

# Семинары по линейным классификаторам

Евгений Соколов

27 октября 2013 г.

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:

$$w = \arg \min_{w \in \mathbb{R}^d} Q(w, X^\ell). \quad (0.1)$$

## 1 Градиент функции

Как правило, оптимизационная задача (0.1) 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:

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

### §1.1 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.

## §1.2 Векторное дифференцирование

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

**Задача 1.1.** *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 1.2.** 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 1.3.** 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 1.4.** 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:

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

■

## 2 Геометрия линейных классификаторов

Уравнение  $\langle w, x \rangle = 0$  задает гиперплоскость вектором нормали  $w$ . Если вектор  $x$  находится по одну сторону этой гиперплоскости, то он относится к классу  $+1$ , иначе к классу  $-1$ .

**Задача 2.1.** Пусть гиперплоскость проходит через начало координат и задана уравнением  $\langle w, x \rangle = 0$ . Покажите, что евклидово расстояние между ней и точкой  $x_0$  равно

$$\frac{|\langle w, x_0 \rangle|}{\|w\|}.$$

**Решение.** Искомое расстояние можно найти как решение задачи условной оптимизации

$$\begin{cases} \|x - x_0\|_2^2 \rightarrow \min_{x \in \mathbb{R}^d} \\ \langle w, x \rangle = 0 \end{cases} \quad (2.1)$$

Здесь мы ищем такую точку на прямой, что расстояние от нее до  $x_0$  минимально; иными словами, мы ищем проекцию. Расстояние от точки до ее проекции на прямую и является расстоянием от точки до прямой.

Выпишем лагранжиан:

$$\mathcal{L} = \|x - x_0\|^2 + \lambda \langle w, x \rangle.$$

Найдем его градиент по  $x$  и приравняем его нулю:

$$\nabla_x \mathcal{L} = 2(x - x_0) + \lambda w = 0.$$

Это условие, вкупе с ограничением  $\langle w, x \rangle = 0$ , является необходимым условием для решения задачи (2.1).

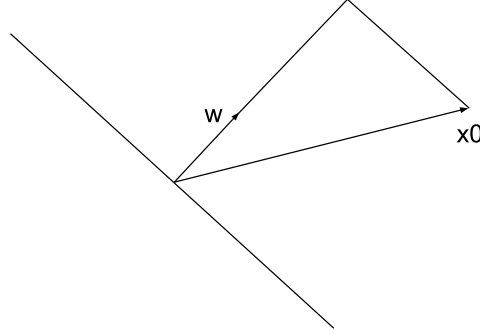


Рис. 1. Вычисление расстояния от точки до прямой.

Выразим отсюда  $x$ :  $x = x_0 - \frac{\lambda w}{2}$ . Подставим его в ограничение:

$$\langle w, x_0 - \frac{\lambda w}{2} \rangle = \langle w, x_0 \rangle - \frac{\lambda}{2} \langle w, w \rangle = \langle w, x_0 \rangle - \frac{\lambda}{2} \|w\|^2 = 0.$$

Отсюда получаем:

$$\lambda = \frac{2\langle w, x_0 \rangle}{\|w\|^2}.$$

Подставим обратно в выражение для  $x$ :

$$x = x_0 - \frac{\lambda w}{2} = x_0 - \langle w, x_0 \rangle \frac{w}{\|w\|^2}.$$

Мы нашли проекцию точки  $x_0$  на прямую. Найдём теперь расстояние:

$$\|x - x_0\|_2 = \|\langle w, x_0 \rangle \frac{w}{\|w\|^2}\| = |\langle w, x_0 \rangle| \frac{\|w\|}{\|w\|^2} = \frac{|\langle w, x_0 \rangle|}{\|w\|}.$$

■

Таким образом, если вектор весов нормирован, то скалярное произведение объекта на этот вектор равно расстоянию от объекта до разделяющей гиперплоскости. Это объясняет, почему величина отступа  $M(x_i) = \langle w, x_i \rangle y_i$  характеризует уверенность классификатора в объекте: чем больше отступ, тем дальше от гиперплоскости расположен объект.

Смысл полученной формулы легко понять, если рассмотреть двумерное пространство. Рассмотрим прямоугольный треугольник, образованный векторами  $w$  и  $x_0$  (см. рис. 1). Для косинуса угла между ними выполнено

$$\cos(w, x_0) = \frac{\|\text{proj}_w(x_0)\|}{\|x_0\|},$$

где  $\text{proj}_w(x_0)$  — проекция  $x_0$  на  $w$ . Длина же данной проекции равна расстоянию от  $x_0$  до прямой с нормалью  $w$ . Значит, расстояние можно записать как

$$\|\text{proj}_w(x_0)\| = \|x_0\| \cos(w, x_0) = \frac{\|x_0\| \|w\| \cos(w, x_0)}{\|w\|} = \frac{|\langle w, x_0 \rangle|}{\|w\|},$$

то есть его можно вычислять по полученной нами формуле.

