Линейные классификаторы и случайные блуждания

Евгений Соколов
ВМК МГУ

13 апреля 2013 г.

Определения

- $\mathbb{X} = \{x_1, \ldots, x_L\}$ — конечное генеральное множество объектов;
- $A = \{a_1, \ldots, a_D\}$ — конечное множество алгоритмов;
- $I(a, x) = [\text{алгоритм } a \text{ ошибается на объекте } x]$ — индикатор ошибки;
- $m(a, X)$ — *число ошибок* алгоритма a на выборке X.
- $\nu(a, X) = m(a, X)/|X|$ — *частота ошибок* алгоритма a на выборке X.
- *Метод обучения* $\mu : 2^{\mathbb{X}} \to A$ по произвольной выборке $X \subset \mathbb{X}$ выбирает алгоритм $a \in A$.
- *Вероятность переобучения:*

$$Q_\varepsilon = \mathsf{P}[\nu(\mu X, \bar{X}) - \nu(\mu X, X) \geqslant \varepsilon] = \frac{1}{C_L^\ell} \sum_{X \sqcup \bar{X}} [\nu(\mu X, \bar{X}) - \nu(\mu X, X) \geqslant \varepsilon]$$

Комбинаторные оценки вероятности переобучения имеют вид

$$B_\varepsilon = \sum_{a \in \mathbb{A}} b(a),$$

где a — вершины графа расслоения-связности.

Рассмотрим более общую задачу.

- $G = (V, E)$ — граф.
- $f : V \to \mathbb{R}$ — функция на его вершинах.
- $\boldsymbol{p} = (p(v))_{v \in V}$ — распределение на вершинах.

Требуется оценить величину

$$F = \mathbb{E}_{\boldsymbol{p}} f = \sum_{v \in V} f(v) p(v).$$

Требуется оценить величину

$$F = \mathbb{E}_{\boldsymbol{p}} f = \sum_{v \in V} f(v)p(v).$$

1. Допустим, мы умеем выбирать вершину v из распределения $\boldsymbol{q} = (q(v))_{v \in V}$ на V.
   - Можно использовать оценку
   $$\hat{F} = \frac{1}{n} \sum_{i=1}^{n} f(v_i) \frac{p(v_i)}{q(v_i)}.$$
   - Оценка является несмещенной и состоятельной.
2. Зачастую такое сэмплирование невозможно (наш случай!).
   - Стартуем из одной из вершин графа, ходим по нему, набирая вершины.
   - Как организовать ходьбу по графу так, чтобы были известны вероятности попадания в каждую вершину?
   - Случайные блуждания!
   - Формализуются с помощью аппарата марковских цепей.

- Последовательность случайных величин $\{V_t \in V \,|\, t = 0, 1, 2, \dots\}$ называется марковской цепью, если

$$\mathbb{P}(V_{t+1} = v \,|\, V_1 = v_1, \dots, V_t = v_t) = \mathbb{P}(V_{t+1} = v \,|\, V_t = v_t).$$

- V_t — вершина, в которой блуждание находится в момент t.
- Переходы возможны только по ребрам.
- Цепь называется *неприводимой*, если из любой ее вершины можно попасть в любую с ненулевой вероятностью:

$$\forall v, u \in V \;\; \exists s: \;\; \mathbb{P}(V_s = v \,|\, V_0 = u) > 0.$$

- Цепь называется *апериодической*, если для всех вершин v выполнено

$$k_v = \gcd\{t \,|\, \mathbb{P}(V_t = v \,|\, V_0 = v) > 0\} = 1.$$

- Вероятности перехода:

$$p_{vu} = \mathbb{P}(V_{t+1} = v \mid V_t = u).$$

- Матрица перехода: $P = (p_{vu})_{v,u \in V}$.
- Стационарное распределение:

$$P^T \pi = \pi.$$

- Если цепь неприводимая и апериодическая, то у нее гарантированно есть стационарное распределение.
- Если цепь неприводимая, но имеет период, то у нее есть стационарное распределение в обобщенном смысле, для которого все дальнейшие результаты также верны.

