Universidad de los Andes Colombia

TÓPICOS AVANZADOS EN ESTADÍSTICA - IIND 4114
Métodos de Aprendizaje Estadístico y Machine Learning
DEPARTAMENTO DE INGENIERÍA INDUSTRIAL
Periodo 2015-20. Profesor: Carlos Valencia

PROYECTO FINAL

- ∼ El informe del proyecto debe ser entregado antes del Lunes 30 de Noviembre a las 5pm
- ∼ La solución puede ser elaborada en grupos de máximo 3 personas
- ∼ Se recomienda no hacer el proyecto de manera indiviudal por la carga de trabajo que requiere
- ∼ Utilice procedimientos explícitos y análisis detallados

I. Introducción y Descripción del Problema.

El proyecto consiste en utilizar datos relaes para predecir la popularidad que tienen ciertos artículos de información que se publican en un medio de divulgación online (`www.mashable.com`). Predecir esta popularidad, permite planear de mejor forma las ediciones dinámicas de ciertos tipos de noticias, anuncios, etc.
La respuesta se mide con el número de *shares* que lo ususarios dan al artículo, la cual se busca predecir a partir de variables que explican las características del contenido o de la publicación. Por ejemplo, algunos predictores explican la longitud del artículo, mientras otros tienen que ver con cuándo fue publicado. A continuación se presenta la descripción de los datos (note que las primeras dos variables no se usan para predecir), donde la última es la variable de respuesta:

```
0. url: URL of the article (non-predictive)
1. timedelta: Days between the article publication and the dataset acquisition (non-predictive)
2. n_tokens_title: Number of words in the title
3. n_tokens_content: Number of words in the content
4. n_unique_tokens: Rate of unique words in the content
5. n_non_stop_words: Rate of non-stop words in the content
6. n_non_stop_unique_tokens: Rate of unique non-stop words in the content
7. num_hrefs: Number of links
8. num_self_hrefs: Number of links to other articles published by Mashable
9. num_imgs: Number of images
10. num_videos: Number of videos
11. average_token_length: Average length of the words in the content
12. num_keywords: Number of keywords in the metadata
13. data_channel_is_lifestyle: Is data channel 'Lifestyle'?
14. data_channel_is_entertainment: Is data channel 'Entertainment'?
15. data_channel_is_bus: Is data channel 'Business'?
16. data_channel_is_socmed: Is data channel 'Social Media'?
17. data_channel_is_tech: Is data channel 'Tech'?
18. data_channel_is_world: Is data channel 'World'?
19. kw_min_min: Worst keyword (min. shares)
20. kw_max_min: Worst keyword (max. shares)
21. kw_avg_min: Worst keyword (avg. shares)
22. kw_min_max: Best keyword (min. shares)
23. kw_max_max: Best keyword (max. shares)
24. kw_avg_max: Best keyword (avg. shares)
```

25. kw_min_avg: Avg. keyword (min. shares)
26. kw_max_avg: Avg. keyword (max. shares)
27. kw_avg_avg: Avg. keyword (avg. shares)
28. self_reference_min_shares: Min. shares of referenced articles in Mashable
29. self_reference_max_shares: Max. shares of referenced articles in Mashable
30. self_reference_avg_sharess: Avg. shares of referenced articles in Mashable
31. weekday_is_monday: Was the article published on a Monday?
32. weekday_is_tuesday: Was the article published on a Tuesday?
33. weekday_is_wednesday: Was the article published on a Wednesday?
34. weekday_is_thursday: Was the article published on a Thursday?
35. weekday_is_friday: Was the article published on a Friday?
36. weekday_is_saturday: Was the article published on a Saturday?
37. weekday_is_sunday: Was the article published on a Sunday?
38. is_weekend: Was the article published on the weekend?
39. LDA_00: Closeness to LDA topic 0
40. LDA_01: Closeness to LDA topic 1
41. LDA_02: Closeness to LDA topic 2
42. LDA_03: Closeness to LDA topic 3
43. LDA_04: Closeness to LDA topic 4
44. global_subjectivity: Text subjectivity
45. global_sentiment_polarity: Text sentiment polarity
46. global_rate_positive_words: Rate of positive words in the content
47. global_rate_negative_words: Rate of negative words in the content
48. rate_positive_words: Rate of positive words among non-neutral tokens
49. rate_negative_words: Rate of negative words among non-neutral tokens
50. avg_positive_polarity: Avg. polarity of positive words
51. min_positive_polarity: Min. polarity of positive words
52. max_positive_polarity: Max. polarity of positive words
53. avg_negative_polarity: Avg. polarity of negative words
54. min_negative_polarity: Min. polarity of negative words
55. max_negative_polarity: Max. polarity of negative words
56. title_subjectivity: Title subjectivity
57. title_sentiment_polarity: Title polarity
58. abs_title_subjectivity: Absolute subjectivity level
59. abs_title_sentiment_polarity: Absolute polarity level
60. shares: Number of shares (target)

