№1

$$L = -\frac{1}{n}\sum_{i=1}^{n}\left(y_i \cdot \log \hat{y}_i + (1-y_i)\cdot \log(1-\hat{y}_i)\right)$$

$\hat{y}_i = \sigma(z_i)$ x_i - вектор столбец признаков i-ого объекта

$z_i = x_i^T w$

$\sigma(x) = \frac{1}{1+e^{-x}}$ пусть d признаков

$\sigma'(x) = \sigma(x)(1-\sigma(x))$ я доказал во 2 ДЗ

$z_i = x_{i0}\cdot w_0 + x_{i1}\cdot w_1 + \ldots + x_{id}\cdot w_d$

Считаем, что $x_{i0} = 1$ $\forall i = 1,\ldots n$

$$\frac{\partial L}{\partial w_j} = -\frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i}{\hat{y}_i}\cdot \hat{y}_i(1-\hat{y}_i)\cdot x_{ij} - \frac{(1-y_i)}{1-\hat{y}_i}\cdot(1-\hat{y}_i)\hat{y}_i \cdot x_{ij}\right) =$$

$$= -\frac{1}{n}\sum_{i=1}^{n} x_{ij}(y_i - y_i\cdot\hat{y}_i - \hat{y}_i + y_i\cdot\hat{y}_i) =$$

$$= -\frac{1}{n}\sum_{i=1}^{n} x_{ij}(y_i - \hat{y}_i) \text{ — производная по } j\text{-ому весу}$$

Тогда
$$\frac{\partial L}{\partial w} = -\frac{1}{n}\underbrace{\begin{pmatrix}\sum_{i=1}^{n} x_{i0}(y_i-\hat{y}_i)\\ \sum_{i=1}^{n} x_{i1}\cdot(y_i-\hat{y}_i)\\ \vdots \\ \sum_{i=1}^{n} x_{id}\cdot(y_i-\hat{y}_i)\end{pmatrix}}_{\overrightarrow{\text{вектор}}} = -\frac{1}{n}\sum_{i=1}^{n} x_i (y_i - \hat{y}_i)$$

я можно записать так

Если $Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n\end{pmatrix}$, $\hat{Y} = \begin{pmatrix}\hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_n\end{pmatrix}$ то $\frac{\partial L}{\partial w} = -\frac{1}{n}\cdot X^T\cdot(Y-\hat{Y})$ ← матричный вид

№2 M моделей

$\hat{y}_m(x) = f(x) + \varepsilon_m(x) \Rightarrow \varepsilon_m(x) = \hat{y}_m(x) - f(x)$

$\hat{y}_{bag}(x) = \frac{1}{M}\sum_{m=1}^{M}\hat{y}_m(x)$

ошибка бэггинга:

$E_{bag} = E[(\hat{y}_{bag}(x) - f(x))^2]$

средняя ошибка всех моделей:

$E_{av} = \frac{1}{M}\sum_{m=1}^{M} E[(\hat{y}_m(x) - f(x))^2]$

$\varepsilon_{m_0}\cdot\varepsilon_{m_1} = 0$, $m_0 \neq m_1$, т.к. ошибки независимы

$E[x+y] = E[x] + E[y]$

$$\varepsilon_{bag}(x) = \hat{y}_{bag}(x) - f(x) = \frac{1}{M}\sum_{m=1}^{M}(f(x)+\varepsilon_m(x)) - f(x) = \frac{1}{M}\sum_{m=1}^{M}\varepsilon_m(x)$$

$E_{av} = \frac{1}{M}\sum_{m=1}^{M} E[\varepsilon_m^2(x)]$

Раскроем скобки, зная $\varepsilon_{m_0}\cdot\varepsilon_{m_1} = 0$

$$E_{bag} = E[\varepsilon_{bag}^2(x)] = E\left[\frac{1}{M^2}\left(\sum_{m=1}^{M}\varepsilon_m(x)\right)^2\right] = \frac{1}{M^2}\left(E\left[\sum_{m=1}^{M}\varepsilon_m^2(x)\right]\right) =$$

$$= \frac{1}{M^2}\sum_{m=1}^{M} E[\varepsilon_m^2(x)] = \frac{1}{M}E_{AV} \blacktriangle$$

№3 Используем часть условия из 2 задачи

Пусть L - выпуклая функция ошибки

Тогда по неравенству Йенсена:

$$L(a_1 x_1 + \dots + a_n x_n) \le a_1 L(x_1) + \dots + a_n L(x_n)$$, a_i - положительные числа

