1. Объяснение целевой переменной t_j

$$t_j = \begin{cases} 1 & \text{если } j = \text{правильный класс (label)} \\ 0 & \text{для всех остальных классов} \end{cases}$$

$$\sum_{j=1}^{K} t_j = 1 \quad \text{(ровно одна единица)}$$

1.1. Пример для 3 классов

Пусть правильный класс - второй ($j = 2$):

$$\mathbf{t} = [t_1, t_2, t_3] = [0, 1, 0]$$

1.2. Связь с функцией потерь

Кросс-энтропия раскладывается как:

$$\begin{aligned} L &= -\sum_{j=1}^{K} t_j \log y_j \\ &= -t_1 \log y_1 - t_2 \log y_2 - t_3 \log y_3 \\ &= -0 \cdot \log y_1 - 1 \cdot \log y_2 - 0 \cdot \log y_3 \\ &= -\log y_2 \end{aligned}$$

1.3. Производная кросс-энтропии

$$\begin{aligned} \frac{\partial L}{\partial y_j} &= -\frac{t_j}{y_j} \\ &= \begin{cases} -\frac{1}{y_j} & \text{если } j = \text{label} \\ 0 & \text{иначе} \end{cases} \end{aligned}$$

2. Производная кросс-энтропии

Исходная функция потерь для одного примера:

$$L = -\sum_{k=1}^{K} t_k \log y_k$$

Где:

$$t_k = \begin{cases} 1 & \text{если } k = \text{label} \\ 0 & \text{иначе} \end{cases}$$

$$\sum_{k=1}^{K} t_k = 1 \quad \text{(так как ровно один правильный класс)}$$

Распишем производную по y_j:

$$\begin{aligned} \frac{\partial L}{\partial y_j} &= \frac{\partial}{\partial y_j} \left(-\sum_{k=1}^{K} t_k \log y_k \right) \\ &= -\sum_{k=1}^{K} t_k \frac{\partial}{\partial y_j} (\log y_k) \end{aligned}$$

Учитывая, что:

$$\frac{\partial}{\partial y_j} (\log y_k) = \begin{cases} \frac{1}{y_j} & \text{при } k = j \\ 0 & \text{при } k \neq j \end{cases}$$

Получаем:

$$\begin{aligned}
\frac{\partial L}{\partial y_j} &= -\sum_{k=1}^{K} t_k \cdot \begin{cases} \frac{1}{y_j} & \text{при } k = j \\ 0 & \text{при } k \neq j \end{cases} \\
&= -t_j \cdot \frac{1}{y_j} - \sum_{k \neq j} t_k \cdot 0 \\
&= -\frac{t_j}{y_j}
\end{aligned}$$

Таким образом:

$$\frac{\partial L}{\partial y_j} = \begin{cases} -\frac{1}{y_j} & \text{если } j = \text{label} \quad (t_j = 1) \\ 0 & \text{иначе} \quad (t_j = 0) \end{cases}$$

3. Полный вывод производной функции softmax

Рассмотрим функцию softmax для класса j:

$$y_j = \frac{e^{z_j}}{\sum_{k=1}^{K} e^{z_k}} = \frac{e^{z_j}}{S}, \quad \text{где } S = \sum_{k=1}^{K} e^{z_k} \tag{1}$$

3.1. Производная $\frac{\partial y_j}{\partial z_j}$ (по своему логиту)

Применяем правило производной частного:

$$\begin{aligned}
\frac{\partial y_j}{\partial z_j} &= \frac{\partial}{\partial z_j} \left(\frac{e^{z_j}}{S} \right) \\
&= \frac{e^{z_j} S - e^{z_j} \frac{\partial S}{\partial z_j}}{S^2} \\
&= \frac{e^{z_j} S - e^{z_j} e^{z_j}}{S^2} \quad \left(\text{так как } \frac{\partial S}{\partial z_j} = e^{z_j}\right) \\
&= \frac{e^{z_j} (S - e^{z_j})}{S^2} \\
&= \frac{e^{z_j}}{S} \cdot \frac{S - e^{z_j}}{S} \\
&= y_j (1 - y_j)
\end{aligned}$$

3.2. Производная $\frac{\partial y_j}{\partial z_k}$ (по чужому логиту, $k \neq j$)

