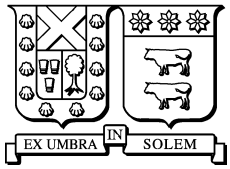
Análisis Inteligente de Datos

Tarea 3

Paulina Aguila - Felipe Flores

14 de julio de 2016



1. Reducción de Dimensionalidad para Clasificación

La reducción de dimensionalidad es un proceso que tiene mucha importancia dentro del análisis de datos, ya que ayuda a la visualización y la exploración de los datos, también reduce el costo computacional del procesamiento de los datos debido a que son menos dimensiones. Otro beneficio importante de la reducción de dimensionalidad, es que reduce significativamente el riesgo de *overfitting* o sobre ajuste del modelo.

En esta primera sección, se trabajará con datos sobre sonidos fonéticos que debe ser identificados con vocales del inglés británico. Los datos se representan en un espacio de 10 características ($d = 10$), en donde 528 registros corresponden a datos de entrenamiento y 462 son datos de prueba. Los autores reportan que el mejor desempeño corresponde a un 56 % de accuracy y se alcanza con un modelo de vecinos más cercanos y una red neuronal artificial de radio basal.

A través del lenguaje de programación Python, se cargan los datos de la fuente [**?**] y se llevan a un dataframe de entrenamiento con 528 registros y un dataframe de prueba con 462 registros.

Para cada conjunto de datos (entrenamiento y test), se deben normalizar los datos. Este es un paso muy importante, ya que permite ajustar la escala de las variables a la varianza de la unidad, lo que hace que los valores de datos que se encuentran ubicados en los extremos, no ejerzan un peso excesivo en la función objetivo.

Utilizando PCA (Análisis de Componentes Principales) se genera una representación en dos dimensiones para el dataset inicial (10 dimensiones). La Figura 7, muestra la clasificación que realiza PCA.

Figura 1: Gráfica que muestra las dos componentes principales de PCA diferenciando con distintos colores las 9 clases.

Utilizando LDA (Linear Discriminant Analysis) se genera una representación en dos dimensiones para el dataset inicial (10 dimensiones). La Figura 2, muestra la clasificación que realiza LDA.

Al analizar las Figuras 7 y 2, se puede observar que en ambas se redujo la dimensionalidad de 10 a 2 componentes. Sin embargo, con PCA se observa que las clases no están separadas lo suficiente como para diferenciarlas, pero al utilizar LDA, se puede ver que las clases se separan más entre sí. Esto se puede deber a que LDA es un método de clasificación que en este caso se utiliza para reducir dimensionalidad y evitar el *overfitting*.

Para el caso de querer clasificar un registro x escogido aleatóriamente considerando solo la probabilidad a priori de cada clase, se debe calcular la probabilidad de ocurrencia de cada clase, luego en base a la ecuación (1), se selecciona la clase que tiene mayor probabilidad.

$$j = \underset{i}{Argmax}(P(y = C_i)) \tag{1}$$

En el siguiente ítem, se utilizaron los métodos de clasificación LDA, QDA y KNN, sin realizar reducción de dimensionalidad. A continuación, la Tabla 1 muestra un resumen con los valores del accuracy para cada uno de estos modelos tanto con los datos de entrenamiento como con los de test.

De la Tabla 1, se puede ver que para el conjunto de datos de train el método que mejor se comporta es QDA, mientras que al aplicarlo a los datos de test el método KNN tiene la mejor accuracy.

Figura 2: Gráfica que muestra las dos componentes principales de LDA diferenciando con distintos colores las 9 clases.

	LDA	**QDA**	**KNN (k=10)**
Train set	0.6837	**0.9886**	0.9318
Test set	0.4524	0.4156	**0.4913**

Tabla 1: Recuadro con el score o accuracy para cada método aplicado a datos de entrenamiento o de prueba.

La Figura 3 que se muestra a continuación, muestra un gráfico en donde se aplica el método KNN variando el valor de k desde 1 a 10, con respecto al score o accuracy. Se puede apreciar que cuando k=7 entonces el método KNN funciona mejor y tiene mejor score.

Para la última parte, se realizó reducción de domiensionalidad con los métodos PCA y LDA, para cada uno de los cuales se fue variando la cantidad de dimensiones desde 1 hasta la máxima 10, además, para cada iteración se clasificó el modelo con LDA, QDA y KNN (k=7), calculando los errores de clasificación para los datos de entrenamiento y para los de test. Estos errores se graficaron obteniéndose los gráficos de las Figuras 4 y 5.

