Sistemas de Aprendizaje Automático

Conforme a contenidos del «Curso de Especialización en Inteligencia Artificial y Big Data»

Universidad de Castilla-La Mancha

Escuela Superior de Informática
Ciudad Real

Presentación del módulo Sistemas de Aprendizaje Automático

Las técnicas de *aprendizaje automático* o *Machine Learning* están adquiriendo gran notoriedad hoy en día de la mano de una digitalización creciente de nuestras sociedades. Estas técnicas se están aplicando en todos los sectores económicos, con objetivos diversos, que van desde el incremento de la eficiencia y eficacia de los procesos productivos a la automatización de actividades económicas.

En este módulo se abordan las principales técnicas para el desarrollo de sistemas de aprendizaje automático. Los **resultados de aprendizaje** perseguidos en este módulo son que los alumnos sepan seleccionar y aplicar los algoritmos más adecuados del aprendizaje supervisado y no supervisado al problema que se intenta resolver, optimizando el resultado del modelo aplicado.

Los **contenidos** básicos del curso y su planificación temporal se muestran en la tabla 1.1

El módulo se desarrollará sobre la herramienta **Scikit-learn** que proporciona una biblioteca para aprendizaje automático de software libre para el lenguaje de programación Python. Esta herramienta se complementará con **Keras** que es una biblioteca para redes neuronales escrita en Python. Está especialmente diseñada para usar redes neuronales profundas. Este software es el adecuado para un modo de trabajo local. Para evitar problemas de versiones se empleará **Google Colab** para ejecutar programas en Python en el navegador. Esta herramienta en la nube permite compartir ficheros fácilmente y evita definir la configuración.

Todos los módulos del curso de especialización en Inteligencia Artificial y Big Data comparten la misma metodología docente. La carga docente de cada semana del módulo es de $0{,}5$ ECTS, desarrollándose los siguientes recursos/actividades didácticas:

- **Clases teóricas y prácticas** mediante 8 vídeos de unos 5 minutos en los que se impartirán de forma no concurrente los contenidos básicos y ejemplos prácticos, con el código fuente para realizar las actividades.

SEMANA 1.	Inteligencia artificial fuerte y débil. Modelización matemática y aprendizaje automático.
SEMANA 2.	Herramientas para el aprendizaje automático. Pre-procesamiento.
SEMANAS 3. 4. y 5.	Algoritmos y herramientas para el aprendizaje supervisado. *kNN, NaiveBayes, DecisionTree, RandonForests, Boosting, Logistic Regression y SVM.*
SEMANAS 6. y 7.	Algoritmos y herramientas para el aprendizaje no supervisado. *K-means, Fuzzy C-means, DBSCAN, algoritmos jerárquicos.*
SEMANA 8.	Modelos de redes neuronales. *Perceptrón, Perceptrón, Multicapa.*
SEMANA 9.	Modelos de redes neuronales profundas. *Red Neuronal Convolucional.*
SEMANA 10.	Validación de modelos.

Tabla 1.1: Planificación temporal de los contenidos

- **Documentación.** Los contenidos de estos vídeos serán recogidos y complementados en un material escritos de aproximadamente 15 páginas/semana.
- **Ejercicios de autoevaluación H5P.** Cada semana los alumnos contendrán unos 4 ejercicios de autoevaluación.
- **Ejercicios propuestos complementarios**.
- **Tutorías**. Todas las semanas se dispondrá de una hora de tutorías para resolver las dudas sobre los materiales desarrollados en dicha semana.

Empleamos como principio metodológico para el módulo el de **aprender a hacer se aprende haciendo**. La evaluación se alinea a esta visión y se realiza a través de tres actividades consistentes en la resolución de un problema de clasificación, uno de regresión y haciendo un análisis cluster.

Caracterización de la Inteligencia Artificial fuerte y débil

1.1. Inteligencia Artificial fuerte

El término de Inteligencia Artificial (IA) en sus orígenes se le asignaba a un sistema hardware-software que operaba del mismo modo que la inteligencia humana. El reconocimiento de los límites de la Inteligencia Artificial ha creado dos subcategorías: la Inteligencia Artificial fuerte y la débil. El concepto de inteligencia en IA fuerte es entendido como *la capacidad del sistema de actuar adecuadamente en un entorno con incertidumbre*. Aquí se enfatiza en la capacidad de abordar la universalidad de las situaciones que pueden aparecer en el dominio de aplicación. Debemos asumir una premisa para pensar que un ordenador pueda simular la mente humana, y es que el pensamiento humano puede ser implementado computacionalmente de un modo formal y no ambiguo. En esta situación el ordenador puede llegar a alcanzar un estado de auto-consciencia y *entender* los objetos del mundo en una forma activa y autónoma.

Esta visión de la IA tiene su origen en la máquina de Turing universal la cual puede resolver todos los problemas que son solucionables por un algoritmo o un método efectivo de computación. Si todos los procesos de resolución de problemas pudiesen ser simplificados a procesos computacionales, entonces una única máquina, con relativamente operaciones simples, podría resolverlos con una adecuada serie de comandos. La máquina requiere una memoria infinita de almacenamiento y que ésta pueda ser tratada.

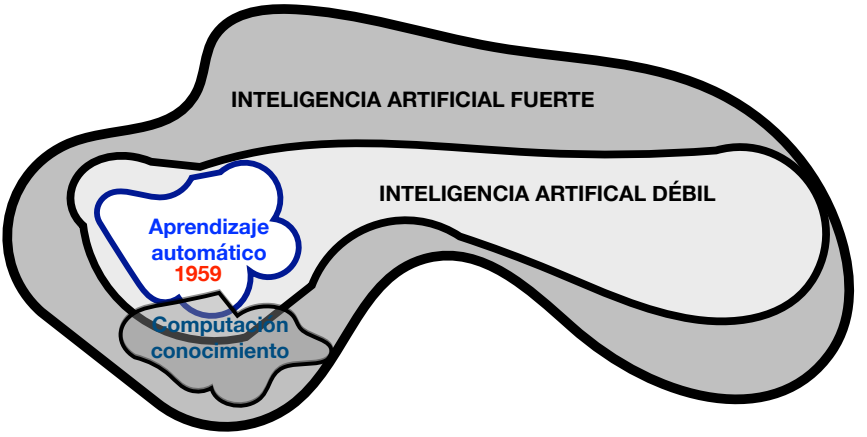

Figura 1.1: Dominio de la inteligencia artificial fuerte, débil y aprendizaje automático.

