Minería de Datos

UTDT

Master in Management + Analytics

Sesión 1

Formalidades de la materia

6 clases teóricas.

- Consistirán de difusión teórica, discusión de ejemplos, y resolución/práctica de problemas.
- Al final de cada clase se proveerá un listado de lecturas obligatorias (importante leerlas).

2 clases prácticas.

- Se repasará cómo hacer uso de los conceptos vistos en las clases teóricas.
- Será un espacio de consultas referidas al trabajo práctico.

Si retomamos las clases presenciales, se pedirá que asistan con su computadora portátil y que tengan instalado el software usado en la materia.

Formalidades de la materia

La nota final seguirá la siguiente ponderación:

- Participación, asistencia y puntualidad (10%)
- Trabajo práctico integrador (50%) (sin recuperatorio)
- Examen final (40%)

Adicionalmente, se debe sacar una nota igual o mayor a 5 tanto en el trabajo práctico como en el final.

Lean el programa de la materia para mayores detalles.

Nota Final	Calificación
>= 9.50	A
9.00 a 9.49	A-
8.00 a 8.99	B+
7.00 a 7.99	B
6.00 a 6.99	B-
5.50 a 5.99	C+
5.00 a 5.49	C
< 5.00	D

Bibliografía:

- Básica
  - James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An introduction to statistical learning*. Vol. 6. New York: springer, 2013.
  - Bramer, Max. *Principles of data mining*. Vol. 180. London: Springer, 2007.
- Avanzada
  - Tan P.N., Steinbach M., and Kumar V., *Introduction to Data Mining*, Pearson, 2005.
  - Alpaydin, Ethem. *Introduction to machine learning*. MIT press, 2014.
  - Friedman, Jerome, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning*. Vol. 1. Springer, Berlin: Springer series in statistics, 2001.

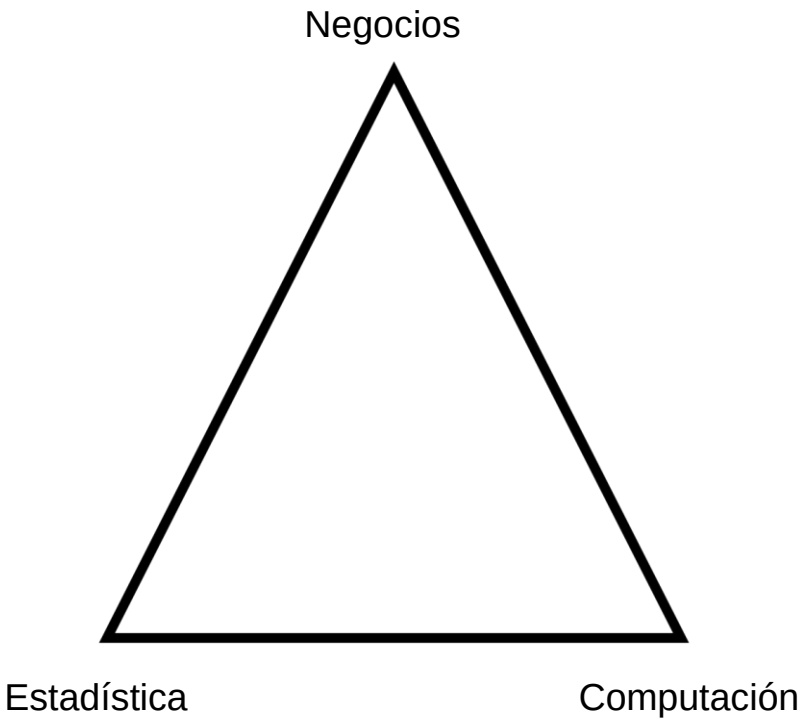
¿Qué es la minería de datos?

"*At a high level, **data science** is a set of fundamental principles that guide the extraction of knowledge from data. Data mining is the extraction of knowledge from data, via technologies that incorporate these principles.*" (Provost & Fawcett, 2013)

¿Qué es la minería de datos?

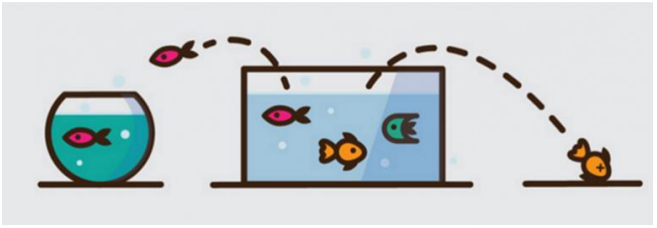
"*Data mining is a business process for exploring large amounts of data to discover meaningful patterns and rules.*" (Gordon & Berry, 2011)

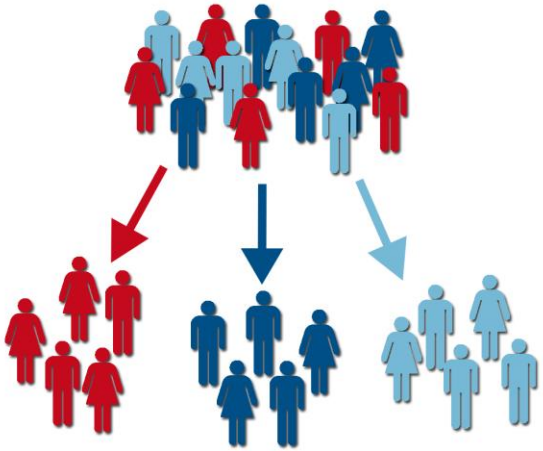
Noten que en este sentido, se requiere de habilidades en (al menos) tres campos:

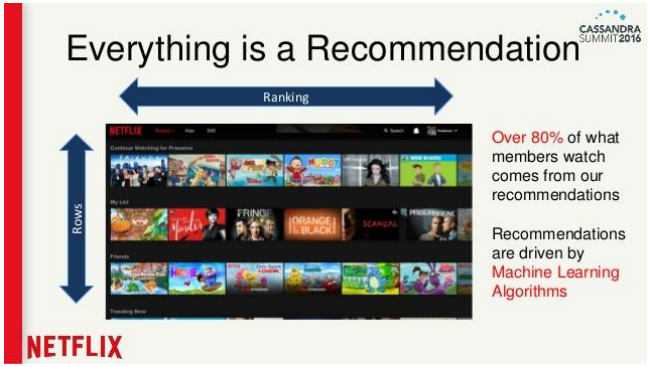


Ejemplo de de aplicaciones en negocios

- Churn (attrition) de clientes.
- Segmentación de clientes.
- Recomendación de productos.
- Publicidades personalizadas.
- Credit scoring.
- Predicción de valor de un cliente.
- Análisis de sentimiento.







Aprendizaje automático (panorama)

Existen **diversas técnicas para descubrir patrones** en grandes volúmenes de datos (por ejemplo, a través de la exploración "manual").

En este curso nosotros nos enfocaremos en utilizar técnicas de aprendizaje automático.

Definición “casi canónica”:

"*A **computer program** is said to learn from **experience E** with respect to some class of **tasks T** and **performance measure P** if its performance at tasks in T, as measured by P, improves with experience E.*" (Mitchell, T., 1997, *Machine Learning*).

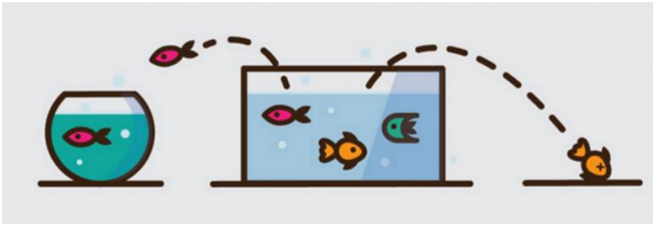
- *Computer program* ~= modelo estadístico.
- *Experience E* ~= datos.
- *Task T* ~= tarea (e.g., predecir el peso de un chico de acá a medio año).
- *Performance measure P* ~= medida de performance (e.g., *logloss*).

Aprendizaje automático (panorama)

Existen tres grandes familias de algoritmos de aprendizaje automático (supervisado, no supervisado, por refuerzos).

1. Supervisado (la que vamos a abordar principalmente en este curso):

"*For each observation of the predictor measurement(s) x_i, i = 1, ..., n there is an associated response measurement y_i*".



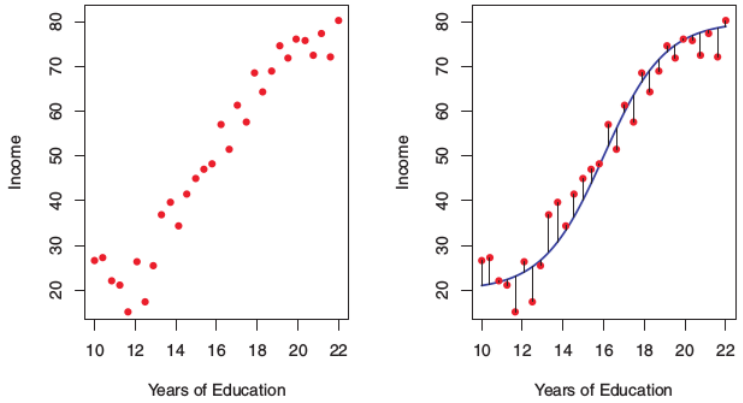


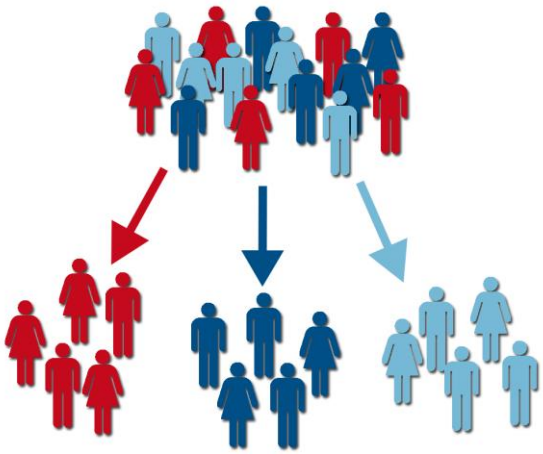
FIGURE 2.2. *The* `Income` *data set.* Left: *The red dots are the observed values of* `income` *(in tens of thousands of dollars) and* `years of education` *for* 30 *individuals.* Right: *The blue curve represents the true underlying relationship between* `income` *and* `years of education`*, which is generally unknown (but is known in this case because the data were simulated). The black lines represent the error associated with each observation. Note that some errors are positive (if an observation lies above the blue curve) and some are negative (if an observation lies below the curve). Overall, these errors have approximately mean zero.*

Aprendizaje automático (panorama)

Existen tres grandes familias de algoritmos de aprendizaje automático (supervisado, no supervisado, por refuerzos).

2. No supervisado (la vamos a abordar en menor medida):

"*Unsupervised learning describes the somewhat more challenging situation in which for every observation i = 1, ..., n, we observe a vector of measurements* x_i *but no associated response* y_i *… We can seek to understand the relationships between the variables or between the observations*".



