Аннотация

Решается задача классификации, рассматривается проблема подбора оптимальных параметров глубокой нейронной сети, в частности существование множества точек глобального минимума функции потерь. Причем известно, что различные оптимумы обладают различным качеством на тестовой выборке. Таким образом проблема подбора оптимальных параметров преобразовывается в проблему выбора глобальных минимумов обладающих лучшей обобщающей способностью.

В работе вводится понятие марджина (зазора), описывающий уверенность модели в предсказании, и дисконтированная функция потерь, которая дополнительно штрафует модель за низкий марджин.

Также предполагается, что минимумы с плохой генерализацией либо перестанут быть глобальными минимумами, либо схлопнутся вовсе.

Этот эффект явным образом подтверждается на игрушечном эксперименте и в задаче классификации изображений.

Ключевые слова: *доменная адаптация, нейронная сеть, GAN, WGAN, функция сходства Адуенко.*

Содержание

1 Введение

Рассматривается задача классификации изображений с помощью глубоких нейронных сетей. Модель нейронной сети представляется как последовательность параметрических нелинейных преобразований. Параметры подбираются с помощью решения задачи минимизации функции потерь.

Проблема такого подхода заключается в том, что размерность входных данных и число объектов обучающей выборки намного меньше, чем размерность пространства параметров нейронной сети. Это значит, что существует множество решений оптимизационной задачи (минимумов). Допустим два скрытых слоя сети представляются, как линейное преобразование с некоторой функцией нелинейности. Тогда можно изменить порядок соответствующих нейронов в этих двух слоях так, чтобы предсказание сети никак не изменилось, но фактически параметры сети получились другие. Таким образом получили вектор параметров, отличный от изначального, но, на котором достигается минимум функционала. Также, если в сети присутствуют нормировки, например, как нормировка по батчу (BatchNorm), то умножение параметров сети на константу не меняет предсказание модели. Более того, можно умножать параметры каждого слоя независимо. Это лишь некоторые примеры, которые показывают неединственность решения оптимизационной задачи. Они не меняют качество модели, потому что сохраняют сеть в том же классе эквивалентности (отношением эквивалентности является совпадение предсказания сети), что и изначальная сеть. И стоит отметить, что таких классов эквивалентности много.

Помимо того, что в пространстве параметром модели существует множество глобальных оптимумов, но также, они качественно отличаются друг от друга, например по ширине, по качеству генерализации и тд. Для оценки генерализации модели используется метрика точности модели на тестовой выборке. Для оценки ширины минимума предлагается использовать среднюю норму стохастического градиента по обучающей выборке.

Техника нормировки по батчу (BatchNorm), описанная в статье [https://arxiv.org/pdf/1502.03167.p
используется для укорения и стабилизации обучения модели. Предлагается дополнительно добавлять внутрь глубокой нейронной сеть промежуточные слои нормализации, которые приводят разнородные данные к единому виду, то есть к нулевому мат. ожиданию и единичной дисперсии. Таким образом метод позволяет использовать гораздо более высокие темпы обучения (lr) и быть менее аккуратными при начальной инициализации весов.

Но такие промежуточные нормализующие слои порождают дополнительные сложности в анализе поведения модели, в частности они делают нейронные сети масштаб-инвариантными, то есть умножение весов на некоторую константу не меняет предсказание модели. Таким образом две сети, отличающиеся только умножением на константу будут иметь одинаковую точность на тесте, но разную ширину минимума. Чтобы компенсировать этот эффект предлагается перед подсчетом шириной

минимума приводить веса сети к единичной норме.

Для дальнейшего изложения введем понятие зазор генерализации (generalization gap), он показывает разницу качества модели на обучающей и тестовой выборке.

В научном сообществе существует гипотеза [.....], что широкие минимумы обладают большей генерализацией. На рис 1 дано интуитивное объяснение этого предположения, на нем изображена кривая функции потерь при линейной интерполяции между двумя минимумами в пространстве весов. причем интерполяция происходит между широким и узким минимумом. Так как поверхность функции потерь на тестовых данных будет немного смещена, это значит, что у широкого минимума зазор генерализации будет меньше, а сама генерализация будет выше, чем у узкого минимума. Таким образом данный эффект основывается на том, что узкие минимумы менее гладкие, поэтому небольшое изменение в данных ведет к сильной просадке качества модели.

