Домашнее задание по материалу 1-го семинара. ММП, весна 2013 21 февраля

В лекциях был описан алгоритм AnyBoost. Предположим, что мы используем дифференцируемую, строго убывающую и выпуклую верхнюю оценку пороговой функции потерь $[yh(x) \leqslant 0] \leqslant L\big(yh(x)\big)$. Напомним, что на t-й итерации сначала ищется новый базовый классификатор

$$h_t = \arg\min_h - \sum_{i=1}^{\ell} \tilde{w}_i^{(t)} y_i h(x_i),$$

где

$$\tilde{w}_i^{(t)} = \frac{w_i^{(t)}}{\sum_{j=1}^{\ell} w_j^{(t)}}; \quad w_i^{(t)} = - \left.\frac{\partial L(z)}{\partial z}\right|_{z=y_i F_{t-1}(x_i)}; \quad F_T(x) = \sum_{t=1}^{T} \alpha_t h_t(x),$$

а затем находится значение веса α_t только что выбранного базового классификатора в виде решения следующей задачи одномерной оптимизации:

$$\alpha_t = \arg\min_\alpha \frac{1}{\ell} \sum_{i=1}^{\ell} L\big(y_i F_{t-1}(x_i) + y_i \alpha h_t(x_i)\big). \tag{1}$$

Задача 1. Докажите, что в случае $h_t \colon \mathbb{X} \to \{-1, +1\}$ и $L(z) = e^{-z}$ описанные шаги алгоритма AnyBoost в точности переходят в алгоритм AdaBoost.

Задача 2. Что вы можете сказать об оптимизируемой функции (1)? Является ли она выпуклой? Вогнутой? Имеет ли она единственный минимум? Достижим ли он? Если да — то в каких случаях?

Задача 3. Рассмотрим алгоритм AnyBoost с произвольной функцией потерь (удовлетворяющей описанным выше свойствам). Чему равна взвешенная ошибка добавленного на t-ой итерации базового классификатора h_t относительно обновленных весов $\tilde{\boldsymbol{w}}^{(t+1)}$?

Определим класс базовых классификаторов, которые мы будем называть «решающими пнями» (decision stumps):

$$h(x) = h(x|i,t) = \begin{cases} 1, & \text{если } x^i < t; \\ -1 & \text{иначе.} \end{cases}$$

Таким образом у таких классификаторов два параметра: номер признака i и пороговое значение этого признака t. Они разбивают пространство на два полупространства гиперплоскостью, ортогональной i-й оси координат.

Взвешенной композицией функций из $\mathcal{H} = \{h\colon \mathbb{X} \to \mathbb{R}\}$ мы будем называть классификатор

$$a(x) = \operatorname{sgn}\left(\sum_{t=1}^{T} \alpha_t h_t(x)\right),$$

где $\alpha_t \in \mathbb{R}$, $h_t \in \mathcal{H}$. В этой композиции корректирующей операцией выступает взвешенная сумма ответов классификаторов, а решающим правилом — функция sgn.

Задача 4. Рассмотрим одномерную обучающую выборку $X^\ell = \{x_1, \ldots, x_\ell\} \subset \mathbb{R}$.
а) Покажите, что при любом векторе ответов $\boldsymbol{y} = \{-1, +1\}^\ell$ обучающей выборки, существует взвешенная композиция из 2ℓ различных решающих пней, не допускающая ошибок на обучающей выборке.
б) Каково в общем случае минимальное число T, такое что для *произвольной* одномерной обучающей выборки из ℓ объектов найдется взвешенная композиция из T решающих пней, не ошибающаяся на этой выборке?

Задача 5. Рассмотрим обучающую выборку $X^\ell = \{x_1, \ldots, x_\ell\} \subset \mathbb{R}^n$, состоящую из ℓ различных объектов. Рассмотрим следующие базовые функции (которые также называются *радиальными*):

$$h_\beta(x|x_i, y_i) = y_i \exp\{-\beta\|x - x_i\|^2\}.$$

Параметром такой базовой функции является объект обучающей выборки x_i с его меткой класса $y_i \in \{-1, +1\}$ а $\beta > 0$ — настраиваемый *гиперпараметр* модели. Покажите, что при любом векторе ответов $\boldsymbol{y} = \{-1, +1\}^\ell$ обучающей выборки, существует взвешенная композиция радиальных функций, которая не допускает ошибок на объектах обучающей выборки.

Задача 6. Попробуйте вывести алгоритм AdaBoost для случая многих классов $\mathbb{R} = \{1, \ldots, K\}$. Настраиваемая вами композиция будет иметь вид

$$a(x) = \arg\max_{k=1,\ldots,K} F_T^k(x) = \arg\max_{k=1,\ldots,K} \sum_{t=1}^{T} \alpha_t [h_t(x) = k].$$

Вам предстоит справиться со сложностью, связанной с заменой «простого» в обращении решающего правила $\operatorname{sgn}(\cdot)$ на $\arg\max$.