Требуется оценить величину

$$F = \mathbb{E}_{\boldsymbol{p}} f = \sum_{v \in V} f(v) p(v).$$

Подойдет ли следующая оценка?

$$\mu_n(f) = \frac{1}{n} \sum_{i=1}^{n} f(V_i).$$

Теорема (Закон больших чисел для марковских цепей)

Пусть $\{V_t\}$ — неприводимая марковская цепь со стационарным распределением $\boldsymbol{\pi}$. Тогда для любого начального распределения $\mathbb{P}(V_0 = v)$ и для любой функции $f : V \to \mathbb{R}$, такой что $\mathbb{E}_{\boldsymbol{\pi}} |f| < \infty$, выполнено

$$\mu_n(f) \xrightarrow[n \to \infty]{\text{a. s.}} \mathbb{E}_{\boldsymbol{\pi}} f.$$

- Итак, оценка

$$\mu_n(f) = \frac{1}{n}\sum_{i=1}^{n} f(V_i)$$

является состоятельной для $\mathbb{E}_{\boldsymbol{\pi}} f$.

- Как перейти от $\mathbb{E}_{\boldsymbol{\pi}}$ к $\mathbb{E}_{\boldsymbol{p}}$?
- Зададим корректирующую функцию

$$w(v) = \frac{p(v)}{\pi(v)}$$

и построим оценку

$$\mu_n(wf) = \frac{1}{n}\sum_{i=1}^{n} w(V_i) f(V_i).$$

- Можно показать, что

$$\mu_n(wf) \xrightarrow[n\to\infty]{\text{a. s.}} \mathbb{E}_{\boldsymbol{p}} f.$$

- Конкретный случай:

$$p_{vu} = \begin{cases} \frac{1}{d(v)}, & \text{если } (v,u) \in E; \\ 0, & \text{иначе.} \end{cases}$$

- Такое блуждание называется *простым*.
- Известно стационарное распределение: $\pi(v) = \frac{\deg(v)}{2|E|}$.
- Весовая функция имеет вид

$$w(v) = \frac{p(v)}{\pi(v)} = \frac{2|E|}{|V|}\frac{1}{\deg(v)}$$

- Не знаем $|V|$ и $|E|$!
- Рассмотрим оценку

$$\hat{F}_{\mathrm{RW}} = \frac{\mu_n(wf)}{\mu_n(w)} = \frac{\sum_{i=1}^n w(V_i)f(V_i)}{\sum_{i=1}^n w(V_i)} = \frac{\sum_{i=1}^n f(V_i)/\deg(V_i)}{\sum_{i=1}^n 1/\deg(V_i)}.$$

- Оценка является состоятельной и несмещенной.

Модификация оценки

- Вклады алгоритмов из разных слоев графа расслоения-связности отличаются на порядки;
- Дисперсия функции f большая, и для оценивания $\mathbb{E}_{\boldsymbol{p}} f$ требуется большая выборка;
- При этом в пределах одного слоя дисперсия мала!
- Новая оценка:

$$\hat{F}'_{\mathsf{RW}} = \frac{1}{|V|} \sum_{m=1}^{L} \hat{f}_m \hat{c}_m,$$

где f_m — среднее значение функции f по m-му слою,
c_m — мощность m-го слоя.

Подробности:

- Пусть $C_1, \ldots, C_L$ — слои графа, $c_1, \ldots, c_L$ — их мощности.
- Преобразуем матожидание:

$$F = \mathbb{E}_{\boldsymbol{p}} f = \frac{1}{|V|} \sum_{v \in V} f(v) = \frac{1}{|V|} \sum_{m=1}^{L} \left(\frac{1}{c_m} \sum_{v \in C_m} f(v) \right) c_m.$$

