Введение в нейронные сети 2

Пальчиков Николай

162

- Нужно оптимизировать функцию ошибки

$$w^* = \arg\min Q(w)$$

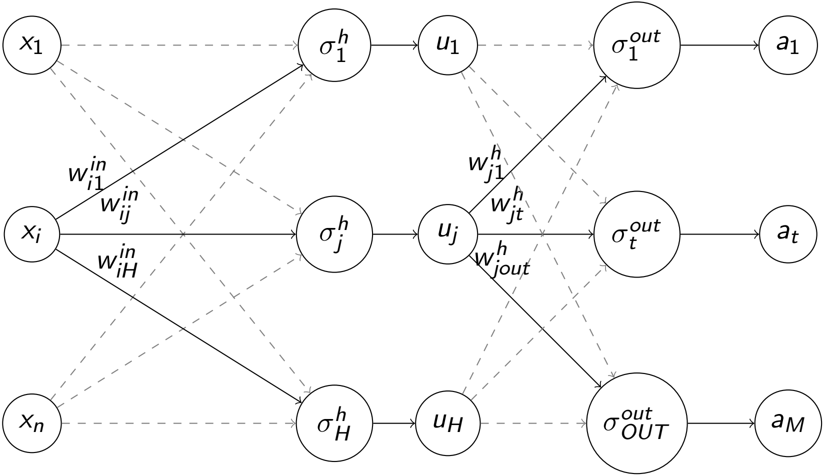
- Используются: стохастические методы (генетический алгоритм, метод отжига)
- Градиентные методы

Градиентные методы

- Пусть w – вектор всех весов в нейронной сети. Привычно, что обычно вреся вычисления сравнимо с количеством весовых парамет
- BACKPROP – алгоритм, позволяющий считать градиент.

Градиентные методы: BACKPROP

Рассмотрим нейронную сеть с единственным скрытым слоем



Градиентные методы: BACKPROP

Обозначения

- w_{jt}^{h} – вес синаптических связей между j-м нейроном скрытого слоя и t-м нейроном выходного слоя
- w_{ij}^{in} – вес синаптических связей между i-м нейроном входного слоя и j-м нейроном скрытого слоя
- σ_j^h – функция активации j-го нейрона скрытого слоя
- σ_t^{out} – функция активации t-го нейрона выходного слоя
- a_t – результат работы t-го нейрона выходного слоя.
- u_j – результат работы j-го нейрона скрытого слоя.
- x_i – i-й входной нейрон нейронной сети

Градиентные методы: BACKPROP

- Зафиксируем один объект x^k. Посчитаем частные производные по результатам вычислених выходных нейронов.
- Тогда
$$a_t(x^k) = \sigma_j^{out}\left(\sum_{j=1}^{H} w_{jt}^h u_j(x^k)\right)$$
И среднеквадратичная ошибка на этом объекте это просто
$$Q(w) = \frac{1}{2}\sum_{t=1}^{M}\left(a_t(x^k) - y_t^k\right)^2$$

Градиентные методы: BACKPROP

- Зафиксируем t
- Тогда
$$\frac{\partial Q(w)}{\partial a_t} = \left(a_t(x^k) - y_t^k\right) = \delta_t^{out}$$
Это ни что иное, как ошибка на t-м выходном нейроне
- Теперь зафиксируем нейрон j на скрытом слое и посчитаем

$$\frac{\partial Q(w)}{\partial u_j} = \sum_{t=1}^{OUT} \left(a_t(x^k) - y_t^k\right) \cdot \sigma_t'^{out} \cdot w_{jt}^h =$$

