Trabajo Final de Práctica

Seminario OS14

Mayo de 2015

Con este trabajo práctico se intenta estudiar el desempeño de distintas familias de clasificadores en función de la dimensión del espacio de características $\mathcal{X}$ y el tamaño n del conjunto $\mathcal{A}_n = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ de datos de entrenamiento disponible para estimar la regla de decisión.

1. 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 d eprobabilidad 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.

2. 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.

a) 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

- KPCA+discriminante lineal y cuadrático
- KPCA+SVC basado en kernels

Cuando incluya KPCA, utilice dos proyecciones. Cuando sea necesario, encuentre el mejor clasificador en cada familia usando validación cruzada. Estime el error de predicción.

b) 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.

c) Repita los puntos a) y b) para $n_{train} = 120, 150, n$, con n la cantidad total de datos. Extraiga conclusiones.

d) Escriba un reporte de no más de dos páginas de extensión con sus conclusiones. Adjunte los códigos de cómputo utilizados.