Análisis de Datos Masivos para el Negocio

Tema 5. Proceso del modelado de datos

Índice

Esquema

5.1. Introducción y objetivos

En el tema «Modelo de proceso de un proyecto orientado a datos» de esta asignatura, se presentaron varios modelos de proceso de proyectos orientados a datos. Estos modelos dividen el proyecto en fases, siendo una de ellas la dedicada al modelado de los datos. En esta fase se utilizan técnicas de la estadística, aprendizaje automático o de la investigación operativa para construir modelos en el dominio del conocimiento que busquen en los datos preparados previamente las respuestas a las preguntas formuladas como objetivos del proyecto. Estas respuestas pueden incluir entre otras, predicciones sobre ventas, almacenaje, beneficios, etc., clasificaciones de clientes, proveedores, activos, etc., o agrupamiento de productos, clientes o empresas. **La construcción de un modelo** es un proceso iterativo que incluye la selección de variables y la ejecución y valoración del modelo. Este tema constituye una presentación general del modelado en base a datos. Los siguientes temas profundizan en las técnicas presentadas en este tema.

Objetivos que se pretenden conseguir:

- Entender qué es un modelo de datos y qué características debe tener.
- Conocer las etapas necesarias para realizar un modelo de datos.
- Diferenciar entre aprendizaje supervisado y no supervisado.

5.2. El modelado de datos

El modelado de datos es una fase de las metodologías de procesado en proyectos orientados a datos. Fases previas han permitido establecer los objetivos del proyecto y la recopilación, preparación y exploración de los datos. El **modelado de los datos** es la fase donde se construye un modelo basado en los datos disponibles que busca el cumplimento de los objetivos establecidos en el proyecto. Entre las funciones que puede cumplir un modelo se tiene entre otras:

- **Predicción.** Capacidad de anticipar el valor de una variable. Por ejemplo, se puede intentar predecir la cantidad de ventas, la demanda de un producto, el stock futuro, etc.

- **Clasificación.** Capacidad para asignar una clase a un objeto de interés. Por ejemplo, dado un nuevo cliente se puede intentar clasificar como solvente o posiblemente no solvente. Dado un currículum se puede clasificar como apto para cierto tipo de trabajo o no. Dada una película, se puede clasificar como afín o no al gusto de un cliente. Dado un producto se puede clasificar como de posible interés o no para un cliente.

- **Clusterización.** Capacidad de dividir un conjunto de objetos de interés en subgrupos relacionados entre sí. Por ejemplo, se puede dividir un conjunto de clientes en subgrupos basados en comportamientos de compras. Se puede usar como una técnica automática de segmentación del mercado.

- **Reducción de la dimensión.** En este caso el modelo busca simplificar la información. Hoy en día, puede haber un exceso de datos. Los modelos de datos también pueden servir para simplificar la cantidad de datos, pero intentando mantener la cantidad de información disponible.

Ideas clave

Un modelo se puede ver como un algoritmo que toma un conjunto de datos, las denominadas variables dependientes (también se las puede llamar características, predictores ovariables de entrada o) los procesa y proporciona un nuevo dato o conjunto de datos de salida denominados variables dependientes, (también se le puede llamar predicción, variables de salida o de respuesta). El tipo de procesamiento de datos depende de la funcionalidad asociada al modelo. Existen muchos tipos de modelos, por ejemplo, entre los modelos de predicción que podemos utilizar se encuentran la regresión lineal, los árboles de decisión o las redes neuronales.

Varias son las preguntas que podemos hacernos: ¿qué datos son los que tenemos que proporcionar como entrada?, ¿qué tipo de algoritmo hemos de aplicar?, ¿cómo debe aprender ese algoritmo?, ¿cómo decidimos que el modelo tiene la calidad suficiente?

Simplificado de modelado de datos

Supongamos que queremos realizar un modelo que prediga los ingresos de una persona. Se dispone de la siguiente información por cada persona: edad, DNI, nivel de estudios y color favorito. Esas cuatro variables es lo que denominamos variables independientes. La variable ingresos es la variable dependiente. Los datos de cada persona en concreto son las observaciones.