Для $k \neq j$:

$$\begin{aligned}
\frac{\partial y_j}{\partial z_k} &= \frac{\partial}{\partial z_k} \left(\frac{e^{z_j}}{S} \right) \\
&= \frac{0 \cdot S - e^{z_j} \frac{\partial S}{\partial z_k}}{S^2} \\
&= \frac{-e^{z_j} e^{z_k}}{S^2} \quad \left(\text{так как } \frac{\partial S}{\partial z_k} = e^{z_k}\right) \\
&= -\frac{e^{z_j}}{S} \cdot \frac{e^{z_k}}{S} \\
&= -y_j y_k
\end{aligned}$$

3.3. Объединение результатов

Таким образом, полная производная функции softmax:

$$\frac{\partial y_j}{\partial z_k} = \begin{cases} y_j (1 - y_j) & \text{при } j = k \\ -y_j y_k & \text{при } j \neq k \end{cases} \tag{2}$$

3.4. Матричная форма Якобиана

Матрица Якоби для всех $j, k \in \{1, ..., K\}$:

$$\mathbf{J} = \begin{pmatrix} y_1(1-y_1) & -y_1y_2 & \cdots & -y_1y_K \\ -y_2y_1 & y_2(1-y_2) & \cdots & -y_2y_K \\ \vdots & \vdots & \ddots & \vdots \\ -y_Ky_1 & -y_Ky_2 & \cdots & y_K(1-y_K) \end{pmatrix} \tag{3}$$

3.5. Применение к кросс-энтропии

Для функции потерь $L = -\log y_{\text{label}}$:

$$\begin{aligned} \frac{\partial L}{\partial z_k} &= \sum_{j=1}^{K} \frac{\partial L}{\partial y_j} \frac{\partial y_j}{\partial z_k} \\ &= -\frac{1}{y_{\text{label}}} \frac{\partial y_{\text{label}}}{\partial z_k} + \sum_{j \neq \text{label}} 0 \cdot \frac{\partial y_j}{\partial z_k} \end{aligned}$$

Подставляя производные softmax:

$$\frac{\partial L}{\partial z_k} = \begin{cases} -\frac{1}{y_k} \cdot y_k(1-y_k) = y_k - 1 & \text{при } k = \text{label} \\ -\frac{1}{y_{\text{label}}} \cdot (-y_{\text{label}}y_k) = y_k & \text{при } k \neq \text{label} \end{cases} \tag{4}$$

Что можно записать компактно:

$$\frac{\partial L}{\partial z_k} = y_k - t_k, \quad \text{где } t_k = \begin{cases} 1 & \text{при } k = \text{label} \\ 0 & \text{иначе} \end{cases} \tag{5}$$

4. Комбинированная производная по логитам

Применяем цепное правило для вычисления $\frac{\partial L}{\partial z_k}$:

$$\begin{aligned} \frac{\partial L}{\partial z_k} &= \sum_{j=1}^{K} \frac{\partial L}{\partial y_j} \frac{\partial y_j}{\partial z_k} \\ &= \underbrace{\frac{\partial L}{\partial y_k} \frac{\partial y_k}{\partial z_k}}_{\text{Случай } j=k} + \underbrace{\sum_{j \neq k} \frac{\partial L}{\partial y_j} \frac{\partial y_j}{\partial z_k}}_{\text{Случай } j \neq k} \end{aligned}$$

4.1. Случай для $j = k$ (собственный логит)

Из предыдущих вычислений имеем:

$$\frac{\partial L}{\partial y_k} = \begin{cases} -\frac{1}{y_k} & \text{если } k = \text{label} \\ 0 & \text{иначе} \end{cases}$$

$$\frac{\partial y_k}{\partial z_k} = y_k(1-y_k)$$

Таким образом:

$$\begin{aligned} \frac{\partial L}{\partial y_k} \frac{\partial y_k}{\partial z_k} &= \begin{cases} -\frac{1}{y_k} \cdot y_k(1-y_k) & \text{если } k = \text{label} \\ 0 \cdot y_k(1-y_k) & \text{иначе} \end{cases} \\ &= \begin{cases} -(1-y_k) & \text{если } k = \text{label} \\ 0 & \text{иначе} \end{cases} \\ &= y_k - t_k \quad \text{где } t_k = \begin{cases} 1 & \text{при } k = \text{label} \\ 0 & \text{иначе} \end{cases} \end{aligned}$$

4.2. Случай для $j \neq k$ (чужие логиты)