Para la Figura 4, se puede ver que la cantidad de dimensiones igual a 7 es óptima ya que reduce el error de clasificación para los datos de test, comportándose de mejor manera el método LDA. Para la Figura 5, se observa que con una cantidad de dimensiones igual a 2 se tiene un menor error para los datos de prueba, siendo el más estable LDA no variando mucho su error al aumentar de dimensionalidad.

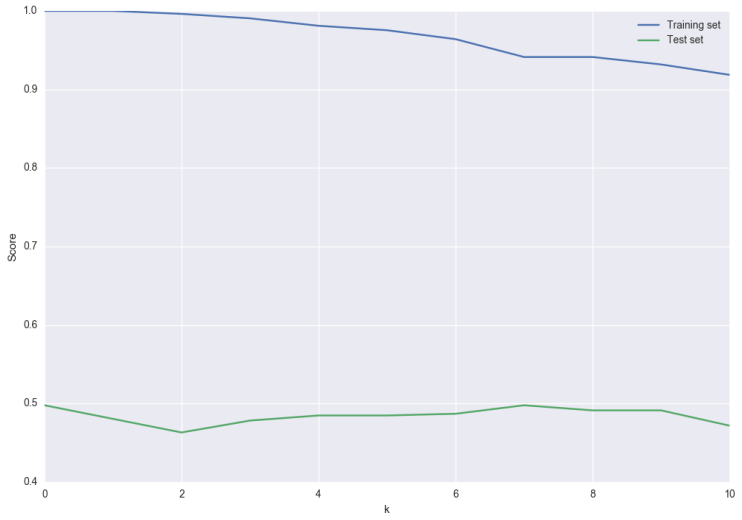


Figura 3: Gráfica que muestra las dos componentes principales de LDA diferenciando con distintos colores las 9 clases.

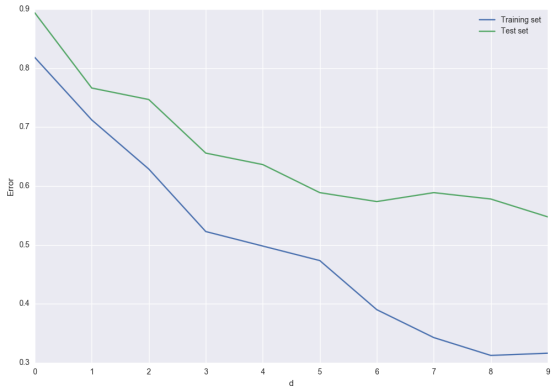
2. Análisis de Opiniones sobre Películas

En esta sección se realiza análisis de sentimiento sobre películas de una Dataset publicado en Kaggle. El dataset es una colección de opiniones de películas etiquetadas con +1 y 0 si la opinión es positiva o negativa respectivamente. Datos de Entrenamiento: 3554 registros. Datos de Test: 3554 registros. El dataset tiene 2 dimensiones. El conjunto de datos tiene 2 clases. De la data de entrenamiento 1784 registros tienen una opinión negativa y 1770 opinión positiva. En el caso de la data de test, 1803 son opiniones positivas y 1751 opiniones negativas. El dataset es tratado convirtiendo todo el texto a minúsculas, eliminando signos de puntuación y palabras sin significado como artículos, pronombres y proposiciones. Además existen dos técnicas que permiten convertir las palabras a su tronco léxico, la lematización y el stemming, técnicas similares, pero con diferentes resultados, a continuación se muestran ejemplos donde se usan ambas técnicas sobre ciertos textos, para poder entender la diferencia entre ambas.