Para estudiar el problema de predicción, la solución se divide en dos estrategias: predecir le número de *shares* como un problema de regresión, o predecir si el artículo es o no popular (categorizando la variable *shares*) como un problema de clasificación. Cada grupo deberá resolver los dos problemas.

II. Datos y Competencia.

Para obtener los datos, y la desripción de las variables, puede ir al sitio:
`https://inclass.kaggle.com/c/advanced-topics-in-statistics-regression-201520-uniandes`
para el problema de regresión, y para el problema de clasificación puede ir a:
`https://inclass.kaggle.com/c/advanced-topics-in-statistics-classification-201520-uniandes`
Podrá obtener un archivo con los datos de entrenamiento (`train.csv`) que incluyen tanto las variables como la respuesta. Además encontrará un archivo con datos de prueba (`test.csv`).

Parte del proyecto incluye participar en una competencia entre todos los grupos. Usted debe encontrar un modelo predictivo en cada caso (regersión y clasificación) y usarlo para predecir en los inputs proporcionados en los datos de prueba (`test.csv`). El desempeño de sus modelos será evaluado automáticamente por el sitio online (`kaggle.com`), y se hará el ranking de los grupos. Más instrucciones sobre el formato de la competencia, pueden ser revisados en los `url` suministrados.

III. Guía para desarrollo de proyecto y elaboración de informe.

En general, cada grupo tiene libertad para desarrollar su proyecto. Tenga en cuenta que el objetivo fundamental es utilizar corrrectamente los conocimientos adquiridos en clase para encontrar un buen modelo predictivo. La evaluación no se basa únicamente en el desempeño de su algoritmo, sino en el buen desarrollo del modelo y la justificación de su uso.
Para encontrar un modelo apropiado, se recomienda que su proyecto contenga las siguientes etapas:

1. **Exploración y Visuaización de datos**. Antes de empezar a estimar funciones, es mejor si aduqiere un conocimiento previo de qué exactamente son los datos, cómo se relacionan entre sí y cómo afectan la variable de respuesta. Algunas referencias que puede usar para hace exploración y visualizaciń de datos las puede encontrar en el archivo anexo (`data-exploration.zip`), o puede revisar el link:
   `http://www.rdatamining.com/docs/data-exploration-and-visualization-with-r`

2. **Selección y Extracción de variables**. Uno de los resultados importantes que salen de la exploración de los datos, determinar qué variables, o qué transformación de ellas se debe usar en cada modelo. Referencias para realizar estos procesos se pueden encontrar en el capítulo 3 del libro *Applied Preditive Modeling* de Kunh y Johnson, el cual recomendamos leer. Tenga en cuenta que en esta etapa se debe usar la intución que se tiene del problema real, para determinar patrones que sean útiles para predecir o explicar la respuesta.

3. **Preparación y limpieza de datos**. Esta parte requiere mirar la validez de algunos datos y decidir que inofrmación no es útil. Los datos suministrados son bastante limpios y no requieren mucha limpieza. Procesos típicos de limpieza incluyen tratamiento de datos perdidos (*missing data*) y datos atípicos (*outliers*).

4. **Evaluación de modelos**. Acá se incluye el ajuste y la evaluación de diferentes modelos predictivos. Recuerde que es muy importante calibrar los modelos correctamente y escoger adecuadamente entre ellos.

5. **Selección de modelo final, análsis y conlcuiones**

IV. Evaluación.

La nota del proyecto será sobre 100 puntos. El informe, con todos los detalles de la solución tendrá un valor de 90 puntos. El resto de la nota dependerá de su desmpeño en la competencia. Para cada problema (regresión y clasificación), habrá 15 puntos máximos disponibles (para el primer lugar). La bonificación para cada grupo será proporcional al ranking que ocupe. Por ejemplo, si un grupo queda justo en la mitad (mediana!!!) en la competencia de clasificación, bonificará 7.5 puntos. Note que la nota máxima que un grupo puede obtener e de 120/100.