Лекция 2

Задачи прогнозирования, Линейная машина, Теоретические методы оценки обобщающей способности,

Лектор – *Сенько Олег Валентинович*

Курс «Математические основы теории прогнозирования»
4-й курс, III поток

Содержание лекции

Пример модели распознавания - Линейная машина

Множество алгоритмов $\widetilde{W} = \{A : \widetilde{X} \to \widetilde{Y}\}$, внутри которого производится поиск оптимального алгоритма прогнозирования, вместе со способом решения оптимизационной задачи будем называть методом прогнозирования или методом распознавания, если прогнозируемая величина принадлежит конечному множеству. В качестве примера рассмотрим известный метод решения задачи распознавания – Линейная машина

Пример модели распознавания - Линейная машина

Метод «Линейная машина» предназначен для решения задачи распознавания с классами $K_1, \ldots, K_L$. Алгоритм распознавания имеет следующий вид. В процессе обучения классам $K_1, \ldots, K_L$ ставятся в соответствие линейные функции от переменных $X_1, \ldots, X_n$:

$$f_1(X_1, \ldots, X_n) = w_0^1 + w_1^1 X_1 + \ldots + w_n^1 X_n$$

$$\ldots\ldots\ldots\ldots\ldots\ldots$$

$$f_L(X_1, \ldots, X_n) = w_0^L + w_1^L X_1 + \ldots + w_n^L X_n.$$

Таким образом алгоритм распознавания задаётся матрицей параметров $\begin{pmatrix} w_0^1 & \dots & w_n^1 \\ \dots & \dots & \dots \\ w_0^L & \dots & w_n^L \end{pmatrix}$

Пусть требуется распознать объект s^* , описание которого задаётся вектором $\boldsymbol{x}^*$. Вычисляются значения функций $f_1, \dots, f_L$ в точке $\boldsymbol{x}^*$. Объект s^* будет отнесён классу K_i, если выполняется набор неравенств

$$f_i(\boldsymbol{x}^*) > f_j(\boldsymbol{x}^*),$$

где $j \in \{1, \dots, L\} \backslash \{i\}$.

Линейная машина

Максимальная точность на выборке $\tilde{S}_t$ соответствует выполнению максимального числа блоков неравенств:

$$f_{J(1)}(\boldsymbol{x}_1) > f_i(\boldsymbol{x}_1), i \in \{1, \ldots, L\} \backslash \{J(1)\} \tag{1}$$

$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$

$$f_{J(m)}(\boldsymbol{x}_m) > f_i(\boldsymbol{x}_m), i \in \{1, \ldots, L\} \backslash \{J(m)\}.$$

Каждый из блоков соответствует одному из объектов выборки $\tilde{S}_t$ и включает $L-1$ неравенств. Таким образом суммарное число неравенств во всех блоках составляет $m(L-1)$. Каждое из неравенств из системы (1) соответствует сравнению оценки вектора $\boldsymbol{x}_r$ за класс $K_{J(r)}$ с оценкой за класс $K_i \neq K_{J(r)}$.

Линейная машина. Обучение

Рассмотрим неравенство t системы, соответствующее блоку с номером r, в котором производится сравнение оценки за класс $K_{J(r)}$ с оценкой за класс K_i. Очевидно, что t и i связаны равенствами:

$$t = (r-1)(L-1) + j, j < J(r)$$

$$t = (r-1)(L-1) + j - 1, j > J(r)$$

Неравенство с номером t можно переписать в виде

$$\sum_{i=1}^{L}\sum_{h=1}^{n} z_h^{it} w_h^i > \sum_{i=1}^{L} w_0^i z_0^{it},$$

При этом $z_h^{it} = x_h r$ и $z_0^{it} = 1$ при $i = J(r)$,
$z_h^{it} = -x_h r$ и $z_0^{it} = -1$ при $i = j$. $z_h^{it} = 0$ и $z_0^{it} = 0$ при $i \neq j$ и $i \neq J(r)$.