1.1.1. Inteligencia Artificial débil

La IA débil intenta construir sistemas para ejecutar eficientemente ciertas tareas intelectuales que un ser humano puede realizar. Este precepto conduce a que el desarrollo de IA débil no requiere implementar exhaustivamente la inteligencia humana para obtener las funcionalidades del sistema deseado. La IA débil implementa agentes que abordan tareas en un específico entorno

El desarrollo de la IA ha ido evolucionando por un visión débil de la misma, marcada por hechos clave como la aparición de las redes neuronales o la introducción del término *machine learning* (aprendizaje automático) en 1959 por Arthur Samuel. No obstante, la investigación en IA para simular el comportamiento humano ha contribuido a crear nuevas disciplinas como la Computación del Conocimiento focalizada en *construir máquinas que tengan habilidades de razonamiento análogas al cerebro humano.* La Figura 1.1 sintetiza las relaciones entre estos conceptos. La relación fundamental es que las técnicas de aprendizaje automático constituyen una parte de la IA débil.

La IA débil ha tenido numerosos éxitos que han llegado a la opinión pública. La empresa IBM construyó el ordenador *Deep Blue* que derrotó al campeón del mundo de ajedrez el 11 de Mayo de 1997. Desarrollaron el ordenador *Watson* que podía procesar lenguaje natural y aprender sin supervisión de documentos. En Febrero de 2011, *Watson* ganó a dos campeones previos en el concurso televisivo estadounidense *Jeopardy*, que trata de preguntas de conocimientos sobre una variedad diversa de temas. Este hecho mostró que la capacidad de *Watson* de entender las preguntas formuladas en lenguaje natural, buscar en bases de datos los hechos relevantes y hallar la respuesta correcta. Otrass empresas como *Gooogle* emplean redes neuronales profundas para tareas de reconocimiento de imágenes La red *GoogLeNet* fue la ganadora del concurso *2014 ImageNet Large-Scale Visual Recognition Challenge*. En el años 2021, *DeepMind*, una empresa comprada

por Google, es capaz de predecir con gran exactitud la estructura de las proteínas que forman un ser humano. Esta información puede resultar esencial para el desarrollo de tratamiento de las enfermedades. Esta herramienta de IA va a cambiar radicalmente la investigación biológica entre otros motivos porque acelera el proceso. Los métodos tradicionales requieren meses o incluso años para dilucidar la forma de una proteína y este sistema lo hace en minutos aunque con ciertos errores.

J. C. R. Licklider fue un pionero de la IA, pronosticando en la década de los años 60 del siglo pasado que la aparición de la IA fuerte no iba a ser inminente y requeriría un periodo interino *de entre 10 y 500 años* en el cual los humanos y los ordenadores existiría una relación simbiótica en *el que las máquinas ayuden de forma efectiva el procesos de pensamiento.* Argumentó que trabajarían con los humanos, existiendo como organismos diferentes, pero viviendo juntos en una asociación íntima que mejoraría ciertas capacidades del pensamiento humano. El reto no el de construir máquinas que simulen la mente humano sino entender el proceso de conocimiento humano y desarrollar máquinas que ayuden/realicen aquellos aspectos de la solución de problemas más difíciles o tediosos para el ser humano.

1.2. Inteligencia artificial fuerte y débil: problemas morales, éticos y sociales

El desarrollo de la IA fuerte se circunscribe a restricciones tecnológicas que limiten y guíen su evolución. Pero no es solamente ese el único aspecto que debe dirigir su evolución sino también los problemas éticos, morales y sociales que aparecen en una nueva realidad. Estados Unidos, Rusia y China apuestan decididamente, sin limitaciones, sobre el uso de la inteligencia artificial (débil). En la Unión Europea se ponen límites a su uso con leyes de protección de datos. El campo de discusión y desarrollo debe delimitarse, y marcar límites éticos donde desarrollar y aplicar la tecnología. En esta sección se plantean situaciones nuevas que pudieran originarse con el desarrollo de la IA (débil o fuerte). El análisis y la discusión de las siguientes situaciones, planteadas como actividades donde el lector contesta a las preguntas que se formulan, evidencia la existencia de posibles riesgos legales y éticos de la aplicación de Inteligencia Artificial y evidencia que estas otras consideraciones deben de estar también presentes en el desarrollo de la misma.

A continuación se plantean dos situaciones para intentar contestar la cuestión ¿Deberíamos tratar a un sistema de IA con consideraciones morales-éticas?.

hitchBOT es un robot de cuerpo cilíndrico, fabricado con un cubo de plástico, con brazos y piernas unidos, y una pantalla LED que mostraba ojos y una boca. Estaba programado para pedir a las personas que lo recogieran y los llevara con ellas. Cruzó con éxito Canada y algunos países europeos. En agosto de 2015, apareció decapitado en Filadelfia. ¿Hicieron algo malo los vándalos al decapitar a HitchBOT más allá de destruir la propiead privada? ¿Tenían la obligación moral de no interferir en su estado normal de funcionamiento y dejarle continuar su viaje por el mundo?

Replicantes de la película *Blade Runner* (1982). Supongamos que un día se pudiera fabricar una "persona"de "plástico". La IA fuerte ha triunfado. Piensa, siente y habla como nosotros. Se diferencia en los materiales, se diferencia en cómo llega a este mundo. Roba, mata, crea, ama, siente, ... como nosotros. ¿Cual es la ley que se le debería aplicar a las máquina? ¿Qué significa *matar* una máquina? ¿Qué derechos morales tendría dicha máquina? ¿Es solo una máquina? Un punto de vista bajo el cual es relevante analizar las respuestas es empleando el argumento aristotélico de equidad: los mismos casos se deberían tratar del mismo modo.

La tecnología cambia el mundo y modifica las relaciones sociales y económicas existentes. Las siguientes dos situaciones va en esa dirección y cuestionan por dos escenarios hipotéticos.

Cambios económicos. Supóngase que el nivel de automatización alcanzado provoca una reducción drástica el mercado de trabajo. Miles de empleos desaparecen y otros nuevos se generan (pero en un número insuficiente para la población). ¿Qué mecanismos de redistribución de rentas se deberían establecer? ¿Deberían pagar impuestos los robots (ciertos tipos)? Las inversiones en tecnología productiva hacen que solo las grandes corporaciones tengan acceso, concentrando la capacidad productiva y poseyendo un poder superior a muchos estados. ¿Qué compromisos sociales se les debe imponer? ¿Se debe circunscribir a impuestos o a otros aspectos?

