Aprendizaje semi-supervisado

Alejandro Cholaquidis[a], Ricardo Fraiman[a] and Mariela Sued [b]
[a] CABIDA and Centro de Matemática,
Facultad de Ciencias, Universidad de la República, Uruguay
[b] Instituto de Cálculo,
Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires

El problema clásico de clasificación tiene como objetivo asignar una etiqueta a un nuevo dato a partir de una muestra $\mathcal{D}^n = (\mathcal{X}^n, \mathcal{Y}^n) = (X^1, Y^1), \ldots, (X^n, Y^n)$ de entrenamiento. Típicamente se asume $\mathcal{D}^n$ iid, y se prueban resultados de consistencia cuando $n \to \infty$. En el contexto de aprendizaje semi-supervisado, la muestra de entrenamiento es pequeña, y se tiene una enorme cantidad, $l \gg n$, de datos para clasificar, $\mathcal{X}_l = X_1, \ldots, X_l$. Esto sucede por ejemplo cuando obtener una muestra ya clasificada es costoso, y sin embargo obtener datos no etiquetados es fácil. El objetivo es usar (si es posible) la enorme cantidad de datos no clasificados, para construir un clasificador que sea mejor (se equivoque menos) que el que se puede construir con la muestra inicial $\mathcal{D}^n$ ya etiquetada. Intuitivamente, $\mathcal{X}_l$ será de ayuda si conocer la distribución de las X aporta información a la clasificación. Si bien no es cierto en general, veremos que bajo ciertas hipótesis si. En la charla propondremos un algoritmo que permite clasificar secuencialmente la muestra $\mathcal{X}_l$, y que asintoticamente (cuando $l \to \infty$ y n es fijo), se comporta como la mejor regla (teórica) posible. Si bien esto requiere imponer hipotesis fuertes sobre la distribución de las X, veremos que las mismas son necesarias, por la dificultad intrínseca del problema.