Теоретическое домашнее задание №4

Задача 1. Пусть даны выборка X, состоящая из 8 объектов, и классификатор $b(x)$, предсказывающий оценку принадлежности объекта положительному классу. Предсказания $b(x)$ и реальные метки объектов приведены ниже:

$$
\begin{aligned}
&b(x_1) = 0.1, \quad y_1 = +1,\\
&b(x_2) = 0.8, \quad y_2 = +1,\\
&b(x_3) = 0.2, \quad y_3 = -1,\\
&b(x_4) = 0.25, \quad y_4 = -1,\\
&b(x_5) = 0.9, \quad y_5 = +1,\\
&b(x_6) = 0.3, \quad y_6 = +1,\\
&b(x_7) = 0.6, \quad y_7 = -1,\\
&b(x_8) = 0.95, \quad y_8 = +1.
\end{aligned}
$$

Постройте ROC-кривую и вычислите AUC-ROC для множества классификаторов $a(x; t)$, порожденных $b(x)$, на выборке X.

Задача 2. Пусть дан классификатор $b(x)$, который возвращает оценку принадлежности объекта x классу $+1$. Отсортируем все объекты по неубыванию ответа классификатора b: $x_{(1)}, \dots, x_{(\ell)}$. Обозначим истинные ответы на этих объектах через $y_{(1)}, \dots, y_{(\ell)}$.

Покажите, что AUC-ROC для данной выборки будет равен вероятности того, что случайно выбранный положительный объект окажется в отсортированном списке не раньше случайно выбранного отрицательного объекта.

Задача 3. Позволяет ли предсказывать корректные вероятности экспоненциальная функция потерь $L(y, z) = \exp(-yz)$?

Задача 4. Рассмотрим постановку оптимизационной задачи метода опорных векторов для линейно разделимой выборки:

$$
\begin{cases}
\frac{1}{2}\|w\|^2 \to \min_{w,b},\\
y_i(\langle w, x\rangle + b) \geq 1, \quad i = \overline{1, \ell},
\end{cases}
$$

а также её видоизменёный вариант для некоторого значения $t > 0$:

$$
\begin{cases}
\frac{1}{2}\|w\|^2 \to \min_{w,b},\\
y_i(\langle w, x\rangle + b) \geq t, \quad i = \overline{1, \ell}.
\end{cases}
$$

Покажите, что разделяющие гиперплоскости, получающиеся в результате решения каждой из этих задач, совпадают.

Задача 5. Вычислите градиент $\frac{\partial}{\partial w} L(x, y; w)$ логистической функции потерь для случая линейного классификатора

$$L(x, y; w) = \log(1 + \exp(-y \langle w, x \rangle))$$

и упростите итоговое выражение таким образом, чтобы в нём участвовала сигмоидная функция

$$\sigma(z) = \frac{1}{1 + \exp(-z)}.$$

При решении данной задачи вам может понадобиться следующий факт (убедитесь, что он действительно выполняется):

$$\sigma'(z) = \sigma(z)(1 - \sigma(z)).$$

Задача 6. Ответьте на следующие вопросы:

1. Почему в общем случае распределение $p(y|x)$ для некоторого объекта $x \in \mathbb{X}$ отличается от вырожденного?

2. Каким важным с точки зрения задачи классификации преимуществом обладает логистическая функция потерь?

3. Почему логистическая регрессия позволяет предсказывать корректные вероятности принадлежности объекта классам?

4. Рассмотрим оптимизационную задачу hard-margin SVM. Всегда ли в обучающей выборке существует объект x_i, для которого выполнено $y_i(\langle w, x_i \rangle + b) = 1$? Почему?

5. С какой целью в постановке оптимизационной задачи soft-margin SVM вводятся переменные ξ_i, $i = \overline{1, \ell}$?