Метод стохастического градиента (Stochastic Gradient)

Достоинства:

1. легко реализуется;
2. применим к любым моделям и функциям потерь;
3. допускает онлайновое (потоковое) обучение;
4. на сверхбольших выборках позволяет получать неплохие решения, даже не обработав все (x_i, y_i);
5. всё чаще применяется для Big Data.

Недостатки:

1. возможно застревание в локальных экстремумах;
2. возможна расходимость или медленная сходимость;
3. возможно переобучение;
4. подбор комплекса эвристик является искусством.

Варианты инициализации весов

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

   эта оценка w оптимальна при квадратичной функции потерь, если признаки некоррелированы, $\langle f_j, f_k \rangle = 0$, $j \neq k$.
4. $w_j := \ln \frac{\sum_i [y_i = +1] f_j(x_i) \; \sum_i [y_i = -1]}{\sum_i [y_i = -1] f_j(x_i) \; \sum_i [y_i = +1]}$;

   эта оценка w оптимальна для задач классификации, $Y = \{-1, +1\}$, если признаки независимы.
5. оценки w_j по небольшой случайной подвыборке объектов;
6. 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 те объекты, на которых была допущена бóльшая ошибка
   (чем меньше M_i, тем больше вероятность взять объект)
   (чем меньше $|M_i|$, тем больше вероятность взять объект);
3. вообще не брать «хорошие» объекты, у которых $M_i > \mu_+$
   (при этом немного ускоряется сходимость);
4. 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*:
$$\mathscr{L}_i\big(w - h\nabla\mathscr{L}_i(w)\big) \to \min_h,$$
позволяет найти *адаптивный шаг* h^*;

   при квадратичной функции потерь $h^* = \|x_i\|^{-2}$;
3. периодически можно делать пробные случайные шаги для «выбивания» 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. Тогда

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

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.