МАТЕМАТИЧЕСКИЕ ОСНОВЫ ТЕОРИИ ПРОГНОЗИРОВАНИЯ

Лектор

Сенько Олег Валентинович

Лекция 2

Методы прогнозирования (распознавания)

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

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 функция от переменных $X_1,\ldots,X_n$

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

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

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

:

Линейная машина

Таким образом алгоритм распознавания задаётся матрицей параметров

$$\begin{pmatrix} w_0^1 \, w_1^1 \, \dots \, w_n^1 \\ \dots\dots\dots\dots\dots \\ w_0^L \, w_1^L \, \dots w_n^L \end{pmatrix}$$

Пусть требуется распознать объект S^*, описание которого задаётся вектором $\mathbf{x}^*$. Вычисляются значения функций $f_1,\dots,f_L$ в точке $\mathbf{x}^*$. Объект S^* будет отнесён классу $K_i, i \in \{1,\dots,L\}$, если выполняется набор неравенств

,

$$f_i(\mathbf{x}^*) > f_j(\mathbf{x}^*),\ j \in \{1,\dots,L\} \setminus \{i\}$$

Линейная машина

Метод обучения

Процесс обучения по выборке $\tilde{S}_t = \{(y_1, \mathbf{x}_1), \ldots, (y_m, \mathbf{x}_m)\}$ состоит в поиске таких значений параметров,

$$\begin{pmatrix} w_0^1 \, w_1^1 \ldots \, w_n^1 \\ \ldots\ldots\ldots\ldots \\ w_0^L \, w_1^L \ldots w_n^L \end{pmatrix}$$

при которых максимальное число объектов $\tilde{S}_t$ оказывается правильно распознанным, Пусть $J(j) = i$, если $y_j = K_i$. Максимальная точность на выборке $\tilde{S}_t$ соответствует выполнению максимального числа блоков неравенств.

Линейная машина

Метод обучения

$$f_{J(1)}(\mathbf{x}_1) > f_j(\mathbf{x}_1),\ j = \{1,\ldots,L\} \setminus \{J(1)\}$$

$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \qquad (1)$$

$$f_{J(m)}(\mathbf{x}_m) > f_j(\mathbf{x}_m),\ j = \{1,\ldots,L\} \setminus \{J(m)\}$$

Каждый из блоков включает $l-1$ неравенств. Таким образом суммарное число неравенств составляет $m(l-1)$.

. Неравенства из системы (1) могут быть записаны в виде

$$\sum_{i=1}^{2L}\sum_{j=1}^{n} z_j^{it} w_j^i > \sum_{i=1}^{2L} z_0^{it} \qquad t \in \{1,\ldots,m(l-1)\} \qquad (2)$$

Линейная машина

Метод обучения

Если в уравнении t системы (2), соответствующим номеру блока r $J(r)=i$, то $z_j^{it}=x_{rj}, z_0^{it}=1$

Если в уравнении t системы (2), соответствующим номеру Блока r $j=i-L$, то $z_j^{it}=-x_{rj}, z_0^{it}=-1$

Во всех остальных случаях коэффициенты z_j^{it} равны 0

Для поиска максимальной совместной подсистемы блоков неравенств системы (1) используется релаксационный алгоритм.

Линейная машина

Метод обучения. Релаксационный алгоритм.

- На начальном этапе каждое из уравнений системы (2)
- нормируется на величину

$$z_j^{it} \sqrt{\sum_{j=1}^{(L*N)} (z_j^{it})^2}$$

- В результате мы переходим к системе неравенств

$$\sum_{i=1}^{2L} \sum_{j=1}^{n} \hat{z}_j^{it} w_j^i > \sum_{i=1}^{2L} \hat{z}_0^{it} \qquad t \in \{1,\ldots,m(l-1)\} \qquad (3)$$

- Релаксационный алгоритм состоит в вычислении релаксационной последовательности матриц параметров w:

$\Delta^{(i)}$

$$\tilde{\mathbf{w}}^{(0)}, \tilde{\mathbf{w}}^{(1)}, \ldots, \tilde{\mathbf{w}}^{(i)}, \ldots$$

Линейная машина

Метод обучения. Релаксационный алгоритм.