Cambios sociales. *Yo soy yo y mis circundantes: Sistemas híbridos hombre-máquina*. En la actualidad el nivel de integración entre la tecnología y el ser humano es muy limitado. Existen dispositivos que corrigen defectos en la visión o la audición de un ser humano. La tecnología puede ir evolucionando y mejorando diferentes capacidades humanas, convirtiendo ciertos desarrollos tecnológicos en artefactos de uso común. En ese futuro, los límites del ser humano se desdibujan y se cuestionará cuales son. ¿Qué límites hombre-máquina se definirán para evaluar capacidades en el trabajo, en el colegio, en el deporte, etc?

En el momento actual la IA ya tiene un impacto en nuestras sociedades. Las siguientes tres situaciones son ejemplos de ello y por tanto la reflexión sobre su uso debe ir de la mano de su desarrollo.

El coche autónomo ya es una realidad. Supóngase un escenario donde los coches autónomos conviven con coches conducidos por particulares. Esta situación se dará en cualquier otro modo de transporte, como metro autónomos, drones, barcos, etc. Supóngase que se produce un accidente en el que está involucrado un coche autónomo, ¿Quien es el responsable? ¿El propietario del vehículo?¿El fabricante del vehículo? ¿El desarrollador del software?

Los sistemas de reconocimiento facial han alcanzado un alto grado de sofisticación y fiabilidad. Estos sistemas permitirían implementar una vigilancia masiva de la ciudadanía. Por motivos de seguridad ¿Queda justificado el uso de estos sistemas? ¿Se podrían implantar pero restringiendo su uso? Por ejemplo, un juez autoriza el visionado de imágenes para la búsqueda de una persona determinada con una finalidad específica. O por el contrario, empleando la despedida de Robe Iniesta al final de un concierto de Extremoduro: *Recordad!!, este es un país libre!!, podéis ir y hacer lo que queráis!! pero que no os vean.*

Los algoritmos de IA (débil) aprenden el mundo tal cual es, con sus sesgos, sus discriminaciones, abusos y perpetuando el *status quo*. En el libro *Armas de destrucción matemática* su autor O'Neil expone ejemplos de como los algoritmos califican a maestros y estudiantes, ordenan currículos, conceden (o niegan) préstamos, evalúan a los trabajadores, se dirigen a los votantes, fijan la libertad condicional y monitorean nuestra salud. Los ciudadanos se sitúan frente a esta nueva tecnología como se hizo en el pasado con la teoría económica. La teoría económica es Ciencia, es inexorable, tiene sus reglas objetivas, incuestionables, que impactan en la vida de las personas sin que estas intervengan u opinen. ¿Qué autoridad o poder de decisión debe tener los algoritmos? ¿Cómo los algoritmos deben rendir cuentas y a quien? ¿Cuales decisiones deben ser justificadas y cómo?

Caracterización de sistemas de aprendizaje automático

1.3. Construcción de modelos de aprendizaje automático

1.3.1. ¿Que es un modelo?

Los métodos de aprendizaje automático son ejemplos notables de la llamada IA débil. En los métodos de aprendizaje automático el dominio de aplicación está delimitado y es conocido. Estos métodos elaboran un **modelo matemático** de esta realidad. Según Marvin Lee Minsky, M es un modelo (matemático) de una realidad R para un observador O si se cumple las siguientes dos condiciones:

i) El modelo M contesta a las preguntas que el observador O hace sobre la realidad R

ii) Los datos que genera el modelo M sobre la realidad R reproducen las observaciones que se dispone del fenómeno R.

La condición i) apunta a que tipo de datos deben registrarse y ser usados. Estos deben estar alineados con los objetivos del estudio. La condición ii) afirma que los modelos no intentan *sintetizar la verdad* sino simplemente reproducir las observaciones. Por ejemplo, la física Newtoniana será un modelo válido para un observador O de la vida cotidiana pero no para un observador O' que analice problemas de astrofísica, ya que esta teoría no es capaz de reproducir las observaciones relativas a la distorsión del espacio y tiempo en torno a un cuerpo gravitacional. En este caso hay que elaborar nuevos modelos como la teoría de la relatividad de Einstein.

La Figura 1.2 esquematiza el concepto de modelo. El aprendizaje automático considera *la realidad* (dominio de aplicación) como una *caja negra* que para una entrada $\mathbf{x}$ genera una salida $\mathbf{y}$. El modelo de aprendizaje automático simula este proceso sin reproducir los mecanismos causales que operan en la realidad. El modelo busca exclusivamente que las predicción del modelo (denotada $\hat{\mathbf{y}}$) y las observaciones sean coincidentes.

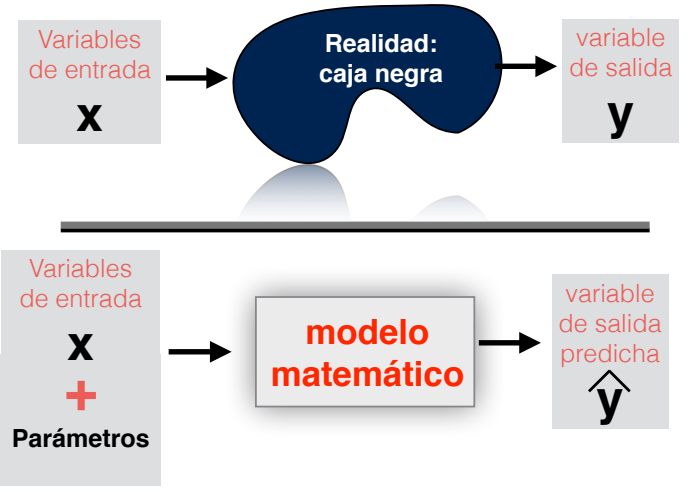

Figura 1.2: Modelización matemática.

De forma sintética un modelo matemático se puede expresar:

$$\mathbf{y} = F_{\boldsymbol{\theta}}(\mathbf{x}, \boldsymbol{\alpha}) \tag{1.1}$$

donde $\mathbf{y}$ es la respuesta del sistema (variable respuesta) que puede ser un número, un vector, una matriz, etc., y $\mathbf{x}$ son los datos de entrada que a su vez puede ser un número, un vector, una matriz, etc. El símbolo $\boldsymbol{\alpha}$ representa un vector de **parámetros** (en el contexto de las redes neuronales se les llaman los pesos). La familia de funciones (modelos) $F_{\boldsymbol{\theta}}$ está parametrizada por un conjunto diferente de parámetros $\boldsymbol{\theta}$ que reciben el nombre de **hiperparámetros**. Esto permite ensayar más de un solo modelo, mejorando el resultado final.