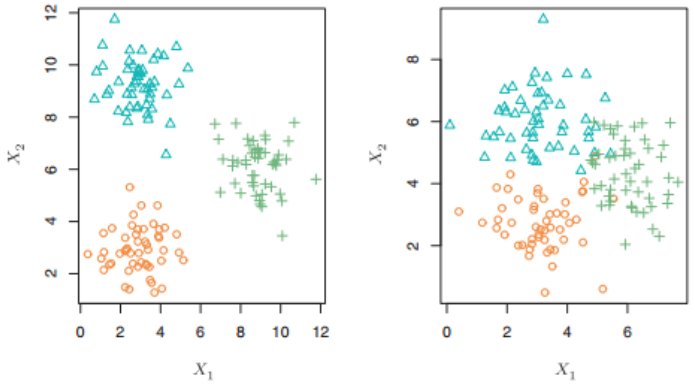


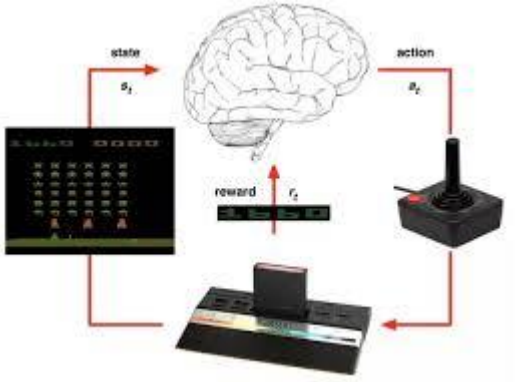
FIGURE 2.8. *A clustering data set involving three groups. Each group is shown using a different colored symbol.* Left: *The three groups are well-separated. In this setting, a clustering approach should successfully identify the three groups.* Right: *There is some overlap among the groups. Now the clustering task is more challenging.*

Aprendizaje automático (panorama)

Existen tres grandes familias de algoritmos de aprendizaje automático (supervisado, no supervisado, por refuerzos).

3. Por refuerzos (no la vamos a abordar):

A grandes rasgos, se enfoca en lograr que agentes maximicen un beneficio esperado mediante la interacción con un ambiente (el cual muchas veces no conocen)... Muchos problemas de la realidad se ajustan a este esquema.



Aprendizaje supervisado

Lo que caracteriza al aprendizaje supervisado es que uno quiere predecir una variable. Puede haber dos tipos de variables a predecir:

- Continua → regresión
- Categórica → clasificación (binaria, multiclase)

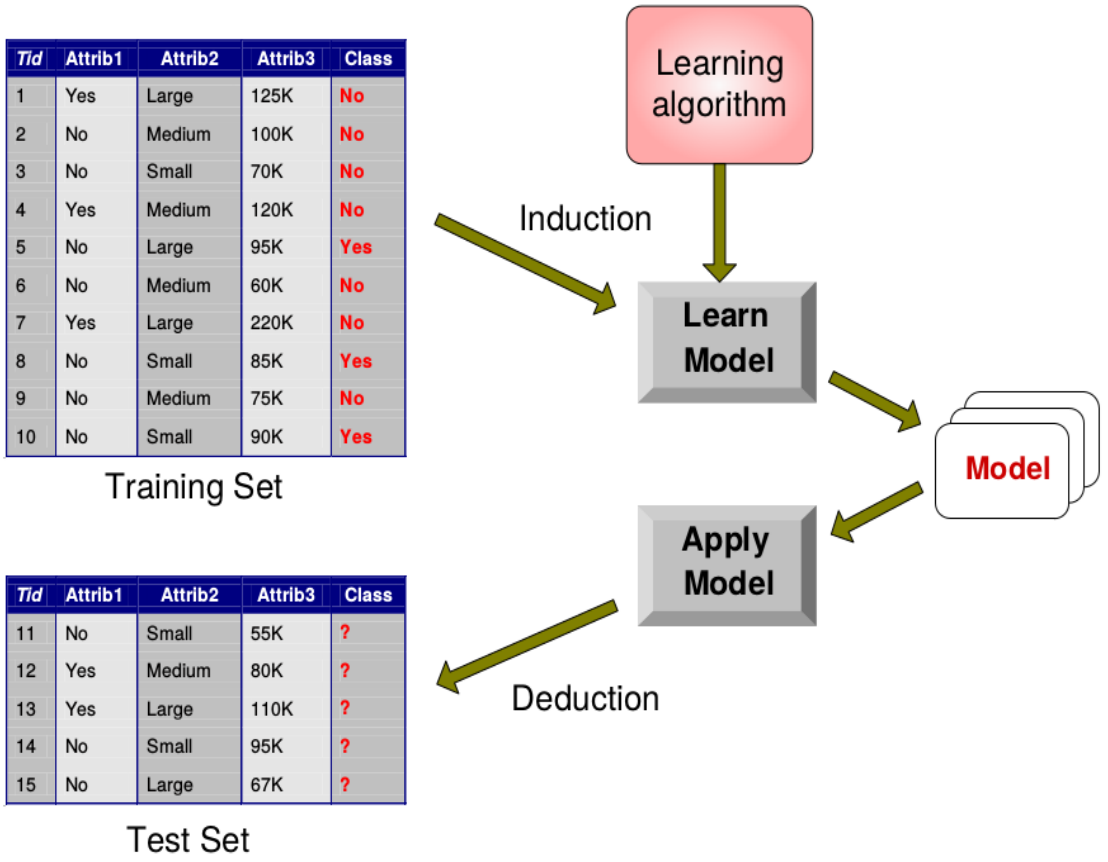
$$Y = f(X)$$

Distintos problemas se atacan con distintos algoritmos (algunos sirven tanto para regresión como clasificación).

¿Qué tipo de problema son los siguientes?