То есть мы получаем систему неравенств:

$$\sum_{i=1}^{L}\sum_{h=1}^{n} z_h^{it} w_h^i > \sum_{i=1}^{L} w_0^i z_0^{it}, t = 1, \ldots, m(L-1) \qquad (2)$$

При этом коэффициенты из множества $\{z_h^{it} \mid i = 1, \ldots, L, h = 1, \ldots, n\}$ однозначно выражаются через t. Для поиска максимальной совместной подсистемы блоков неравенств системы (2) используется релаксационный алгоритм На начальном этапе каждое из уравнений системы (2) нормируется на величину $D_t = \sqrt{\sum_{i=0}^{L}\sum_{h=0}^{n}(z_h^{it})^2}$

Линейная машина. Обучение

В результате от системы неравенств (2) мы переходим к системе

$$\sum_{i=1}^{L}\sum_{h=1}^{n}\hat{z}_h^{it}w_h^i > \sum_{i=1}^{L}\hat{z}_h^{0t}, t = 1,\ldots,m(L-1) \qquad (3)$$

где $\hat{z}_h^{it} = z_h^{it}/D_t$,$h = 0,\ldots,n, i = 1,\ldots,L$ Релаксационный алгоритм состоит в вычислении релаксационной последовательности матриц искомых коэффициентов $\{w_h^j \mid j = 1,\ldots,n; h = 1,\ldots,n\}$:

$$\widetilde{\mathbf{W}^0}, \widetilde{\mathbf{W}^1}, \ldots, \widetilde{\mathbf{W}^k}, \ldots$$

Линейная машина. Обучение

При этом на итерации k производится коррекция матрицы $\widetilde{\mathbf{W}}^k$, полученных на предыдущей итерации

$$\widetilde{\mathbf{W}}^{k+1} = \widetilde{\mathbf{W}}^k + \mu^k \times \Delta^k,$$

где скалярая величина μ^k и матрица Δ^k вычисляются по невыполненным неравенствам из системы (3). Пусть $\widetilde{I}^{((k)}$ - множество неравенств, которые остались невыполненными на итерации k-1. Тогда $\Delta^{(k)} = \sum_{t \in \widetilde{I}^{((k)}} d_t$, где d_t- матрица размерности $(n+1)L$, в позиции (h, j) которой стоит коэффициент перед w_h^j в уравнении с номером t из системы (3).

Линейная машина. Обучение

Коэффициент μ_k пропорционален суммарной величине нарушения неравенств из набора $\widetilde{I}^{((k)}$, нормированной на сумму квадратов коэффициентов матрицы $\Delta^{(k)}$

$$\mu_k = \frac{\sum_{t \in \widetilde{I}^k} \{\sum_{i=1}^{L} \hat{z}_0^{it} - \sum_{i=1}^{L} \sum_{h=1}^{n} \hat{z}_h^{it} w_h^i\}}{\sum_{i=1}^{L} \sum_{h=1}^{n} (\Delta_{ij})^2} \tag{4}$$

Линейная машина. Обучение

Процесс поиска решений. Задаётся произвольная начальная точка. В начале каждой итерации подсчитывается число полностью выполненных блоков неравенств. Если оно максимально относительно всех предыдущих итераций, то текущее приближение $\widetilde{\mathbf{W}^k}$ запоминается как лучшее на данный момент решение. Процесс продолжается до выполнения одного из критериев остановки:

- Отсутствие невыполненных блоков неравенств;
- Число итераций превысило некоторую заранее заданную величину;
- В течение нескольких итераций число полностью выполненных блоков неравенств не изменяется.

Линейная машина. Пример

Имеется задача распознавания с 3-я классами и 2-я признаками. Предполагается, что с использованием метода ЛМ для каждого класса найдены линейные разделяющие функции:

- $f_1(X_1, X_2) = 4 + 2X_1 - X_2$;
- $f_2(X_1, X_2) = -2 + X_1 - 3X_2$;
- $f_3(X_1, X_2) = 1 + X_1 - 2X_2$.

