**Mayo 2015**

# Trabajo Final de Práctica

## Seminario OS14 - Pattern Recognition

Angel Cancio [1*],

**Resumen**
Con este trabajo práctico se intenta estudiar el desempeño de distintas familias de clasificadores en función de la dimensión del espacio de características $\chi$ y el tamaño $n$ del conjunto $An = (x_1, y_1), ..., (x_n, y_n)$ de datos de entrenamiento disponible para estimar la regla de decisión.

[1] *ITeDA, Mendoza, Argentina*
*__Contacto__: angel.cancio@gmail.com

## Indice

## Actividades previas

En función de lo trabajado en clases, para desarrollar la actividad propuesta será necesario haber completado y así disponer de códigos computacionales que implementen los siguientes procedimientos:

- Clasificador de $K$ vecinos más cercanos, con posibilidad de variar el valor de $K$.
- Clasificador por estimación de densidades de tipo Parzen, con posibilidad de variar el parámetro característico del kernel usado.
- Clasificador suponiendo que las distribuciones de probabilidad de ambas clases son Gaussianas (discriminante lineal y cuadrático)
- Clasificador por máquinas de vector soporte (SVM), con posibilidad de variar el parámetro característico del kernel usado y el valor de C.
- Procedimiento de extracción de características usando KPCA.
- Procedimiento general de validación cruzada para estimar errores de predicción.

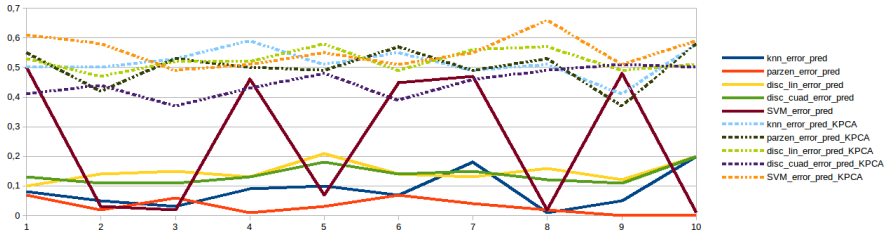
## Actividades a desarrollar

El archivo `datosOS14.mat` contiene datos correspondientes a un problema de clasificación de vocales. Se han extraído solamente dos vocales, para trabajar con clasificación binaria.

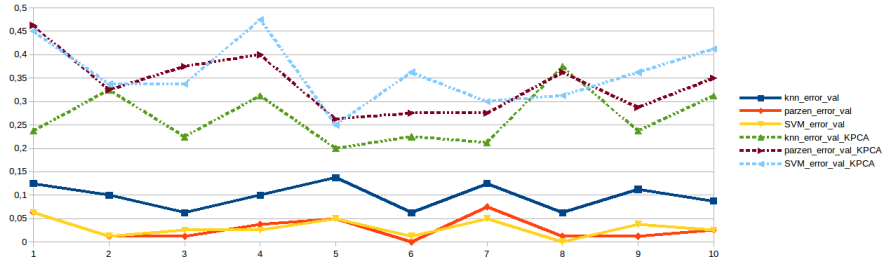
1. Extraiga al azar un subconjunto de entrenamiento de $n_{train} = 80$ datos en total. Reserve el resto de las muestras para estimar el error de predicción. Utilice el conjunto de entrenamiento obtenido para entrenar un clasificador automático de las vocales del problema, usando las siguientes estrategias:
   - k-vecinos más cercanos
   - estimación de densidades tipo Parzen
   - discriminante lineal y cuadrático
   - SVC basado en kernels
   - $KPCA + k-$vecinos más cercanos
   - $KPCA+$estimación de densidades tipo Parzen
2. Repita el procedimiento anterior 10 veces. Reporte el error de predicción promedio para cada tipo de clasificador, usando los resultados de las 10 repeticiones. Reporte también el desvío estándar de los resultados. ¿Encontró mucha variabilidad entre las distintas repeticiones en los valores de los parámetros óptimos de los clasificadores (cantidad de vecinos, ancho de banda de kernels, etc)? Comente.
3. Repita los puntos 1) y 2) para $n_{train} = 120$, 150, $n$, con $n$ la cantidad total de datos. Extraiga conclusiones.
4. Escriba un reporte de no más de dos páginas de extensión con sus conclusiones. Adjunte los códigos de cómputo utilizados.

## Desarrollo

En el gráfico siguiente se puede observar la estimacion del error cometido durante las 10 repeticiones para cada clasificador, antes y despues de utilizar KPCA sobre los datos.



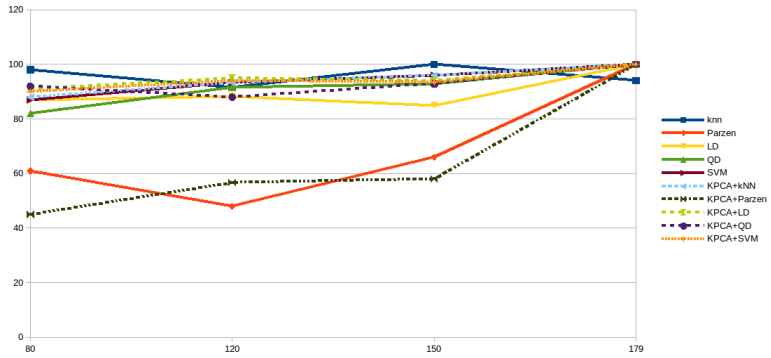

En el gráfico siguiente se puede observar la variacion del error error cometido durante la validacion cruzada para las 10 repeticiones para cada clasificador, antes y despues de utilizar KPCA sobre los datos.