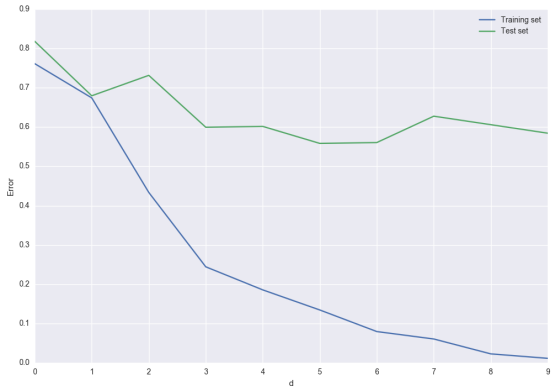
Stemming / sin tratamiento / lematice:
love eat cake / I love to eat cake / love eat cake
love eat cake / I love eating cake / love eating cake
love eat cake / I loved eating the cake / loved eating cake
love eat cake / I do not love eating cake / love eating cake
n't love eat cake / I don't love eating cake / n't love eating cake

Luego se cortan las frases y se realiza un conteo de las palabras contenidas, al ver que palabras son las más recurrentes para los diccionarios formados para el set de entrenamiento y el de prueba: Para el set de entrenamiento se tienen: charact, comedi, director, doe, even, feel, film, get, good, ha, hi, like, look, make, movi, much, one, perform, stori, thi, time, way, well, work.

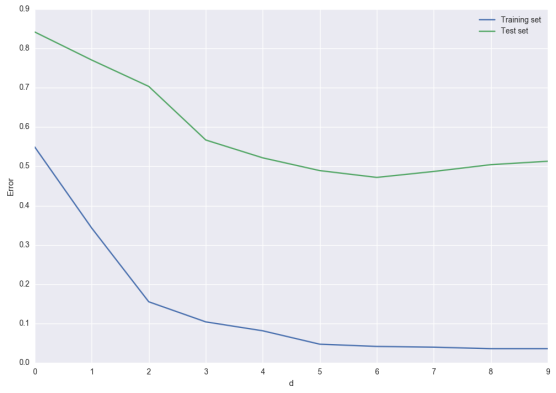
Para el set de prueba: charact, comedi, director, doe, even, film, good, ha, hi, like, make, movi, much, one, perform, stori, thi, time, well, work.



(a) LDA

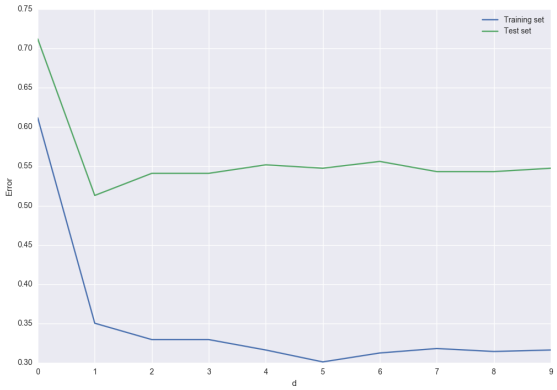


(b) QDA

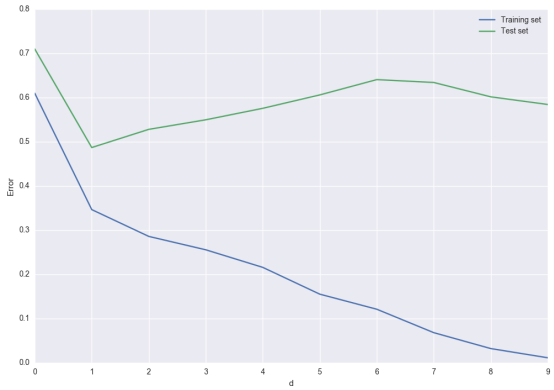


(c) KNN (k=7)

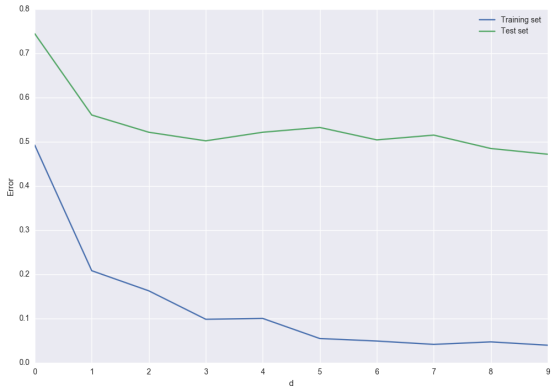
Figura 4: Gráficas de los errores de clasificación haciendo reducción de dimensionalidad con PCA con d dimensiones.



(a) LDA



(b) QDA



(c) KNN (k=7)

Figura 5: Gráficas de los errores de clasificación haciendo reducción de dimensionalidad con LDA con d dimensiones.

