Метод стохастического градиента

Градиентный метод численной минимизации

Минимизация эмпирического риска (регрессия, классификация):

$$Q(w) = \sum_{i=1}^{\ell} \mathscr{L}_i(w) \to \min_w .$$

Численная минимизация методом *градиентного спуска*:

$$w^{(0)} := \text{начальное приближение;}$$

$$w^{(t+1)} := w^{(t)} - h \cdot \nabla Q(w^{(t)}), \qquad \nabla Q(w) = \left(\frac{\partial Q(w)}{\partial w_j}\right)_{j=0}^{n},$$

где h — *градиентный шаг*, называемый также *темпом обучения*.

$$w^{(t+1)} := w^{(t)} - h \sum_{i=1}^{\ell} \nabla \mathscr{L}_i\bigl(w^{(t)}\bigr).$$

Идея ускорения сходимости:

брать (x_i, y_i) по одному и сразу обновлять вектор весов.

Алгоритм SG (Stochastic Gradient)

Вход: выборка X^ℓ, темп обучения h, темп забывания λ;
Выход: вектор весов w;

инициализировать веса $w_j,\ \ j = 0, \ldots, n$;
инициализировать оценку функционала:
    $\bar{Q} := \frac{1}{\ell} \sum_{i=1}^{\ell} \mathscr{L}_i(w)$;
повторять
    выбрать объект x_i из X^ℓ случайным образом;
    вычислить потерю: $\varepsilon_i := \mathscr{L}_i(w)$;
    сделать градиентный шаг: $w := w - h\nabla \mathscr{L}_i(w)$;
    оценить функционал: $\bar{Q} := \lambda\varepsilon_i + (1-\lambda)\bar{Q}$;
пока значение $\bar{Q}$ и/или веса w не сойдутся;

Robbins, H., Monro S. A stochastic approximation method // Annals of Mathematical Statistics, 1951, 22 (3), p. 400–407.

Откуда взялась рекуррентная оценка функционала?

Проблема: вычисление оценки Q по всей выборке $x_1, \dots, x_\ell$ намного дольше градиентного шага по одному объекту x_i.

Решение: использовать приближённую рекуррентную формулу.

Среднее арифметическое:

$$\bar{Q}_m = \frac{1}{m}\varepsilon_m + \frac{1}{m}\varepsilon_{m-1} + \frac{1}{m}\varepsilon_{m-2} + \dots$$

$$\bar{Q}_m = \frac{1}{m}\varepsilon_m + \left(1 - \frac{1}{m}\right)\bar{Q}_{m-1}$$

Экспоненциальное скользящее среднее:

$$\bar{Q}_m = \lambda\varepsilon_m + (1-\lambda)\lambda\varepsilon_{m-1} + (1-\lambda)^2\lambda\varepsilon_{m-2} + \dots$$

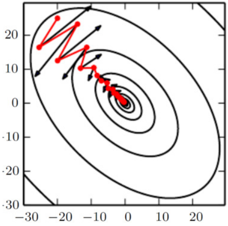
$$\bar{Q}_m = \lambda\varepsilon_m + (1-\lambda)\bar{Q}_{m-1}$$

Параметр λ — *темп забывания* предыстории ряда.

Метод накопления инерции (momentum)

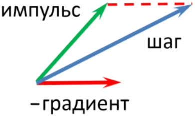
Momentum — экспоненциальное скользящее среднее градиента по последним $\approx \frac{1}{1-\gamma}$ итерациям [Б.Т.Поляк, 1964]:

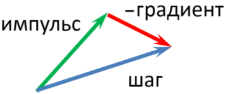
$$v := \gamma v + (1-\gamma)\mathscr{L}_i'(w)$$
$$w := w - hv$$



NAG (Nesterov's accelerated gradient) — стохастический градиент с инерцией [Ю.Е.Нестеров, 1983]:

$$v := \gamma v + (1-\gamma)\mathscr{L}_i'(w - h\gamma v)$$
$$w := w - hv$$





Варианты инициализации весов

1. $w_j := 0$ для всех $j = 0, \dots, n$;
2. небольшие случайные значения:
   $w_j := \text{random}\left(-\frac{1}{2n}, \frac{1}{2n}\right)$;
3. $w_j := \frac{\langle y, f_j \rangle}{\langle f_j, f_j \rangle}$, $\; f_j = \bigl(f_j(x_i)\bigr)_{i=1}^{\ell}$ — вектор значений признака.

   Эта оценка w оптимальна, если
   1) функция потерь квадратична и
   2) признаки некоррелированы, $\langle f_j, f_k \rangle = 0$, $\; j \neq k$.
4. обучение по небольшой случайной подвыборке объектов;
5. мультистарт: многократные запуски из разных случайных начальных приближений и выбор лучшего решения.