El problema de **entremaniento** (ajuste) del modelo consiste en encontrar, para un valor dado de los hiperparámetros $\boldsymbol{\theta}$, el vector de parámetros $\boldsymbol{\alpha}$ que mejor describa las observaciones existentes.

Ejemplo. Un modelo para relacionar dos variables $\mathbf{x}$ e $\mathbf{y}$ puede ser una función polinomial. En este caso se tiene la siguiente expresión de la variable respuesta en función de la variable de entrada, parámetros e hiperparámetro:

$$\mathbf{y} = F_{\boldsymbol{\theta}}(\mathbf{x}, \boldsymbol{\alpha}) = \alpha_0 + \alpha_1 x + \alpha_2 x^2 + \ldots + \alpha_\theta x^\theta \tag{1.2}$$

En este ejemplo el grado del polinomio θ es el hiperparámetro del modelo. Este parámetro permite considerar una variedad amplia de modelos. Por ejemplo, si tomamos $\theta = 1$ estamos considerando una recta, si elegimos $\theta = 2$ estaríamos analizando parábolas. Fijado el valor de θ se trata de encontrar los coeficientes del polinomio (problema de entrenamiento) que mejor describan los datos. La Figura 1.3 muestra dos ejemplos. Supóngase que tenemos tres datos (cuadrados rojos), si nos planteamos un modelo lineal $F_{\boldsymbol{\theta}}(\mathbf{x}, \boldsymbol{\alpha}) = \alpha_0 + \alpha_1 \mathbf{x}$ para describir los datos (gráfica de la derecha) tendríamos que ajustar dos parámetros (α_0, α_1) para encontrar la recta *más cercana a los puntos*. Si decidimos buscar una parábola tendríamos un modelo con tres parámetros $F_{\boldsymbol{\theta}}(\mathbf{x}, \boldsymbol{\alpha}) = \alpha_0 + \alpha_1 \mathbf{x} + \alpha_2 \mathbf{x}^2$. Si ajustamos los tres parámetros adecuadamente podemos encontrar una parábola que pase por los tres puntos. Cuando buscamos un modelo que reproduzca exactamente las observaciones se denomina un problema de interpolación mientras que si lo que se desea es que se describa aproximadamente el conjunto de observaciones se denomina un problema de aproximación.

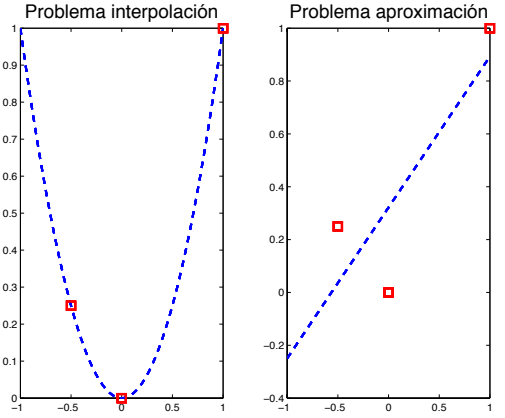

Figura 1.3: Modelo cuadrático y lineal.

1.3.2. Principales técnicas para desarrollar modelos de aprendizaje automático

El aprendizaje automático incluye un conjunto de técnicas que permiten que los ordenadores aprendan a realizar unas tareas determinadas. Algunas de estas técnicas se basan en un **aprendizaje deductivo**, que parte de reglas generales para obtener conclusiones particulares. En estos casos se especifican mecanismos causales en la construcción del modelo $F_{\boldsymbol{\theta}}$. Un ejemplo de estas técnicas es el razonamiento basado en casos que es usado en los **sistemas expertos**. Los sistemas expertos es un modelo de inteligencia artificial que intenta imitar el comportamiento de un ser humano experto en alguna temática.

Otras técnicas, como árboles de decisión o redes neuronales artificiales, emplean un **aprendizaje inductivo**, ya que a partir de la observación y el análisis de ejemplos concretos se desarrollan modelos que explican dichos datos y que permiten llevar a cabo una generalización. En estos casos los modelos F_{θ} actúan como *cajas negras*. En este módulo vamos a estudiar exclusivamente este último tipo de modelos.

1.4. Clasificación de sistemas de aprendizaje automático

La clasificación de los problemas de aprendizaje automátio se basan en dos criterios: i) la posibilidad o no de conocer la variable respuesta $\boldsymbol{y}$ y ii) la tipología de la variable respuesta $\boldsymbol{y}$. Por este motivo comenzaremos estableciendo los tipos de variables que se pueden considerar en un problema.

1.4.1. Tipología de las variables

Una clasificación de las variables en función de su naturaleza es:

- **Variables cuantitativas.** Los datos toman valores numéricos. A su vez se distingue entre **variables continuas** que toman valores en un intervalo de números reales o **variables discretas**, que toman un número finito de valores numéricos. Ejemplos de variables continuas son la edad, ingresos, altura, etc y como variables discretas el número de hijos, o el número de días realizando un tratamiento.

- **Variables ordinales.** Los datos expresan relación de orden entre las observaciones. Por ejemplo podemos considerar un *ranking* en ciertos tratamientos como el mejor (1), el segundo mejor (2), el tercer mejor (3), etc. Otro ejemplo son las variables que ordenan las preferencias de ciertos usuarios.

- **Variables cualitativas.** En este caso se expresa una cualidad de un objeto. Estas variables solo pueden tomar un conjunto de valores que no miden ninguna magnitud determinada. Por ejemplo, los sujetos de un experimento pueden ser hombre o mujer. Estos datos los podemos codificar dándole el valor 1 a los hombres y 2 a las mujeres. Esta variable no representa que el valor 2 es el doble del valor 1 (si fueran datos cuantitativos) o que unos son los segundos y los otros los primeros (datos ordinales) sino que son dos categorías diferentes. Este es un ejemplo de **variables binarias** en las que solo se pueden tomar dos valores.

1.4.2. Aprendizaje supervisado y no supervisado

Existe una gran cantidad de algoritmos de aprendizaje automático y corremos el riesgo de perdernos en un mar de siglas que inicialmente no aportan más que confusión. Es por esto necesario crear un mapa mental para organizarlos, en función del tipo de aprendizaje que utilizan y del tipo de problema que abordan.

Esta taxonomía de los algoritmos de aprendizaje automático se basa en la naturaleza de las variables del problema. Supongamos que tenemos un modelo como el de la ecuación (1.2) y denotamos por $\mathbf{x}$ e $\mathbf{y}$ respectivamente la variable de entrada y de salida, cada una puede ser de cualquier tipo de los anteriormenete presentados (cuantitativa, ordinal o cualitativa).