La primera pregunta que nos hacíamos es ¿qué datos son los que tenemos que proporcionar como entrada? Parece sensato que, para predecir los ingresos de una persona, las variables más relacionadas son la edad y nivel de estudios, mientras que el DNI y color favorito no parecen muy relacionados. Ese proceso es lo que se denomina elección de variables dependientes, y lo que nos interesa es un proceso que realice dicha selección de forma automática, no manual.

La segunda pregunta ¿qué tipo de algoritmo hemos de aplicar? Consiste en decidir qué tipo de estructura tiene el modelo de predicción (función lineal, no lineal, compleja, etc.). Se ha de tomar una decisión en ese sentido.

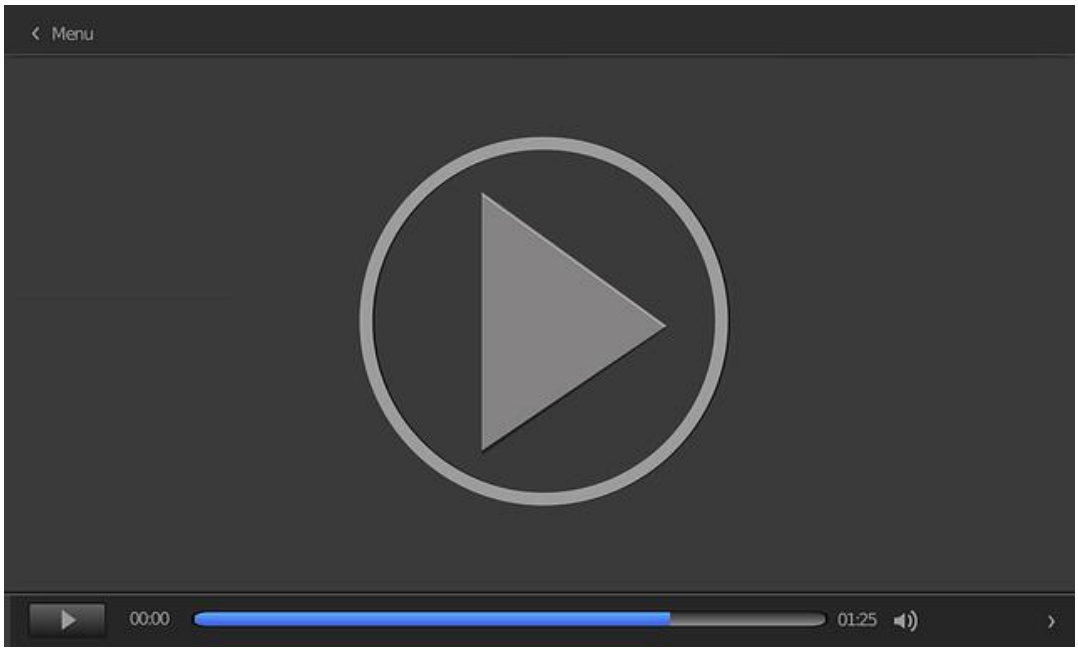
La tercera pregunta ¿cómo debe aprender ese algoritmo? Puede tener una respuesta sencilla, podemos darle ejemplos para que aprenda. Se necesita un conjunto de observaciones que incluyan edad, nivel de estudios e ingresos de un conjunto de personas. Estos son los denominados datos de entrenamiento que se utilizarán para ajustar o sintonizar el modelo de predicción elegido.

Finalmente, ante la pregunta ¿cómo decidimos que el modelo tiene la calidad suficiente? Una respuesta puede ser tomar un segundo conjunto de observaciones. Este segundo conjunto de datos debe ser diferente al conjunto de entrenamiento. La calidad del modelo se puede medir comparándolas predicciones de ingreso por persona que obtiene con los ingresos reales que se indican en este segundo conjunto de datos. Si se «equivoca poco», el modelo lo podemos dar por bueno, si se «equivoca mucho» tenemos que plantearnos si volver a iterar desde la primera pregunta.

Las etapas de la creación de un modelo ilustradas en el ejemplo anterior se formalizan en la siguiente sección.

