Tarea 3 Análisis Inteligente de Datos

Juan Ávalo, Bastien Got

July 13, 2016

Contents

1 Reducción de Dimensionalidad para Clasificación

La idea del problema es de clasificar las once vocales del inglés británico para poder reconocer éstas vocales independamente de la persona que habla. Los cadidatos dijeron 11 palabras cada uno, que corresponden a las once vocales. Luego, la clasificación será probada con nuevos candidatos.

1. En el caso de los datos de entrenamiento, hay 48 registros que corresponden al numero de veces distintas que cada palabra fue registrada. Hay 42 registros en el dataset de prueba. Por cada registro, la persona dijo 11 palabras que corresponden a las 11 vocales de la idioma.

2. Normalizamos los datos para que una de las variables no tenga un mayor peso que las otras solo por tener valores mas grandes en valor absoluto.

3. Elección de paleta : **Espectral**

   Porque es la paleta que tiene la divesidad de colores más grande y por lo tanto es más fácil de ver donde se situan los datos de cada clase.

   En el codigo dado en la tarea mclases tiene 9 elementos, sin embargo el numero de clases es 11 por el numero de vocales.

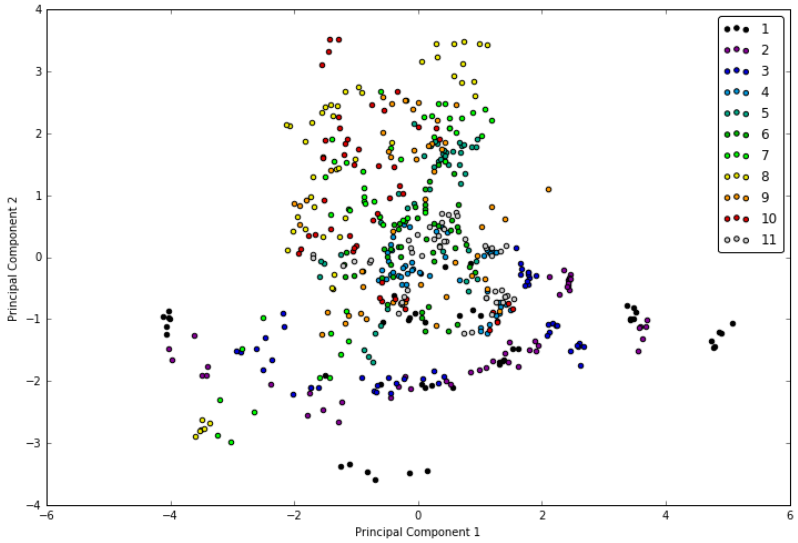

Figure 1: *PCA*.

4. La paleta es la misma que en la pregunta anterior.

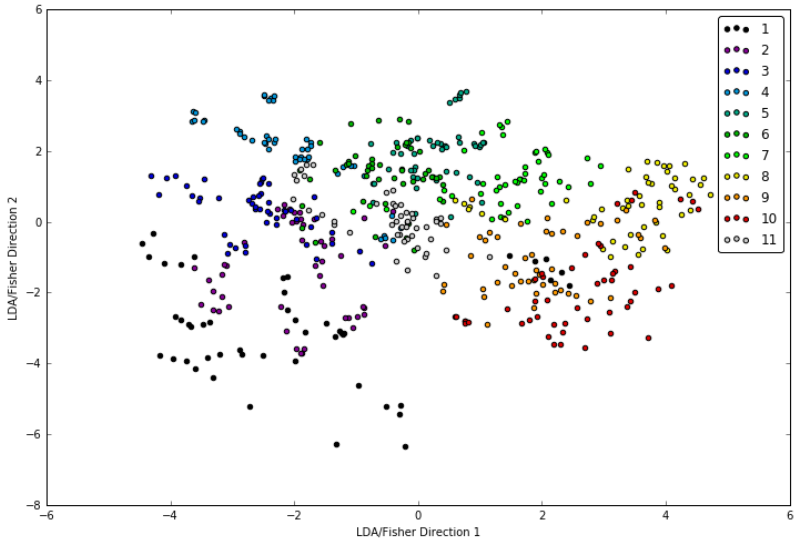

Figure 2: *LDA*.

5. Cualitativamente parece que el metodo de reducción de dimensionalidad LDA es más interesante que el PCA ya que las clases son más separadas con LDA, o dicho de otra manera por cada clase los datos son más agrupados usando LDA.