El modelador siempre dispone de un conjunto de observaciones $\{\mathbf{x}_i\}_{i=1}^N$ donde $\mathbf{x}_i$ es la variable de entrada de la observación i mientras que no siempre se dispone de las etiquetas $\mathbf{y}_i$ asociada a la observación i.

El **aprendizaje supervisado** supone que partimos de un conjunto de datos etiquetado previamente, es decir, conocemos $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^N$ para el conjunto de datos que disponemos. El aprendizaje **no supervisado** parte de datos no etiquetados previamente, esto es, exclusivamente dispone de la información $\{\mathbf{x}_i\}_{i=1}^N$. Dentro de estas dos situaciones aparecen tres tipos de problemas (ver la Figura 1.4):

- **Clustering**. En este problema se busca patrones dentro de los datos. Esto es, encontrar subconjuntos de observaciones que son similares entre sí. Matemáticamente se formularía como determinar las etiquetas $\mathbf{y}$ de los datos de modo que si dos observaciones i y j son similares le asociemos la misma etiqueta, $y_i = y_j$. Una dificultad de este problema es que no se parte de un conjunto previo de etiquetas del cual aprender. Un ejemplo de este tipo de problemas es la segmentación de mercados. Si se realiza un análisis *cluster* sobre los clientes de una determinada empresa se puede encontrar patrones de clientes y con ellos ofrecer productos especialmente diseñadas para estos segmentos.

- **Regresión**. En este problema se predice el valor de una variable continua $\mathbf{y}$ conocida el valor de la variable $\mathbf{x}$. En este problema la variable $\mathbf{x}$ se suele denominar **regresor** o **variable explicativa** mientras que $\mathbf{y}$ se denomina **variable respuesta** o **variable dependiente**. Un ejemplo de este tipo de problema aparece en los sistemas de recomendación que emplean plataformas de streaming como Spotify, YouTube y Netflix. Estos sistemas estiman funciones de regresión $y = f(\mathbf{x}_1, \mathbf{x}_2)$ que permiten estimar el tiempo de consumo de un usuario (y) para un usuario que ha consumido el producto 1 con características $\mathbf{x}_1$ e inicia el consumo del segundo producto con características $\mathbf{x}_2$.

- **Clasificación**. En este problema se debe predecir el valor de la variable cualitativa $\mathbf{y}$ a partir de la variable $\mathbf{x}$. En clasificación la variable $\mathbf{x}$ se denomina **características** o **atributos** mientras que $\mathbf{y}$ se denomina **etiqueta**. El coche autónomo se basa en tres pilares: IoT, técnicas de aprendizaje automático

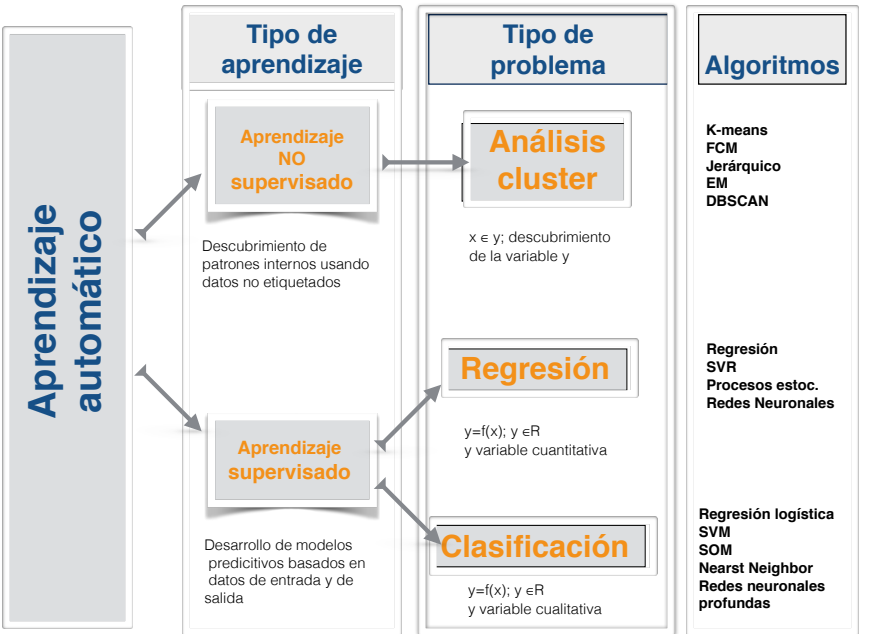

Figura 1.4: Aprendizaje supervisado y no supervisado.

aplicado a Big Data y conexión a internet a tiempo real. En la construcción de *los ojos* del vehículo se requiere resolver el problema de clasificar las imágenes que rodean al vehículo y poder así determinar que tipo de objeto está delante, detrás o a los lados del vehículo.

La figura 1.4 muestra la clasificación de los modelos de aprendizaje automático. Para cada tipo de problemas se han propuesto multitud de algoritmos. No existe un método que sea mejor que otro para todas las circunstancias por lo que es necesario disponer de una *caja de herramientas* con las que poder ensayar diferentes soluciones. A modo de ilustración de lo que queremos decir, el algoritmo *K-means* es rápido y funciona bien para problemas de *clustering* en los que los patrones se pueden separar linealmente (entre los grupos existentes se pueden intercalar hiperplanos que los separe). Este método constituye por tanto una buena alternativa en muchas aplicaciones pero si el problema no es linealmente separable el algoritmo *K-means* no sería capaz de identificar los patrones adecuadamente y por tanto habría que recurrir a técnicas alternativas como algoritmos basados en densidad (DBSCAN).

1.5. Toma de decisiones basada en modelos

La primera cuestión que se le exige a un modelo es que sea capaz de contestar a las preguntas del modelador. Este requisito indica que los modelos se elaboran con un propósito. En muchos casos se emplean para la ayuda a la toma de decisiones. En esta sección analizamos los modelos de aprendizaje automático como herramienta para el apoyo a la toma de decisiones pero nos centramos en los **modelos de optimización**.

La teoría de la optimización es esencial en el aprendizaje automático ya que establece los métodos de entrenamiento de los modelos de aprendizaje. En esta sección abordamos brevemente la formulación de modelos de optimización (no la resolución) porque, además de ser parte de los fundamentos matemáticos del aprendizaje automático, es una técnica esencial en la toma de decisiones basada en modelos. En los modelos de optimización hay que formalizar todos los elementos que intervienen en la toma de decisiones, por este motivo le denominamos modelo de **caja blanca**, en contraposición con modelos tipo **caja negra** (como los de aprendizaje automático) empleados en la toma de decisiones.