1. Predecir la cantidad de ventas de líneas telefónicas en un día dado.
2. Predecir si un cliente en particular se dará de alta dada nuestra campaña de marketing.
3. Predecir si un día determinado tendremos o no más de 200 líneas nuevas vendidas.

Aprendizaje supervisado



$$Y = f(X)$$

Veamos un primer algoritmo de aprendizaje supervisado.

Naïve Bayes

Bayes ingenuo (algoritmo de clasificación) modela la probabilidad de que una observación *i* con determinadas características (x_i) pertenezca a una determinada clase *k*, es decir: $P(C_k \mid x_i)$.

Desarrollemos la expresión:

$P(C_k \mid x_i) = P(C_k , x_i) / P(x_i)$ *por probabilidad condicional*

$P(C_k \mid x_i) = (P(x_i \mid C_k) * P(C_k)) / P(x_i)$ *Teorema de Bayes*

$P(C_k \mid x_i) = (P(v_1 = x_{i1} \wedge v_2 = x_{i2} \wedge \ldots v_q = x_{iq} \mid C_k) * P(C_k)) / P(v_1 = x_{i1} \wedge v_2 = x_{i2} \wedge \ldots v_q = x_{iq})$

Naïve Bayes

$$P(C_k \mid x_i) = (P(v_1 = x_{i1} \wedge v_2 = x_{i2} \wedge \ldots v_q = x_{iq} \mid C_k) * P(C_k)) \; / \; P(v_1 = x_{i1} \wedge v_2 = x_{i2} \wedge \ldots v_q = x_{iq})$$

Irrelevante (C_k no interviene!)

$$P(C_k \mid x_i) \propto (P(v_1 = x_{i1} \wedge v_2 = x_{i2} \wedge \ldots v_q = x_{iq} \mid C_k) * P(C_k))$$

Complejo de estimar

Para solucionar este problema el algoritmo hace un supuesto "ingenuo": las variables poseen independencia condicional (P(A, B | C) = P(A | C) * P (B | C)).

$$P(C_k \mid X_i) \propto (P(v_1 = x_{i1} \mid C_k) * P(v_2 = x_{i2} \mid C_k) * \ldots * P(v_q = x_{iq} \mid C_k)) * P(C_k))$$

Likelihood Prior

Al conjunto de supuestos usados para definir un modelo de aprendizaje se lo conoce como “sesgo inductivo” (Alpaydin, pp. 38).

Naïve Bayes

¿Cómo se pueden calculan las probabilidades condicionales para variables categóricas?

Véamoslo para Give Birth y para Can Fly

Name	Give Birth	Can Fly	Live in Water	Have Legs	Class
human	yes	no	no	yes	mammals
python	no	no	no	no	non-mammals
salmon	no	no	yes	no	non-mammals
whale	yes	no	yes	no	mammals
frog	no	no	sometimes	yes	non-mammals
komodo	no	no	no	yes	non-mammals
bat	yes	yes	no	yes	mammals
pigeon	no	yes	no	yes	non-mammals
cat	yes	no	no	yes	mammals
leopard shark	yes	no	yes	no	non-mammals
turtle	no	no	sometimes	yes	non-mammals
penguin	no	no	sometimes	yes	non-mammals
porcupine	yes	no	no	yes	mammals
eel	no	no	yes	no	non-mammals
salamander	no	no	sometimes	yes	non-mammals
gila monster	no	no	no	yes	non-mammals
platypus	no	no	no	yes	mammals
owl	no	yes	no	yes	non-mammals
dolphin	yes	no	yes	no	mammals
eagle	no	yes	no	yes	non-mammals

Naïve Bayes

Name	Give Birth	Can Fly	Live in Water	Have Legs	Class
human	yes	no	no	yes	mammals
python	no	no	no	no	non-mammals
salmon	no	no	yes	no	non-mammals
whale	yes	no	yes	no	mammals
frog	no	no	sometimes	yes	non-mammals
komodo	no	no	no	yes	non-mammals
bat	yes	yes	no	yes	mammals
pigeon	no	yes	no	yes	non-mammals
cat	yes	no	no	yes	mammals
leopard shark	yes	no	yes	no	non-mammals
turtle	no	no	sometimes	yes	non-mammals
penguin	no	no	sometimes	yes	non-mammals
porcupine	yes	no	no	yes	mammals
eel	no	no	yes	no	non-mammals
salamander	no	no	sometimes	yes	non-mammals
gila monster	no	no	no	yes	non-mammals
platypus	no	no	no	yes	mammals
owl	no	yes	no	yes	non-mammals
dolphin	yes	no	yes	no	mammals
eagle	no	yes	no	yes	non-mammals

Give Birth	Can Fly	Live in Water	Have Legs	Class
yes	no	yes	no	?

$$p(C_k)\prod_{j=1}^{q} p(x_{ij} \mid C_k)$$

$$P(A \mid M) = \frac{6}{7} \times \frac{6}{7} \times \frac{2}{7} \times \frac{2}{7} = 0.06$$

$$P(A \mid N) = \frac{1}{13} \times \frac{10}{13} \times \frac{3}{13} \times \frac{4}{13} = 0.0042$$

$$P(A \mid M)P(M) = 0.06 \times \frac{7}{20} = 0.021$$

$$P(A \mid N)P(N) = 0.004 \times \frac{13}{20} = 0.0027$$

*P(A | M) * P(M) > P(A | N) * P(N)* → Se predice mamífero

Naïve Bayes

Detalles:

El algoritmo se puede adaptar a atributos continuos. Estrategias posibles:

- Discretizando atributos categóricos (ya vamos a ver más sobre esto en la clase de ingeniería de atributos).
- Efectuando estimaciones de densidad.
  - Paramétricas: por ej. asumiendo normalidad (Tan, pp. 233)
  - No paramétricas, por ej. utilizando estimadores de tipo kernel (Alpaydin, pp. 186). MUY COSTOSO (en cómputo y espacio)

Naïve Bayes

Detalles:

¿Qué sucede en el ejemplo anterior con *P(lives in water = sometimes | mammal)*?