   Podemos verificarlo utilizando el metodo del K-means: se debe calcular la distancia de cada punto con el promedio de los puntos de su clase. El metodo más adecuado es el que minimiza la suma de los cuadros de estas distancias

6. Cada palabra es dicha el mismo numero de veces : 48 por los datos de pruebas. Lo que significa que una palabra tiene la misma probabilidad a priori de pertenecer a cada una de las clases : 1/11 0.091.

7. La tabla siguiente muestra la puntuación de cada modelo sobre los datos de entrenamiento y los datos de prueba.

Modelo	LDA	QDA	10-NeighborsClass
Entrenamiento	0.68	**0.98**	0.93
Prueba	0.45	0.42	**0.49**

El modelo más apropriado para los datos de entrenamiento es QDA con una puntuación de 0.98. El modelo que se comporta lo mejor sobre los datos de prueba es el K-NeighborsClasifier con $K = 10$ que nos da una puntuación de 0.49.

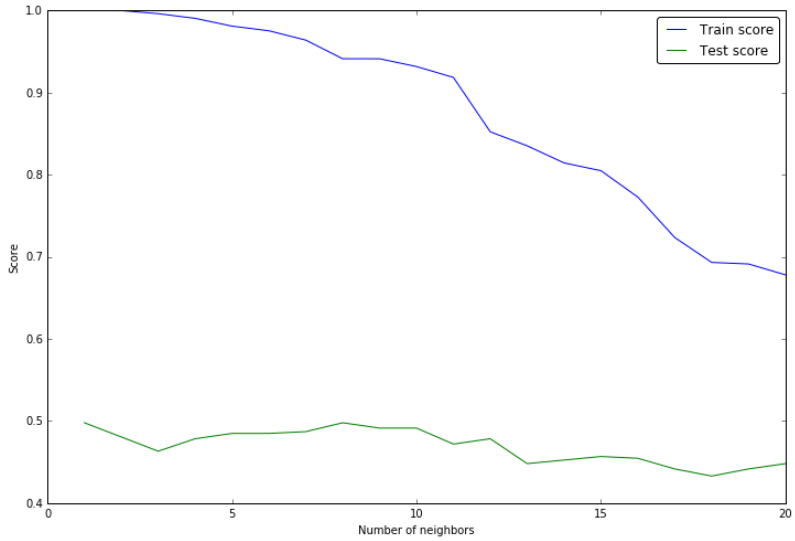

Figure 3: Score de K-Neighbors.

El score de entrenamiento va bajando cuando K aumenta, aunque el score de prueba se queda entre 0.45 y 0.5. El mejor que podemos obtener es un score de prueba de 0.498 con $K = 1$ y $K = 8$.

8. En esta pregunta utilizamos PCA para reducir la dimensionalidad. El mejor modelo es el que minimiza el error de prueba y por lo tanto encontramos que QDA en aplicado al dataset reducido a una dimensión es lo más interesante, con un error de prueba de 0.77.

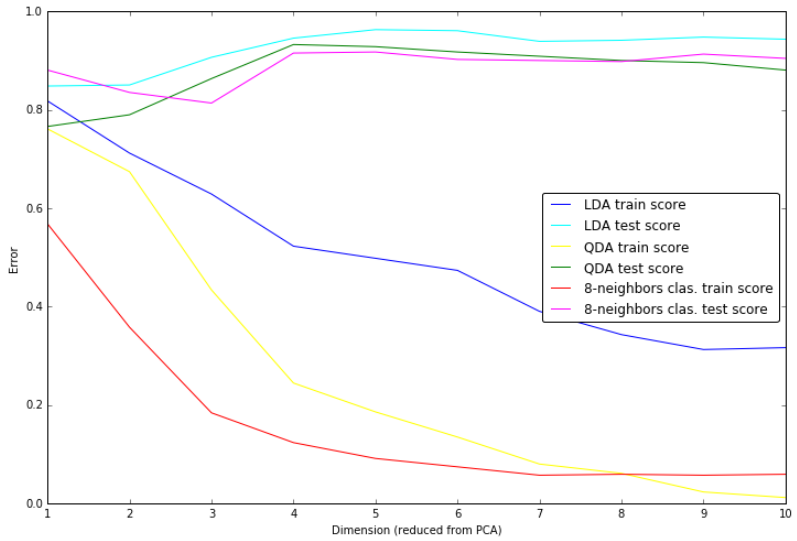

Figure 4: Errores de entrenamiento LDA, QDA, K-NN usando PCA.

9. Utilizando LDA para reducir la dimensionalidad del dataset, encontramos que el modelo más interesante todavía es QDA, pero esta vez con un error de prueba de 0.78.