Область, где одновременно выполняются неравенства

- $f_1(X_1, X_2) > f_2(X_1, X_2)$;
- $f_1(X_1, X_2) > f_3(X_1, X_2)$;

соответствует классу 1.

Линейная машина. Пример

Последняя система эквивалентна неравенствам

- $6 + X_1 + 2X_2 > 0$ (I),
- $3 + X_1 + X_2 > 0$ (II).

Данные неравенства задают граничные прямые на плоскости, которые обозначены римскими цифрами (I) и (II) соответственно. Область на плоскости, соответствующая классу 1, обозначена красными квадратиками. Предположим, что точка на плоскости не принадлежит классу 1. Тогда она принадлежит классу 2, если выполняется неравенство:

$$f_1(X_1, X_2) > f_3(X_1, X_2),$$

которое эквивалентно неравенству $X_2 < -3$. Область на плоскости, соответствующая классу 2, обозначена зелёными треугольниками. Область, соответствующая классу 3 обозначена синими кружками.

Линейная машина.

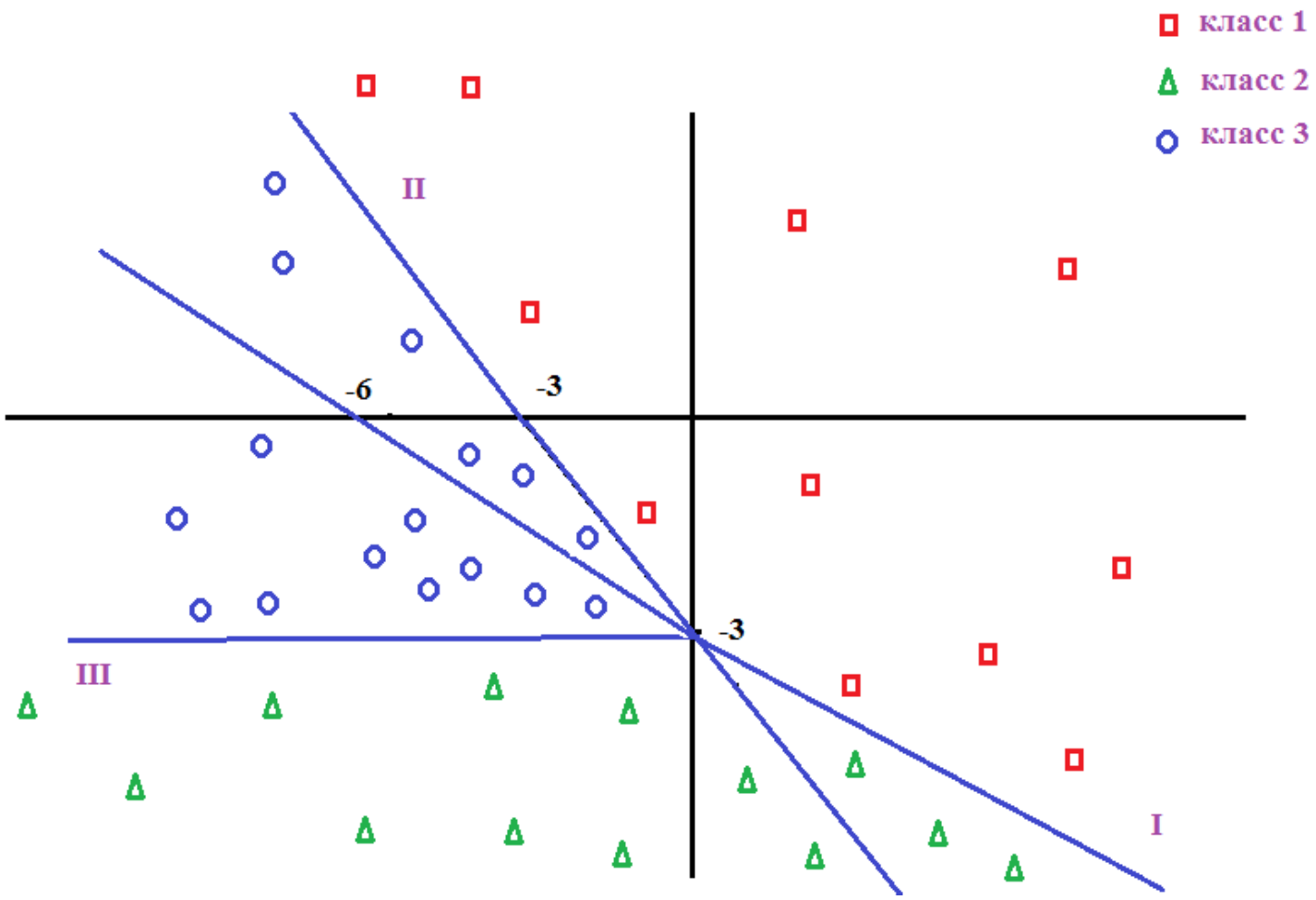

Рис 1. Пример распознавания с помощью метода - Линейная машина

Теоретические подходы к исследованию обобщающей способности

Обобщающая способность (ОС) алгоритма прогнозирования может быть эффективно оценена по выборке данных с помощью методов:

- оценивание ОС на новой контрольной выборке
- Кросс-проверка
- Скользящий контроль