Варианты порядка предъявления объектов

Возможны варианты:

1. *перетасовка объектов* (shuffling):
   попеременно брать объекты из разных классов;
2. чаще брать объекты, на которых ошибка больше:
   чем меньше M_i, тем больше вероятность взять объект;
3. чаще брать объекты, на которых уверенность меньше:
   чем меньше $|M_i|$, тем больше вероятность взять объект;
4. вообще не брать «хорошие» объекты, у которых $M_i > \mu_+$
   (при этом немного ускоряется сходимость);
5. вообще не брать объекты-«выбросы», у которых $M_i < \mu_-$
   (при этом может улучшиться качество классификации);

Параметры μ_+, μ_- придётся подбирать.

Варианты выбора градиентного шага

1. сходимость гарантируется (для выпуклых функций) при
$$h_t \to 0, \quad \sum_{t=1}^{\infty} h_t = \infty, \quad \sum_{t=1}^{\infty} h_t^2 < \infty,$$
в частности можно положить $h_t = 1/t$;
2. *метод скорейшего градиентного спуска*:
$$\mathscr{L}_i\bigl(w - h\nabla\mathscr{L}_i(w)\bigr) \to \min_h,$$
позволяет найти *адаптивный шаг* h^*;
При квадратичной функции потерь $h^* = \|x_i\|^{-2}$.
3. пробные случайные шаги для «выбивания» итерационного процесса из локальных минимумов;
4. метод Левенберга-Марквардта (второго порядка)

Диагональный метод Левенберга-Марквардта

Метод Ньютона-Рафсона, $\mathscr{L}_i(w) \equiv \mathscr{L}(\langle w, x_i\rangle y_i)$:

$$w := w - h\big(\mathscr{L}_i''(w)\big)^{-1}\nabla\mathscr{L}_i(w),$$

где $\mathscr{L}_i''(w) = \left(\frac{\partial^2\mathscr{L}_i(w)}{\partial w_j \partial w_{j'}}\right)$ — гессиан, $n\times n$-матрица

Эвристика. Считаем, что гессиан диагонален:

$$w_j := w_j - h\left(\frac{\partial^2\mathscr{L}_i(w)}{\partial w_j^2} + \mu\right)^{-1}\frac{\partial\mathscr{L}_i(w)}{\partial w_j},$$

h — темп обучения, можно полагать $h = 1$

μ — параметр, предотвращающий обнуление знаменателя.

Отношение h/μ есть темп обучения на ровных участках функционала $\mathscr{L}_i(w)$, где вторая производная обнуляется.

Проблема переобучения

Возможные причины переобучения:

- слишком мало объектов; слишком много признаков;
- линейная зависимость (мультиколлинеарность) признаков:
  пусть построен классификатор: $a(x, w) = \operatorname{sign}\langle w, x\rangle$;
  мультиколлинеарность: $\exists u \in \mathbb{R}^n$: $\forall x_i \in X^\ell \;\; \langle u, x_i\rangle = 0$;
  неединственность решения: $\forall \gamma \in \mathbb{R} \quad a(x, w) = \operatorname{sign}\langle w + \gamma u, x\rangle$.

Проявления переобучения:

- слишком большие веса $|w_j|$ разных знаков;
- неустойчивость дискриминантной функции $\langle w, x\rangle$;
- $Q(X^\ell) \ll Q(X^k)$;

Основной способ уменьшить переобучение:

- регуляризация (сокращение весов, weight decay);

Регуляризация (сокращение весов)

Штраф за увеличение нормы вектора весов:

$$\widetilde{\mathscr{L}}_i(w) = \mathscr{L}_i(w) + \frac{\tau}{2}\|w\|^2 = \mathscr{L}_i(w) + \frac{\tau}{2}\sum_{j=1}^{n} w_j^2 \to \min_w .$$

Градиент:

$$\nabla\widetilde{\mathscr{L}}_i(w) = \nabla\mathscr{L}_i(w) + \tau w.$$

Модификация градиентного шага:

$$w := w(1 - h\tau) - h\nabla\mathscr{L}_i(w).$$

Методы подбора коэффициента регуляризации τ:

1. скользящий контроль;
2. стохастическая адаптация;
3. двухуровневый байесовский вывод.

SG: Достоинства и недостатки

Достоинства:

1. легко реализуется;
2. легко обобщается на любые $g(x, w)$, $\mathscr{L}(a, y)$;
3. легко добавить регуляризацию
4. возможно динамическое (потоковое) обучение;
5. на сверхбольших выборках можно получить неплохое решение, даже не обработав все (x_i, y_i);
6. подходит для задач с большими данными

Недостатки:

1. подбор комплекса эвристик является искусством (не забыть про переобучение, застревание, расходимость)