Informática y comunicaciones
ies azarquiel

Curso especialización inteligencia artificial y big data

Capitulo 2: APRENDIZAJE SUPERVISADO

Profesor: Sebastián Rubio Valero

Septiembre 2025

Sistemas de aprendizaje automático

1 Vectores y etiquetas

El aprendizaje supervisado es una familia de algoritmos que utiliza un conjunto de patrones etiquetados, llamado conjunto de entrenamiento y denotado como $\{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$, para aprender a hacer predicciones sobre nuevos datos.

$$\{(\mathbf{x}_i, y_i)\}_{i=1}^{N}, \tag{1}$$

Ejemplo

$\mathbf{x}_i = [1, 3, 1]$, es un vector de tres característica, por ejemplo número de salones, dormitorios y baños
$\mathbf{x}_i = [1]$, es un vector de una características
Un conjunto de entrenamiento sería
$\{([1, 2, 3], 1); ([2, 2, 1], 2); ([2, 1, 2], 1)\}$
El elemento x_2^1, hace referencia a la característica 1 del segundo vector, es decir 2.
Este conjunto de entrenamiento se puede denotar como sigue:$\{(\mathbf{x}_i, y_i)\}_{i=1}^{3}$

Si las características toman valores continuos el problema puede resolverse por modelos de regresión, en cambio si toma valores discretos será de clasificación.

Cuando las características no son valores numéricos, por ejemplo las características de un correo spam $x_i = [$"$alumno123@azarquiel.es$","$Regalo$","$Enhorabuena.....$"]$. Hay que convertir las caracteristicas a valores numéricos. Una forma es mediante un diccionario en la que a cada palabra se le asigna un número, de esta forma entrenamos un modelo NLP (Procesado de lenguaje natural) que se encargará de todo el proceso.

2 Modelos

Hoy en día existe muchos modelo de aprendizaje automático supervisado. En este curso es imposible abarcarlos todos, así estudiaremos los marcados con *.

2.1 Nivel Básico

- **Regresión Lineal** (**Linear Regression*)
- **Regresión Logística** (**Logistic Regression*)
- **Análisis Discriminante Lineal (LDA)** (*Linear Discriminant Analysis*)

- **Vecinos más Cercanos (k-NN)** (**k-Nearest Neighbors*)
- **Naïve Bayes** (**Gaussian, Multinomial, Bernoulli*)

2.2 Nivel Intermedio

- **Árboles de Decisión** (**Decision Trees*)
- **Máquinas de Vectores de Soporte (SVM)** (**Support Vector Machines*)
- **Ensembles Simples** (**Bagging, Random Forest*)
- **Boosting Básico** (**AdaBoost, Gradient Boosting inicial*)

2.3 Nivel Avanzado

- **XGBoost** (*Extreme Gradient Boosting*)
- **LightGBM** (*Light Gradient Boosting Machine*)
- **CatBoost** (*Categorical Boosting*)
- **Redes Neuronales Artificiales (ANN)** (*Artificial Neural Networks*)
- **SVM con Kernels No Lineales** (*RBF, Polinomial*)

2.4 Nivel Experto

- **Redes Neuronales Profundas (DNN)** (*Deep Neural Networks*)
- **Redes Convolucionales (CNN)** (*Convolutional Neural Networks*)
- **Redes Recurrentes (RNN, LSTM, GRU)** (*Recurrent Neural Networks*)
- **Transformers** (*BERT, GPT, etc.*)

3 Criterios de Clasificación

- **Interpretabilidad**: Facilidad de explicación.
- **Hiperparámetros**: Cantidad de ajustes necesarios.
- **Matemáticas Involucradas**: Complejidad teórica.
- **Capacidad de Generalización**: Necesidad de datos y fine-tuning.

4 Entrenamiento de un modelo

Cuando queremos solucionar un problema aplicando un modelo de aprendizaje automático hay que tener en especial consideración dos aspectos, primero los datos de entrenamiento y segundo la elección del modelo a entrenar.

4.1 Sesgo

Con respecto a los datos de entrenamiento la cantidad de datos utilizados afecta a la bondad del modelo obtenido, de esta forma si queremos un sistema que para identificar los coches de una foto cuanto mayor número de imágenes de entrenamiento mejores resultados obtendremos del modelo. Otro factor importante en lo que respecta a los datos de entrenamiento es la calidad de los mismos. Supongamos que queremos entrenar un modelo para reconocer perros y gatos, si el conjunto de datos de entrenamiento tienen mas perros que gatos, el modelo estará sesgado hacia los perros,lo que se conoce como **sesgo muestral**, así que ante cualquier otra imagen que analice hay más posibilidades que me diga que es un perro que un gato. Otra situación puede ser, usando el mismo ejemplo, que las fotos de los perros estén todas al aire libre, y la de los gatos en el interior de una casa. El modelo será entrenado con este condicionante, de modo que cuando le pasemos un gato al aire libre, pude que el resultado obtenido no sea el correcto, diciendo que es un perro. Por lo tanto los datos de entrenamiento deben representativos de las dos partes, y en las misma condiciones. Ahora la pregunta que surge es, ¿si tengo un millón de fotos de perros y gatos, quien se encarga de clasificarlos, un conjunto de personas,?, la respuesta es otros modelos de machine learning. También hay cierta colaboración de los internautas de forma oculta, por ejemplo cuando nos piden que para demostrar que no somos un robot, seleccionemos las imagenes donde aparecen gatos. El **ruido muestral**, es el ruido debido a la propia muestra, para evitar este problema se aconseja que el tamaño de la muestra de entrenamiento se grande, esto atenúa este problema

4.2 Sobreajuste

El sobreajuste se produce cuando el modelo entrenado se ajusta demasiado a los datos de entrenamiento, esto puede dar lugar a un rendimiento muy pobre cuando se enfrenta a otros datos no utilizados durante el entrenamiento, en inglés se le conoce como **overfitting**. Cuando sucede esto, lo que realmente ha ocurrido es que el modelo ha memorizado los datos de entrenamiento, así que ante un dato de la vida real podemos obtener dos resultados diferentes en sucesivas veces. Para evitar el sobreajuste

existe diversas técnicas de regularización, que dependerán del modelo a entrenar.

4.3 Outlier

Antes de entrenar nuestro modelo es aconsejable, estudiar el conjunto de datos de entrenamiento para detectar los datos que están fuera del patrón general. Estos datos pueden afectar significativamente al modelo obtenido.
Esta anomalía en los datos se debe a diferentes factores, por ejemplo una inputación de datos no correcta, una mala lectura de un sensor, etc.

Los modelos deben ser reentrenados con cierta regularidad, bien por que los datos de entrenamiento son obsoletos, por que algunos nuevos datos influyen en la eficacia del modelo, incluso cuando los datos no se ven alterados con el paso del tiempo, por ejemplo por que las cámaras han modificado la resolución, sensores, mapa de colores, etc. Cuando surgen nuevos datos, por ejemplo nuevas leyes, o cambios en la regulación de medicamentos, tenemos que entrenar el modelo con los datos antiguos, más los datos nuevos. Estos cambios puden durar horas e incluso diás. ¿Pero que pasa con los datos cambia de forma dinámica con el paso del tiempo, por ejemplo la bolsa?, para solucionar estos casos utilizamos modelo de aprendizaje online El análisis y tratamiento de los datos se estudia más afondo en Sistemas de Big Data