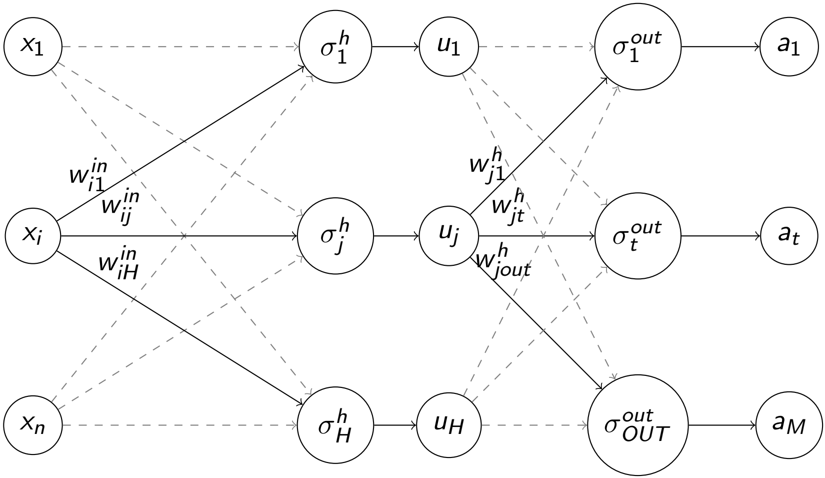
$$= \sum_{q=1}^{OUT} \delta_t^{out} \cdot \sigma_t'^{out} \cdot w_{jt}^h = \delta_j^h$$

δ_j^h – (назовём по аналогии) ошибка на j-м нейроне выходного слоя

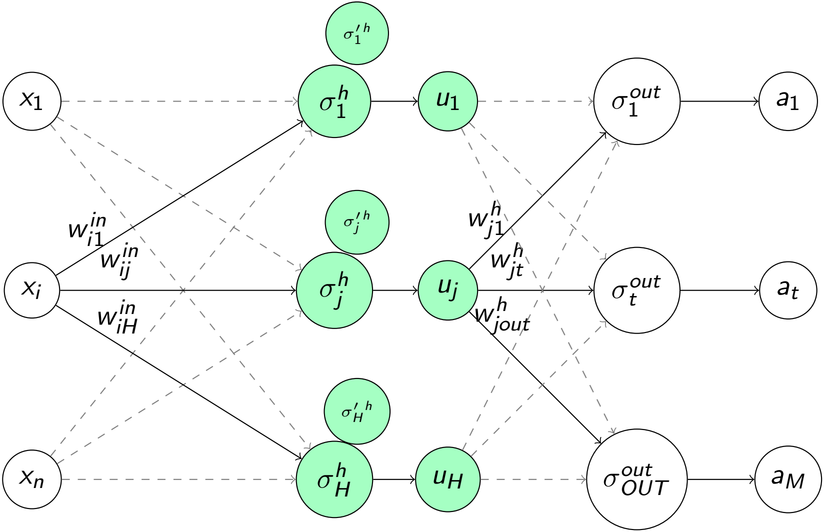
Градиентные методы: BACKPROP

Заметим, что ошибки на скрытом слое вычисляется через ошибки на последнем слое. Так можно вычислить ошибки для выходов всех слоев (а заодно и частные производные функционала ошибки по выходу каждого нейрона), отсюда и название метода