5.3. El proceso del modelado de datos

Vídeo *Proceso de modelado de datos.*



Accede al vídeo:

https://unir.cloud.panopto.eu/Panopto/Pages/Embed.aspx?id=36a0f4cc-5aa5-4e96-9c2c-b15d00aa2c52

El modelado de datos se divide a su vez en varias etapas:

- Selección de modelo y variables de dependientes.
- Entrenamiento del modelo.
- Validación del modelo.
- Aplicación del modelo a nuevos datos.

El proceso de modelado no es una secuencia pura, ya que normalmente requerirá varias iteraciones o repeticiones de las tres primeras etapas. El objetivo es entrenar un modelo que cumpla los objetivos fijados en los datos de validación. Se puede iterar el proceso hasta alcanzar ese nivel de precisión deseado. Los modelos obtenidos no tienen por qué ser únicos. Se pueden obtener varios modelos y concatenarlos o combinarlos. Concatenar varios modelos significa usar la salida de un modelo para que sea la entrada de otro modelo. Combinar modelos significa entrenarlos de forma individual para después usar como modelo final la combinación de sus salidas. Esta última técnica se la denomina *ensemble.*

Selección de la familia de modelos

Esta es una de las etapas más importantes del modelado de datos ya que las predicciones se realizan usando el modelo que procesa las variables de entrada. Actualmente se pueden encontrar muchas familias de modelos a usar. Los siguientes temas profundizan en los tipos de modelos disponibles. La elección de una familia de modelos no es definitiva, ya que, si cierto modelo no obtiene los resultados deseados, se puede cambiar por otro o incluso de familia. Lo normal en un proyecto es probar con diferentes opciones. Por ejemplo, si estamos en un problema de predicción, se puede optar por modelos basados en regresión lineal, modelos de árboles de decisión, modelos basados en máquinas de soporte vectorial u otras opciones.

Las familias de modelos pueden estar parametrizadas, esto es, tienen un conjunto de parámetros usados para seleccionar un modelo concreto en la familia. **El número de parámetros usados define el tamaño de la familia correspondiente.** Cuantos más parámetros tiene una familia, más modelos contiene. Es importante resaltar que para cada proyecto concreto no se conoce a priori la mejor familia ni modelo, dentro de la familia, a aplicar.

Además, cada familia de modelos tiene asociada una *complejidad o flexibilidad.* Por ejemplo, normalmente, cuanto mayor es el número de parámetros de una familia, mayor es la complejidad o flexibilidad. La flexibilidad proporciona capacidad de adaptación. Normalmente, una mayor flexibilidad implica una mayor capacidad de adaptación y por tanto mayores posibilidades de que el modelo pueda funcionar correctamente. Por otro lado, una mayor simplicidad en el modelo representa un mayor grado de *interpretabilidad.* A veces no solamente nos interesa un modelo por la predicción que proporciona, también nos interesa interpretar por qué obtiene esa predicción, que nos explique por qué genera esos resultados. Históricamente, los modelos estadísticos han buscado la interpretabilidad, mientras que los basados en aprendizaje automático la precisión. Por otro lado, la interpretabilidad no solo depende del número de parámetros, sino también de las relaciones matemática que los enlazan. Por ejemplo, un modelo con una relación lineal de sus parámetros es más interpretable que uno basado en complejas relaciones no lineales. Un modelo basado en regresión lineal suele ser más interpretable que un modelo basado en redes neuronales. Por regla general, a un mismo nivel de precisión, se suele preferir el modelo menos complejo.

En la figura 1 se puede observar la relación entre flexibilidad e interpretabilidad de varias familias de modelos entre las que se encuentran mínimos cuadrados lineales, máquinas de soporte vectorial, etc.

Figura 1. Flexibilidad vs interpretabilidad de varias familias de modelos. Fuente: James, G., Witten, D., Hastie, T. y Tibshirani, R. (2013).

Selección de variables independientes

Una vez seleccionado el modelo, o previamente a esa elección, hay seleccionar las variables independientes a utilizar. Básicamente hay que identificar qué variables independientes influyen en las variables dependientes. Claramente, si la influencia de una variable independiente es nula, no se tiene que usar en el modelo.

