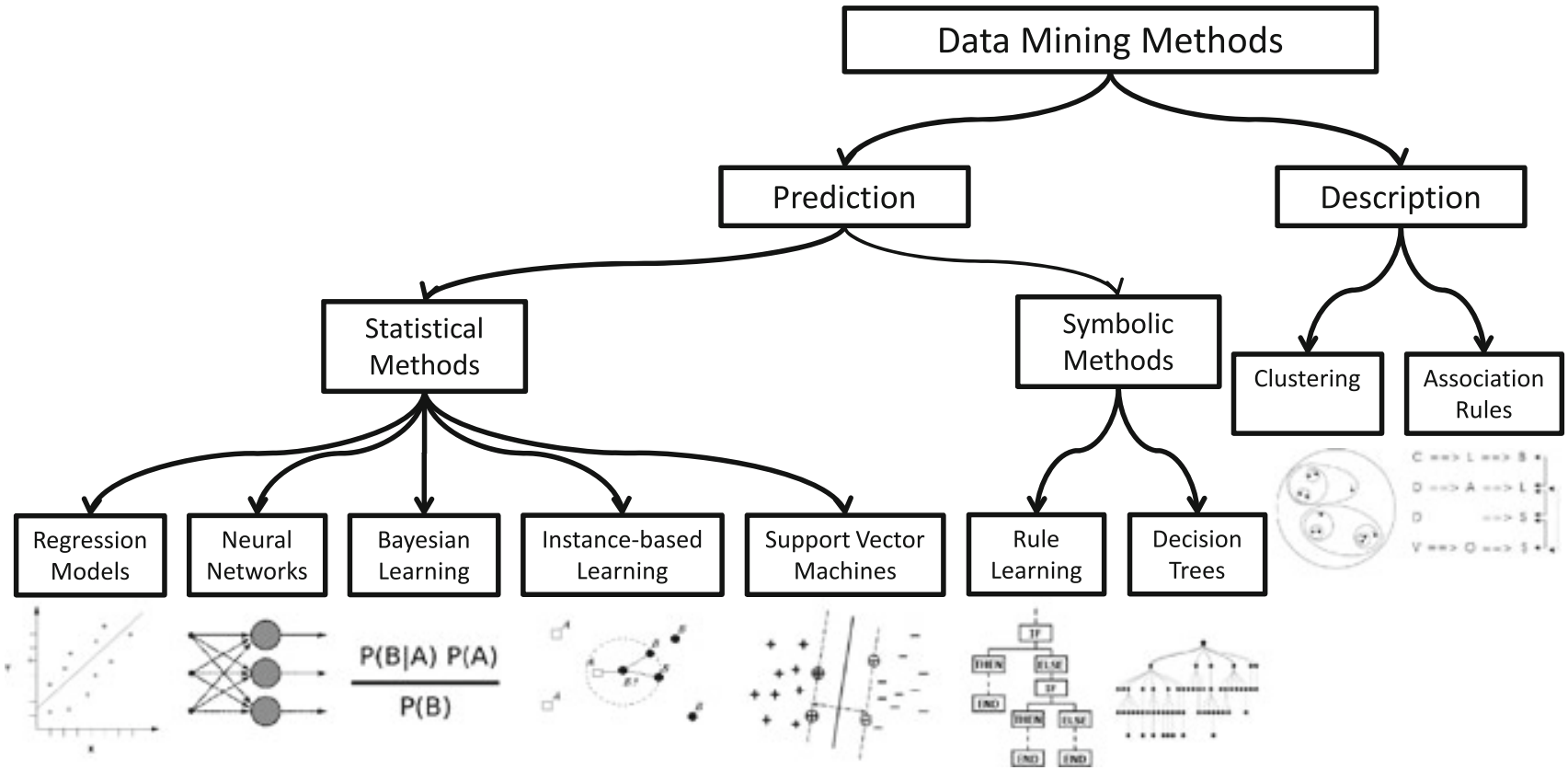
Bloque 1. Diseño y explotación de un almacén de datos