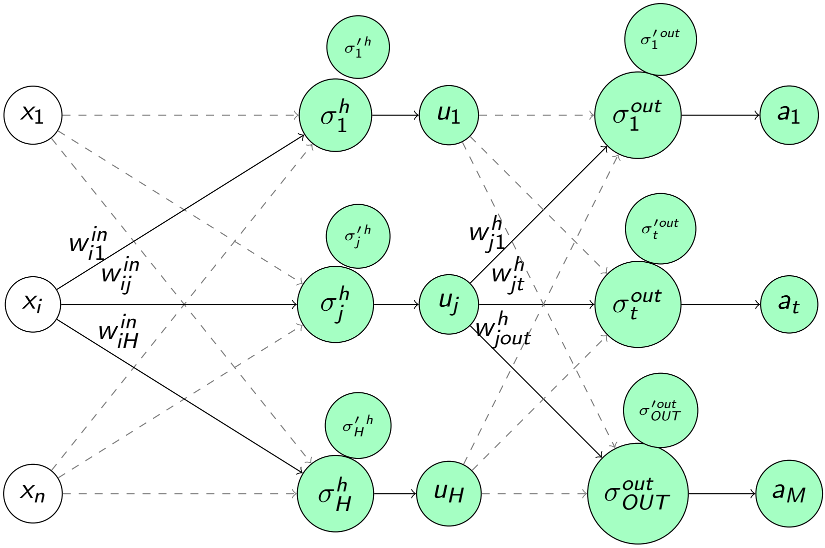
Градиентные методы: BACKPROP



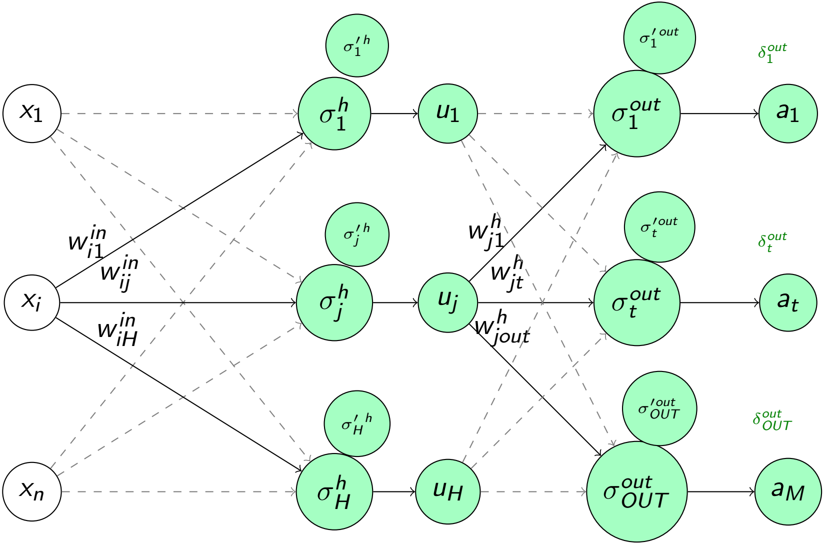
Градиентные методы: BACKPROP



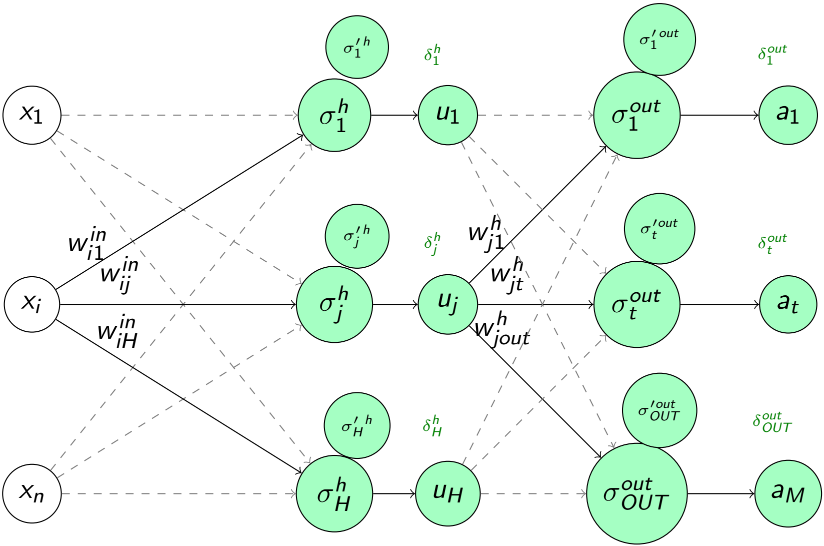
Градиентные методы: BACKPROP



Градиентные методы: BACKPROP



Градиентные методы: BACKPROP



Градиентные методы: BACKPROP

- Теперь легко вычислить градиент.