Existen varias técnicas. A continuación, se exponen algunos de ellos:

- **Aplicar técnicas de reducción de la dimensión.** En este caso se aplican técnicas que permiten simplificar los datos a usar, pero intentando mantener la información que contienen. Un ejemplo de este tipo de técnicas es el análisis de componentes principales. Es un tipo de técnica de aprendizaje no supervisado que se expondrá en mayor profundidad en el tema «Técnicas de aprendizaje no supervisado» de esta asignatura. Las técnicas de reducción de la dimensión se pueden aplicar de forma independiente al modelo seleccionado.

- **Elección de un subconjunto.** Esta opción consiste en diseñar una estrategia para

elegir un subconjunto de variables independientes del total.

- La primera estrategia es probar todos los subconjuntos posibles y quedarnos con el subconjunto que mejores resultados proporcioneusando el modelo elegido. Esta técnica es sistemática, pero no es realista cunado se tiene un número grande de variables independientes, pues el número de subconjuntos posibles crece de forma desmesurada.

Estrategia sistemática

Con tres variables (a,b,c) se tienen siete subconjuntos posibles: {(a), (b), (c), (a,b), (a,c), (b,c), (a,b,c)}. Si se tiene cuatro variables (a,b,c,d) se pueden formar hasta quince subconjuntos posibles: {(a), (b), (c), (d), (a,b), (a,c), (a,d), (b,c), (b,d), (c,d), (a,b,c), (a,b,d), (a,c,d), (b,c,d), (a,b,c,d)}. El número de subconjuntos es dos elevado al número de variables por lo que crece mucho más rápidamente que el número de variables.

- La segunda estrategia consiste en aplicar una metodología incremental o decremental en los subconjuntos construidos. En este caso se consideran únicamente una parte pequeña del total de subconjuntos posibles. Por tanto, no es una técnica sistemática y puede llegar a una solución buena, pero quizás no la mejor.

Estrategia decremental

Considerando que tenemos cuatro variables (a,b,c,d), una posible estrategia decremental consiste en probar los subconjuntos {(a,b,c), (a,b,d), (a,c,d), (b,c,d)} y quedarnos con el mejor candidato. Supongamos que es (a,b,c). El paso siguiente es probar los subconjuntos únicamente de ese candidato. En este caso serían {(a,b), (a,c), (b,c)}. Si alguno de los nuevos candidatos mejora a (a,b,c) lo tomamos como nuevo candidato, si

no, el ganador es (a,b,c). Supongamos que el subconjunto (a,b) mejora (a,b,c). En este caso probaríamos con otro decremento, esto es los candidatos {(a),(b)}, y finalmente tomaríamos como subconjunto ganador al mejor entre (a), (b) y (a,b,c).

- La tercera opción es los que se denomina **regularización del modelo**. Esta técnica consiste en modificar el entrenamiento del modelo para detectar qué variables son las que influyen menos en las predicciones obtenidas. Aquellas variables independientes con poca o nula influencia en las predicciones son las que se pueden descartar. El principio en el que se basa la regularización es en mantener la simpleza del modelo y su capacidad de predicción.

Entrenamiento del modelo

Una vez seleccionada la familia de modelos y las variables independientes a usar, el siguiente paso es lo que se denomina **entrenar el modelo.** Básicamente consiste en seleccionar un modelo de toda la familia disponible. Si la familia está parametrizada, consiste en elegir un valor concreto para los parámetros.

Una de las formas más sencillas de entrenar un modelo es utilizar ejemplos. Para ello es necesario un conjunto de datos de entrenamiento que incluya observaciones de las variables independientes y dependientes. El **entrenamiento** básicamente es un proceso de búsqueda de los parámetros que hacen que el modelo correspondiente relacione correctamente las variables independientes y dependientes. Entrenar un modelo es extraer de la familia el modelo que «mejor replica» los ejemplos incluidos en los datos de entrenamiento. Sin embargo, este «mejor modelo» puede no sea la mejor elección. Una estrategia mejor es combinar datos de entrenamiento y validación.

Validación del modelo

Un buen modelo es aquel que tiene un gran poder predictivo y generaliza bien ese poder predictivo a nuevas observaciones, observaciones no usadas en entrenamiento. Que un modelo tenga gran poder predictivo en los datos de entrenamiento es bueno, pero no significa que pueda generalizarlos a nuevas observaciones.