Однако большой интерес представляют теоретические методы оценки обобщающей способности, которые позволили бы ответить на вопросы: Будет ли обладать достаточной обобщающей способностью алгоритм прогнозирования, найденный внутри некоторой модели $\tilde{M} = \{A : \tilde{X} \to \tilde{Y}\}$? Какие требования необходимо предъявить к $\tilde{M}$, чтобы обеспечить эффективное обучение?
Ответы на данные вопросы даёт теория Вапника-Червоненкиса

Теоретические подходы к исследованию обобщающей способности

Далее будет рассмативается задача распознавания. Предположим, что по обучающей выборке $\widetilde{S}_t$ внутри модели $\widetilde{M}$ найден оптимальный алгоритм A_{opt} с минимальной долей ошибок на $\widetilde{S}_t$ - $\nu_{err}(A_{opt})$. - Достижение высокой обучающей способности соответствует низкой доле ошибок на всей генеральной совокупности или, иными словами, низкой вероятности ошибок для алгоритма A_{opt}. . Теория Вапника-Червоненкиса устанавливает условия, которым должна удовлетворять $\widetilde{M}$ для гарантированной сходимости частоты ошибки оптимального обученного алгоритма к вероятности этой ошибки при возрастании объёма обучающей выборки

Теория Вапника-Червоненкиса

Пусть k - число ошибочных классификаций, сделанных на обучающей выборке $\widetilde{S}_t$ длины m некоторым агоритмом A. Частота ошибок $\nu_{err}(A) = k/m$ распределена по биномиальному закону

$$P[\nu_{err}(A)] = C_m^k [p_{err}(A)]^k [1 - p_{err}(A)]^{m-k}$$

где $p_{err}(A)$ - вероятность ошибочной классификации для алгоритма A.