1.5.1. Modelos tipo caja blanca: modelos de optimización

Supongamos que tenemos que tomar una decisión. Esta decisión se debe concretar en la determinación de unos valores de las variables esenciales del problema. El primer elemento del problema de decisión es por tanto un vector de **variables de decisión** $\mathbf{x}$. El segundo elemento es un criterio que nos permite evaluar la bondad de cada una de nuestras decisiones $\mathbf{x}$, este se denomina **función objetivo** $f(\mathbf{x})$. Finalmente, como no siempre es posible decidir lo que se quiera si no que existe unas limitaciones en nuestras decisiones tenemos que considerar restricciones a los valores de nuestras variables de decisión. Esto se indica diciendo que nuestras decisiones deben pertenecer a un conjunto de decisiones factibles $\mathbf{x} \in X$. Este conjunto X se le denomina **región factible**. Dependiendo del significado de la función objetivo se puede desear maximizar (producción, ventas, utilidad, capacidad, etc.) o minimizar (costes, consumos, emisiones, etc.) su valor. En el primer caso se dice que se tiene un **problema de maximización** y en el segundo uno de **minimización**. Estos tres elementos del proceso de decisión aparecen en un modelo de optimización: i) variables de decisión $\mathbf{x}$, ii) función objetivo $f(\mathbf{x})$ y iii) región factible X y se formula matemáticamente:

$$\begin{aligned} \text{Minimizar} \quad & f(x) \\ \text{Sujeto a:} \quad & x \in X \end{aligned}$$

Resolver el anterior problema de optimización significa encontrar una solución factible $\mathbf{x}^* \in X$ que sea mejor que el resto de soluciones, esto es, que se cumpla $f(\mathbf{x}^*) \leq f(\mathbf{x})$ para cualquier otra solución $\mathbf{x} \in X$. La solución $\mathbf{x}^*$ se denomina **solución óptima**. Un problema de optimización puede tener múltiples soluciones óptimas.

Si se sabe resolver un problema de minimización se sabe resolver uno de maximización y viceversa. Por ese motivo se plantea como referencia uno de minimización.

Ejemplo. Planificación de la producción. Supóngase que se tienen dos máquinas M_1 y M_2 con las que se fabrica dos tipos de productos A y B. La siguiente tabla muestra el consumo en hora en cada máquina para la producción de cada tipo de artículo:

	M_1	M_2
A	1	5
B	2	1

El beneficio obtenido de la venta de cada producto tipo A es de 75 euros y del producto B de 50 euros. Un modelo de optimización nos permite formular el problema de planificar la producción (un ejemplo de problema de decisión), es decir, de determinar el número de unidades de cada producto que conviene fabricar diariamente para maximizar el beneficio. Determinemos los tres elementos del modelo de optimización:

- Variables de decisión. Denotamos por x_A y x_B las variables de decisión y representan el número de unidades que se va a fabricar diariamente de cada producto. El subíndice hace referencia al tipo de producto.
- Función objetivo. El beneficio es el criterio que nos permite evaluar si una decisión es bueno o mala. Este se expresa $f(x_A, x_B) = 75x_A + 50x_B$.
- Restricciones. Existe una capacidad de producción definida por el hecho que las máquinas no pueden estar funcionando más de 24 horas al día. Si analizamos la máquina M_1 y teniendo en cuenta los consumos en esta máquina (ver la tabla anterior) tendremos que el número de horas empleadas en la producción es $x_A + 2x_B$ y esta cantidad debe ser menor o igual a las 24 horas del día. La máquina M_2 impone su propia restricción de capacidad. Además, las variables x_A y x_B tienen naturaleza entera, tienen que ser productos enteros para ser válidos para su venta.

Finalmente obtenemos el siguiente modelo de optimización cuya solución nos proporciona la producción óptima:

$$\begin{aligned} \text{Maximizar} \quad & f(x_A, x_B) = 75x_A + 50x_B \\ \text{Sujeto a:} \quad & x_A + 2x_B \leq 24 (\text{capacidad horaria diaria máquina } M_1) \\ & 5x_A + x_B \leq 24 (\text{capacidad horaria diaria máquina } M_2) \\ & x_A, x_B \geq 0 \text{ valores enteros} \end{aligned}$$

Clasificación de problemas de optimización

En el ejemplo anterior la región factible se puede expresar, tras simples operaciones aritméticas, mediante la expresión $g(\mathbf{x}) \leq \mathbf{0}$ donde

$$g(\mathbf{x}) = \begin{bmatrix} x_A + 2x_B - 24 \\ 5x_A + x_B - 24 \\ -x_A \\ -x_B \end{bmatrix}, \tag{1.3}$$

el símbolo $\mathbf{0}$ representa un vector columna con cuatro 0 y el operador $\leq$ se aplica componente a componente del vector.

La región factible de un problema de optimización se puede expresar empleando funciones como la anterior $g(\mathbf{x})$, conduciendo la la formulación general de un problema de optimización como:

$$\begin{array}{ll} \text{Minimizar} & f(\mathbf{x}) \\ \text{Sujeto a:} & g(\mathbf{x}) \leq \mathbf{0} \\ & h(\mathbf{x}) = \mathbf{0} \end{array}$$

Los problemas de optimización se clasifican por el tipo de variables: **continuos** o **enteros**. En el caso anterior las variables toman valores enteros, se fabrica unidades de cada producto. Según el tipo de funciones se clasifican en: **lineales** o **no lineales**. En el caso anterior tanto la función objetivo f como las función g que define la región factible son funciones lineales (un coeficiente por una variable mas otro coeficiente por otra variable, así sucesivamente). Este tipo de problemas se conoce como un **problema de programación lineal entera**. Otro tipo de clasificación es si aparecen o no restricciones. En el primer caso se denomina **problema de optimización con restricciones** y en el segundo **problema de optimización irrestringida**. Otras clasificaciones se basan en las propiedades matemáticas que cumplen las funciones f, g, y h hablándose de optimización diferenciable, convexa, etc. Cada tipología de problema de optimización posee métodos específicos de resolución.