Cuando se entrena un modelo hay dos conceptos que es importante tener en cuenta, el balance entre el sesgo y la varianza (bias/variance) y el sobre-entrenamiento *(overfitting).*

- Una familia de modelos pequeña (pocos parámetros) o poco flexible puede generar modelos con cierto sesgo, pero poca variabilidad. La consecuencia es que los modelos generados, ante nuevas observaciones, pueden tener un error de predicción sistemático, pero casi siempre del mismo tamaño. Por otro lado, una familia de modelos amplia (con un gran número de parámetros) y por tanto flexible, puede generar un modelo, que, con las nuevas observaciones, obtenga un error muy pequeño o también muy grande. La variabilidad es mayor. Por supuesto, el mejor caso es elegir una familia que genere modelos con sesgo nulo y varianza pequeña, pero como ya se comentó antes, para un problema concreto la familia de modelos adecuada no se conoce a priori.

> Sesgo versus varianza
>
> Vamos a incluir un pequeño ejemplo para entender el sesgo y la varianza. Supongamos que nuestro problema es adivinar un número natural entre 1 y 6. Para ellos usamos dos modelos de predicción diferentes. El primero es un dado con seis caras. El dado puede tener equivocaciones pequeñas (acertar incluso) o bien equivocaciones muy grandes (predecir un seis mientras que el número a acertar era un uno, un error de cinco unidades). Por otro lado, el dado siempre tiene opciones de acertar, ósea, predecir

bien. Este dado de seis caras es lo que hemos llamado **familia amplia**. Por otro lado, el segundo modelo predictor es un dado de dos caras que solo tiene como opciones los números tres y cuatro. En este caso el dado no siempre tiene opciones de acertar, por ejemplo, cuando el número oculto es uno, dos, cinco o seis. Esto es lo que se denomina **sesgo de error,** pues no se puede eliminar. A cambio el error cometido por este segundo dado nunca excede las tres unidades. Este segundo dado representaría una familia acotada o simple de modelos.

- Por otro lado, una familia amplia de modelos (con muchos parámetros) puede llegar a tener un gran poder predictivo en los datos de entrenamiento, pero tener un rendimiento mediocre con observaciones nuevas no incluidas en el conjunto de entrenamiento. Este efecto es el que se llama **sobre-entrenamiento.**

Con el fin de evitar los dos efectos antes mencionados y generalizar mejor el poder predictivo de los modelos, se aplica la etapa de validación del modelo. Existen varias opciones para este fin:

- La primera opción es tomar todas las observaciones disponibles y dividirlas en dos conjuntos, el conjunto de entrenamiento y el conjunto de validación. En este caso, el modelo seleccionado es el modelo ajustado con datos de entrenamiento, pero tiene el mayor poder predictivo en datos de validación.

- La segunda opción se denomina **validación cruzada con k-iteraciones** *(k-fold crossvalidation).* En este caso se divide el conjunto total de observaciones en k subconjuntos. En cada iteración se toma uno de los subconjuntos como observaciones de validación y el resto de los subconjuntos como observaciones de entrenamiento. Se selecciona el modelo que mejor se comporta en media en las *k*-iteraciones.

- En el punto extremo, si se tienen *N*observaciones, se puede utilizar la técnica *leave-*

one-out que básicamente consiste en realizar *N* iteraciones, siendo en cada iteración el conjunto de validación una sola observación, constituyendo el resto de las observaciones el conjunto de entrenamiento.