При этом $\tilde{\mathbf{w}}^{(i+1)} = \tilde{\mathbf{w}}^{(i)} + \kappa^{(i)} * \Delta^{(i)}$, где скаляр $\kappa^{(i)}$ и матрица $\tilde{\Delta}^{(i)}$ вычисляются по невыполненным неравенствам из системы

Пусть $\tilde{I}^{(i)}$- множество неравенств невыполненных на i- ой итерации. Тогда $\tilde{\Delta}^{(i)} = \sum_{t \in \tilde{I}^{(i)}} \tilde{d}_t$, где $\tilde{d}_t$ - матрица

размерности $(n+1) \times L$, в позиции (i, j) которой стоит коэффициент перед w_j^i в t-ом уравнении системы (3).

Линейная машина

Метод обучения. Релаксационный алгоритм.

- Коэффициент $\kappa^{(i)}$ пропорционален суммарной величине нарушения неравенств из набора $\tilde{I}^{(i)}$, нормированной на сумму квадратов коэффициентов матрицы $\tilde{\Delta}^{(i)}$

$$\kappa^{(i)} = \frac{\sum\limits_{t \in \tilde{I}^i} \{ \sum\limits_{i=1}^{2L} \hat{z}_0^{it} - \sum\limits_{i=1}^{2L} \sum\limits_{j=1}^{n} \hat{z}_j^{it} w_j^i \}}{\sum\limits_{i=1}^{L} \sum\limits_{j=1}^{n+1} (\Delta_{ij})^2}$$

Линейная машина

Метод обучения. Релаксационный алгоритм.

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 $\tilde{\mathbf{w}}^{(i)}$ запоминается как лучшее на данный момент решение.

Процесс продолжается до выполнения одного из критериев остановки.

Линейная машина

Метод обучения. Релаксационный алгоритм.

- Критерии остановки

1) Отсутствие невыполненных неравенств

2) Число итераций превысило некоторую заранее заданную величину

3) В течение нескольких итераций число полностью выполненных блоков неравенств не изменяется

Линейная машина.
Задача

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

$$f_1(x_1, x_2) = 4.0 + 2x_1 - x_2$$

$$f_2(x_1, x_2) = -2.0 + x_1 - 3x_2$$

$$f_3(x_1, x_2) = 1.0 + x_1 - 2x_2$$

Требуется изобразить на двумерной диаграмме области, соответствующие отнесению классам !, 2 и 3