Рис. 1: Flat minima results in better generalization compared to sharp minima. Pruning neural ODEs flattens the loss around local minima. Figure is reproduced from Keskar et al. (2017).

Например, в [что-то про ... https://arxiv.org/pdf/2006.15081.pdf] авторы визуализировали с помощью t-SNE

В [статья про ширину минимумов, например https://arxiv.org/pdf/1906.03291.pdf] авторы исследовали данную проблему, показали, что различные минимумы обладают различной генерализацией (обобщающей способность т.е. качеством на тестовой выборке), а широкие минимумы обладают лучшей генерализацией, чем узкие. Связывают это с тем, что узкие минимумы менее гладкие, поэтому небольшое изменение в данных ведет к сильной просадке качества модели. Например, известно, что опти-

мизатор SGD [] избегает узких минимумов [].

Глубокие ансамбли, предложенные Lakshminarayanan и др. [[2017], представляют из себя ансамбль глубоких нейронных сетей обученных из различных начальных инициализаций. В [https://arxiv.org/pdf/1912.02757.pdf] авторы эмперически показали, что глубокие ансамбли являются хорошим подходом для повышения точности, за счет того, что модели захватывают различные моды в пространстве весов, в то время как вариационные байесовские методы, как правило, фокусируются на одной моде. Различные, ортогональные веса ансамблей позволяют в совокупности более аккуратно выучивать обучающую выборку.

Бустинг - метод построения ансамбля, основанный на идеи, что каждая следующая обучаемая модели стремится компенсировать ошибку всех предыдущих моделей. Но такой способ применим для слабых алгоритмов, т.е. таких, которые не могут правильно распознать всю обучающую выборку. Глубокие нейронные сети могут обучиться до стопроцентной точности, поэтому классические бустинги как [adaboost, gradboost и тд] для нейронных сетей неприменимы. Мы предлагаем строить бустинг на глубоких нейронных сетях используя марджин в качестве ошибки, которую будут компенсировать следующие модели.

В данной работе предлагается метод регуляризации модели (дисконтированная функция потерь), который сглаживает поверхность функции потерь, тем самым остаются только широкие минимумы. Также рассматривается ансамблирование моделей, обученных с помощью предлагаемого методы и различные эвристики бустинга.

Определение 1 *Функционал $L(f, g)$ назовем дисконтированной функцией потерь, g - значение гэпа:*

$$L(f, g) = \frac{1}{N} \sum_i - \log \frac{e^{X_{i,y_i} - g}}{e^{X_{i,y_i} - g} + \sum_{j \neq y_i} e^{X_{i,j}}},$$

Причем при $g = 0$, т.е. функционал $L(f, 0)$ превращается в функцию потерь кросс-энтропии [].

Поверхность функции потерь по всей выборке это среднее значение функций потерь на каждом из объектов. Такой функционал "двигает"кривую, соответствующую одному объекта в сторону точки оптимума, на рис. 2 это изображено с помощью стрелок. Поэтому, если в пространстве весов модели существует узкий глобальный минимум для кросс-энтропии, то для дисконтированной функции потерь эта точка будет локальным минимумом.