Aplicación de los datos de validación

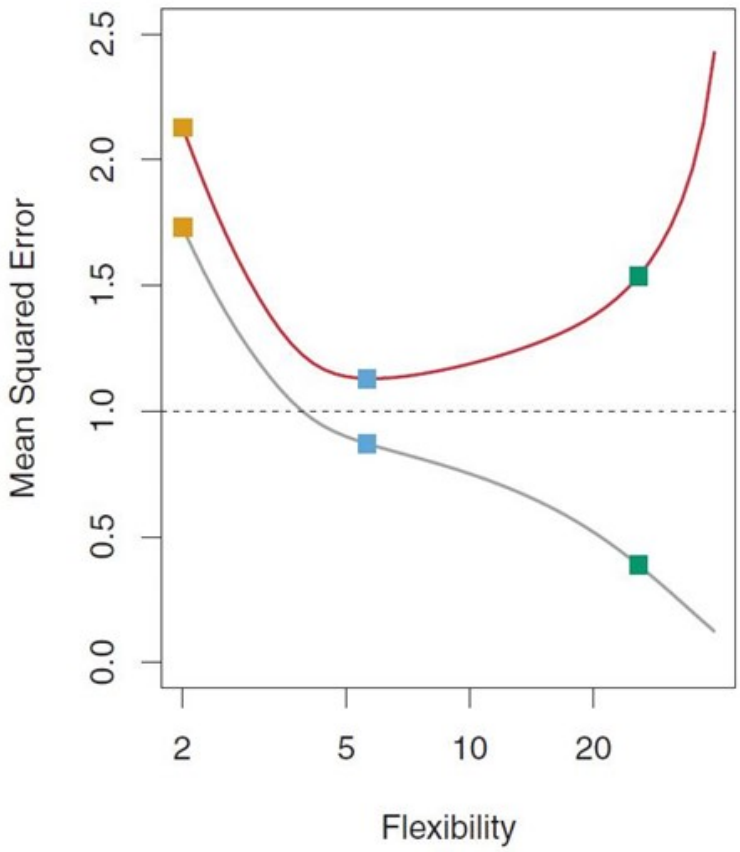


Figura 2. Ejemplo de sobre-entrenamiento. Fuente:James, G., Witten, D., Hastie, T. yTibshirani, R. (2013).

La figura 2 muestra la capacidad predictiva de tres modelos (naranja, azul y verde). El eje horizontal indica el nivel de flexibilidad de cada modelo, siendo el verde el más flexible. Por otro lado, el eje vertical muestra la capacidad predictora (cuanto más cercano a cero el error de predicción, mayor capacidad predictora). La línea gris indica capacidad predictora en

datos de entrenamiento, y la roja en datos de validación. El modelo naranja tiene capacidad predictora mediocre en entrenamiento y validación. El azul tiene capacidad de predicción buena en entrenamiento y validación. El modelo verde tiene muy buena capacidad de predicción en entrenamiento, pero mediocre en validación. Por tanto, aunque el modelo verde es mejor en entrenamiento, se selecciona el modelo azul que es capaz de generalizar la capacidad de predicción.

5.4. Aprendizaje automático

El aprendizaje automático (*machine learning*) proporciona la habilidad de aprender de los computadores sin realizar una programación explícita de las acciones a realizar. Por ejemplo, cuando entrenamos un modelo, no indicamos de antemano el modelo a usar, es el computador el que selecciona el modelo usando los ejemplos incluido en el conjunto de datos de entrenamiento.

Existe básicamente dos tipos de aprendizaje automático, el aprendizaje supervisado y el no supervisado.

Aprendizaje supervisado

El aprendizaje supervisado es aquel que **se basa en la utilización de observaciones de ejemplo u observaciones previamente etiquetadas.** En nuestro ejemplo del predictor de los ingresos que tienen las personas, para entrenar el modelo se utiliza un conjunto de observaciones de entrenamiento. En estas observaciones la variable dependiente (los ingresos) son conocidas, por lo que sirven de ejemplo para entrenar el modelo.

Por tanto, dado un conjunto de datos u observaciones etiquetadas (ejemplos de los resultados que queremos obtener) los métodos de aprendizaje supervisado entrenan un modelo que sea capaz de reproducir esas observaciones y generalizar ese comportamiento a nuevas observaciones ya fuera del conjunto de entrenamiento.

La disponibilidad de datos etiquetados puede ser un problema para el proyecto, ya que a veces o no es sencillo conseguirlos, o bien son datos muy escasos.

El aprendizaje supervisado es muy utilizado en problemas de predicción y clasificación, ya que la precisión suele ser uno de los factores importantes para estos casos.

Aprendizaje no supervisado

Las técnicas de aprendizaje no supervisado se aplican a datos no etiquetados. Dado un conjunto de datos, estas técnicas **intentan estudiar la estructura interna, latente, que tienen esos datos.** Esta estructura no suele ser observable directamente, pero se puede extraer de forma automática aplicando estos algoritmos.

