

# Семинары по линейным классификаторам

Евгений Соколов  
sokolov.evg@gmail.com

13 ноября 2016 г.

## 1 Основные определения

Пусть  $X \subset \mathbb{R}^d$  — пространство объектов,  $Y = \{-1, +1\}$  — множество допустимых ответов,  $X^\ell = (x_i, y_i)_{i=1}^\ell$  — обучающая выборка. Каждый объект  $x \in X$  описывается вещественным вектором  $(x_1, \dots, x_d) \in \mathbb{R}^d$ .

*Линейный классификатор* определяется следующим образом:

$$a(x, w) = \text{sign}(\langle w, x \rangle + b) = \text{sign}\left(\sum_{j=1}^d w_j x_j + b\right),$$

где  $w \in \mathbb{R}^d$  — вектор весов,  $b \in \mathbb{R}$  — сдвиг (bias).

Если не сказано иначе, мы будем считать, что среди признаков есть константа,  $x_0 = 1$ . В этом случае нет необходимости вводить сдвиг  $b$ , и линейный классификатор можно задавать как

$$a(x, w) = \text{sign}\langle w, x \rangle.$$

Обучение линейного классификатора заключается в поиске вектора весов, на котором достигается минимум некоторого функционала качества:

$$w = \arg \min_{w \in \mathbb{R}^d} Q(w, X^\ell). \quad (1.1)$$

Наиболее логичным функционалом для задачи классификации является число неверно классифицированных объектов:

$$Q(w, X^\ell) = \sum_{i=1}^{\ell} \left[ y_i (\langle w, x_i \rangle + b) < 0 \right] \rightarrow \min_w.$$

У такого функционала, однако, есть большой недостаток — он не является дифференцируемым, из-за чего поиск оптимального вектора весов  $w$  становится крайне трудной задачей. Чтобы преодолеть эту проблему, оптимизируют гладкую верхнюю оценку на данный функционал:

$$Q(w, X^\ell) = \sum_{i=1}^{\ell} \left[ y_i (\langle w, x_i \rangle + b) < 0 \right] \leq \sum_{i=1}^{\ell} L\left(y_i (\langle w, x_i \rangle + b)\right) \rightarrow \min_w. \quad (1.2)$$

В качестве оценки  $L(M)$  можно использовать, например, логистическую функцию потерь  $L(M) = \log(1 + e^{-M})$ .

## 2 Градиент функции

Как правило, оптимизационная задача (1.2) решается с помощью градиентных методов (или же методов, использующих как градиент, так и информацию о производных более высокого порядка).

Градиентом функции  $f : \mathbb{R}^d \rightarrow \mathbb{R}$  называется вектор его частных производных:

$$\nabla f(x_1, \dots, x_d) = \left( \frac{\partial f}{\partial x_j} \right)_{j=1}^d.$$

### §2.1 Свойства градиента

Градиент является направлением наискорейшего роста функции, а антиградиент (т.е.  $-\nabla f$ ) — направлением наискорейшего убывания. Это ключевое свойство градиента, обосновывающее его использование в методах оптимизации.

Докажем данное утверждение. Пусть  $v \in \mathbb{R}^d$  — произвольный вектор, лежащий на единичной сфере:  $\|v\| = 1$ . Пусть  $x_0 \in \mathbb{R}^d$  — фиксированная точка пространства. Скорость роста функции в точке  $x_0$  вдоль вектора  $v$  характеризуется производной по направлению  $\frac{\partial f}{\partial v}$ :

$$\frac{\partial f}{\partial v} = \frac{d}{dt} f(x_{0,1} + tv_1, \dots, x_{0,d} + tv_d) \Big|_{t=0}.$$

Из курса математического анализа известно, что данную производную сложной функции можно переписать следующим образом:

$$\frac{\partial f}{\partial v} = \sum_{j=1}^d \frac{\partial f}{\partial x_j} \frac{d}{dt} (x_{0,j} + tv_j) = \sum_{j=1}^d \frac{\partial f}{\partial x_j} v_j = \langle \nabla f, v \rangle.$$

Распишем скалярное произведение:

$$\langle \nabla f, v \rangle = \|\nabla f\| \|v\| \cos \varphi = \|\nabla f\| \cos \varphi,$$

где  $\varphi$  — угол между градиентом и вектором  $v$ . Таким образом, производная по направлению будет максимальной, если угол между градиентом и направлением равен нулю, и минимальной, если угол равен 180 градусам. Иными словами, производная по направлению максимальна вдоль градиента и минимальна вдоль антиградиента.

