uno de los tipos de hipótesis ilustrados en la Figura 1 se pide:

- identificar los parámetros de la hipótesis y describir el espacio de hipótesis H;
- pensar un algoritmo para encontrar una hipótesis y describir su sesgo inductivo.

Ejercicio 2.5. Completar el notebook `notebook_2_titanic.ipynb` disponible en la sección de Descargas. En este ejercicio, deberán descargar y explorar el contenido del archivo `titanic-train.csv` que contiene datos de pasajeros del naufragio del transatlántico Titanic en 1912, incluyendo edad, sexo, clase del pasaje y supervivencia a la tragedia, entre otros. Para una descripción completa, ver `https://www.kaggle.com/c/titanic/data`. Completar el notebook de manera de clasificar a los pasajeros en supervivientes y no supervivientes tan bien como sea posible. No está permitido usar ninguna técnica de Aprendizaje Automático, por ahora. El objetivo es conseguir un buen porcentaje de aciertos sobre estos datos (ver final del notebook).

*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).

Figura 1:



3. Repaso de probabilidades

Recomendamos revisar los apuntes de la materia Probabilidad y Estadística para resolver los siguientes ejercicios: `http://cms.dm.uba.ar/academico/materias/verano2018/probabilidades_y_estadistica_C/apunte-probaC-2017-2C.pdf`

Ejercicio 3.1. Explique por qué los siguientes eventos son independientes de a pares pero no independientes entre todos. Dadas 2 monedas,

- la primera moneda es cara;
- la segunda moneda cara;
- las dos monedas son iguales.

Ejercicio 3.2. Demostrar el teorema de Bayes: dados dos eventos A y B,

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

Ejercicio 3.3. Demostrar el teorema de Probabilidad Total: dados una partición A_i del espacio muestral tal que $P(A_i) > 0$ para todo i, y un evento B:

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

Ejercicio 3.4. Supongamos que la probabilidad de que un paciente tenga una forma determinada del virus de herpes es $P(herpes) = 0{,}008$. Se tiene un test con una sensibilidad de $0{,}98$ (es decir, $P(\oplus \mid herpes) = 0{,}98$) y una especificidad de $0{,}97$ (es decir, $P(\ominus \mid \neg herpes) = 0{,}97$), donde $\oplus$ y $\ominus$ representan los resultados positivo y negativo del test, respectivamente. Si un paciente se realiza el test y le da resultado positivo, ¿cuál es la probabilidad de que realmente tenga ese virus de herpes? Es decir, se pide calcular $P(herpes \mid \oplus)$.

Ejercicio 3.5.

(a) Escribir en Python una función $f(n, a, b)$ con n, a, b naturales tal que imprima en pantalla n números reales aleatorios en el intervalo $[a, b]$;

(b) Escribir en Python una función $g(n, m, v)$ imprima en pantalla n números reales aleatorios muestreando una distribución normal con media m y varianza v.

Plotear luego histogramas para $f(1000, 10, 20)$ y $g(1000, 3, 9)$.