

AAI - Tema 1: Introducción al aprendizaje automático

Proceso de generación de un modelo

El proceso de generación de un modelo se puede resumir en:

1. Definición del conjunto de datos
2. Preprocesamiento
3. Separación de conjunto de datos
4. Entrenamiento
5. Validación y ajuste de hiperparámetros
6. Puesta en producción

Definición del conjunto de datos

- Identificación de la variable respuesta

Preprocesado

- Revisión de valores faltantes
- Codificación de variables categóricas
- Escalado (si fuera necesario)

Separación de conjuntos de datos

Se separan los datos entre entrenamiento y validación, siguiendo la norma 80-20%.

Entrenamiento

Se entrena la muestra.

Validación

En la fase de validación se ajustan los hiperparámetros.

Si se usaran datos del conjunto de prueba para ajustar hiperparámetros el modelo puede ajustarse en exceso a los datos de prueba y la estimación del error de generalización será demasiado optimista.

Apareció en: AAI.EX.20200201.T.3

Métodos de validación:

- Validación cruzada (CV)

- Validación cruzada de k pliegos (k-fold)

Ver también métricas usadas.

Test

El test sirve para valorar

Puesta en producción

Explotación del modelo.

Métricas usadas

- Clasificación
 - Basadas en matriz de confusión
 - * Exactitud/*accuracy*
 - * Precisión/*Precision*
 - * Sensibilidad/*Recall*
 - * F1
 - * Especificidad/*Specificity* (menos)
 - Basadas en probabilidades
 - * ROC
 - * AUC-ROC
 - * Log-loss (cross-entropy)
- Regresión
 - Basadas en error
 - * Mean Absolute Error (MAE)
 - * Mean Squared Error (MSE)
 - * RMSE
 - * Median Absolute Error
 - Basadas en varianza explicada
 - * R^2
 - * Adjusted R^2

Métricas de clasificación

Exactitud o *accuracy* es la proporción de predicciones correctas.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$$

Precisión o *precision* predice como positivo, cuántos son correctos.

$$Precision = \frac{TP}{TP+FP}$$

Se usa cuando los falsos positivos son costosos.

La sensibilidad o *recall* indica de los positivos reales cuántos detecta el modelo:

$$Recall = \frac{TP}{TP+FN}$$

Se usa cuando los falsos negativos son costosos.

La puntuación F1 o *F1 score* es la media armónica entre precision y recall. Se usa cuando las clases están muy desbalanceadas.

$$F1 = 2 \frac{Precision \cdot Recall}{Precision + Recall}$$

Es la medida armónica entre precision y recall.

AUC-ROC se usa en clasificación binaria. Un clasificador perfecto tiene valor 1, uno igual al azar (inservible) vale 0.5 y menor acertaría menos que el azar.