В частности $a_i = \frac{1}{n}$ $\qquad \sum_{i=1}^{n} a_i = 1$

$$L\left(\frac{x_1 + \dots + x_n}{n}\right) \le \frac{L(x_1) + \dots + L(x_n)}{n} \qquad \varepsilon_{bag} = \frac{1}{M}\sum_{m=1}^{M} \varepsilon_m(x)$$

Ошибка бэгинга: используем: $f(x) \ge g(x) \Rightarrow E[f(x)] \ge E[g(x)]$

$$E_{bag} = E[L(\varepsilon_{bag}(x))] = E\left[L\left(\frac{\sum_{m=1}^{M} \varepsilon_m(x)}{M}\right)\right] \le E\left[\sum_{m=1}^{M} L(\varepsilon_m(x))\right]\cdot\frac{1}{M} =$$

$$= \frac{1}{M}\sum_{m=1}^{M} E[L(\varepsilon_m(x))] = E_{av}$$ Итого $E_{bag} \le E_{av}$ ▲

№4

$$L(y, \hat{y}_k(x)) = -\frac{1}{n}\sum_{i=1}^{n}\left(y_i \log \hat{y}_k(x_i) + (1-y_i)\log(1-\hat{y}_k(x_i))\right)$$ берем производную по прогнозу

$$S(x_i) = -\frac{\partial L(y, \hat{y}_k(x_i))}{\partial \hat{y}_k} = +\frac{1}{n}\left(\frac{y_i}{\hat{y}_k(x_i)} - \frac{1-y_i}{1-\hat{y}_k(x_i)}\right) = \frac{+1}{n}\left(\frac{y_i - \hat{y}_k(x)}{\hat{y}_k(x_i)(1-\hat{y}_k(x_i))}\right)$$

искомый остаток

- s - вектор остатков для n объектов подвыборки $x^{(k+1)}$
- Мы уже построили модель из k деревьев, теперь надо добавить $k+1$: мы его строим минимизируя функцию потерь MSE $(b_{k+1}(X^{(k+1)}) - S)^T (b_{k+1}(x^{(k+1)}) - S) \to \min_{b_{k+1}}$ (на подвыборке $x^{(k+1)}$)

b_{k+1} - прогноз $k+1$ дерева. • Далее обновляем композицию алгоритмов:

$$\hat{y}_{k+1}(x^{(k+1)}) = \hat{y}_k(x^{(k+1)}) + \gamma_{k+1} \cdot b_{k+1}(x^{(k+1)})$$

Далее подбираем оптимальный $\gamma_{k+1} \Rightarrow$

$\Rightarrow$ ~~$\gamma_{k+1} = \arg\min_{\gamma} \sum_{i=1}$~~, $\gamma_{k+1} = \arg\min_{\gamma_t} L(y; \hat{y}_k(x^{(k+1)}) + \gamma_t b_{k+1}(x^{(k+1)}))$

композиция алгоритмов обновлена, можем добавлять ещё деревья пока не достигнем какого-то критерия останова

Функция для обучения нового дерева:

$$(b_{k+1}(X^{(k+1)}) - S)^T (b_{k+1}(X^{(k+1)}) - S) \to \min_{b_{k+1}}$$ или $$\sum_{i=1}^{n}\left((b_{k+1}(x_i^{(k+1)}) - S_{k+1}(x_i)\right)^2 \to \min_{b_{k+1}}$$

№5 n объектов - выборка

метод бутстрапа с повторением: n раз берем объекты из выборки (можно повторяющиеся) и получаем подвыборку

возьмём какой-то элемент, вероятность, что он не будет в подвыборке: $\left(\frac{n-1}{n}\right)^n \xrightarrow[n\to\infty]{} \frac{1}{e}$ (из Матана-1)

т.к. n большой

(вероятность выбора этого элемента на каждом отдельном шаге равна $\frac{1}{n}$, значит не выбора: $\frac{n-1}{n} = 1-\frac{1}{n}$, ни на одном шаге не выбираем)

Итого, вероятность, что какой-то элемент будет в выборке: $1-\frac{1}{e}$, но т.к. события „выбора элементов" независимы друг от друга, то в среднем в подвыборке будет $n\cdot(1-\frac{1}{e})$ элементов, а процент: $\frac{n\cdot(1-\frac{1}{e})}{n}\cdot 100\% \approx 63\%$ ▲