Ejercicio. En esta actividad se va a formular un problema de decisión mediante un modelo de programación lineal. Este problema se le conoce con el nombre de **problema de la dieta**. Este problema fue formulado y resuelto por George Stigler en 1947, motivado por encontrar una dieta para el ejercito americano que contuviese los requerimientos nutricionales básicos y que fuese lo más económica posible. Vamos a plantear una versión sencilla del mismo. Supóngase que un nutricionista establece una dieta especial basada en tres productos (arroz, pescado y verduras frescas) que han de combinarse de manera que cumplan una serie de requisitos mínimos en cuanto a proteínas y calorías. Estos mínimos se sitúan en 3 unidades de proteínas y en 4000 calorías. Supóngase además que los productos básicos tienen las siguientes características por kilogramo: el arroz contiene 1 unidad de proteína y 2000 calorías, el pescado tiene 3 unidades de proteínas y 3000 calorías y, por ultimo, las verduras frescas poseen 2 unidades de proteínas y 1000 calorías. Si los precios de los tres productos básicos son respectivamente de 1,5, 7 y 2,5 euros el kilogramo, plantea un modelo de optimización para decidir la composición óptima de la dieta. Debemos encontrar la combinación de productos que cubriendo las necesidades mínimas suponga un menor coste (por kilogramo).

- Variables de decisión. Denotamos por x_A, x_P y x_V las variables de decisión que representan respectivamente la cantidad (en kilogramos) de arroz, pescado y verduras que deben contener la dieta diaria.
- Función objetivo. La minimización del coste del menú es el objetivo perseguido en este problema. $f(x_A, x_P, x_V) = 1{,}5x_A + 70x_P + 2{,}5x_V$.
- Restricciones. Hay tres restricciones. La primera es que se cumplan los requerimientos de proteínas, la segunda de calorías y la tercera debido a la naturalez de las variables, las cantidades tiene que ser no negativas.

Finalmente el modelo de optimización para encontrar la dieta óptima es:

$$\begin{array}{ll} \text{Minimizar} & f(x_A, x_P, x_V) = 1{,}5x_A + 7x_P + 2{,}5x_V \\ \text{Sujeto a:} & x_A + 3x_P + 2x_V \geq 3 \\ & 2000x_A + 3000x_P + 1000x_V \geq 4000 \\ & x_A, x_P, x_V \geq 0 \end{array}$$

1.5.2. Modelos tipo caja negra: aprendizaje automático

Observar que los modelos de optimización especifican la función objetivo y las restricciones. Estas funciones modelan explícitamente los mecanismos causales del problema, esto es, definen las relaciones que existen entre las variables del problema. Los modelos de aprendizaje automático aprenden el mundo en su estado actual y no explican los mecanismos que operan entre las variables. Estos modelos relacionan las entradas con las salidas. Los modelos de aprendizaje automático son adecuados para la toma de decisiones en entornos en los que se preservan los mecanismos causales que operaban cuando fueron entrenados.

La toma de decisiones basada en modelos de aprendizaje automático se basa exclusivamente en el valor $\mathbf{y}$ generado por el modelo para la situación de interés definida por la variable $\mathbf{x}$. Por ejemplo, se puede plantear un modelo de clasificación binaria para predecir la devolución de un préstamo bancario en función de unas características $\mathbf{x}$ del solicitante del préstamo. En este problema la variable respuesta $\mathbf{y}$ toma los valores $0/1$ en función de si devuelve el préstamo o entra en mora. Un nuevo usuario $\mathbf{x}'$ será evaluado mediante el modelo de clasificación y se tomará la decisión de concederle el préstamo en función de la predicción $\mathbf{y}'$ obtenida por el modelo. El modelo no explica como llega a la conclusión solamente proporciona el resultado final. Este modelo ha sido validado con datos del pasado y seguirá siendo válido mientras las condiciones del pasado sigan siendo válidas. Por ejemplo, una situación disruptiva como una pandemia hace que el modelo pierda su validez. Hay que remarcar que no se impone que las condiciones sean estacionarias, no cambiantes, sino que los mecanismos que rigen la evolución del sistema se mantengan. Ahondando en esta cuestión, si asumimos que se trata de una situación de recuperación económica sostenida estas tasas de crecimiento deben mantenerse en el tiempo, esto no quiere decir que la situación económica siga inmutable, sino que la forma en que lo hace sea constante.

1.5.3. Modelos tipo caja gris: método híbrido optimización-aprendizaje automático

Un método alternativo consiste en usar un modelo de optimización en la toma de decisiones pero usando modelos de aprendizaje automático para especificar algunas relaciones entre las variables, ya sea para definir la función objetivo o algunas de las restricciones.

El enfoque híbrido permite un mayor control sobre los mecanismos que operan y por tanto introducir situaciones no existente en la realidad actual. Además se es más conscientes de las hipótesis que se están asumiendo, de las restricciones efectivas del problema y del objetivo que se está persiguiendo.

Ilustremos esta metodología con el siguiente ejemplo.

Ejemplo. Supóngase que estamos ante un problema de fijación de precios de productos. Una relación fundamental es modelar una función de demanda que permita calcular el número de unidades vendidas (d) en función del precio (p) del producto. Esta relación se puede plantear mediante un modelo de regresión $d = F_{\boldsymbol{\theta}}(p)$ e introducida dentro de un modelo de optimización mediante la relación $h(p) = F_\theta(p) - d = 0$. Finalmente el problema a resolver sería:

$$\begin{array}{rl} \text{Maximizar} & f(p,d) = p \cdot d \\ \text{Sujeto a:} & d - F_\theta(p) = 0 \\ & p \geq 0 \end{array}$$

En este modelo de optimización las expresiones de las funciones están definidas por el modelador (la función objetivo) y generadas por técnicas de aprendizaje automático (la restricción de igualdad). Notar también que la función objetivo representa los ingresos generados por la venta del producto.

Ejercicio. Considerar una bodega que compra uva para la elaboración de vino. La decisión que se desea tomar es el precio de compra de la uva para maximizar el beneficio de la bodega. Este problema tiene elementos con incertidumbre. El primero es el precio de venta del vino. El segundo es la cantidad de uva que podrá adquirir en función del precio pagado a los agricultores y la cosecha de uva (ley de oferta/demanda). Asumiendo ciertas constantes para transformar kilogramos de uva a litros de vino, asumimos que existe un precio de compra p_c y un precio de venta p_v de la uva (una vez transformada la uva en vino). Asumimos que existe N zonas vitícolas que entran en competencia con la zona vitícola donde está situada la bodega. Suponemos conocida una base de datos histórica con el precio medio de la compra de la uva para la zona i (p_i) y la cosecha obtenida medida en millones de kilos para dicha zona (q_i). Supongamos también que tenemos una estimación para la campaña actual de los valores $(p_1, q_1, p_2, q_2, \ldots, p_N, q_N)$. Supongamos además que con todos los datos históricos hemos sido capaces de ajustar los siguientes dos modelos de regresión:

$$p_v = P(p, p_1, q_1, p_2, q_2, \ldots, p_N, q_N) \text{ precio de venta de la uva (vino)} \quad (1.4)$$

$$d = D(p, p_1, q_1, p_2, q_2, \ldots, p_N, q_N) \text{ kilogramos de uva adquirido por la bodega} \quad (1.5)$$

La bodega tiene una capacidad de vinificación C que no puede ser sobrepasada. Supongamos que se tiene unos costes variables lineales $k \cdot q$ y un coste fijo K_F. El problema de decisión se pude formular mediante el siguiente modelo de optimización:

$$\begin{array}{rl} \text{Maximizar} & f(p) = p_v \cdot d - p \cdot d - k \cdot d - K_T \\ \text{Sujeto a:} & p_v = P(p, q, p_1, q_1, p_2, q_2, \ldots, p_N, q_N) \\ & d = D(p, q, p_1, q_1, p_2, q_2, \ldots, p_N, q_N) \\ & d \leq C \\ & p \geq 0 \end{array}$$

Listado de problemas propuestos.

