CS57800 Statistical Machine Learning

Homework 2 Solutions

1 Foundations

1. Given n boolean variables $(x_1, \dots, x_n)$, we define our target classification function $f(.)$ as $f(.) = 1$ if at least 2 variables are active. For $n = 4$ show how this function can be represented as (1) Boolean function (2) Linear function.

   Solution:

   (1) Boolean function $= (x_1 \wedge x_2) \vee (x_1 \wedge x_3) \vee (x_1 \wedge x_4) \vee (x_2 \wedge x_3) \vee (x_2 \wedge x_4) \vee (x_3 \wedge x_4)$

   (2)

   $$f(\mathbf{X}) = \begin{cases} 1, \text{if } \sum\limits_{i=1}^{4} x_i \geq 2; \\ 0, \text{otherwise.} \end{cases}$$

2. Let CON_B be the set of all different monotone conjunctions defined over n boolean variables. What is the size of CON_B ?

   Solution:

   Since CON_B is monotone, it's size is 2^n .

3. Suppose there are N points x_i in $\mathcal{R}^p$, with labels $y_i \in \{-1, 1\}$. Denote a hyperplane by a function $f(x) = \beta_1' x + \beta_0 = 0$, or $\beta' x^* = 0$, where $x^* = (x, 1)$ and $\beta = (\beta_1, \beta_0)$. Let $u_i = x_i^* / \|x_i^*\|$ and $y_i \beta^{*'} u_i \geq 1, \forall i$ (β^* is the final separating parameter vecter). Given a current β_o, the perceptron algorithm identifies a point u_i that is misclassified, and produces the update $\beta_n \leftarrow \beta_o + y_i u_i$. Show that

   $$\|\beta_n - \beta^*\|^2 \leq \|\beta_o - \beta^*\|^2 - 1,$$

   and hence that the perceptron converges to a separating hyperplane.

   Solution:

   From $\beta_n = \beta_o + y_i u_i$, we have that

   $$\beta_n - \beta^* = \beta_o - \beta^* + y_i u_i.$$

   We take the squared norm of both sides and we got

   $$\|\beta_n - \beta^*\|^2 = \|\beta_o - \beta^*\|^2 + y_i^2 \|u_i\| + 2y_i \langle (\beta_o - \beta^*), u_i \rangle.$$

Since $y_i = \pm 1$ and $\|u_i\|^2 = 1$ we have that $y_i^2 \|u_i\|^2 = 1$ for the second term on the right-hand-side. Since the "point" y_i, u_i was misclassified by the vector β_o we have $y_i \beta_o u_i < 0$. And since β^* is the final separating parameter vector we have $y_i \beta^* u_i > 1$. Therefore,

$$2y_i\langle(\beta_o - \beta^*), u_i\rangle \leq 2(0-2) = -2.$$

Thus we have just shown that

$$\|\beta_n - \beta^*\|^2 \leq \|\beta_o - \beta^*\|^2 + 1 - 2 = \|\beta_o - \beta^*\|^2 - 1.$$

4. Suggest a mistake bound algorithm for learning Boolean conjunctions (*hint: recall the elimination algorithm for monotone conjunctions*). Show that your algorithm is a mistake bound algorithm for Boolean conjunctions.

   Solution:

   Boolean conjunctions are of the form $\langle(x_1, x_2, \ldots, x_n, \neg x_1, \neg x_2, \ldots, \neg x_n), y\rangle$ where $y = 1$ or 0. Initialize hypothesis is $h = x_1 \wedge \neg x_1 \wedge \cdots \wedge x_n \wedge \neg x_n$. One elimination algorithm could be: for all examples in X, eliminate literals that are not active and keeps eliminating until a sufficient target function is learned. For every mistake, we remove at least one unnecessary literal from the conjunctions. Since we have $2n$ literals at the beginning, and at least one literal in the conjunctions will be remained in the end, the total number of mistakes is at most $2n1$, which is $O(n)$. Since this is polynomial in the size of the hypothesis space, this is a mistake bound algorithm.

5. Given a linearly separable dataset consisting of 50000 positive examples and 50000 negative examples, we train two linear classifier using the perceptron algorithm. We provide the first classifier with a sorted dataset in which all the positive examples appear first, and then the negative examples appear. The second classifier is trained by randomly selecting examples at each training iteration. (1) Will both classifiers converge? (2) what will be the training error of each one of the classifiers?

   Solution:

   (1) By the perceptron convergence theorem, both classifiers should converge if given enough iterations. The first classifier might converge quicker than the second, but both should converge.

   (2) Assuming we run there is a sufficient number of iterations, both classifiers should have 0% training error since the data is linearly separable.

6. Let N denote the set of iterations at which the Perceptron algorithm makes an update when it sees a sequence of training instances $x_1, ..., x_n \in \mathcal{R}^p$. Also, we have a condition that the initial weight vector is a $\overrightarrow{\mathbf{0}}$ vector. Then, the following inequality holds:

$$\|\sum_{i \in N} y_i x_i\| \leq \sqrt{\sum_{i \in N} \|x_i\|^2}$$

   Try to prove this.

Solution:

$$\begin{aligned}
\| \sum_{i \in N} y_i x_i \| &= \| \sum_{i \in N} (w_{i+1} - w_i) \| \\
&= \| w_{n+1} \| \\
&= \sqrt{\sum_{i \in N} \| w_{i+1} \|^2 - \| w_i \|^2} \\
&= \sqrt{\sum_{i \in N} \| w_i + x_i y_i \|^2 - \| w_i \|^2} \\
&= \sqrt{\sum_{i \in N} \underbrace{2 x_i y_i}_{\leq 0} + \| x_i \|^2} \\
&\leq \sqrt{\sum_{i \in N} \| x_i \|^2}
\end{aligned}$$