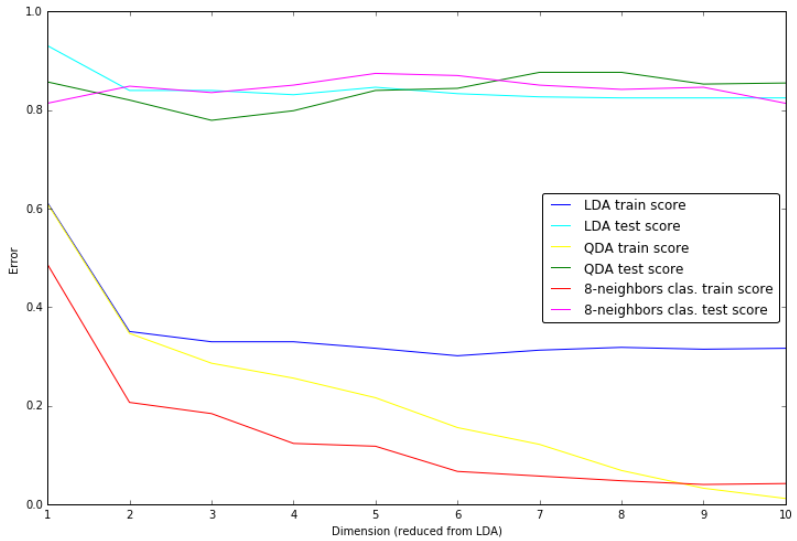

Figure 5: Errores de entrenamiento LDA, QDA, K-NN usando LDA.

2 Reconocimiento de Texto

El problema de ésta parte es reconocer si un comentario de un sitio web de películas es favorable o desfavorable.

2.1 Preprocesamiento de los datos.

1. Tanto el set de entrenamiento como el de pruebas tienen **3554** datos. Los datos se separan en textos positivos y negativos de acuerdo a la tabla:

Sets	Positivos	Negativos
Entrenamiento	1770	1784
Prueba	1803	1751

2. Se creó una función que extrae palabras usando *stemming* y quitando *stopwords* si así es pedido. Sobre los resultados obtenidos se puede observar con los ejemplos usados que *stemming* lo que hace es hacer que una palabra tome una pseudoraiz. Además, el proceso de quitar *stopwords* efectivamente quita palabras como "I" o "to".

   Como ejemplo, las frases: "I love to eat cake" y "I love eating cake" ambas se reducen a "love", "eat" y "cake".

   Pero un ejemplo más interesante es el de aplicar *stemming* sobre palabras como "absolutely" y "dislike", las cuales se traducen en "absolut" y "dislik". Ninguna de las pseudoraices es una palabra del inglés verdadera, pero van a ser útiles para saber que palabras están relacionadas o no.

3. También se creó una función análoga a la del punto anterior, pero lematizando. Para poder lograr ésto no se pudo usar la función `WordNetLemmatizer` directamente como estaba en los ejemplos.

   Lematización involucra hacer un análisis sintáctico de las palabras, por lo que la función usada pide marcar cada palabra con la posición dentro de la oración que tiene (verbo, sustantivo, adjetivo, adverbio), la cual se obtuvo mediante la función `pos_tag`.

   Los efectos en las palabras de ejemplo son aparentes. En casos como el de "I love to eat cake" el lematizador las reduce de la misma forma que usando *stemming*. La diferencia se nota al usar las palabras "dislike" y "absolutely", las cuales se mantienen iguales. O con palabras como "are" e "is" las cuales se reducen a "to be".

4. Se generaron cuatro representaciones vectoriales para los dos conjuntos de datos. La razón de ésto es porque se necesita extraer palabras con *stemming* y con *lemmatize*, con *stopwords* y sin ellas.

   La representación del texto consiste en resumir cada comentario a un vector binario con todo el vocabulario obtenido de todos los mensajes. Si el mensaje tiene una palabra dentro del vocabulario, el valor de la variable correspondiente a esa palabra es **1**. Sino es **0**.

   Luego las etiquetas son **0** si el mensaje es negativo, y **1** si es positivo.

   Considerando como se tratan los datos, se puede rankear las palabras que globálmente se encontraron por frecuencia. Un ejemplo de ello es la siguiente tabla:

Frecuencia	Palabra
115	way
125	get
127	well
128	much
129	work
143	even
143	time
145	comedy
163	character
169	good
176	story
246	one
254	like
264	make
481	movie
573	film

5. El evaluador de desempeño considera las siguientes medidas:

   - La precisión del modelo sobre los datos de entrenamiento.
   - La precisión del modelo sobre los datos de prueba.
   - La *precisión*, esto es, el porcentaje de datos bien clasificados dentro de todos los datos clasificados.
   - El *recall* el cual es el porcentaje de los datos seleccionados bien clasificados dentro de los datos de su clase.
   - El *f1-score*, el cual es la media armónica entre la precisión y el recall.
   - El support, que cuenta cuantos datos de cada clase hay.