Luego se realiza análisis de los resultados entregados por cuatro clasificadores distintos, con sus respectivos accuracy:
Clasificador Bayesiano Ingenuo Binario:

- Stemming: Training accuracy=0.942881; Test accuracy= 0.747819
- Lematización: Training accuracy=0.877603; Test accuracy=0.681677

Clasificador Bayesiano Multinomial:

- Stemming: Training accuracy=0.942319; Test accuracy=0.749789
- Lematización: Training accuracy=0.879291; Test accuracy=0.685055

Regresión Logística Regularizada:

- Stemming:
  - C=0.01: Training accuracy= 0.782217; Test accuracy=0.690684
  - C=0.1: Training accuracy=0.880135; Test accuracy=0.731213
  - C=10: Training accuracy=0.999719; Test accuracy=0.725303
  - C=100: Training accuracy=1; Test accuracy=0.719111
  - C=1000: Training accuracy=1; Test accuracy=0.711793
- Lematización:
  - C=0.01: Training accuracy=0.720597; Test accuracy=0.638615
  - C=0.1: Training accuracy=0.806978; Test accuracy=0.666479
  - C=10: Training accuracy=0.964547; Test accuracy=0.649310
  - C=100: Training accuracy=0.984806; Test accuracy=0.623980
  - C=1000: Training accuracy=0.989589; Test accuracy=0.614692

SVM

- Stemming:
  - C=0.01: Training accuracy=0.873382; Test accuracy=0.729243
  - C=0.1: Training accuracy=0.981992; Test accuracy=0.731213
  - C=10: Training accuracy=1; Test accuracy=0.701942
  - C=100: Training accuracy=1; Test accuracy=0.700535
  - C=1000: Training accuracy=1; Test accuracy=0.700535
- Lematización:
  - C=0.01: Training accuracy=0.801913; Test accuracy=0.663946
  - C=0.1: Training accuracy=0.907991; Test accuracy=0.666198
  - C=10: Training accuracy=0.984524; Test accuracy=0.610751
  - C=100: Training accuracy=0.987057; Test accuracy=0.606248
  - C=1000: Training accuracy=0.983118; Test accuracy=0.603152

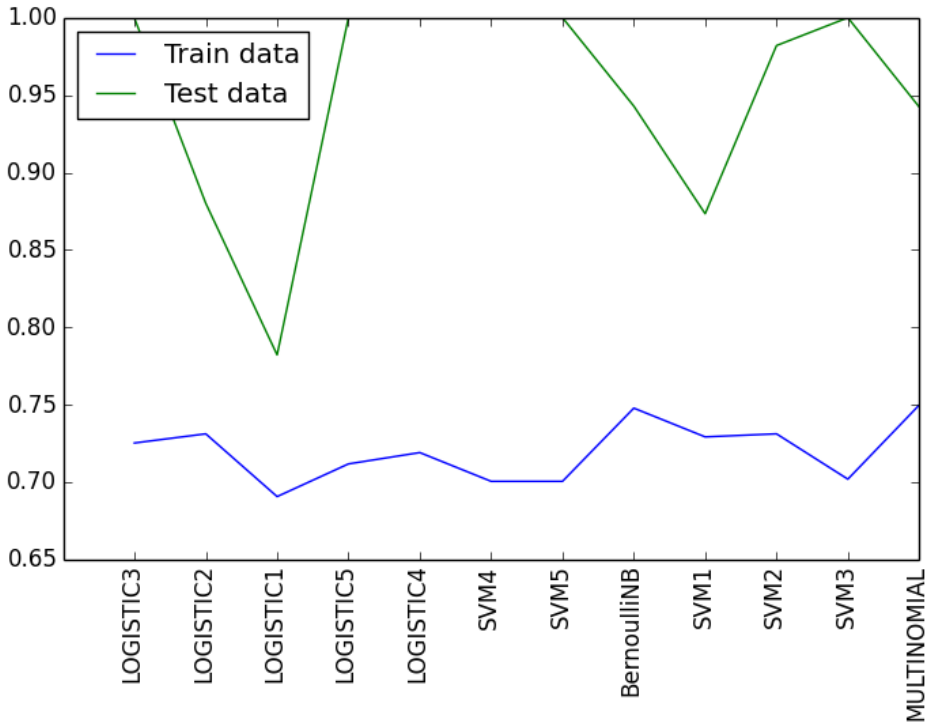


Figura 6: Gráfica de accuracy de distintos clasificadores para data de test y de entrenamiento, con data preprocesada con técnica stemming.