- Новая оценка:

$$\hat{F}'_{\mathsf{RW}} = \frac{1}{|V|} \sum_{m=1}^{L} \hat{f}_m \hat{c}_m;$$

$$\hat{f}_m = \frac{\sum_{i=1}^{n} \frac{f(V_i)}{\deg(V_i)} [m(V_i) = m]}{\sum_{i=1}^{n} \frac{1}{\deg(V_i)} [m(V_i) = m]};$$

$$\hat{c}_m = \frac{|V|}{n} \sum_{i=1}^{n} \left(\frac{1}{\deg(V_i)} \sum_{u \in N(V_i)} [m(u) = m] \right).$$

Frontier Sampling

Вход: Граф $G = (V, E)$; число итераций N;
набор стартовых вершин $P = (v^1, \ldots, v^s)$;
Выход: Выборка вершин графа $v_1, v_2, \ldots, v_N$;

для $i = 1, \ldots, N$
  выбрать $v \in P$ с вероятностью $\frac{\deg(v)}{\sum_{u \in P} \deg(u)}$;
  с вероятностью $\frac{1}{2}$
    выбрать вершину v' из равномерного
    распределения на $\{v' \in V \mid (v, v') \in E\}$;
    $v_i := v'$;
  иначе
    $v' := v$; $v_i := v$;
  Заменить в P вершину v на v';

Основная задача: построение непереобученных композиций линейных классификаторов вида

$$a(x) = \operatorname{sign} \sum_{i=1}^{p} \operatorname{th}\langle w_i, x\rangle.$$

Каждый базовый классификатор строится в подпространстве малой размерности.

Пусть известна некоторая оценка вероятности переобучения

$$\mathsf{P}[\nu(\mu X, \bar{X}) - \nu(\mu X, X) \geqslant \varepsilon] \leqslant \eta(\varepsilon).$$

Обратив ее, можно оценить частоту ошибок на контроле: с вероятностью не менее $1 - \eta$

$$\nu(\mu X, \bar{X}) \leqslant \nu(\mu X, X) + \varepsilon(\eta).$$

Величину в правой части неравенства будем использовать в качестве критерия при отборе признаков.

Построение базовых алгоритмов

- Базовый алгоритм — линейный классификатор, настроенный методом SVM;
- Классификатор строится по отобранным признакам;
- Отбор признаков: жадное наращивание с минимизацией комбинаторного критерия:

$$\nu(\mu X, X) + \varepsilon(1/2) \to \min .$$

- Для вычисления комбинаторной оценки на каждом наборе признаков используем случайные блуждания.