Для $j \neq k$:

$$\frac{\partial L}{\partial y_j} = \begin{cases} -\frac{1}{y_j} & \text{если } j = \text{label} \\ 0 & \text{иначе} \end{cases}$$

$$\frac{\partial y_j}{\partial z_k} = -y_jy_k$$

Следовательно:

$$\sum_{j\neq k} \frac{\partial L}{\partial y_j}\frac{\partial y_j}{\partial z_k} = \sum_{j\neq k} \begin{cases} -\frac{1}{y_j}\cdot(-y_j y_k) & \text{если } j = \text{label} \\ 0\cdot(-y_j y_k) & \text{иначе} \end{cases}$$
$$= \sum_{j\neq k} \begin{cases} y_k & \text{если } j = \text{label} \\ 0 & \text{иначе} \end{cases}$$
$$= \begin{cases} y_k & \text{если label} \in \{j | j \neq k\} \\ 0 & \text{иначе} \end{cases}$$
$$= y_k \cdot (1 - \delta_{k=\text{label}})$$

где $\delta_{k=\text{label}}$ - индикаторная функция Кронекера:

$$\delta_{k=\text{label}} = \begin{cases} 1 & \text{при } k = \text{label} \\ 0 & \text{иначе} \end{cases}$$

4.3. Комбинирование результатов

Объединяя оба случая:

$$\frac{\partial L}{\partial z_k} = \underbrace{(y_k - t_k)}_{\text{Случай } j=k} + \underbrace{y_k(1-\delta_{k=\text{label}})}_{\text{Случай } j\neq k}$$
$$= y_k - t_k + y_k - y_k\delta_{k=\text{label}}$$
$$= 2y_k - t_k - y_k\delta_{k=\text{label}}$$

Однако при более внимательном рассмотрении:

- Если $k = \text{label}$:
$$\frac{\partial L}{\partial z_k} = (y_k - 1) + 0 = y_k - 1$$
- Если $k \neq \text{label}$:
$$\frac{\partial L}{\partial z_k} = 0 + y_k = y_k$$

Таким образом, окончательный результат:

$$\frac{\partial L}{\partial z_k} = y_k - t_k, \quad \text{где } t_k = \begin{cases} 1 & \text{при } k = \text{label} \\ 0 & \text{иначе} \end{cases} \tag{6}$$

4.4. Интерпретация результата

Полученная производная имеет интуитивно понятный вид:

- Если k - правильный класс ($t_k = 1$), то градиент равен $(y_k - 1)$
- Для неправильных классов ($t_k = 0$) градиент равен просто y_k

Это означает, что:

- Если вероятность правильного класса мала ($y_k \ll 1$), градиент будет большим по модулю отрицательным числом
- Для неправильных классов с высокой вероятностью градиент будет большим положительным числом

5. Полный вывод производной по весам

5.1. Производная логита по весам

Рассмотрим выражение для логита z_j:

$$z_j = \sum_{i=1}^{n} W_{ij}x_i + b_j \tag{7}$$

где:

- W_{ij} - вес между входным нейроном i и выходным нейроном j
- x_i - значение i-го входного нейрона
- b_j - смещение для нейрона j

Производная логита по весу:

$$\frac{\partial z_j}{\partial W_{ij}} = \frac{\partial}{\partial W_{ij}}\left(\sum_{k=1}^{n} W_{kj}x_k + b_j\right) = x_i \quad \text{(так как только один член суммы содержит } W_{ij}\text{)} \tag{8}$$

5.2. Применение цепного правила

Используем ранее полученную производную $\frac{\partial L}{\partial z_j} = y_j - t_j$ и применяем цепное правило:

$$\frac{\partial L}{\partial W_{ij}} = \frac{\partial L}{\partial z_j} \cdot \frac{\partial z_j}{\partial W_{ij}} = (y_j - t_j) \cdot x_i \tag{9}$$

В итоге получается:

$$\frac{\partial L}{\partial W_{ij}} = \frac{\partial L}{\partial y_j} \cdot \frac{\partial y_j}{\partial z_j} \cdot \frac{\partial z_j}{\partial W_{ij}} = -\frac{1}{y_j} \cdot y_j(t_j - y_j) \cdot x_i = (y_j - 1) \cdot x_i \tag{10}$$

$$\frac{\partial L}{\partial b_j} = \frac{\partial L}{\partial y_j} \cdot \frac{\partial y_j}{\partial z_j} \cdot \frac{\partial z_j}{\partial b_j} = -\frac{1}{y_j} \cdot y_j(t_j - y_j) \cdot x_i = (y_j - 1) \cdot 1 \tag{11}$$

