Introducción a la Inteligencia Artificial

Clase 1.3. Iris Dataset. Terminología básica de aprendizaje automático

Departamento de Computación, FCEFQyN, Universidad Nacional de Río Cuarto, 2024

Iris Dataset

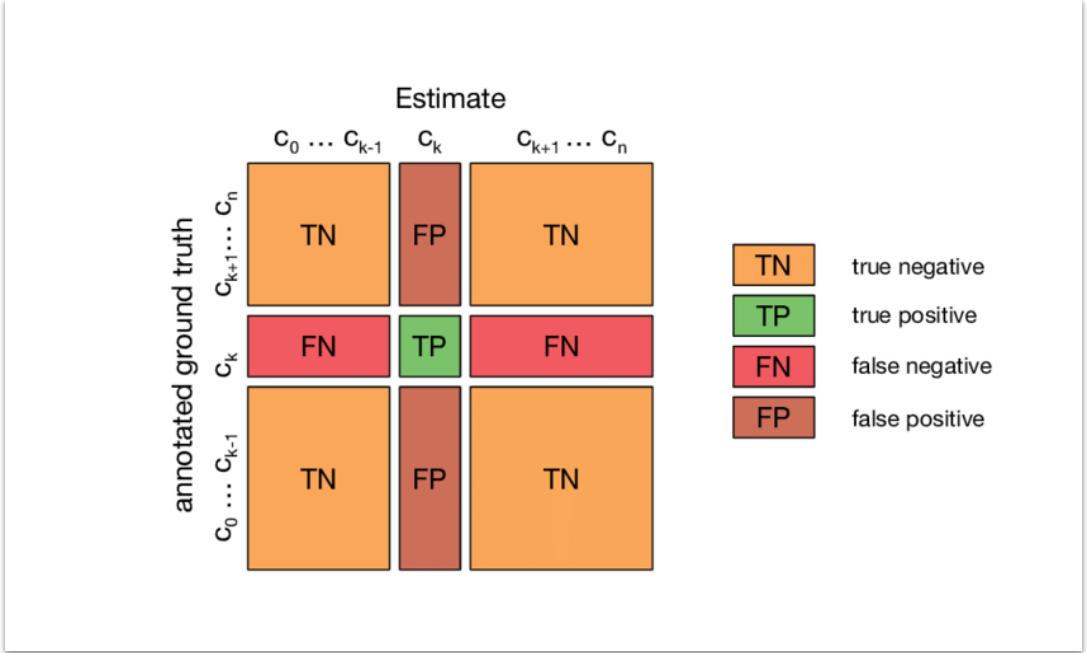
- Base de datos de la flor de iris
- Consta de un número de registros para los cuales de midió largo y ancho de pétalos y sépalos, y se caracterizó (manualmente) la especie
  - Importante: son datos etiquetados", i.e., cada registro de información indica a qué especie realmente pertenece
- Tres especies posibles:
  - Setosa, Virginica, Versicolor
- Es un dataset clásico (simple) en machine learning

Clasificación

- La clasificación es una de las clases de problemas más importantes en machine learning:
  - Dadas una serie de características K de individuos o items, se quiere predecir el valor de una característica adicional k’
    - k’ debe ser *discreta*

Métricas de evaluación

- Consideremos un problema de clasificación P, que intenta predecir la clase ci para items arbitrarios de una población, y sea Prog un clasificador.
- Los clasificadores suelen no ser perfectos, es decir, predicen clases incorrectas para algunos individuos de la población
  - Prog(x) no coincide con la clase correcta que corresponde a x
- La precisión (accuracy) de un clasificador se suele medir, para cada clase ck, en relación a la tasa de
  - True positives: casos en los cuales la clase ck predicha coincide con la real
  - False positives: casos en los cuales la clase predicha ck no coincide con la clase del individuo correspondiente
  - True negatives: casos en los cuales la clase predicha no es ck, y efectivamente los individuos correspondientes no son de clase ck
  - False negatives: casos en los cuales la clase predicha por el clasificador no es ck, pero los individuos correspondientes corresponden a la clase ck



Entrenamiento vs Evaluación

- Cuando los clasificadores son basados en aprendizaje automático, éstos se construyen a partir de datos.
- Los datos provienen de un dataset (anotado)
- Para evaluar la precisión (accuracy) de un modelo, se suele separar los datos en conjuntos disjuntos de datos de entrenamiento, y datos de evaluación
    - Una tasa usual es usar 70% de los datos para entrenamiento, y 30% para evaluación