Построение композиции методом ComBoost

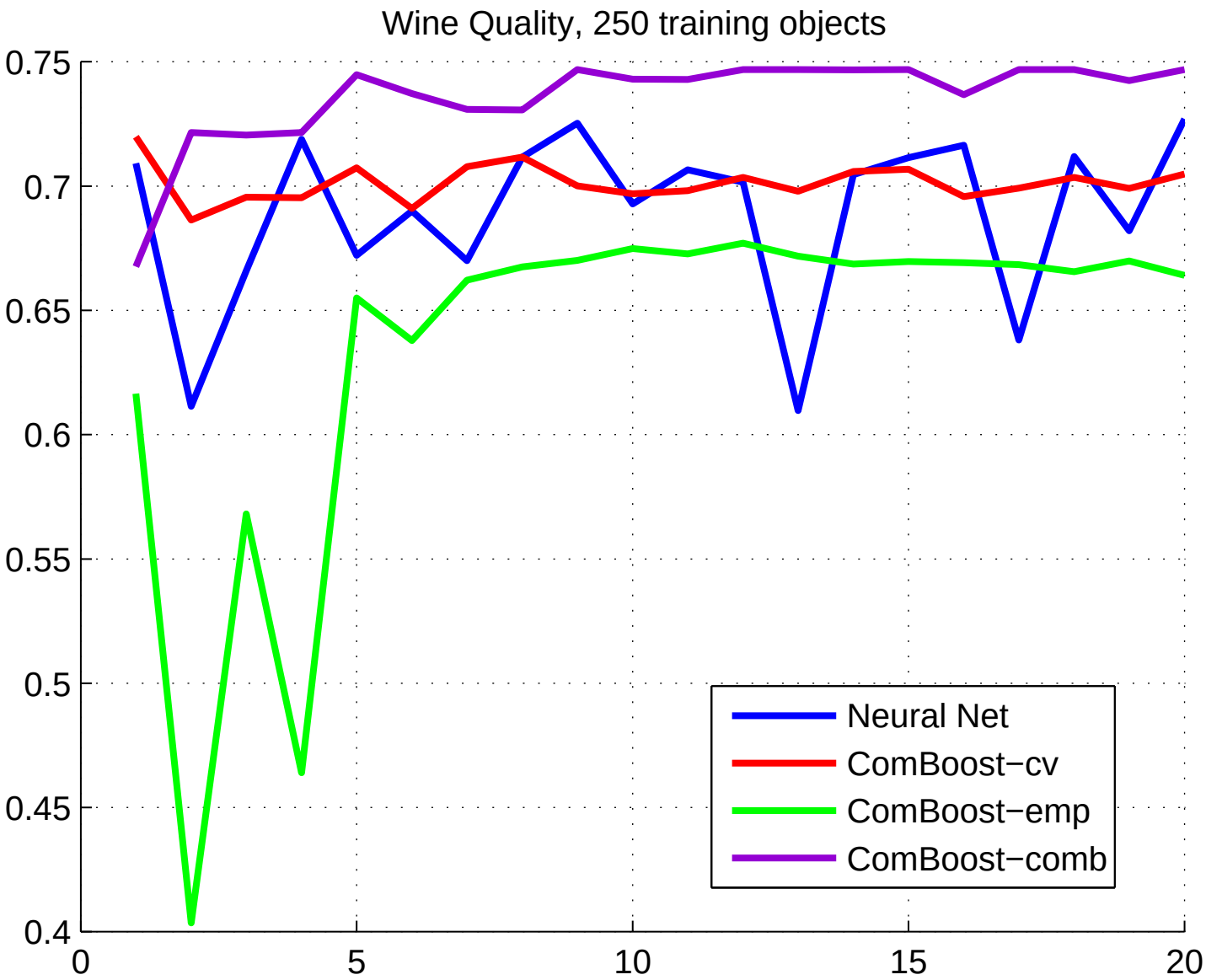
Вход: Выборка X; параметры T, ℓ_0, ℓ_1;
Выход: Базовые линейные классификаторы $b_1, \ldots, b_T$

1: Инициализировать веса и отсутпы:
$w_i = 1$, $M_i = 0$ для всех $i = 1, \ldots, \ell$;
2: **для** $t = 1, \ldots, T$, пока не выполнен критерий останова
3: Обучить базовый алгоритм (SVM по отобранным признакам):
$b_t := \operatorname{argmin}_b Q(b, X, W)$;
4: Обновить значения отступов:
$M_i = M_i + y_i b_t(x_i)$ для всех $i = 1, \ldots, L$;
5: упорядочить выборку X по возрастанию отступов M_i;
6: Отобрать объекты для обучения следующего базового алгоритма:
$w_i = [\ell_0 \leqslant i \leqslant \ell_1]$;

	Wine Quality	Statlog	Waveform	Faults
Оценка, основанная на эмпирическом риске $\nu(\mu, \mathbb{X})$	64,70	84,92	84,78	73,39
Двухслойная нейронная сеть, настроенная методом backpropagation	72,06	85,41	86,79	74,51
Оценка (1), в которой обращается комбинаторная оценка [Воронцов, 2010], вычисленная с помощью случайных блужданий	69,48	86,26	85,77	77,81
Оценка скользящего контроля, вычисленная по 100 случайным разбиениям	71,06	85,26	86,38	75,76
Оценка (1), в которой обращается комбинаторная оценка с истоками, вычисленная с помощью случайных блужданий	74,68	86,75	86,91	74,03

$$Q_c = \nu(a_0, X) + \varepsilon\left(\tfrac{1}{2}\right) \qquad (1)$$

Зависимость ошибки от длины композиции



Конец!

Вопросы?