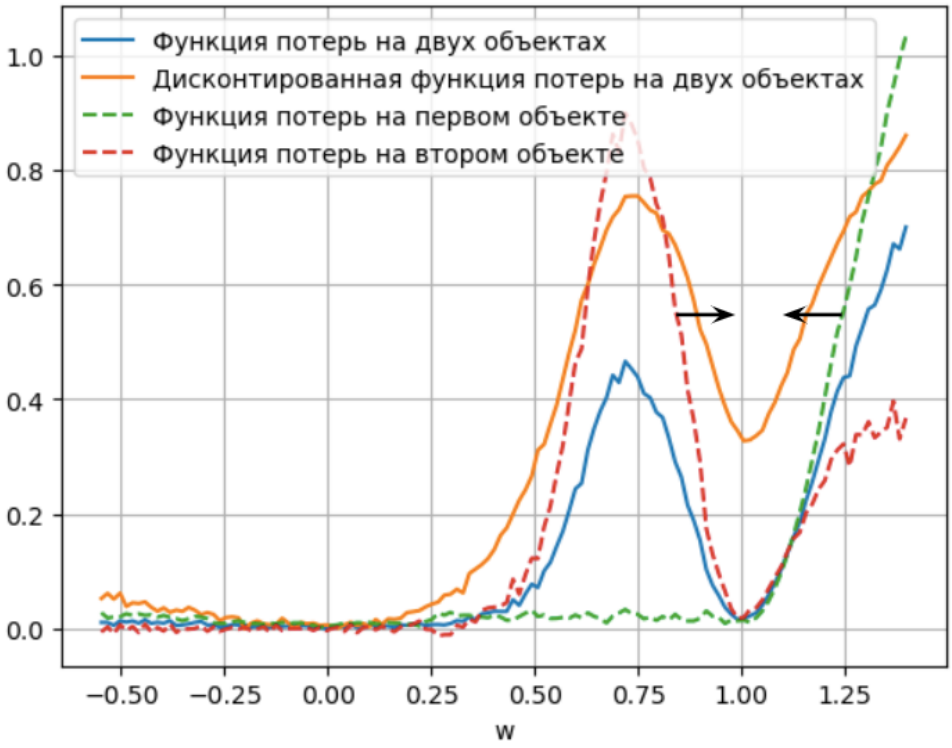
Функция потерь на двух объектах
Дисконтированная функция потерь на двух объектах
Функция потерь на первом объекте
Функция потерь на втором объекте
1.0
0.8
0.6
0.4
0.2
0.0
−0.50
−0.25
0.00
0.25
0.50
0.75
1.00
1.25
w

Рис. 2

1.1 Обзор литературы

В задаче верификации лиц популярны так называемые metric Learning методы. Их суть заключается в том, что используется модель, получающая векторное представление изображения. В качестве меры сходства изображений используют косинусное расстояние. Во время обучения модель поощряют изображениям одного класса строить векторные представления более похожие друг на друга (косинусное сходство должно быть большое), а вектора изображений из разных классов должны быть далеки друг от друга. Таким образом строиться функция потерь, которая отражает данную логику.

В статье [https://arxiv.org/pdf/1703.09507.pdf] авторы ввыли дополнительно дополнительные обучаемые параметры, W, которые описывают центроиды соответствующих классов. Таким образом предположение о том, что изображения одного класса должны быть похожи переформулируется в виде того, что вектора соответствующие изображениям одного класса должны быть близки к одному определенному вектору и далеки от всех остальных векторов матрицы W.

Затем исследователи заметили, что можно использовать дополнительный гиперпараметр m, который описывает, насколько вектора соответствующие изображениям одного класса должны быть ближе к одному определенному вектору, чем близость к всем остальным векторам матрицы W. Данная идея была реализована, например, в [CosFace: https://arxiv.org/pdf/1801.09414.pdf]. Итоговая функция потерь описывается как:

$$L = \frac{1}{N}\sum_i -\log\frac{e^{s(cos(\theta_{y_i,i})-m}}{e^{s(cos(\theta_{y_i,i})-m} + \sum_{j\neq y_i} e^{s(cos(\theta_{j,i})}}$$

$$\begin{aligned} W &= \frac{W^*}{\|W^*\|}, \\ x &= \frac{x^*}{\|x^*\|}, \\ cos(\theta_{j,i}) &= W_j^T x_i. \end{aligned}$$

Данная процедура имеет очень похожий вид, что и предлагаемый дисконтированная функция потерь, но имеет иную идею.

2 Теоретическая часть

Для дальнейшего повествования введем математические описание.