Name	Give Birth	Can Fly	Live in Water	Have Legs	Class
human	yes	no	no	yes	mammals
python	no	no	no	no	non-mammals
salmon	no	no	yes	no	non-mammals
whale	yes	no	yes	no	mammals
frog	no	no	sometimes	yes	non-mammals
komodo	no	no	no	yes	non-mammals
bat	yes	yes	no	yes	mammals
pigeon	no	yes	no	yes	non-mammals
cat	yes	no	no	yes	mammals
leopard shark	yes	no	yes	no	non-mammals
turtle	no	no	sometimes	yes	non-mammals
penguin	no	no	sometimes	yes	non-mammals
porcupine	yes	no	no	yes	mammals
eel	no	no	yes	no	non-mammals
salamander	no	no	sometimes	yes	non-mammals
gila monster	no	no	no	yes	non-mammals
platypus	no	no	no	yes	mammals
owl	no	yes	no	yes	non-mammals
dolphin	yes	no	yes	no	mammals
eagle	no	yes	no	yes	non-mammals

Naïve Bayes

Detalles:

¿Qué sucede en el ejemplo anterior con *P(lives in water = sometimes | mammal)*?

La probabilidad condicional estimada es 0, entonces la clase tendrá probabilidad estimada nula de ocurrir (esto no parece buena idea…).

Por esto se suele modificar la forma de estimar las probabilidades de la siguiente manera (suavizado laplaciano o aditivo):

$$P(v_j = x_{ij}|C_k) = \frac{n_{x_{jk}} + \alpha}{n_k + \alpha K}$$

Valores comunes:

- *K = # valores distintos de la variable x* (3 para "*Live in Water*").
- *α = 1* (add-one smoothing, ¿Por qué?).

Naïve Bayes

Probémoslo en R. Veamos el primer bloque de código.

Naïve Bayes

Noten que el modelo puede predecir tanto sobre los datos usados para entrenar como sobre los separados antes de entrenar.

Esto da lugar a dos performance.

¿Cuál métrica nos interesa optimizar?

¿Creen que sirve para algo la performance en training?

Naïve Bayes

Puntos a favor:

- Simple de entender y de implementar
- Se entrena fácilmente, incluso anda bien con datasets pequeños
- Es muy rápido.
- Es relativamente insensible a atributos irrelevantes.
- Puede manejar de manera simple valores nulos (revisen cómo lo hace la librería que usamos en R).
- Una vez entrenado, ocupa poco espacio (salvo que se usen estimaciones no paramétricas de densidad en atributos continuos).

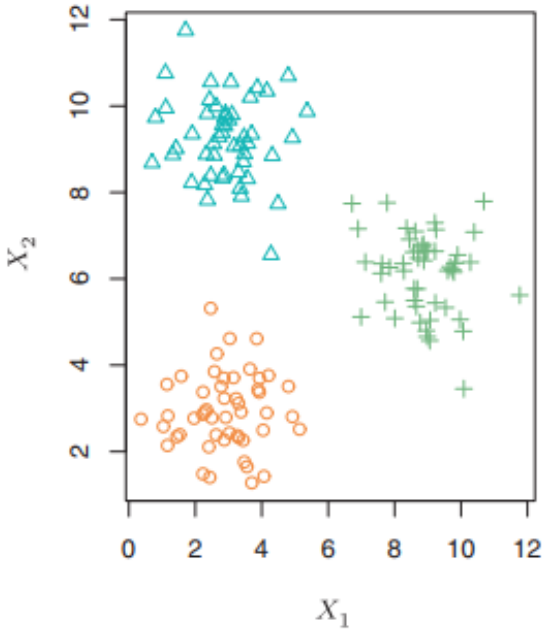
Puntos en contra:

- El supuesto de independencia condicional puede ser irreal.
- No suele tener gran performance (si algo anda peor que bayes ingenuo, muy probablemente algo esté mal).
- No descubre relaciones complejas, de modo que no aprovecha bien grandes volúmenes de datos.

K-nearest neighbors

Veamos un modelo “más complejo” que bayes ingenuo: el modelo de k-vecinos más cercanos (*k-nearest neighbors, knn*).

Para entenderlo, debemos acostumbrarnos a ver los registros como puntos en el espacio de atributos (*feature space*).



Noten que se puede proponer alguna medida de distancia que contemple todo par de observaciones.

K-nearest neighbors

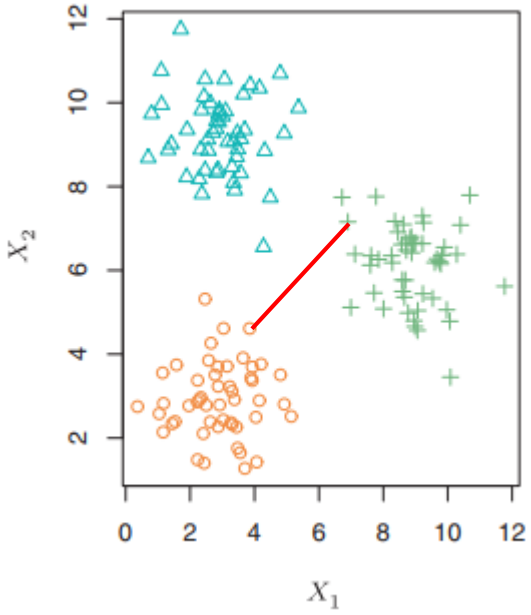
La medida de distancia más común de usar es la distancia euclídea:

$$\mathrm{d}(\mathbf{p},\mathbf{q}) = \mathrm{d}(\mathbf{q},\mathbf{p}) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \cdots + (q_n - p_n)^2}$$

$$= \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2}.$$

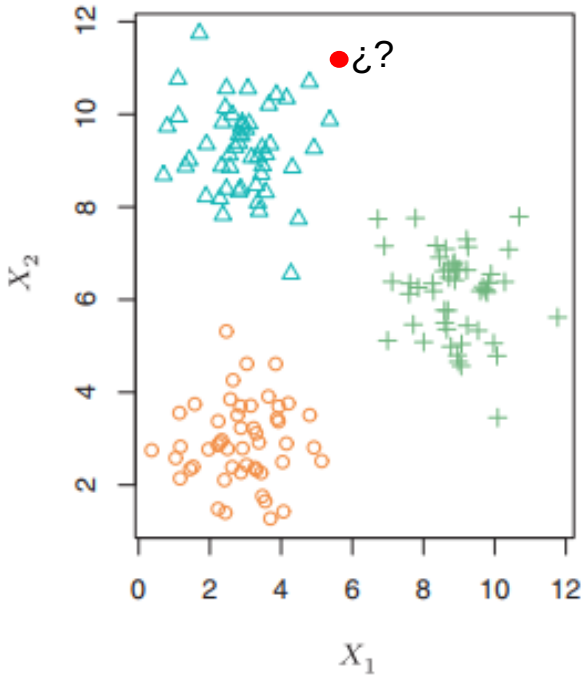
Algunas otras medidad de distancias/similitud alternativas son:

- Manhattan distance (parecida a euclidea).
- Mahalanobis distance (atributos normales).
- Cosine distance (embeddings).
- Levenshtein distance (secuencias).
- Hamming distance (atributos categóricos).



K-nearest neighbors

¿Supongamos que llega una nueva observación cuya clase no conocemos, qué clase parece razonable asignarle?



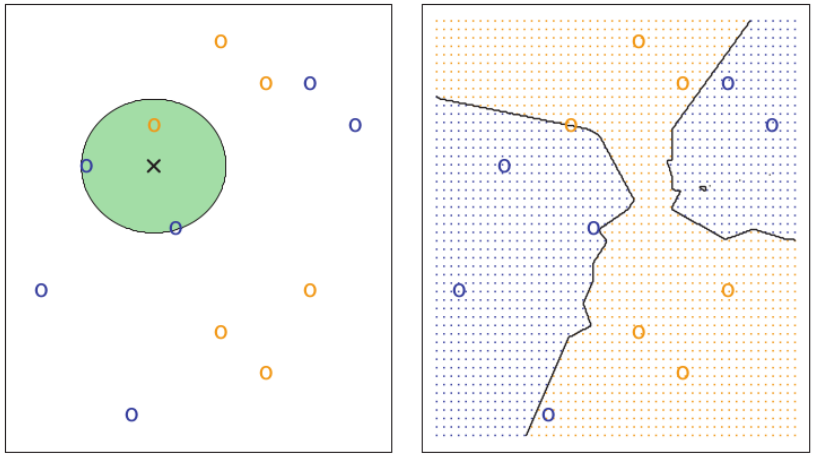
K-nearest neighbors

Dada una observación a clasificar (x_0) y un valor de K:

1. Se identifican los k vecinos (N_0) más cercanos de x_0.
2. Se estima la probabilidad condicional de pertenecer a la clase j como la fracción de observaciones de N_0 que pertenecen a j.

$$\Pr(Y = j|X = x_0) = \frac{1}{K} \sum_{i \in \mathcal{N}_0} I(y_i = j)$$

3. Se asigna como clase de x_0 a aquella clase para la cual se obtuvo la mayor probabilidad condicional.



K-nearest neighbors

El valor de *K* tiene un efecto dramático en el comportamiento del clasificador.

- Valores de *K* pequeños hacen que la frontera de decisión sea muy flexible.
- Valor de *K* grandes hacen que la frontera sea menos flexible.