El aprendizaje no supervisado se puede aplicar al problema del agrupamiento o clusterización. Dado un conjunto de observaciones, el objetivo es crear subconjuntos de esas observaciones. Los miembros de esos subconjuntos deben estar fuertemente relacionados, pero los subconjuntos entre sí deben estar débilmente relacionados. En la figura 3 se ilustra el proceso de clusterización. En la primera subfigura se presentan las observaciones no etiquetadas las cuales se quieren clusterizar. En este caso, cada observación es un punto de los representados. Cada observación tiene dos valores asociados (por lo que estamos en dimensión dos). El objetivo es dividir el conjunto de datos en varios subconjuntos cumpliendo las relaciones antes expuestas. En este caso, se le indica al algoritmo de clusterización que se desean dos subconjuntos. El resultado que el algoritmo devuelve se puede apreciar en la segunda figura. El algoritmo claramente ha identificado los dos subconjuntos que componen las observaciones originales. El algoritmo también proporciona un representante de casa clúster o subconjunto, el denominado centroide.

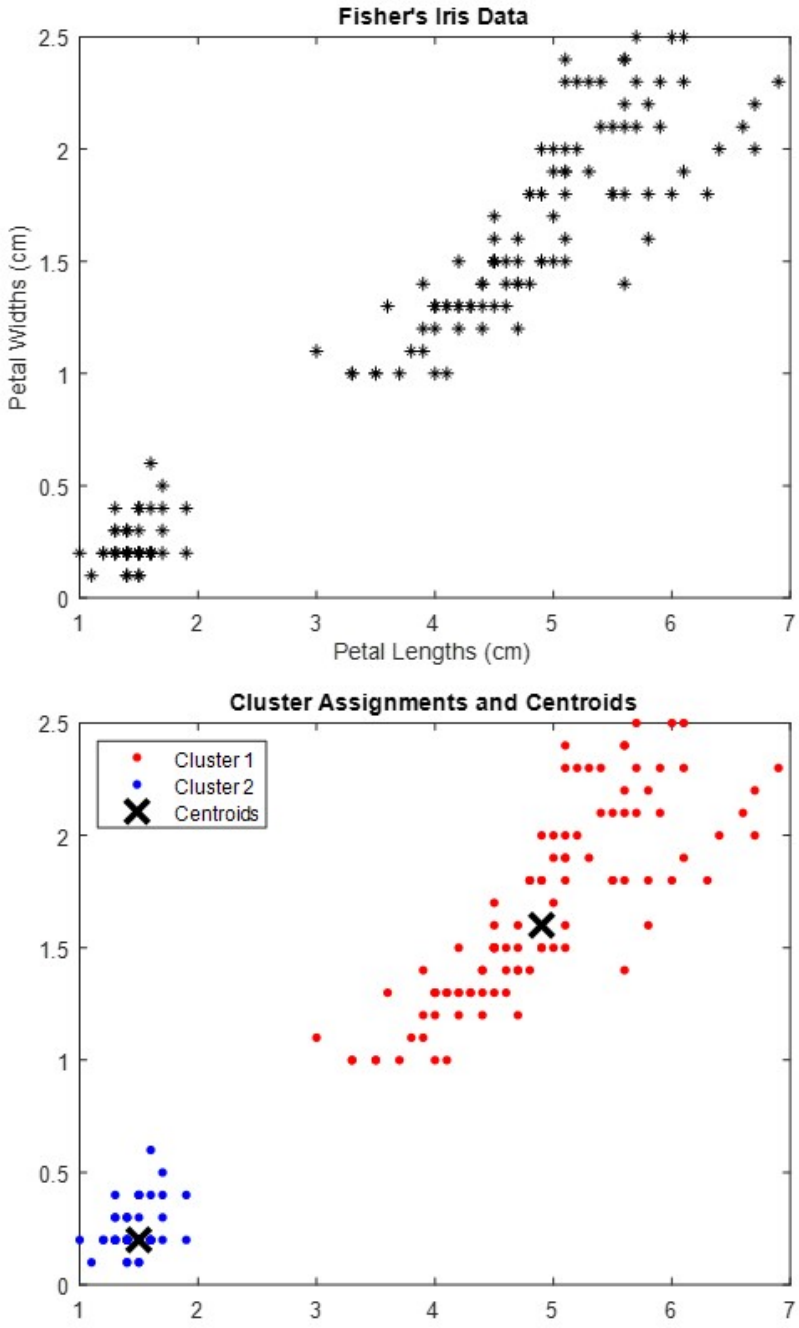


Figura 3. Ejemplo de clusterización. Fuente: https://es.mathworks.com/matlabcentral/fileexchange/48768-simple-example-and-generic-function-for-kmeans-clustering