Покажем теперь, что градиент ортогонален линиям уровня. Пусть  $x_0$  — некоторая точка,  $S(x_0) = \{x \in \mathbb{R}^d \mid f(x) = f(x_0)\}$  — соответствующая линия уровня. Разложим функцию в ряд Тейлора на этой линии в окрестности  $x_0$ :

$$f(x_0 + \varepsilon) = f(x_0) + \langle \nabla f, \varepsilon \rangle + o(\|\varepsilon\|),$$

где  $x_0 + \varepsilon \in S(x_0)$ . Поскольку  $f(x_0 + \varepsilon) = f(x_0)$  (линия уровня же), получим

$$\langle \nabla f, \varepsilon \rangle = o(\|\varepsilon\|).$$

Поделим обе части на  $\varepsilon$ :

$$\left\langle \nabla f, \frac{\varepsilon}{\|\varepsilon\|} \right\rangle = o(1).$$

Устремим  $\|\varepsilon\|$  к нулю. При этом вектор  $\frac{\varepsilon}{\|\varepsilon\|}$  будет стремиться к касательной к линии уровня в точке  $x_0$ . В пределе получим, что градиент ортогонален этой касательной.

## §2.2 Векторное дифференцирование

При аналитическом вычислении градиента крайне полезны формулы векторного дифференцирования. Выведем простейшие из них.

**Задача 2.1.** Покажите, что

$$\nabla_x \langle a, x \rangle = a.$$

**Решение.** Найдем производную по  $j$ -й координате:

$$\frac{\partial}{\partial x_j} \langle a, x \rangle = \frac{\partial}{\partial x_j} \sum_{k=1}^d a_k x_k = a_j.$$

Значит, градиент равен  $a$ . ■

**Задача 2.2.** Покажите, что

$$\nabla_x \|x\|_2^2 = 2x.$$

**Решение.** Найдем производную по  $j$ -й координате:

$$\frac{\partial}{\partial x_j} \|x\|_2^2 = \frac{\partial}{\partial x_j} \sum_{k=1}^d x_k^2 = 2x_j.$$

Значит, градиент равен  $2x$ . ■

**Задача 2.3.** Покажите, что

$$\nabla_x \langle Ax, x \rangle = (A + A^T)x,$$

где  $A \in \mathbb{R}^{d \times d}$ .

**Решение.** Распишем интересующую нас функцию:

$$\begin{aligned} \langle Ax, x \rangle &= \sum_{j=1}^d (Ax)_j x_j = \sum_{j=1}^d \left( \sum_{k=1}^d a_{jk} x_k \right) x_j = \\ &= \sum_{j=1}^d \sum_{k=1}^d a_{jk} x_j x_k = \sum_{j=1}^d a_{jj} x_j^2 + \sum_{j \neq k} a_{jk} x_j x_k. \end{aligned}$$

Найдем частную производную по  $i$ -й координате:

$$\begin{aligned} \frac{\partial}{\partial x_i} \langle Ax, x \rangle &= \frac{\partial}{\partial x_i} \sum_{j=1}^d a_{jj} x_j^2 + \frac{\partial}{\partial x_i} \sum_{j \neq k} a_{jk} x_j x_k = \\ &= \frac{\partial}{\partial x_i} \sum_{j=1}^d a_{jj} x_j^2 + \frac{\partial}{\partial x_i} \left( \sum_{j \neq i} a_{ij} x_i x_j + \sum_{j \neq i} a_{ji} x_i x_j \right) = \\ &= 2a_{ii} x_i + \sum_{j \neq i} a_{ij} x_j + \sum_{j \neq i} a_{ji} x_j = \sum_{j=1}^d a_{ij} x_j + \sum_{j=1}^d a_{ji} x_j = \\ &= (Ax)_i + (A^T x)_i \end{aligned}$$