В качестве нейронной сети понимается параметрическая функция $f(x, \theta)$, где x - тензор изображения, $\theta \in \mathbb{R}^t$ - вектор параметров модели. Причем используется масштабинвариантная сеть, т.е. $\forall \alpha > 0 \ \forall x \mapsto f(x, \alpha\theta) = f(x, \theta)$. $f(\cdot, \theta) : \mathbb{R}^{\mathrm{h}\times\mathrm{w}\times 3} \to \mathbb{R}^{\mathrm{K}}$, где K - число классов классификации. $\mathcal{X} = \{(x, y)|x \in \mathbb{R}^{\mathrm{h}\times\mathrm{w}\times 3}, y -$ меток истинного класса$\}$ - множество объектов (изображений).

Функция $f(x, \theta)$ представляет из себя последовательность параметрических преобразований (для экспериментов используется архитектура ResNet []), причем последние слои представляют из себя слой нормировки по батчу [] и линейный слой.

Определение 2 *Нормировкой по батчу (BN)* $\mathbb{R}^{n\times m} \to \mathbb{R}^{n\times m}$ *назовется преобразование вида:*

$$\mathrm{BN}(x) = \frac{x - \mathbb{E}[x]}{\mathrm{Var}[x]} \tag{1}$$

$\mathbb{E}$ *и* Var *берется поканально, т.е. вдоль второй размерности входной матрицы* x*.*

Таким образом:

$$f(\cdot, \theta) = W \circ \mathrm{BN} \circ \ldots \tag{2}$$

Определение 3 *Марджином модели* $f(\cdot, \theta)$ *для* $(x, y) \sim \mathcal{X}$ *назовем* m*:*

$$m(\theta, x, y) = f(\cdot, \theta)_y - \max_{i \neq y} f(\cdot, \theta)_i \tag{3}$$

Для краткости обозначим $m(\theta, x, y) = m_i$*, если* x *это* i*-ый объект множества* $\mathcal{X}$*.*

Средний марджин модели $f(\cdot, \theta)$ на множестве $\mathcal{X}$ соответственно:

$$\bar{m} = \frac{1}{|\mathcal{X}|} \sum_{(x,y)\sim\mathcal{X}} m(\theta, x, y)$$

Определение 4 *Функцией SoftMax назовем преобразование из* $\mathbb{R}^n$ *в единичный симплекс на* $\mathbb{R}^n$ *вида:*

$$SoftMax(\mathbf{v})_i = \frac{e^{\mathbf{v}_i}}{\sum_j e^{\mathbf{v}_j}}$$

На практике $\mathbf{v}_i$ не может равняться минус бесконечности, поэтому каждая координата вектора $SoftMax(\mathbf{v}) > 0$, сумма всех координат равна единице. Таким образом, без ограничений на область определения логарифма, можем ввести эквивалентное определение дисконтированной функции потерь:

Определение 5 *Функционал $L(f, g)$ назовем дисконтированной функцией потерь, g - значение гэпа:*

$$L(f, g) = \frac{1}{|\mathcal{X}|} \sum_{(x,y)\sim\mathcal{X}} -\log\Big(SoftMax\big(f(x) - \text{onehot}_y \cdot g\big)\Big)_y,$$

где onehot_y *- вектор нулей, с единицей на y-ой координате.*

Причем при $g = 0$, т.е. функционал $L(f, 0)$ превращается в функцию потерь кросс-энтропии [].

$$L(f, g) = \frac{1}{|\mathcal{X}|} \sum_{(x,y)\sim\mathcal{X}} -\log\Big(SoftMax\big(f(x) - \text{onehot}_y \cdot g\big)\Big)_y,$$

Определение 6 *Точностью (Acc.) предсказания модели f на множестве объектов $\mathcal{X}$ назовем:*

$$Acc(f, \mathcal{X}) = \frac{1}{|\mathcal{X}|} \sum_{(x,y)\sim\mathcal{X}} \text{I}\{\arg\max(f(x)) = y\},$$

где I - *индикаторная функция, принимает значение* 1 *при истинном значении аргумента.*

Заметим, что если $m(\theta, x, y) > 0 \Leftrightarrow \text{I}\{\arg\max(f(x)) = y\} = 1$