где $t_j = 1$ при $j = label$

5.3. Учет L2-регуляризации

Функция потерь с L2-регуляризацией:

$$L_{\text{total}} = L_{\text{CE}} + \lambda \sum_{i,j} W_{ij}^2 \tag{12}$$

Производная регуляризационного члена:

$$\frac{\partial}{\partial W_{ij}} \left(\lambda \sum_{k,l} W_{kl}^2 \right) = 2\lambda W_{ij} \tag{13}$$

5.4. Итоговая производная

Комбинируя оба слагаемых:

$$\frac{\partial L_{\text{total}}}{\partial W_{ij}} = \frac{\partial L_{\text{CE}}}{\partial W_{ij}} + \frac{\partial L_{\text{reg}}}{\partial W_{ij}} \tag{14}$$

$$= (y_j - t_j) \cdot x_i + 2\lambda W_{ij} \tag{15}$$

5.5. Интерпретация результата

- Первое слагаемое $(y_j - t_j) \cdot x_i$:
  - Пропорционально ошибке предсказания $(y_j - t_j)$
  - Усиливается входным значением x_i
- Второе слагаемое $2\lambda W_{ij}$:
  - Линейно зависит от величины веса
  - Стремится уменьшить абсолютное значение весов (регуляризация)

5.6. Пример вычисления

Пусть для веса W_{23}:

- $y_3 = 0.8$, $t_3 = 1$ (ошибка 0.2)
- $x_2 = 0.6$
- $\lambda = 0.001$
- $W_{23} = 0.5$

Тогда:

$$\frac{\partial L}{\partial W_{23}} = (0.8 - 1) \cdot 0.6 + 2 \cdot 0.001 \cdot 0.5 \tag{16}$$

$$= -0.2 \cdot 0.6 + 0.001 \tag{17}$$

$$= -0.12 + 0.001 \tag{18}$$

$$= -0.119 \tag{19}$$

6. Полный вывод производной по смещениям

6.1. Производная логита по смещению

Рассмотрим выражение для логита z_j:

$$z_j = \sum_{i=1}^{n} W_{ij} x_i + b_j \tag{20}$$

Производная логита по смещению:

$$\frac{\partial z_j}{\partial b_j} = \frac{\partial}{\partial b_j} \left(\sum_{i=1}^{n} W_{ij} x_i + b_j \right) \tag{21}$$

$$= 1 \quad \text{(так как } b_j \text{ входит линейно)} \tag{22}$$

6.2. Применение цепного правила

Используем ранее полученную производную $\frac{\partial L}{\partial z_j} = y_j - t_j$ и применяем цепное правило:

$$\frac{\partial L}{\partial b_j} = \frac{\partial L}{\partial z_j} \cdot \frac{\partial z_j}{\partial b_j} \tag{23}$$

$$= (y_j - t_j) \cdot 1 \tag{24}$$

$$= y_j - t_j \tag{25}$$

6.3. Особенности смещений

- Смещения b_j не имеют регуляризационного члена (обычно L2-регуляризация не применяется к смещениям)
- Обновление смещений зависит только от ошибки предсказания $(y_j - t_j)$
- В отличие от весов, не зависит от входных значений x_i

6.4. Интерпретация результата

- Если $y_j > t_j$ (модель переоценивает вероятность класса):
  - Производная положительна $(y_j - t_j > 0)$
  - Смещение b_j будет уменьшаться
- Если $y_j < t_j$ (модель недооценивает вероятность класса):
  - Производная отрицательна $(y_j - t_j < 0)$
  - Смещение b_j будет увеличиваться

6.5. Пример вычисления

Рассмотрим два случая для класса $j = 3$:

Случай 1 (недооценка):

- $y_3 = 0.4$ (предсказанная вероятность)
- $t_3 = 1$ (истинное значение)

$$\frac{\partial L}{\partial b_3} = 0.4 - 1 = -0.6 \tag{26}$$

Смещение будет увеличено (при положительной скорости обучения).

Случай 2 (переоценка):

- $y_3 = 0.9$
- $t_3 = 0$

$$\frac{\partial L}{\partial b_3} = 0.9 - 0 = 0.9 \tag{27}$$

Смещение будет уменьшено.