$$\frac{\partial Q(w)}{\partial w^h_{jt}} = \frac{\partial Q(w)}{\partial a_t}\frac{\partial a_t}{\partial w^h_{jt}} = \delta^{out}_t \cdot \sigma'^{out}_t \cdot u_j(x^k)$$

$$\frac{\partial Q(w)}{\partial w^{in}_{ij}} = \frac{\partial Q(w)}{\partial u_j}\frac{\partial u_j}{\partial w^h_{ij}} = \delta^h_j \cdot \sigma'^h_j \cdot x^k_i$$

Градиентные методы: BACKPROP

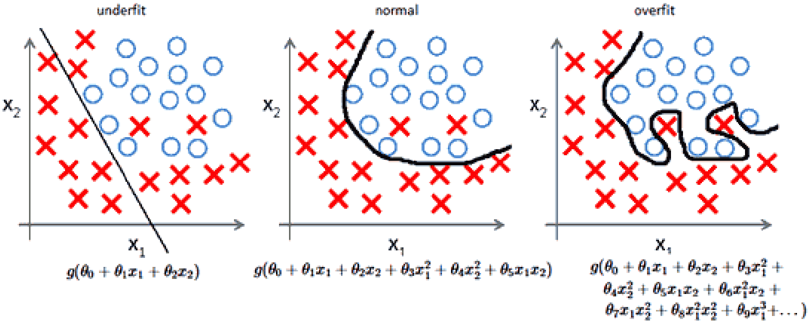
Достоинства

- Высокая эффективность: на одном объекте градиент посчитается за $O(Hn + H \cdot OUT)$.
- Через каждый нейрон проходит информация только о связных с ним нейронах, так что процесс параллелится

Недостатки

- Застревание в локальных минимумах
- Функция активации должна быть дифференциируемой

Оптимизация структуры нейронной сети



Выбор числа слоев

- Линейная разделимость – один слой
- Нелинейная разделимость – два слоя
- Более сложные области – три слоя

Выбор количества нейронов в скрытом слое

- Визульный способ – когда фичей мало
- Оптимизация по отложенной выборке
- Динамическое добавление нейронов: сначала обучается при заведомо недостаточном количестве нейронов, затем нейроны добавляются по одному.

Прореживание: optimal brain damage

- Идея – нужно удалить некоторые наиболее незначимые синаптические связи, более формально – те, от зануления которых $Q(w)$ вырастет меньше всего
- Каждый раз для каждой синаптической связи считать изменение общей ошибки при ее удалении считать непростительно затратно

Прореживание: optimal brain damage

- Зафиксируем вектор весов w, приблизим значение $Q(w+\delta)$ рядом Тейлора в точке w

$$Q(w+\delta) = Q(w) + \sum_i g_i \delta_i + \frac{1}{2} \sum_i h_{ii} \delta_i^2 + \frac{1}{2} \sum_{i \neq j} h_{ij} \delta_i \delta_j + o(||\delta^2||)$$

Где δ_i – i-я компонента приращения, g_i – i-я компонента градиента, h_{ij} – компонента Гессиана (матрицы вторых производных)

Прореживание: optimal brain damage

Сделаем несколько сильных предположений

- Гессиан диагонален
- w – точка локального минимума

Тогда

$$Q(w + \delta) \approx \frac{1}{2} \sum_i h_{ii} \delta_i^2$$

И любое зануление w_i будет приводить к неубыванию $Q(w)$, а именно к увеличению на $s_i = h_{ii} w_i^2$. Последнюю величину называют salience

Прореживание: optimal brain damage

1. Тренировка с помощью градиентных методов (чтобы сойтись в локальный минимум)
2. Вычислить вторые производные h_{ii} для каждого параметра с помощью backprop
3. Вычислить salience для каждого параметра
4. Удалить d связей с наименьшим salience
5. Перейти к шагу 1