- 1 Limpieza e integración de datos (python/sklearn + pandas)
- 2 Técnicas de transformación y reducción: selección de características e instancias, imputación de datos (python/sklearn + pandas)
- 3 Visualización de datos de negocio mediante técnicas de escalado multidimensional, Sammon, PCA, LFA y LLE (python/sklearn + pandas)

Métodos de minería de datos



Métodos predictivos de minería de datos

- **Modelos de regresión** (lineal, cuadrática, logística): con atributos numéricos. Dificultades en tratamiento de datos perdidos o outliers, datos redundantes o interdependientes

- **Redes neuronales** (MLP, RBFN, LVQ): Caso particular de la regresión

- **Aprendizaje bayesiano**: Clasificación estadística, es una estimación no paramétrica de una función de densidad

- **Aprendizaje basado en instancias**: Se almacenan los ejemplos y se define una distancia entre ellos para determinar qué miembros de la base de datos están más cercanos a un nuevo ejemplo (lazy learning). Depende de la distancia, los ejemplos elegidos, la implementación de la búsqueda del ejemplo más cercano. Poco tolerantes al ruido, alto uso de almacenamiento. Se mejora con técnicas de reducción de datos.

Métodos predictivos de minería de datos

- **Máquinas de vectores de soporte**: similares a las redes neuronales, requieren un menor preprocesamiento, más robustas ante outliers.

- Métodos simbólicos:

  - **Aprendizaje de reglas**: Algoritmos de tipo divide y vencerás: AQ, CN2, RIPPER, PART, FURIA, etc.

  - **Arboles de decision**: También de tipo divide y vencerás, estructurados jerárquicamente. Cada separación usa variable, y el resultado final puede expresarse mediante reglas lingüísticas: CART, C4.5, PUBLIC, etc.

Métodos descriptivos de minería de datos

- **Clustering:** Los ejemplos pueden dividirse en grupos naturales o clusters. Dependen de una distancia. Pueden ser jerárquicos (aglomerativos, divisivos, iterativos) o basados en particiones (k-means, COBWEB, SOM)

- **Reglas de asociación:** Relaciones de asociación entre los datos, como las existentes en transacciones comerciales. También pueden usarse para detectar patrones secuenciales, si los datos son discretos (Apriori, etc.)

Aprendizaje supervisado

- Los métodos predictivos suelen ser nombrados “aprendizaje supervisado”
- Descubren la relación entre las variables de entrada (o atributos, o características) y las variables de salida (o clase)
- Se usa un conjunto de entrenamiento y el objetivo es predecir la clase de salida de ejemplos aún no vistos.
- El conjunto de entrenamiento consta de tuplas que comprende un vector de atributos nominales o categóricos (sin orden) o numéricos (con orden).
- Los problemas básicos son **clasificación** y **regresión**, y dentro de este último las **series temporales** tienen técnicas propias que hacen uso de la fuerte dependencia entre las variables de entrada que se obtienen al transformar una serie en un conjunto de tuplas.

Aprendizaje no supervisado

- Se buscan regularidades, irregularidades, relaciones, similaridades o asociaciones en las entradas.

- Además del clustering y las reglas de asociación,

  - **Minería de patrones** (patrones frecuentes, patrones infrecuentes o negativos, minería distribuida o incremental, excepciones, etc.)

  - **Detección de outliers** (detección de anomalías)

Otros paradigmas

- Otros paradigmas:
  - **Aprendizaje no balanceado** (distribución anómala de los datos)
  - **Aprendizaje multi-instancia** (cada ejemplo es una bolsa de instancias)
  - **Clasificación multi-etiqueta** (cada instancia tiene múltiples etiquetas)
  - **Aprendizaje semi-supervisado** (algunas instancias sin etiquetar)
  - **Descubrimiento de subgrupos** (o contrast set mining, emergent pattern mining)
  - **Transfer learning** (dataset shift, la distribución de entrenamiento es diferente de la de test)
  - **Data stream learning** (datos se obtienen secuencialmente en el tiempo)

sklearn

- Abrir tutorial en tutorialSklearn.ipynb