Otro de los ámbitos donde es posible aplicar las técnicas de aprendizaje no supervisado es el problema de la reducción de la dimensión. Dado un conjunto de

datos compuesto de muchas variables, el objetivo es estudiar si es posible disminuir el conjunto de variables o características pero sin perder información. Los métodos de aprendizaje no supervisado se tratarán con mayor profundidad en el tema «Técnicas de aprendizaje no supervisado» de esta asignatura.

5.5. Referencias bibliográficas

James, G., Witten, D., Hastie, T. y Tibshirani, R. (2013). *An introduction to statistical learning – with applications in R.* (Vol. 103). New York: Springer.

Bias y Variance

Mallick, S. (2017). Bias-Variance Tradeoff in Machine Learning. Learn OpenCV.

Este documento realiza una explicación de los conceptos bias, variance y sobre-entrenamiento que es de gran interés para el alumno.

Accede al artículo a través del aula virtual o desde la siguiente dirección web: https://www.learnopencv.com/bias-variance-tradeoff-in-machine-learning/

Test

1. Si queremos estimar la demanda de un producto el próximo mes, necesitamos un modelo de:

A. Predicción.

B. Clasificación.

C. Clusterización.

D. Reducción de la dimensión.

2. Si en una empresa que proporciona servicios de *streaming* queremos dividir los clientes según sus gustos, podemos usar modelos de:

A. Predicción.

B. Clasificación.

C. Clusterización.

D. Reducción de la dimensión.

3. El proceso de modelado de datos es:

A. Una secuencia de cuatro etapas.

B. Un proceso iterativo de cuatro etapas.

C. Una técnica manual de predicción.

D. Un proceso de clasificación y adecuación de datos.

4. Un buen modelo es aquel:

A. Que obtiene un buen resultado en datos de entrenamiento.

B. Tiene gran capacidad explicativa.

C. Mantiene buena precisión de predicción en nuevas observaciones.

D. Basado en una familia altamente parametrizada.

5. Para evitar el sobreentrenamiento de un modelo:

A. Se deben usar pocos datos de entrenamiento.

B. Se deben usar técnicas de validación de resultados.

C. Se debe alcanzar un gran poder predictivo en los datos de entrenamiento.

D. Se debe evitar la generalización de la capacidad de predicción del modelo.

6. ¿Cuál de las siguientes técnicas busca reducir el número de variables sin perder información esencial?

A. Clasificación supervisada.

B. Regularización.

C. Análisis de componentes principales (PCA).

D. Sobreentrenamiento.

7. ¿Qué representa el bias (sesgo) en el entrenamiento de un modelo?

A. El nivel de error en los datos de validación.

B. La capacidad de adaptación a los datos de entrenamiento.

C. La tendencia sistemática a cometer errores por simplificación del modelo.

D. La dispersión del error en distintas observaciones.

8. ¿Qué ocurre cuando un modelo tiene alta varianza?

A. Aprende poco de los datos de entrenamiento.

B. Generaliza bien con nuevos datos.

C. Produce siempre la misma predicción.

D. Se ajusta demasiado a los datos de entrenamiento y falla en generalizar.

9. ¿Cuál de las siguientes afirmaciones es propia del aprendizaje no supervisado?

A. Se entrena el modelo con etiquetas conocidas.

B. Se utilizan algoritmos como regresión lineal o árboles de decisión.

C. Se busca inferir patrones sin tener una variable objetivo.

D. Se valida con datos de entrenamiento.

10. ¿Qué es una técnica de ensemble en modelado de datos?

A. Una forma de regularizar el modelo base.

B. Una técnica para eliminar variables redundantes.

C. Un método que combina varios modelos para mejorar la predicción.

D. Una técnica de análisis no supervisado.