Claramente, se aprecia que sin importar el clasificador utilizado, para nuestro dataset, los mejores resultados se obtienen preprocesando la data con la técnica de stemming, además, se tiene que los clasificadores que arrojaron mejores resultados son los Bayesianos, obteniéndose resultados muy similares entre ambos clasificadores (Multinomial y Binario). Adempas cabe destacar que al usar StopWords, se elimanron palabras claves como Not, las cuales cambian completamente el sentido de una opinión.
A continuación, se muestran textos aleatorios de la data de prueba para ambos métodos Bayesianos, y la probabilidad de que sea una opinión negativa o positiva respectivamente:

- Clasificador Bayesiano Ingenuo Binario:
  - [0.98958049 0.01041951] each scene wreaks of routine ; the film never manages to generate a single threat of suspense .
  - [0.9561989 0.0438011] the sort of movie that gives tastelessness a bad rap .
  - [0.81647136 0.18352864] i don't feel the least bit ashamed in admitting that my enjoyment came at the expense of seeing justice served , even if it's a dish that's best served cold .
  - [0.1751346 0.8248654] the film can depress you about life itself .
- Clasificador Bayesiano Multinomial:

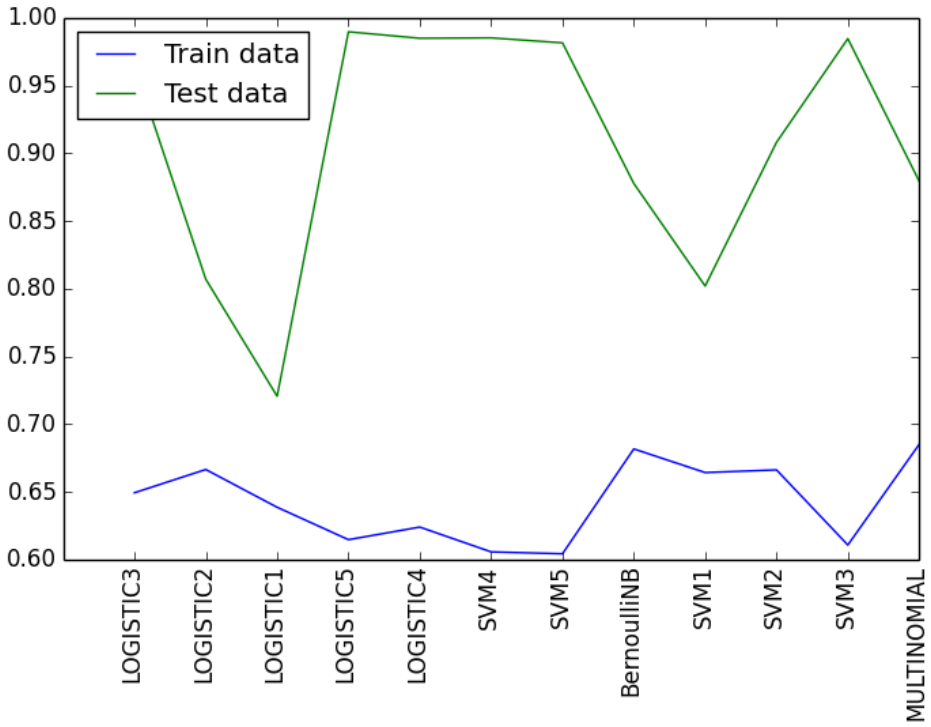


Figura 7: Gráfica de accuracy de distintos clasificadores para data de test y de entrenamiento, con data preprocesada con técnica lematización.

- [0.29397133 0.70602867] this mistaken-identity picture is so film-culture referential that the final product is a ghost .
- [0.6821908 0.3178092] . . . standard guns versus martial arts cliche with little new added .
- [0.50196961 0.49803039] priggish , lethargically paced parable of renewal .
- [0.64955789 0.35044211] brainy , artistic and muted , almost to the point of suffocation .

Los resultados son bastante buenos, pero claramente, es difícil determinar cuando una persona utiliza sarcasmo, o la neutralidad de una opinión, por lo que quizás sería bueno probar no solo con positivo y negativo, si no que generar un rango más amplio de opiniones, agregando al menos la clase neutra.