¿Cómo se puede resolver un problema de maximización a partir de saber resolver un problema de minimización?

Formula un modelo de optimización irrestringida para el problema de encontrar una recta $y = a+bx$ que mejor ajusta a un conjunto de datos (x_i, y_i) desde $i = 1, \ldots, N$. El criterio de ajuste es que minimice la suma de los errores al cuadrado de los datos.

Supóngase que N personas deben realizar J tareas. Se quiere realizar una asignación de personas a las tareas de modo que se maximice la eficiencia total en su realización. Se pide que se formule un modelo de optimización que apoye a la toma de decisiones. Vamos a suponer los siguientes aspectos del problema:

1. No todas las tareas se consideran igualmente importantes y su relevancia es proporcional al peso ω_j.
2. El tiempo requerido para realizar la tarea j es t_j.
3. Supóngase que se han ajustado J modelos de regresión $y = f_j(\mathbf{x})$ que estiman la eficacia en la realización de la tarea como un índice entre 0 y 1 para la tarea j y para un individuo cuyas características viene definida por el vector de características $\mathbf{x}$.
4. Las características de las N personas vienen definidas por los vectores $\mathbf{x}_i$ para $i = 1, \ldots, N$.
5. Cada persona puede dedicar 8 horas a la realización de las tareas.

Formula un modelo de optimización para decidir la asignación que maximiza la eficacia en la realización del conjunto de tareas.

Bibliografía

[AGK82] D. Art, R. Gnanadesikan, and J. Kettenring. Data-based metrics for cluster analysis. *Utilias Mathematica*, (21A):75–99, 1982.

[Aro50] N. Aroszajn. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68(3):337–404, 1950.

[BEF84] J.C. Bezdek, R. Ehrlich, and W. Full. Fcm: The fuzzy c-means clustering algorithm. *Computers and Geosciences*, 10(2-3):191–203, 1984.

[BL97] Michael J. Berry and Gordon Linoff. *Data Mining Techniques: For Marketing, Sales, and Customer Support*. John Wiley & Sons, Inc., New York, NY, USA, 1997.

[CM96] R. Cheng and G.W. Milligan. Measuring the influence of individual data points in cluster analysis. *Journal of Classification*, (13):315–335, 1996.

[Cox57] D. R. Cox. Note on grouping. *Journal of the American Statistical Association*, 52:543–547, 1957.

[DDW02] E. Dimitriadou, S. Dolni?ar, and A. Weingessel. An examination of indexes for determining the number of clusters in binary data sets. *Psychometrika*, 67(1)(1):137–160, 2002.

[GG92] A. Gersho and R.M. Grey. *Vector quantizacion and signal compression*. Kuwer Academic, Boston, 1992.

[GH10] J. González-Hernández. *Representing Functional Data in Reproducing Kernel Hilbert Spaces with Applications to Clustering, Classification and Time Series Problems*. PhD thesis, Department of Statistics, Universidad Carlos III, Getafe, Madrid, 2010.

[HBV01] Maria Halkidi, Yannis Batistakis, and Michalis Vazirgiannis. On clustering validation techniques. *J. Intell. Inf. Syst.*, 17(2-3):107–145, December 2001.

[KR90] L. Kaufman and P. Rousseeuw. *Finding groups in data: An introduccion to cluster*. Wiley, New York, 1990.

[Mac67] J. MacQueen. Some methods of classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, 1. L. M. Le Cam and J. Neyman (Eds.). Berkeley, CA: University of California Press*, pages 281–297, 1967.

[MC85] G.W. Milligan and M.C. Cooper. An examination of procedures for determining the numbers of clusters in a data set. *Psychometrika*, 50:159–179, 1985.

[MC88] G.W. Milligan and M.C. Cooper. A study of standardization of variables in cluster analysis. *Journal of Classification*, (5):181–204, 1988.

[MC12] F. Murtagh and P. Contreras. Algorithms for hierarchical clustering: An overview. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 2(1):86–97, 2012.

[Mer09] J. Mercer. Functions of positive and negative type and their connection with the theory of integral equations. *Philosophical Transactions of the Royal Society of London. Series A*, 209:415–446, 1909.

[Mil80] G.W. Milligan. An examination of the effect of six types of error perturbation on fifteen clustering algorithms. *Phichometrika*, (45):181–204, 1980.

[Spa80] *Cluster analysis algorithms for data reduction and classification of objects*. Wiley, New York, 1980.

[SS71] A.J. Scott and M.J. Symons. Clustering methods based on likelihood ratio criteria. *Biometrics*, (27):387–398, 1971.

[Ste03] D. Steinley. Local optima in k-means clustering: What you don't know may hurt you. *Psychological Methods*, 8(3):294–304, 2003.

[Ste06] D. Steinley. K-means clustering: A half-century synthesis. *British Journal of Mathematical and Statistical Psychology*, 59(1):1–34, 2006.

[Sym81] M.J. Symons. Clustering criteria and multivariate normal mixtures. *Biometrics*, (37):35–43, 1981.

[TWH01] R. Tibshirani, G. Walther, and T. Hastie. Estimating the number of clusterin a data set via the gap statistic. *Journal of the Royal Statistical Society B*, (63):411–423, 2001.

[VCH10] L. Vendramin, R.J.G.B. Campello, and E.R. Hruschka. Relative clustering validity criteria: A comparative overview. *Statistical Analysis and Data Mining*, 3(4):209–235, 2010.

[Win87] M.P. Windham. Parameter modification for clustering criteria. *Journal of Classification*, (4):191–214, 1987.

[XC07] J. Xia and M. Chen. A nested clustering technique for freeway operating condition classification. *Computer-Aided Civil and Infrastructure Engineering*, 22(6):430–437, 2007.