Как описывалось ранее, для оценки ширины минимума предлагается использовать среднюю норму стохастического градиента, т.е.:

$$\frac{1}{|\mathcal{X}|} \sum_{(x,y)\sim\mathcal{X}} \Big\|\nabla_\theta - \log SoftMax\big(f(x, \bar{\theta})\big)\Big\|, \tag{4}$$

где $\bar{\theta} = \theta/\|\theta\|$

Лемма 1 *Пусть дана функция $f(x)$ вида (2), и $f \not\equiv 0$ тогда $\forall g > 0$ и $\alpha > 0$ верно:*

$$\lim_{\alpha\to\infty} |L(\alpha f, g) - L(\alpha f, 0)| = 0$$

Доказательство. Обозначим $\mathbf{v} = f(x)$, тогда для доказательства достаточно показать, что

$$\left|\frac{e^{\alpha\mathbf{v}_y - g}}{e^{\alpha\mathbf{v}_y - g} + \sum_{j\neq y} e^{\alpha\mathbf{v}_j}} - \frac{e^{\alpha\mathbf{v}_y}}{e^{\alpha\mathbf{v}_y} + \sum_{j\neq y} e^{\alpha\mathbf{v}_j}}\right| \xrightarrow[\alpha\to\infty]{} 0 \tag{5}$$

С учетом того, что $g/\alpha \to 0$, при $\alpha \to \infty$, перепишем требуемое 5 и докажем утверждение:

$$\left| \frac{e^{\mathbf{v}_y - g/\alpha}}{e^{\mathbf{v}_y - g/\alpha} + \sum_{j \neq y} e^{\mathbf{v}_j}} - \frac{e^{\mathbf{v}_y}}{e^{\mathbf{v}_y} + \sum_{j \neq y} e^{\mathbf{v}_j}} \right| \xrightarrow[\alpha \to \infty]{} 0$$

■

Из описанной выше леммы видно, что при искусственном увеличении нормы весов последнего слоя W из (2) уменьшается вклад, который достигается дисконтированной функцией потерь при ненулевом значении гэпа. Для компенсации данного эффекта предлагается замораживать веса последнего слоя. Такая модификация нейронной сети используется во всех последующих рассуждениях и экспериментах.

2.1 Анализ свойств дисконтированной функции потерь

Для дальнейшего изучения свойств дисконтированной функции потерь, докажем несколько теорем. Покажем, что существует максимальное значение марджина, который может показывать модель, и как следствие получим, что существует предельное значение гэпа в дисконтированной функции потерь, после которого все функции потерь ведут себя эквивалентно.

Лемма 2 *Пусть дана функция $f(x)$ вида (2), с замороженными W - весами последнего слоя. $\mathcal{X} = \{(x, y) | x \in \mathbb{R}^{\mathrm{h} \times \mathrm{w} \times 3}, y \in \{-1, 1\}\}$ - множество объектов бинарной классификации, причем выборка равновесна, т.е. $\mathrm{P}_{(x,y) \sim \mathcal{X}}(y = 1) = 1/2$. Тогда максимальное значение среднего марджина, которое может получить функция:*

$$\bar{m} = \|W_{[:,0]} - W_{[:,1]}\|_1 \leq \|W\|_1$$

Доказательство.

Так как решается задача бинарной классификации, следовательно $W \in \mathbb{R}^{d \times 2}$. Введем вектор $\mathbf{w} \in \mathbb{R}^d$, $\mathbf{w}_j = W_{j,0} - W_{j,1}$. Также верно, что маржин из выражения (3) переписывается как $m_i = f(x_i)_{y_i} - f(x_i)_{-y_i} \Rightarrow m_i = y_i \cdot X_i^T \mathbf{w}$.

Пусть X это матрица предсказаний функции $f(x)$ до применения последнего слоя на множестве объектов $\mathcal{X}$. Для доказательства леммы сформулируем оптимизационную задачу:

$$\frac{1}{N}\sum_i y_i \cdot X_i^T \mathbf{w} \to \max \tag{6}$$

$$\forall j \mapsto \ \frac{1}{N}\sum_i X_{ij} = 0, \tag{7}$$

$$\forall j \mapsto \ \frac{1}{N}\sum_i X_{ij}^2 = 1, \tag{8}$$

Где $N = |\mathcal{X}|$ - размер выборки, выражение (6) описывает максимальное среднее значение марджина, (7) добавляет ограничение на нулевое матожидание, а (8) на единичную дисперсию. Такие ограничения появляются из-за того, что X является выходом слоя BN.