6.6. Визуализация в матричной форме

Для всех смещений сразу:

$$\frac{\partial L}{\partial \mathbf{b}} = \mathbf{y} - \mathbf{t} \tag{28}$$

где:

- $\mathbf{b} = [b_1, ..., b_K]^T$ - вектор смещений
- $\mathbf{y} = [y_1, ..., y_K]^T$ - вектор предсказаний
- $\mathbf{t} = [t_1, ..., t_K]^T$ - вектор целевых значений

7. Полный вывод итоговых формул обновления

7.1. Общая схема градиентного спуска

Правило обновления параметров в градиентном спуске:

$$\theta \leftarrow \theta - \eta \cdot \frac{\partial L}{\partial \theta} \tag{29}$$

где:

- θ - оптимизируемый параметр (вес или смещение)
- η - скорость обучения (learning rate)
- $\frac{\partial L}{\partial \theta}$ - градиент функции потерь по параметру

7.2. Обновление весов

Используем полученную ранее производную:

$$\frac{\partial L}{\partial W_{ij}} = (y_j - t_j)x_i + 2\lambda W_{ij} \tag{30}$$

Формула обновления весов:

$$W_{ij} \leftarrow W_{ij} - \eta \cdot [(y_j - t_j)x_i + 2\lambda W_{ij}] \tag{31}$$

$$= W_{ij} - \eta(y_j - t_j)x_i - 2\eta\lambda W_{ij} \tag{32}$$

$$= (1 - 2\eta\lambda)W_{ij} - \eta(y_j - t_j)x_i \tag{33}$$

7.3. Физический смысл обновления весов

- Первое слагаемое $(1 - 2\eta\lambda)W_{ij}$:
  - Уменьшает абсолютное значение веса (эффект L2-регуляризации)
  - Коэффициент $2\eta\lambda$ определяет силу "забывания"веса
- Второе слагаемое $-\eta(y_j - t_j)x_i$:
  - Корректирует вес в зависимости от ошибки предсказания
  - Пропорционально входному сигналу x_i
  - Направление зависит от знака ошибки $(y_j - t_j)$

7.4. Обновление смещений

Используем полученную производную:

$$\frac{\partial L}{\partial b_j} = y_j - t_j \tag{34}$$

Формула обновления смещений:

$$b_j \leftarrow b_j - \eta \cdot (y_j - t_j) \tag{35}$$

7.5. Особенности обновления смещений

- Смещения обновляются только на основе ошибки $(y_j - t_j)$
- Не зависит от входных значений x_i
- Обычно не применяется регуляризация к смещениям

7.6. Примеры обновления параметров

Пример 1 (Обновление веса):

- Начальное значение: $W_{12} = 0.5$
- $y_2 = 0.8$, $t_2 = 1$, $x_1 = 0.6$
- $\eta = 0.1$, $\lambda = 0.001$

$$\begin{aligned} W_{12} &\leftarrow 0.5 - 0.1\left[(0.8 - 1) \cdot 0.6 + 2 \cdot 0.001 \cdot 0.5\right] && (36)\\ &= 0.5 - 0.1\left[-0.12 + 0.001\right] && (37)\\ &= 0.5 + 0.1 \cdot 0.119 && (38)\\ &= 0.5119 && (39) \end{aligned}$$

Пример 2 (Обновление смещения):

- Начальное значение: $b_2 = 0.2$
- $y_2 = 0.3$, $t_2 = 1$
- $\eta = 0.1$

$$\begin{aligned} b_2 &\leftarrow 0.2 - 0.1 \cdot (0.3 - 1) && (40)\\ &= 0.2 - 0.1 \cdot (-0.7) && (41)\\ &= 0.2 + 0.07 && (42)\\ &= 0.27 && (43) \end{aligned}$$

7.7. Векторизованная форма

Для всех параметров сразу:

$$\mathbf{W} \leftarrow \mathbf{W} - \eta\left[(\mathbf{y} - \mathbf{t})\mathbf{x}^T + 2\lambda\mathbf{W}\right] \quad (44)$$

$$\mathbf{b} \leftarrow \mathbf{b} - \eta(\mathbf{y} - \mathbf{t}) \quad (45)$$

где:

- $\mathbf{W}$ - матрица весов (размерность $n \times K$)
- $\mathbf{b}$ - вектор смещений (размерность $K \times 1$)
- $\mathbf{x}$ - входной вектор (размерность $n \times 1$)
- $\mathbf{y}$ - вектор предсказаний (размерность $K \times 1$)
- $\mathbf{t}$ - вектор целей (размерность $K \times 1$)