Вероятность выполнения неравенства $\mid \nu_{err}(A) - p_{err}(A) \mid > \varepsilon$ задаётся равенством

$$P\{\mid \nu_{err}(A) - p_{err}(A) \mid > \varepsilon\} = \tag{5}$$

$$= \sum_{|k'/m - p_{err}(A)| > \varepsilon} C_m^{k'} [p_{err}(A)]^{k'} [1 - p_{err}(A)]^{(m-k')}$$

Для оценки сверху вероятности $P\{|\ \nu_{err}(A) - p_{err}(A)\ | > \varepsilon\}$ при больших m может быть использована интегральная теоремы Муавра–Лапласа: при $m \to \infty$ оказывается справедливым неравенство

$$P\{|\ \nu_{err}(A) - p_{err}(A)\ | > \varepsilon\} \leq \frac{2\sigma}{\sqrt{2m\pi}\varepsilon} e^{\frac{-\varepsilon^2 m}{2\sigma^2}},$$

где $\sigma^2 = [1 - p_{err}(A)]p_{err}(A)$. Поскольку $p_{err} \in [0, 1]$, то нетрудно показать, что $\sigma^2 < \frac{1}{4}$. В результате убеждаемся в справедливости неравенств при $m \to \infty$

$$P\{|\ \nu_{err}(A) - p_{err}(A)\ | > \varepsilon\} \leq \frac{1}{\sqrt{2m\pi}\varepsilon} e^{-2\varepsilon^2 m}, \tag{6}$$

Нетрудно видеть, что правая часть неравенства (5) быстро стремится к 0 при $m \to \infty$.

Таким образом, для каждого отдельного алгоритма распознавания на обучающей выборке частота ошибки быстро сходится к вероятности ошибки при $m \to \infty$. На самом деле в процессе обучения оценивается большое число всевозможных алгоритмов модели $\tilde{M}$. Алгоритмы с минимальной частотой ошибки могут соответствовать как раз очень высоким отклонения частот от вероятностей. Достижение высокой обобщающей способности гарантируется при выполнении условия равномерной сходимости:
при произвольном $\varepsilon > 0$

$$P\{\max_{A \in \widetilde{M}}[|\ \nu_{err}(A) - p_{err}(A)\ |] > \varepsilon\} \to 0 \tag{7}$$

при $m \to \infty$.

Обозначим как $\widetilde{A}_{\varepsilon m}$ событие, заключающееся в выполнении для алгоритма A неравенства $\mid \nu_{err}(A) - p_{err}(A) \mid > \varepsilon$ на обучающей выборке $\widetilde{S}_t$ длины m . Тогда, принимая во внимание неравенство Буля, получаем

$$P\{\max_{A\in\widetilde{M}} \mid \nu_{err}(A) - p_{err}(A) \mid > \varepsilon\} = P(\cup_{A\in\widetilde{M}}\widetilde{A}_{\varepsilon m}) \leq \sum_{A\in\widetilde{M}} P(\widetilde{A}_{\varepsilon m}) \quad (8)$$

Принимая во внимание неравенство (8) и (6) получаем

$$P\{\max_{A\in\widetilde{M}} \mid \nu_{err}(A) - p_{err}(A) \mid > \varepsilon\} \leq \sum_{A\in\widetilde{M}} \frac{1}{\sqrt{2m\pi\varepsilon}} e^{-2\varepsilon^2 m} \quad (9)$$

Сначала рассмотрим случай когда модель $\widetilde{M}$ конечна и содержит N различных алгоритмов. Тогда очевидно

$$P\{\max_{A\in\widetilde{M}} \mid \nu_{err}(A) - p_{err}(A) \mid > \varepsilon\} \leq \frac{N}{\sqrt{2m\pi\varepsilon}} e^{-2\varepsilon^2 m} \tag{10}$$

В теории Вапника-Червоненкиса предлагается использовать для оценки разнообразия модели $\widetilde{M}$ конечное множество $\widetilde{M}_dif \subseteq \widetilde{M}$, обладающее следующими свойствами:

- множества ошибок для любых двух различных алгоритмов из $\widetilde{M}_{dif}$ на обучающей выборке не совпадают;
- мощность любого конечного подмножества $\widetilde{M}$, обладающего первым свойством, не превышает мощности $\widetilde{M}_{dif}$.