2.2 Modelos a probar.

Al dataset procesado mediante las combinaciones de stemming, lemmatize y stopword removal se le aplicaron cuatro modelos de clasificación: "Naive Bayes", "Multinomial Naive Bayes", "Logistic Regression" y "Linear SVC".

En general los resultados eran prácticamente todos con un f1-score mayor a 0.70. Se notó en general que, sobre el comportamiento de las predicciones, éstas eran más certeras mientras más vocabulario asociado a textos favorables o desfavorables tenían.

Ninguno de los tipos de modelo se comporta bien con frases sarcásticas o con críticas que tienen muchas palabras que en otro contexto serían positivas.

Casi todos los modelos entregaban, a parte de la predicción concreta, un par de valores que corresponde a la probabilidad de que el texto sea positivo o negativo. Si ambos valores se parecían, indica que el modelo no tiene certeza de que clase de texto es. Por otro lado, si uno de los valores es muy grande indica una casi absoluta certeza en el resultado.

Los resultados generales para todos los modelos son resumidos por el siguiente gráfico, tomando como medida el mayor f1-score obtenido por cada modelo.

Curiosamente, el modelo más simple con la menor cantidad de preprocesamiento es la que resulto mejor evaluada usando la métrica de f1-score y los modelos más complicados resultaron peor evaluados.

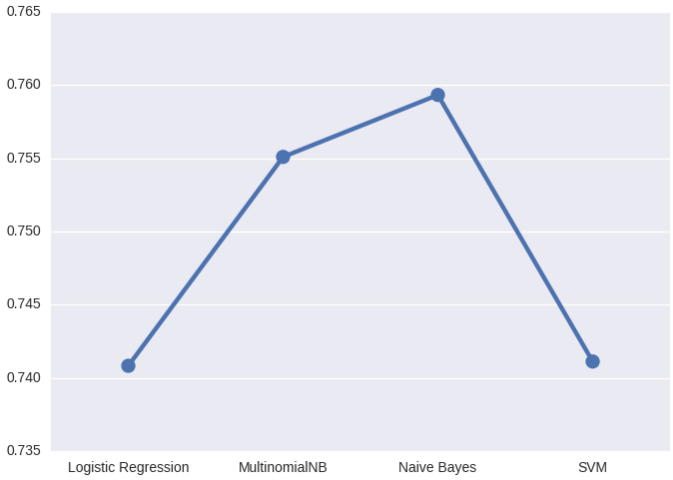

Figure 6: Resultados generales.

Viendo más en detalle, los resultados para los distintos modelos se encuentran en los siguientes gráficos. El resumen para todos ellos es que las predicciones son mejores mientras menos preprocesamiento haya.

En el caso del uso de *stopwords* se asume que las palabras que son omitidas mediante éste paso, para efectos de clasificar sí son importantes. Dejar de lado palabras como "I" puede cambiar el contexto de la frase y con esto cambiar el sentimiento que se quería transmitir.

Sobre la lematización se puede atribuir los problemas a dificultades en obtener los *tags* semánticos apropiados para cada palabra, más el hecho de que los comentarios de internet de por si pueden tener palabras que no están en el diccionario de WordNetLemmatizer, por lo cual no descubriría relaciones que stemming si encuentra.

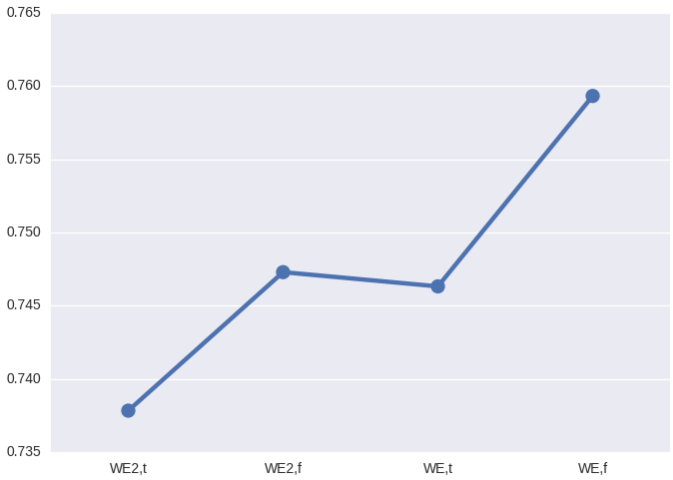

Figure 7: *Naive Bayes*.