Получаем:

$$\nabla_x \langle Ax, x \rangle = Ax + A^T x = (A + A^T)x.$$

■

**Задача 2.4.** Покажите, что

$$\nabla_x \|Ax + b\|_2^2 = 2A^T(Ax + b).$$

Здесь  $x \in \mathbb{R}^d$ ,  $A \in \mathbb{R}^{m \times d}$ ,  $b \in \mathbb{R}^m$ .

**Решение.** Распишем норму:

$$\begin{aligned} \|Ax + b\|_2^2 &= \langle Ax + b, Ax + b \rangle = \langle Ax, Ax \rangle + 2\langle Ax, b \rangle + \langle b, b \rangle = \\ &= \langle A^T Ax, x \rangle + 2\langle x, A^T b \rangle + \langle b, b \rangle. \end{aligned}$$

Воспользуемся уже полученными нами формулами векторного дифференцирования:

$$\begin{aligned} \nabla_x \|Ax + b\|_2^2 &= \nabla_x \langle A^T Ax, x \rangle + \nabla_x 2\langle x, A^T b \rangle + \nabla_x \langle b, b \rangle = \\ &= (A^T A + A^T A)x + 2A^T b = 2A^T Ax + 2A^T b = \\ &= 2A^T(Ax + b). \end{aligned}$$

■

### 3 Методы оптимизации

Пусть  $Q(w)$  — функционал, представимый в виде суммы  $n$  функций:

$$Q(w) = \sum_{i=1}^n q_i(w).$$

В таком виде, например, может быть представлен функционал в линейных методах (1.2). Отдельные функции  $q_i(w)$  будут соответствовать ошибкам на отдельных объектах.

Наиболее известным является метод *градиентного спуска* (full gradient, FG) [1], в котором выбирается начальное приближение  $w^0$ , и затем до сходимости делаются шаги по антиградиенту:

$$w^k = w^{k-1} - \eta_k \nabla Q(w^{k-1}).$$

Если функционал  $Q(w)$  выпуклый, гладкий и имеет минимум  $w^*$ , то имеет место следующая оценка сходимости:

$$Q(w^k) - Q(w^*) = O(1/k).$$

Если функционал состоит из большого числа слагаемых (т.е.  $n$  велико), то градиентный спуск может оказаться слишком трудоемким. В этих случаях можно воспользоваться методом *стохастического градиента* (stochastic gradient) [2]:

$$w^k = w^{k-1} - \eta_k \nabla q_{i_k}(w^{k-1}),$$

где  $i_k$  — случайно выбранный номер слагаемого из функционала. Для выпуклого и гладкого функционала может быть получена следующая оценка:

$$\mathbb{E} [Q(w^k) - Q(w^*)] = O(1/\sqrt{k}).$$

Таким образом, метод стохастического градиента имеет менее трудоемкие итерации по сравнению с полным градиентом, но и скорость сходимости у него существенно меньше.

Недавно был предложен метод *среднего стохастического градиента* (stochastic average gradient) [3], который сочетает в себе быстроту итераций стохастического градиента и высокую скорость сходимости полного градиента. Перед началом итераций в нем выбирается начальное приближение  $w^0$ , и инициализируются вспомогательные переменные  $y_i^0$ , соответствующие градиентам слагаемых функционала:

$$y_i^0 = \nabla q_i(w^0), \quad i = 1, \dots, n.$$

На  $k$ -й итерации выбирается случайное слагаемое  $i_k$  и обновляются вспомогательные переменные:

$$y_i^k = \begin{cases} \nabla q_i(w^{k-1}), & \text{если } i = i_k; \\ y_i^{k-1} & \text{иначе.} \end{cases}$$