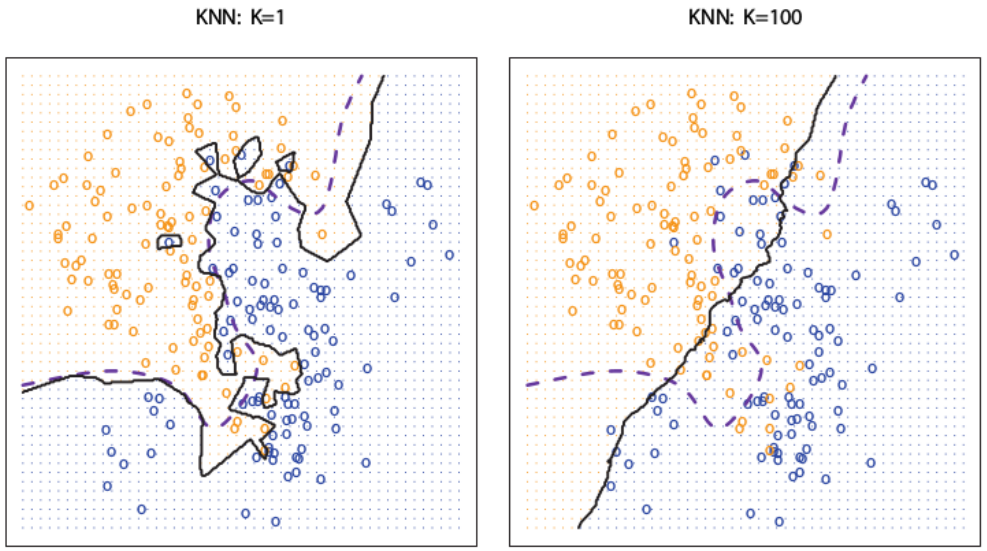
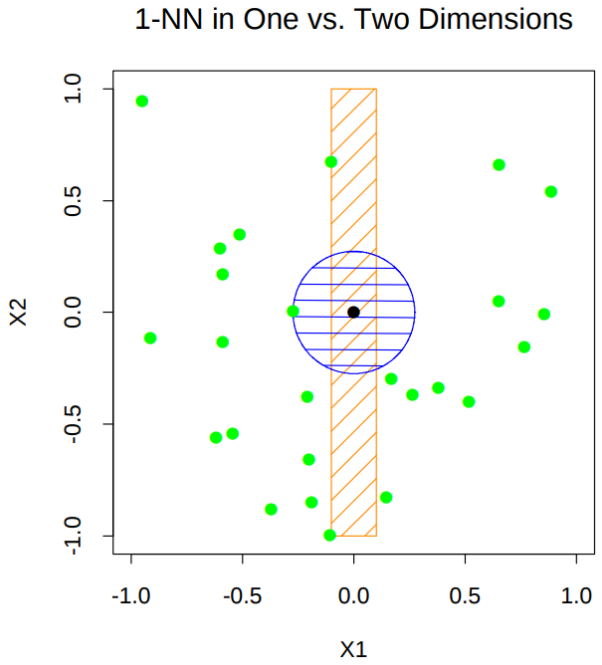


FIGURE 2.16. *A comparison of the KNN decision boundaries (solid black curves) obtained using $K = 1$ and $K = 100$ on the data from Figure 2.13. With $K = 1$, the decision boundary is overly flexible, while with $K = 100$ it is not sufficiently flexible. The Bayes decision boundary is shown as a purple dashed line.*

K-nearest neighbors

Vecinos más cercanos es un método no paramétrico. No asume distribuciones subyacentes en los datos (datos generados por funciones de densidad que pueden parametrizarse con un número finito de parámetros), sólo asume que inputs similares tendrán outputs similares.

Maldición de la dimensionalidad: en altas dimensiones es más difícil encontrar ejemplos cercanos (vean el ej. 4 de la Sección 4.7 de ISLR).



K-nearest neighbors

Consideraciones:

- Comúnmente las medidas de distancia se ven afectadas por las unidades de medida de las variables, por este motivo las variables suelen ser reescaladas/estandarizadas.

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)} \qquad x' = \frac{x - \bar{x}}{\sigma}$$

- En vecinos más cercanos, no hay una manera plenamente satisfactoria de trabajar con atributos categóricos, dos opciones comunes son:
  - Dar un puntaje de 0 si entre ambas observaciones el atributo tiene el mismo valor, y 1 si no lo tiene.
  - Usar one-hot-encoding.

#	Color
0	Red
1	Green
2	Blue
3	Red
4	Blue

→

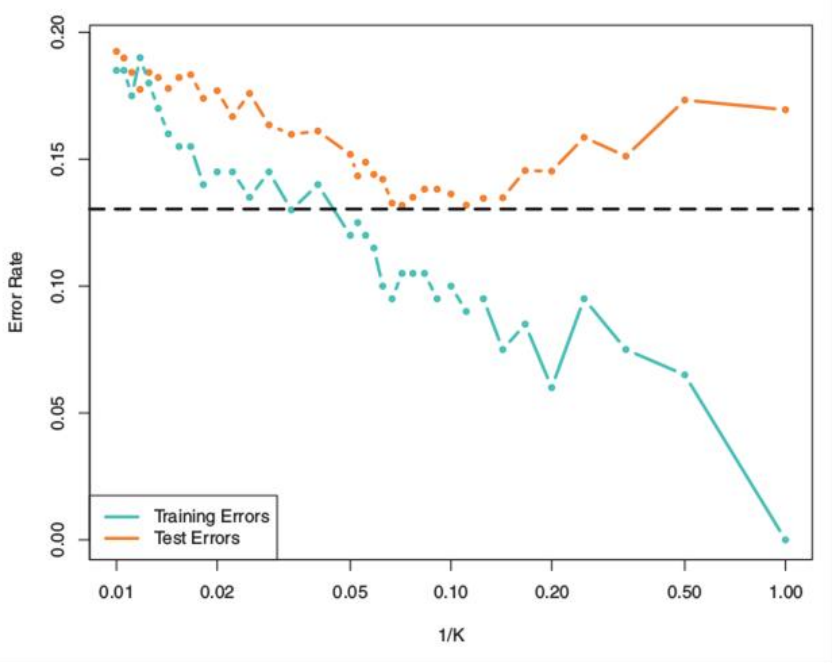
#	Red	Green	Blue
0	1	0	0
1	0	1	0
2	0	0	1
3	1	0	0
4	0	0	1

K-nearest neighbors

Probémoslo en R. Veamos el segundo bloque de código.

K-nearest neighbors

Generalmente la performance de un algoritmo depende de parámetros que se definen antes que el algoritmo comience a aprender de los datos, pero que influyen cómo se aprende los mismos (por ej.: *K* en *K*-nearest neighbors y *α* en naïve bayes con suavizado laplaciano).



A estos inputs se los conoce como "hiperparámetros". Cómo elegir valores buenos para los mismos va a ser uno de los ejes de este curso.

K-nearest neighbors

Probemos diferentes valores de *K* en R. Veamos el tercer bloque de código.

K-nearest neighbors

El algoritmo puede adaptarse fácilmente a problemas de regresión (algo que ocurre con muchos algoritmos).