El siguiente gráfico muestra como mejora la clasificación de los datos conforme se aumenta el tamaño de la muestra.

También se observa que KNN con K=1 es invariante respecto al tamaño de la muestra, pero en terminos de rendimiento computacional KNN se encarece al aumentar la dimensionalidad.

El clasificador de ventana de parzen se encuetra fectado por la independencia de las dos clases.




## Concluciones

Despues de haber probado los distintos clasificadores (kNN, Parzen, Discriminador Lineal, Discriminador Cuadratico) sin aplicar la transformacion por analisis de componente principales y luego aplicando la transformacion utilizando analisis de componente principales (PCA y KPCA), y se llega a la conclucion de que es importante tener en cuenta tres puntos: 1) Es necesario realizar una normalizacion y estandarización de los datos antes de realizar la clacificación (preprocesamiento), 2) Es muy importante realizar una validación cruzada para ajustar los parámetros del modelo del clasificador ya que hay una importante variación según el set de entrenamiento utilizado. 3) Y por último el uso de transformaciones de espacio como PCA, KPCA o inclusive los discriminadores de Fisher LDA y QDA permiten reducir la dimencionalidad de los datos y eventualmente encontrar una proyección óptima para poder visualizarlos.

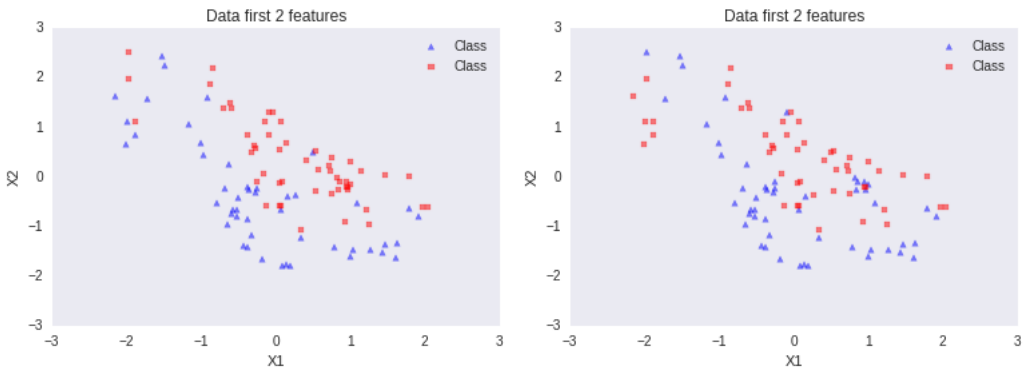
Con las transformaciones espaciales: se provee conjunto relevante de características al clasificador (mejora de desempeño particularmente en clasificadores simples). Se reduce la redundancia y se recuperan las características latentes más significativas de cada característica. Se genera mayor comprensión del proceso de generación de los datos.

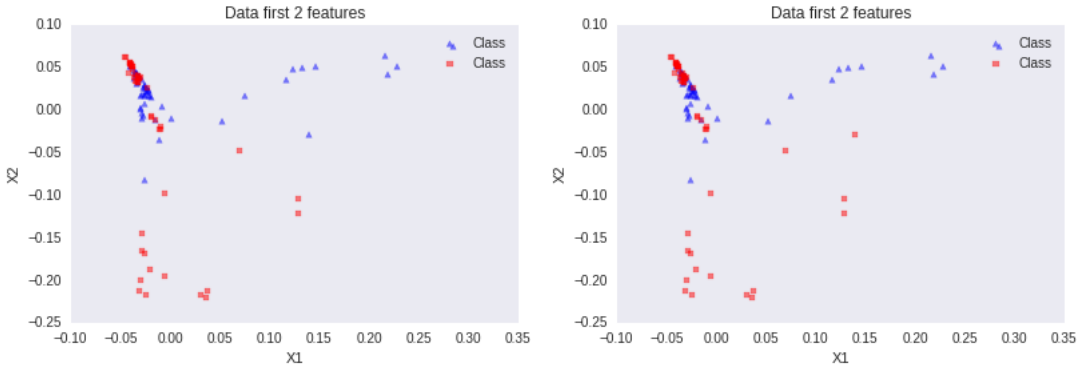
## Apendice 1

**Figure 1.** Proyeccion de datos originales usando los primeros dos features


**Figure 2.** Proyección despues de aplicar KPCA a los datos originales (primeros dos features).


En la imagen se puede ver la clasificcacion usando Discriminante Lineal (izquierda) y Dirciminante Cuadrático (derecha) antes y despues de aplicar KPCA.






Por ultimo la representación de los datos en una matriz de confución muestra la superpocision de las distribuciones entre todas las caracteristicas de las N observaciones.