Иными словами, пересчитывается один из градиентов слагаемых. Наконец, делается градиентный шаг:

$$w^k = w^{k-1} - \eta_k \sum_{i=1}^n y_i^k.$$

Данный метод имеет такой же порядок сходимости для выпуклых и гладких функционалов, как и обычный градиентный спуск:

$$\mathbb{E} [Q(w^k) - Q(w^*)] = O(1/k).$$

### §3.1 Выбор параметров

Параметрами градиентного спуска являются начальное приближение  $w^0$  и темп обучения (или длина шага)  $\eta_t$ . Выбор начального приближения был подробно обсужден на лекции, мы же сосредоточимся на выборе темпа обучения.

Если на каждом шаге выбирать оптимальный темп обучения, то есть полагать его равным решению задачи

$$Q(w^t - \eta_t \nabla Q(w^t)) \rightarrow \min_{\eta_t},$$

то получим метод *наискорейшего градиентного спуска*.

Покажем, что если градиент функционала ограничен по норме, т.е.  $\|\nabla Q\| \leq D$ , то необходимым условием сходимости градиентного спуска к решению является

$$\sum_{t=0}^{\infty} \eta_t = \infty.$$

Расписывая выражение для вектора весов  $w^{t+1}$  на  $(t+1)$ -м шаге, получим

$$w^{t+1} = w^0 - \sum_{s=0}^t \eta_s \nabla Q(w^s).$$

Оценим расстояние между  $w^{t+1}$  и  $w^0$ :

$$\|w^{t+1} - w^0\| = \left\| \sum_{s=0}^t \eta_s \nabla Q(w^s) \right\| \leq \sum_{s=0}^t \eta_s \|\nabla Q(w^s)\| \leq D \sum_{s=0}^t \eta_s.$$

Предположим, что ряд шагов  $\sum_{t=0}^{\infty} \eta_t$  сходится, тогда все его частичные суммы ограничены некоторой константой  $S$ . Получаем, что

$$\|w^{t+1} - w^0\| \leq DS,$$

то есть расстояние между начальным приближением и *любой* точкой, полученной итерационным процессом, ограничено. Значит, если начальное приближение будет отстоять от решения больше, чем на  $DS$ , то градиентный спуск не сойдется к решению.

В пакете `vowpal wabbit`, реализующем настройку и применение линейных моделей, используется следующая формула для шага в градиентном спуске:

$$\eta_t = \lambda d^k \left( \frac{s_0}{s_0 + t} \right)^p,$$

где  $\lambda$ ,  $d$ ,  $s_0$  и  $p$  — параметры. Через  $d$  обозначен номер прохода по обучающей выборке, который в данный момент производится методом оптимизации. На практике достаточно настроить параметр  $\lambda$ , а остальным присвоить разумные значения по умолчанию:  $s_0 = 1$ ,  $p = 0.5$ ,  $d = 1$ .

**Масштабирование признаков.** Для градиентного спуска крайне важно, чтобы признаки имели одинаковый масштаб. Если это не так, то скорость сходимости метода значительно уменьшается.

Также проблемы могут возникнуть, если один из признаков принимает большие значения, а у функции потерь имеется горизонтальная асимптота. Тогда производная функции потерь на большом значении скалярного произведения  $\langle w, x \rangle$  будет близка к нулю, и градиентный спуск застрянет на данном значении вектора параметров  $w$ . Это явление называется «параличем» оптимизационного метода. Чтобы избежать его, следует нормировать все признаки.

## Список литературы

- [1] *Cauchy, M. A.* (1847). Méthode générale pour la résolution des systèmes d'équations simultanées. // Comptes rendus hebdomadaires des séances de l'Académie des sciences, 25, p. 536-538.
- [2] *Robbins, H., Monro S.* (1951). A stochastic approximation method. // Annals of Mathematical Statistics, 22 (3), p. 400-407.
- [3] *Schmidt, M., Le Roux, N., Bach, F.* (2013). Minimizing finite sums with the stochastic average gradient. // Arxiv.org.