En vez de tomar la clase mayoritaria en N_0, se toma el promedio de la variable a predecir.

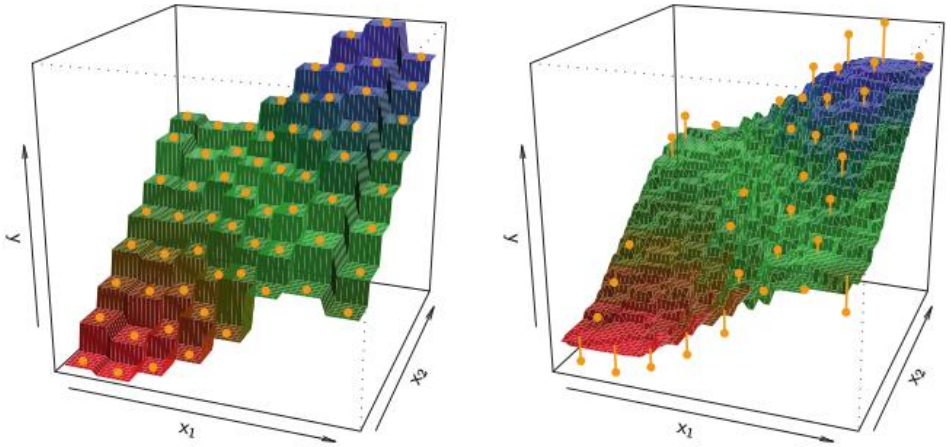


FIGURE 3.16. *Plots of $\hat{f}(X)$ using KNN regression on a two-dimensional data set with 64 observations (orange dots).* Left: $K = 1$ *results in a rough step function fit.* Right: $K = 9$ *produces a much smoother fit.*

$$\hat{f}(x_0) = \frac{1}{K} \sum_{x_i \in \mathcal{N}_0} y_i$$

¿Puede adaptarse naïve bayes a regresión?

K-nearest neighbors

¿Cuándo hace el trabajo más duro Naïve Bayes, al aprender de los datos o al clasificar un nuevo registro? ¿y K-nearest neighbors?

K-nearest neighbors

¿Cuándo hace el trabajo más duro Naïve Bayes, al aprender de los datos o al clasificar un nuevo registro? ¿y K-nearest neighbors?

Ambos algoritmos pertenecen a dos familias distintas.

- *Lazy*: no computan el modelo cuando se les dan los datos de entrenamiento y posponen el grueso del cómputo a cuando se los usa para predecir.

- *Eager*: computan el modelo al recibir los datos de entrenamiento, y guardan los patrones aprendidos de los mismos para luego usarlos para predecir.

¿Qué guarda Naïve Bayes para clasificar instancias futuras? ¿y K-nearest neighbors?

¿Qué desventaja ven a los algoritmos *lazy*?

K-nearest neighbors

Puntos a favor:

- Simple de implementar.
- Permite incorporar medidas de distancia que tengan sentido en el dominio que se está trabajando.
- Maneja de una manera natural problemas multi-clase.
- Puede llegar a tener un buen desempeño de disponer de suficientes datos representativos.

Puntos en contra:

- El problema de buscar los vecinos puede ser costoso computacionalmente.
- Almacenar el modelo puede resultar muy costoso.

Lecturas recomendadas para los temas vistos hoy

Noción de aprendizaje automático/estadístico

- ISLR (Cap 2)

Noción de aprendizaje supervisado

- ISLR (Cap 2)

Naïve Bayes

- Bramer (cap 3), Tan (Secciones 5.3.1, 5.3.2, 5.3.3)

K-nearest neighbors

- Bramer (cap 3), ISLR (Cap 2, Sección 3.5)

Práctica de laboratorio

Para hacer:

El dataset "*bankruptcy_data_red.txt*" contiene datos de empresas, algunas de las cuales entraron en bancarrota (los mismos son un subconjunto de los datos aquí disponibles).

- Separe al azar el dataset en un conjunto de entrenamiento y en otro conjunto que no se use para entrenar. Se pide que este segundo conjunto tenga 300 observaciones.
- Entrene un modelo de bayes ingenuo y evalúe la performance (accuracy) en los dos conjuntos de datos.
- Responda: ¿dado este dataset, tiene sentido la estrategia de add-one smoothing? ¿por qué? (vea qué tipo de variables tiene este dataset)
- Entrene y evalue modelos de k-vecinos más cercanos con distintos valores de k. Realice este ejercicio con las variables escaladas y no escaladas. Evalúe la performance en los dos conjuntos de datos.
- Responda: ¿dados sus experimentos, cuál es el valor óptimo de *K*?
- Responda: ¿dados sus experimentos, conviene escalar las variables?
- Repita los ejercicios anteriores de punta a punta (incluido el de partir los datos en dos conjuntos), ¿los resultados obtenidos y valores óptimos de *K* se mantienen inalterados?

Práctica teórica

¿Si en Bayes ingenuo al usar suavizado aditivo se aumenta el valor de α haciendo que tienda a infinito, a qué valor tiende $P(C_k \mid x_i)$? Justifique su respuesta.

¿Si se usa vecinos cercanos para regresión, a medida que aumenta el número de vecinos a considerar por el algoritmo, a qué valor tiende la predicción? Justifique su respuesta.

¿Por qué Bayes ingenuo no es bueno captando interacciones complejas entre variables? Justifique su respuesta.

¿Si usted tuviera la necesidad de, una vez entrenado el modelo, predecir muy rápido sobre nuevas observaciones, elegiría el modelo de bayes ingenuo o el de vecinos más cercanos? Justifique su respuesta.