Trabajo Práctico 4: Métodos Supervisados Avanzados

Introducción

Evaluaremos el funcionamiento de clasificadores como gradient boosting, random forests y support vector machine en datasets con distintas distribuciones y alterando sus parámetros para obtener distintos resultados.

Se pide entregar un notebook con código funcionando y con comentarios y análisis correspondientes a cada resultado obtenido.

Ejercicio 1

Uno de los usos más comunes del algoritmo de Naive-Bayes es para la clasificación de texto. En este punto vamos a usar un dataset conocido como “20 Newsgroups Dataset”[1] que se compone de 18000 posts de newsgroups[2] de 20 tópicos distintos.

Se le entrega código para vectorizar cada post según las ocurrencias de cada palabra de un diccionario dado, lo que permitirá luego entrenar un clasificador discreto multinomial (pag. 180 del libro de Mitchell).
Escriba código para entrenar este clasificador sobre los datos correspondientes y evaluarlo sobre los conjuntos de validación y test. Evalúe distintas combinaciones del largo del diccionario de palabras (entre 1000 y 4000 por lo menos) y del parámetro alfa (órdenes de magnitud, de 1 a 0.0001), buscando el mínimo en validación.
¿Hay sobreajuste? ¿El comportamiento en validación es representativo del conjunto de test? Calcule una matriz de confusión para el conjunto de test. ¿Hay alguna particularidad que merece atención especial?

Ejercicio 2

Resuelva el problema de las espirales-anidadas usando k-nn. Utilice el datasets de "espirales con ruido" que se entrega junto a éste trabajo práctico. Hay una versión "original" y otra que tiene agregadas dos variables que contienen ruido uniforme.
Realice gráficas de las predicciones sobre el conjunto de test, y gráficas de errores vs número de vecinos. Compare el resultado con el obtenido con árboles de decisión, los dos métodos sobre las dos versiones del dataset.

Otra variante de k-nn que se suele utilizar es usar en la votación a todos los patrones que estén a una distancia menor a un dado valor D del patrón que se quiere clasificar, en lugar de usar un

[1] https://scikit-learn.org/stable/datasets/real_world.html#newsgroups-dataset
[2] https://en.wikipedia.org/wiki/Usenet_newsgroup

número fijo k. El único parámetro del algoritmo, ahora, es la distancia máxima D, la que se optimiza utilizando un conjunto de validación.
Implemente código que optimice el valor de D de forma razonable en función de los datos de entrenamiento, utilizando un conjunto de validación si es necesario. Lo más interesante de este punto es discutir y definir cómo encontrar el valor de D (sin ayuda externa). Repita la primer parte del ejercicio (predicciones y curvas de error vs D).

Ejercicio 3

Evalúe el efecto de la complejidad del clasificador en boosting.
Use los dos datasets adjuntos que se detallan, un dataset es el problema de las espirales anidadas con ruido, el otro es el conocido problema "diagonal".
Controle la complejidad de los arboles poniendo un máximo a la profundidad de los mismos (ejemplo en el código).
Utilice 200 árboles para cada ensemble y estime el error de clasificación en test en función de la profundidad máxima para valores de ésta de 1 a 20.

Grafique y analice el resultado en cada caso.

Ejercicio 4

Evalúe el efecto de la cantidad de features evaluadas a cada paso para Random Forest (*max_features* en sklearn.ensemble.RandomForestClassifier). Use el dataset RRL que se adjunta (la variable a predecir es "Tipo"), cambiando *max_features* como fracción del total de features en potencias de 1/2 (en este caso, 69, 34, 17, 8, 4, 2, 1).
Utilice 1000 árboles para cada ensemble y estime el error de clasificación OOB en función de la fracción utilizada, como promedio de 5 corridas para cada valor de *max_features*.

Grafique y analice el resultado en cada caso.

Ejercicio 5

Aplique las Support Vector Machines (SVM) con kernels RBF y Polinomial (sklearn.svm.SVC) sobre el dataset Lampone para predecir la clase (variable n_tipo), con una metodología adecuada para seleccionar los parámetros internos (C y gamma en ambos, degree además en kernel polinomial) y estimar el error.

Trabajo Práctico 4: Métodos Supervisados Avanzados

Ejercicio 6 (Opcional, 1 punto)

Busque un dataset que considere interesante, aplicando alguno de los métodos de clustering discutidos y alguno de los métodos que determina la cantidad de clusters presentes (obligatorios XGboost y SVM).
Analice y explique los resultados obtenidos.