Trabajo práctico 1

El objetivo de este trabajo práctico es analizar características y particularidades de la utilización de algoritmos de clasificación para algunos casos especiales.

El presente trabajo deberá elaborarse en grupos de 2 personas. Se cuenta con 3 semanas para la entrega.

El material básico para la elaboración del presente trabajo se encuentra en las transparencias, en el libro de Mitchell, documentación de scikit-learn y demás referencias de la materia. Podrá utilizarse cualquier otra fuente siempre que esté correctamente referenciada.

Se debe producir un informe que contenga los resultados de los experimentos que se describen a continuación, que deben considerarse a modo orientativo y no limitativo.

Para los ejercicios 1) y 2) se pide obtener un dataset cuyo tamaño esté en el orden de 5 a 10 atributos x 1000 ejemplos, preferiblemente extraídos de problemáticas reales. Pueden usarse dataset públicos[1] pero sería preferible que fueran de interés de los integrantes del grupo. Cuando se lo requiera, se tomará un 20% de datos para validación.

1) Árboles de Decisión. Sobreajuste y ruido

Se pide:

- Ejecutar corridas de IDT variando la función de poda para crear árboles de diferente tamaño.
- Graficar la performance sobre el conjunto de entrenamiento y sobre el conjunto de validación en función del tamaño del árbol.
- Analizar el fenómeno de sobreajuste sobre los resultados obtenidos.
- Perturbar el set de entrenamiento induciendo diferentes porcentajes de ruido sobre la clase variando de 0% a 50%. Analizar el comportamiento de los árboles

2) Vecinos Más Cercanos. Sobreajuste y curse of dimensionality

Se pide:

- Ejecutar corridas de IBK variando el tamaño de la vecindad. Utilizar las funciones de peso de las distancias según la cercanía a la consulta[2].

[1] UC Irvine Machine Learning Repository - http://archive.ics.uci.edu/ml/
[2] weights=['uniform', 'distance'] en sklearn

- Graficar la performance sobre el conjunto de entrenamiento y sobre el conjunto de validación en función del tamaño de la vecindad.
- Analizar el fenómeno de sobreajuste sobre los resultados obtenidos.
- Perturbar el set de entrenamiento agregando variables con valores random variando de 0 a 20. Analizar el comportamiento de los modelos según la alta-dimensionalidad de atributos irrelevantes.

3) Naive Bayes. Text Mining

Se pide:

- Construir un clasificador de tópicos para noticias periodísticas utilizando Naive Bayes.
- Tomar el dataset de notas periodísticas. El dataset cuenta con 16582 notas periodísticas de los últimos 3 meses. Tomar como clase, el campo *topic*.

Cantidad de notas	Tópico
919	Cultura
3734	Deportes
1738	Economía
1832	Espectáculos
1686	Internacionales
3231	Política
1132	Seguridad
1877	Sociedad
433	Tecnología

- Presentar la matriz de confusión del clasificador.
- Las palabras stop-words[3] son palabras que carecen de semántica y son usualmente filtradas en tareas de procesamiento de lenguaje natural, text mining, IR, etc. Contienen a las palabras funcionales, como los artículos, conjunciones, etc.; y palabras muy utilizadas en diversos dominios de información.
  Se pide construir un modelo de clasificación de notas filtrando las stopwords del vocabulario. Comparar la performance de ambos modelos.

[3] https://en.wikipedia.org/wiki/Stop_words

- ★ Evaluar distintas funciones de codificación del vocabulario (TFIDF, stemming, lematización, n-grams, etc).

Informe

El documento a entregar debe cumplir con los requisitos normales de un “informe de investigación” y debe incluir como mínimo: objetivo, resultados esperados, metodologías utilizadas, análisis de resultados, conclusiones, ideas para futuras investigaciones, bibliografía, referencias. Se pide que la entrega sea utilizando el soporte IPython Notebook[4] para agilizar la documentación de las pruebas y comparaciones realizadas, el dataset y los scripts desarrollados.

Por tratarse de un trabajo de investigación netamente práctico, las conclusiones deben ser la resultante de la elaboración de las pruebas realizadas. La información obtenida de referencias externas puede y debe ser tomada como insumo, pero nunca como conclusión.

La evaluación se basará en la elaboración y precisión del contenido, la calidad del informe y la exposición de los resultados más sobresalientes. Todos los integrantes del grupo obtendrán la misma calificación. La entrega fuera del plazo tiene una penalización de un punto por semana.

A modo estimativo, el informe no deberá tener más que (el equivalente a) a 5 páginas a espacio simple en Arial 11, sin contar los anexos digitales. En el encabezado se deben enumerar los integrantes del grupo responsable del informe.

[4] http://ipython.org/notebook.html