

Machine Learning

Tarea #2

Entrega: 17 de Octubre de 2014, 11:59 p.m.

1. Para este problema utilice el conjunto de datos de la tarea 1. Divida el conjunto de datos apropiadamente en una parte para entrenamiento y una parte para validación.
 - a) Implemente (en MATLAB, o en su lenguaje favorito de programación) el algoritmo AdaBoost utilizando como clasificadores base el conjunto de funciones indicadoras¹ de semiplanos alineados con los ejes (a.k.a stumps). Estas son reglas de clasificación muy simples que asignan la etiqueta basado en si una característica del vector de entrada super o no un umbral. Es decir, si $\mathcal{S} \subset \mathbb{R}^d$ y $\mathbf{x} \in \mathcal{S}$, $\mathbf{x} = [x_1, \dots, x_d]$, el conjunto de clasificadores base es:

$$\mathcal{H} = \{I_{\{\mathbf{x} \in \mathcal{S}: x_i \leq C\}}, C \in \mathbb{R}, i = 1, \dots, d\} \cup \{I_{\{\mathbf{x} \in \mathcal{S}: x_i \geq C\}}, C \in \mathbb{R}, i = 1, \dots, d\}$$

- b) Corra por lo menos 500 iteraciones de AdaBoost² (usando los datos de entrenamiento). Grafique error en el conjunto de entrenamiento y error en el conjunto de validación contra número de clasificadores base. Comente sus resultados.
 - c) Grafique la distribución acumulativa de márgenes para 5, 50, 100, 200 y 500 iteraciones de AdaBoost³. Analice los resultados obtenidos.
 - d) Escriba un reporte detallado con sus resultados. Incluya su código en el apéndice.
 - e) Envíe con su tarea las etiquetas obtenidas en el conjunto de prueba `xtest` para el clasificador combinado obtenido con 50, 100, 200 y 500 iteraciones de Adaboost.

¹Escalice estas funciones para que tomen valores en $\{-1, 1\}$.

²Note que Adaboost asume que el clasificador débil retorna una hipótesis que minimiza el *error pesado* con respecto a los valores actuales de los pesos D_i (y no simplemente el error en los datos).

³Recuerde que el margen está definido como $m = yf(x)$ donde $f(x)$ es la salida del clasificador combinado antes de aplicar el limitador duro. Para calcular la distribución acumulativa, normalice los coeficientes del clasificador combinado a 1 y calcule el porcentaje de márgenes menores a δ para $\delta \in [-1, 1]$.