Число таких алгоритмов задаётся коэффициентом разнообразия $\mathbf{\Delta}(\widetilde{M}, \widetilde{S}_t)$, который определяется как число способов, которыми $\widetilde{S}_t$ может быть разбита на две подвыборки алгоритмами из модели $\widetilde{M}$. Для оценок наличия равномерной сходимости при обучении по модели используется функция роста $\mu(\widetilde{M}, m)$: максимальное значение коэффициентов разнообразия на множестве Ω_m всевозможных обучающих выборок длины m:

$$\mu(\widetilde{M}, m) = \max_{\widetilde{S}_t \in \Omega_m} \mathbf{\Delta}(\widetilde{M}, \widetilde{S}_t) \tag{11}$$

Учитывая, что число отличных друг от друга алгоритмов в указанном ранее смысле ограничено сверху функцией роста, получаем верхнюю оценку вероятности выполнения неравенства $\mid \nu_{err}(A) - p_{err}(A) \mid > \varepsilon$.

$$P\{\max_{A \in \widetilde{M}} \mid \nu_{err}(A) - p_{err}(A) \mid > \varepsilon\} \leq \frac{\mu(\widetilde{M}, m)}{\sqrt{2m\pi}\varepsilon} e^{-2\varepsilon^2 m} \tag{12}$$

Функция роста обладает следующим замечательным свойством. Существует два типа моделей распознавания. Для произвольной модели первого типа $\widetilde{M}_1$ при любом заданном m существует такая выборка $\widetilde{S}^*$, содержащая m объектов, что произвольное разбиение $\widetilde{S}^*$ на два подмножества может быть реализовано алгоритмами из $\widetilde{M}_1$. Иными словами при произвольном m справедливо равенство $\mu(\widetilde{M}, m) = 2^m$. Для произвольной модели второго типа $\widetilde{M}_2$ существует такое натуральное $m_/$, что отсутствуют выборки, разделимые на два произвольных подмножества алгоритмами из $\widetilde{M}_2$. Иными словами существует такое натуральное m, что

$$\mu(\widetilde{M}_2, m) < 2^m. \tag{13}$$

Предположим, что m'- минимальное m, при котором справедливо неравенство (13). В этом случае считается, что ёмкость $\widetilde{M}_2$ конечна и равна $m_* = m' - 1$.

Считается также, что ёмкость произвольной модели первого типа является бесконечной.
Было показано, что для произвольной модели второго типа $\widetilde{M}_2$ при произвольном $m > m_*$ для функции роста справедливо ограничение сверху

$$\mu(\widetilde{M}_2, m) \leq 1.5 \frac{m^{(m_*-1)}}{(m_* - 1)!} \tag{14}$$

Из следует, что $\mu(\widetilde{M}_2, m)$ ограничена сверху полиномом степени $m_* - 1$. Однако при произвольных вещественном $\alpha > 0$ и натуральном $k > 0$ для произвольного положительного полинома $Pol(m, k)$ степени k от аргумента m

$$\lim_{m \to \infty} \frac{Pol(m, k)}{e^{\alpha m}} \to 0. \tag{15}$$

Из (13) и (14) следует, что при произвольном $\varepsilon > 0$

$$\lim_{m \to \infty} \frac{\mu(\widetilde{M}_2, m)}{e^{2m\varepsilon^2}\sqrt{2\pi m}} \to 0. \tag{16}$$

Из стремления к 0 правой части неравенства (12) следует

$$P\{\max_{A\in\widetilde{M}} \mid \nu_{err}(A) - p_{err}(A) \mid > \varepsilon\} \to 0$$

при $m \to \infty$,
что означает выполнение условия равномерной сходимости. Таким образом для любой модели, имеющей конечную ёмкость, получение алгоритмов, обладающих обобщающей способностью является гарантированным при достаточно больших объёмах обучающих выборок. Бесконечная ёмкость модели не позволяет сделать вывод о наличии обобщающей способности даже при очень больших объёмах обучающей выборки.