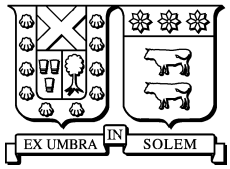
Análisis Inteligente de Datos

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 1, 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 1 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.

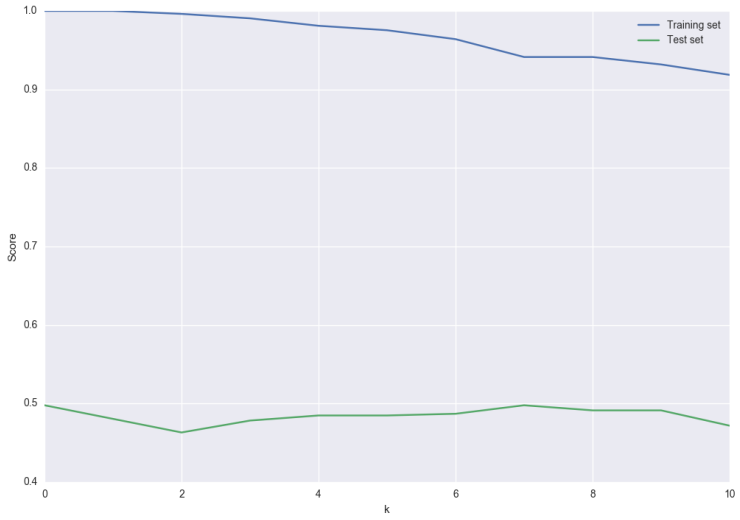
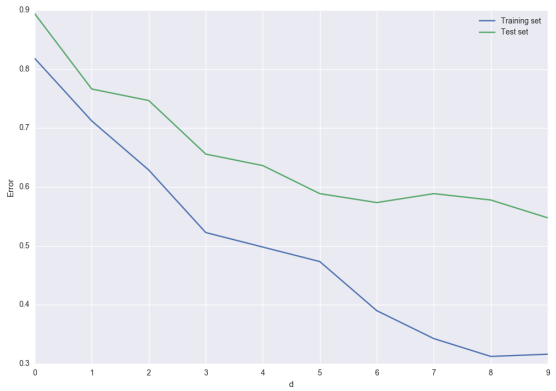


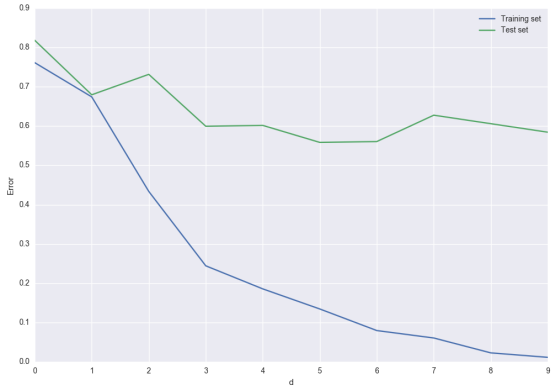
Figura 3: Gráfica que muestra las dos componentes principales de LDA diferenciando con distintos colores las 9 clases.

2. Análisis de Opiniones sobre Películas

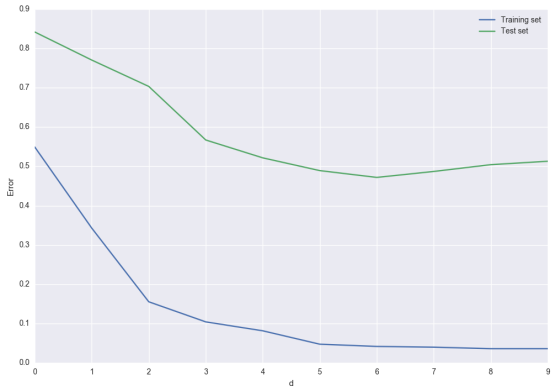
PIPE



(a) LDA

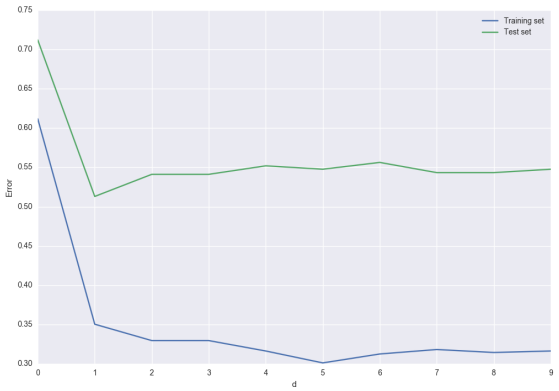


(b) QDA

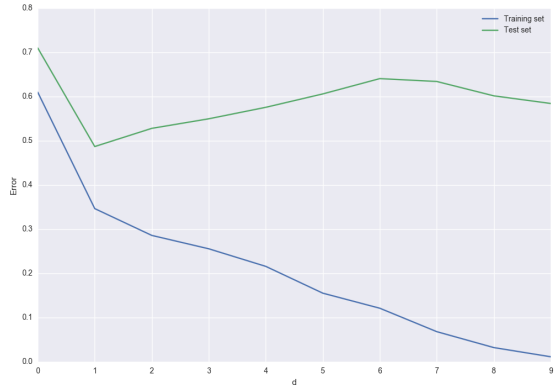


(c) KNN (k=7)

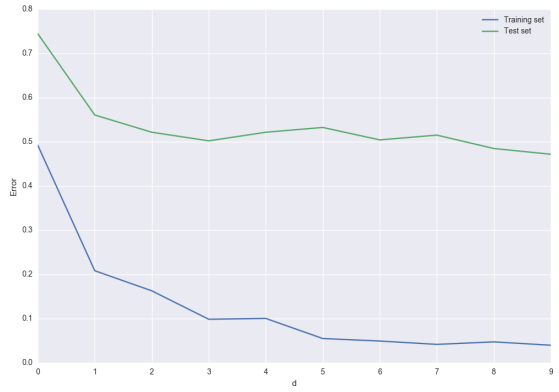
Figura 4: Gráficas de los errores de clasificación haciendo reducción de dimensionalidad con PCA con d dimensiones.



(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.