Для аналитического решения воспользуемся условием Каруша-Куна-Таккера, введем двойственные переменные $\mu, \nu \in \mathbb{R}^d$:

$$\nabla_X \Big[\sum_i \big(- y_i \cdot X_i^T \mathbf{w}\big) + \sum_j \mu_j \sum_i X_{ij} + \sum_j \nu_j \big(\sum_i X_{ij}^2 - c\big) \Big] = 0$$

Упростим выражение и сгруппируем слагаемые:

$$\nabla_X \sum_{i,j} \big(\nu_j X_{ij}^2 + X_{ij}(\mu_j - y_i \mathbf{w}_j) - \nu_j c\big) = 0$$

Поэтому:

$$X_{ij}^* = \frac{y_i \mathbf{w}_j - \mu_j}{2\nu_j} \tag{9}$$

Подставим полученное значение X^* из (9) в (7):

$\forall j \mapsto \ \frac{1}{N}\sum_i \frac{y_i \mathbf{w}_j - \mu_j}{2\nu_J} = 0 \Rightarrow \mathbf{w}_j \sum_i y_i = N \cdot \mu_j \Rightarrow \mu_j = \mathbf{w}_j \cdot \frac{\sum_i y_i}{N} = 0$, т.к. выборка $\mathcal{X}$ равновесна.

Подставим полученное значение X^* из (9) в (8):

$\forall j \mapsto \ \frac{1}{N}\sum_i \big(\frac{y_i \mathbf{w}_j - \mu_j}{2\nu_J}\big)^2 = 1 \Rightarrow \nu_j = \frac{|\mathbf{w}_j|}{2}\sqrt{\frac{\sum_i y_i^2}{N}} = \frac{|\mathbf{w}_j|}{2}$.

Таким образом максимальное значение среднего марджина получается после подстановки всех найденных значений в (6):

$$\frac{1}{N}\sum_i y_i \cdot X_i^T \mathbf{w} = \frac{1}{N}\sum_{ij} y_i X_{ij} \mathbf{w}_j = \frac{1}{N}\sum_{ij} y_i \frac{y_i \mathbf{w}_j}{|\mathbf{w}_j|}\mathbf{w}_j = \sum_j |\mathbf{w}_j| = \|\mathbf{w}\|_1$$

■

Теорема 3 *Пусть дана функция $f(x)$ вида (2), с замороженными W - весами последнего слоя. $\mathcal{X} = \{(x, y)|x \in \mathbb{R}^{\mathrm{h}\times\mathrm{w}\times 3}, y \in \{1, 2 \dots K\}\}$ - множество объектов многоклассовой классификации, причем выборка равновесна, т.е. $\mathrm{P}_{(x,y)\sim\mathcal{X}}(y = k) = 1/K$. Тогда максимальное значение среднего марджина, которое может получить функция:*

$$\bar{m} \leq \frac{2}{K}\|W\|_1.$$

Доказательство.

Для каждого класса зададим пару, например, для класса k парным классом будет $k' = k + 1 \mod K$. Таким образом:

$$m_i = f(x)_k - \max_{j\neq k} f(x)_j \leq f(x)_k - f(x)_{k'},$$

Из доказательства леммы 2 известно, что $f(x)_k - f(x)_{k'} \leq \|W_{[:,k]} - W_{[:,k']}\|_1$. Таким образом:

$$\bar{m} = \frac{1}{N}\sum_i f(x_i)_{y_i} - \max_{j\neq y_i} f(x_i)_j \leq$$

$$\frac{1}{N}\sum_i f(x_i)_{y_i} - f(x_i)_{y_i+1 \mod K} \leq$$

$$\frac{1}{N}\sum_i \|W_{[:,y_i]} - W_{[:,y_i+1 \mod K]}\|_1 =$$

$$\frac{1}{N\cdot K}\sum_i \|W_{[:,1]} - W_{[:,2]}\|_1 + \dots + \|W_{[:,K-1]} - W_{[:,K]}\|_1 + \|W_{[:,K]} - W_{[:,1]}\|_1 \leq$$

$$\frac{2}{N\cdot K}\sum_{i,k} \|W_{[:,k]}\|_1 = \frac{2}{K}\|W\|_1$$

При последнем переходе используется неравенство треугольника.