### 3 Методы оптимизации

Простейшим способом оптимизации функционала ошибки  $Q(w, X^\ell)$  является градиентный спуск и его модификация, стохастический градиентный спуск. В градиентном спуске выбирается начальное приближение  $w^{(0)}$ , и затем до сходимости делаются шаги по антиградиенту:

$$w^{(t+1)} = w^{(t)} - \eta_t \nabla Q(w^{(t)}).$$

Если функционал представляет собой сумму по всем объектам ( $Q(w, X^\ell) = \sum_{i=1}^{\ell} L(w, x_i)$ ), то для его оптимизации можно использовать стохастический градиентный спуск, в котором на каждом выбирается случайный объект выборки, и в градиенте оставляется лишь соответствующее ему слагаемое. Данный метод позволяет успешно находить минимум функционала качества даже при очень большом числе признаков и объектов ( $> 10^5$ ).

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*.

**Задача 3.1.** Рассмотрим функционал, оптимизируемый на одном шаге стохастического градиентного спуска:

$$Q(w) = (\langle w, x_i \rangle - y_i)^2.$$

Шаг итерационного процесса имеет вид

$$w^{(t+1)} = w^{(t)} - \eta_t (\langle w^{(t)}, x_i \rangle - y_i) x_i$$

(убедитесь, что он действительно такой). Покажите, что в наискорейшем спуске длина шага выбирается как  $\eta_t = \frac{1}{\|x_i\|^2}$ .

**Решение.** Длина шага выбирается как решение задачи

$$(\langle w^{(t)} - \eta_t (\langle w^{(t)}, x_i \rangle - y_i) x_i, x_i \rangle - y_i)^2 \rightarrow \min_{\eta_t}.$$

Найдем производную и приравняем ее нулю:

$$\begin{aligned} \frac{d}{d\eta_t} (\langle w^{(t)} - \eta_t (\langle w^{(t)}, x_i \rangle - y_i) x_i, x_i \rangle - y_i)^2 &= \\ &= \frac{d}{d\eta_t} (\langle w^{(t)}, x_i \rangle - \eta_t (\langle w^{(t)}, x_i \rangle - y_i) \langle x_i, x_i \rangle - y_i)^2 = \\ &= 2 (\langle w^{(t)}, x_i \rangle - \eta_t (\langle w^{(t)}, x_i \rangle - y_i) \langle x_i, x_i \rangle - y_i) (\langle w^{(t)}, x_i \rangle - y_i) \langle x_i, x_i \rangle = \\ &= 0 \end{aligned}$$

Выразим отсюда  $\eta_t$ :

$$\eta_t = \frac{(\langle w^{(t)}, x_i \rangle - y_i) \|x_i\|^2}{(\langle w^{(t)}, x_i \rangle - y_i) \|x_i\|^4} = \frac{1}{\|x_i\|^2}.$$

■

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.

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

Часть текста про масштабирование сырая и будет дописана в будущем, а сейчас используйте на свой страх и риск! :)

## 4 Регуляризация

В некоторых случаях (признаков больше чем объектов, коррелирующие признаки) оптимизационная задача  $Q(w) \rightarrow \min$  может иметь бесконечное число решений, большинство которых являются переобученными и плохо работают на тестовых данных. Данная проблема устраняется путем добавления *регуляризатора* к функционалу:

$$Q_{\tau}(w) = Q(w) + \tau R(w).$$

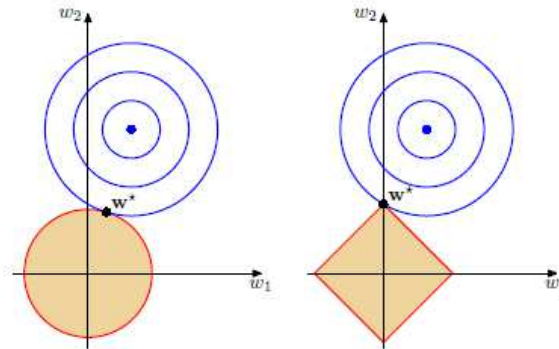


Рис. 2. Линии уровня функционала качества, а также ограничения, задаваемые  $L_2$  и  $L_1$ -регуляризаторами.

Наиболее распространенными являются  $L_2$  и  $L_1$ -регуляризаторы:

$$\|w\|_2 = \sum_{i=1}^d w_i^2,$$

$$\|w\|_1 = \sum_{i=1}^d |w_i|.$$

Особенностью  $L_1$ -регуляризатора является то, что он зануляет часть весов, осуществляя тем самым отбор признаков. Попробуем понять, почему это так.

Можно показать, что если функционал  $Q(w)$  является выпуклым, то задача безусловной минимизации функции  $Q(w) + \tau\|w\|_1$  эквивалентна задаче условной оптимизации

$$\begin{cases} Q(w) \rightarrow \min_w \\ \|w\|_1 \leq C \end{cases}$$

для некоторого  $C$ . На рис. 2 изображены линии уровня функционала  $Q(w)$ , а также множество, определяемое ограничением  $\|w\|_1 \leq C$ . Решение определяется точкой пересечения допустимого множества с линией уровня. В большинстве случаев эта точка будет лежать на одной из вершин ромба, что соответствует решению с одной зануленной компонентой.