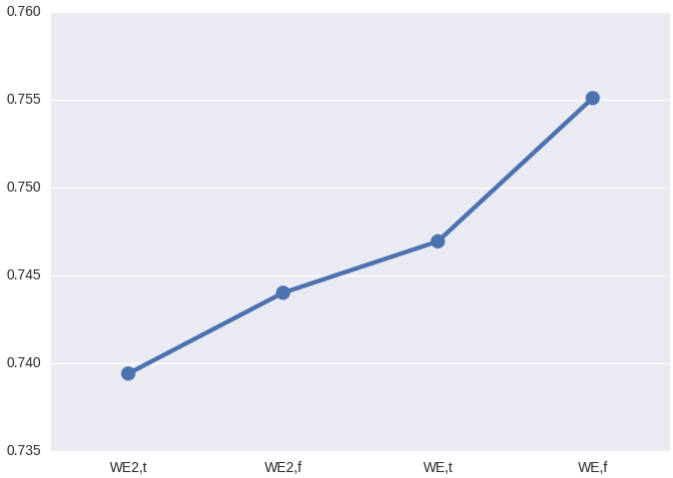

Figure 8: *multinomial naive bayes.*

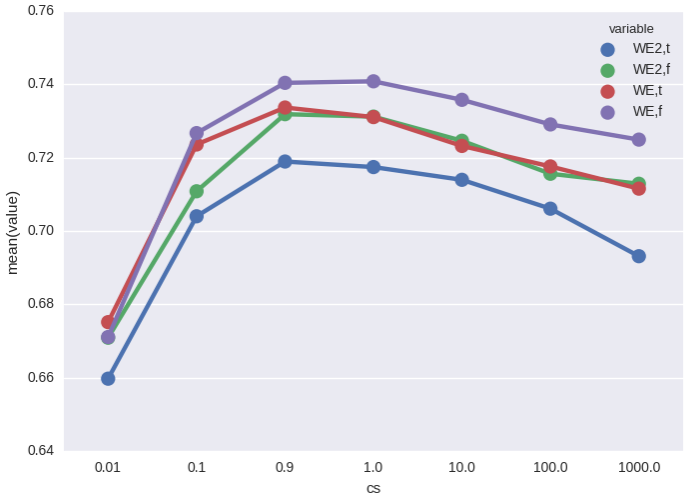

Figure 9: *Logistical Regression.*

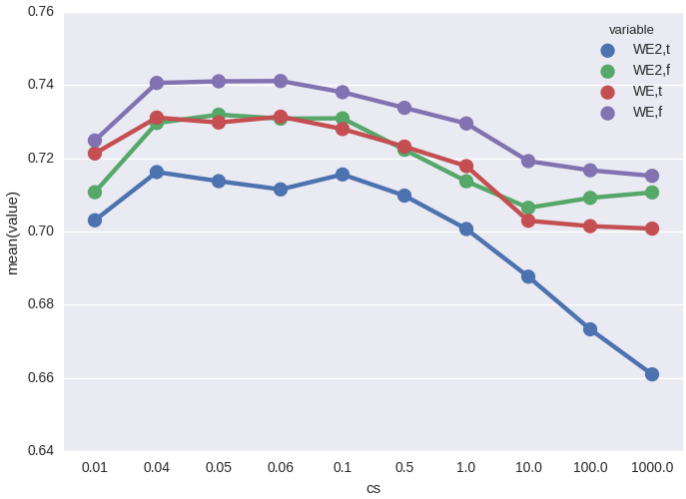

Figure 10: *Linear SVM.*

Hay que hacer ciertas observaciones al respecto del parámetro `C` de *Logistical Regression* y *Linear SVM*. Los gráficos anteriores muestran valores aproximados de los mejores parámetros que se pueden elegir, los cuales son $C_{logit} = 0.9$ y $C_{svm} = 0.05$.

Además, analizando los valores de las probabilidades de obtener un label en cada texto escogido por la función de test, se puede ver en el caso de *Logistical Regression* que aumentar `C` hace que las predicciones sean mas determinantes. O sea, el modelo resultante tiene menos dudas al momento de decidir a que label le corresponde cada texto. Viceversa si `C` disminuye, para efectos del modelo los textos parecen ser más ambiguos.

Usando la función *Linear SVM* no se puede saber si el modelo resultante sigue la misma tendencia con las probabilidades, ya que no son entregadas por la función correspondiente. Usando *SVM* con un kernel lineal, lo cual debería ser equivalente salvo detalles de implementación, permite encontrar las probabilidades, las cuales siguen el mismo patrón.