■

Из теоремы 3 следует, что средний марджин, который получает модель ограничен, поэтому следующую теорема имеет смысл:

Теорема 4 *Для достаточно больших g_1 и g_2, для некоторой функции $f(x)$ вида (2), с замороженными W - весами последнего слоя и множества объектов $\mathcal{X}$ верно:*

$$L(f, g_1) - g_1 \approx L(f, g_2) - g_2 \tag{10}$$

Доказательство.

Пусть X это матрица предсказаний функции $f(x)$ на множестве объектов $\mathcal{X}$, т.е. $X_i = f(x_i)$ для $(x_i, y_i) \sim \mathcal{X}$, $N = |\mathcal{X}|$, тогда выражение (10) переписывается как:

$$\frac{1}{N}\sum_i -\log\frac{e^{X_{y_i,i}-g_1}}{e^{X_{y_i,i}-g_1}+\sum_{j\neq y_i}e^{X_{j,i}}} - g_1 \approx \frac{1}{N}\sum_i -\log\frac{e^{X_{y_i,i}-g_2}}{e^{X_{y_i,i}-g_2}+\sum_{j\neq y_i}e^{X_{j,i}}} - g_2 \quad (11)$$

Покажем, что для каждого i верно:

$$e^{X_{y_i,i}-g_1} \ll e^{X_{j,i}}, \text{ для некоторого } j \neq y_i \quad (12)$$

Для этого достаточно в качестве j взять $j = \arg\max_{j\neq y_i} X_{j,i}$. Из теоремы 3 следует, что $m_i = X_{y_i,i} - X_{j,i}$ ограничен сверху некоторой константой C, т.е. $X_{y_i,i} - X_{j,i} \leq C < g_1$, для достаточно большого g_1. Таким образом получаем (12).

Разложим в левой части выражения (11) логарифм на сумму лагорифмов:

$$\begin{aligned}
&-\log\frac{e^{X_{y_i,i}-g_1}}{e^{X_{y_i,i}-g_1}+\sum_{j\neq y_i}e^{X_{j,i}}} - g_1 = \\
&\qquad = -\log e^{X_{y_i,i}-g_1} + \log\big(e^{X_{y_i,i}-g_1} + \sum_{j\neq y_i} e^{X_{j,i}}\big) - g_1 = \\
&\qquad\qquad = -X_{y_i,i} + \log\big(e^{X_{y_i,i}-g_1} + \sum_{j\neq y_i} e^{X_{j,i}}\big) \approx \\
&\qquad\qquad\qquad \approx -X_{y_i,i} + \log\big(\sum_{j\neq y_i} e^{X_{j,i}}\big)
\end{aligned}$$

Последнее приближение выводится из (12). Правая части выражения (11) оценивается соответствующим образом, откуда следует доказательство теоремы.

■

Таким образом мы показали, что при достаточно больших значениях гэпа дисконтированные функции потерь, совпадают с точностью до константы. Важным следствием является, что в таком случае градиенты у таких функций совпадают, а значит и алгоритм стохастического градиентного спуска [] будет обновлять веса эквиволентным образом, независимо от значения гэпа.

2.2 Бустинг

2.3 Постановка задачи

Оптимальные параметры нейронной сети f подбираются методом стохастического градиентного спуска [] при минимизации:

$$L(f(\cdot, \theta), g) + \lambda \|\theta\|_2^2 \to \min_{\theta},$$

где λ - коэффициент L_2 регуляризации весов ($\|\theta\|_2^2$).

3 Вычислительный эксперимент