Линейная машина.
Решение задачи

Область, где одновременно выполняются неравенства

$$1)\ f_1(x_1, x_2) > f_2(x_1, x_2)$$

$$2)\ f_1(x_1, x_2) > f_3(x_1, x_2)$$

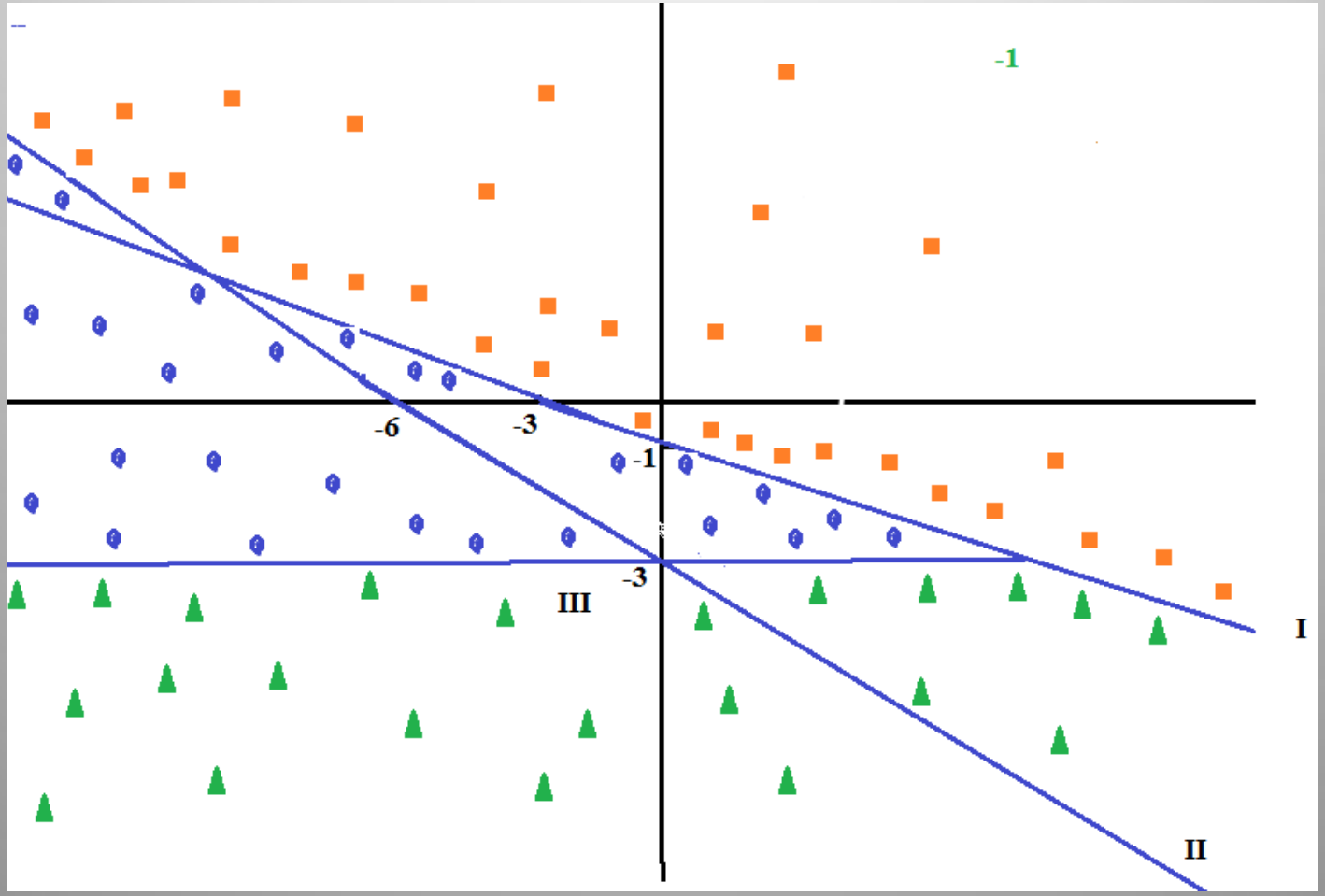
Соответствует классу 1.

Неравенства 1 и 2 эквивалентны неравенствам

$$1')\ 6 + x_1 + 2x_2 > 0$$

$$2')\ 3 + x_1 + 3x_2 > 0$$

Линейная машина.
Решение задачи



Теоретические подходы к исследованию обобщающей способности

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

А) оценивание ОС на новой контрольной выборке

Б) Кросс-проверка

В) Скользящий контроль

Теоретические подходы к исследованию обобщающей способности

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

Теория Вапника-Червоненкиса

Далее будет рассматривается задача распознавания.

Предположим, что по обучающей выборке $\tilde{S}_t$ найден алгоритм A^o с минимальной долей ошибок на $\tilde{S}_t$ - $\nu_{err}(A^o)$

Достижение высокой обучающей способности соответствует низкой доле ошибок на всей генеральной совокупности или, иными словами, низкой вероятности ошибок для алгоритма A^o .

Теория Вапника-Червоненкиса устанавливает условия гарантированной сходимости частоты ошибки к её вероятности при возрастании объёма обучающей выборк

Теоретические подходы к исследованию обобщающей способности

Теория Вапника-Червоненкиса

Пусть k - число ошибочных классификаций, сделанных на обучающей выборке длины m некоторым агоритмом A .

Частота ошибок $\nu_{err}(A)$ распределена по биномиальному закону

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

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

Теоретические подходы к исследованию обобщающей способности

Теория Вапника-Червоненкиса

Вероятность выполнения неравенства

$$|\nu_{err}(A) - p_{err}(A)| > \varepsilon$$

задаётся суммой

$$\sum_{|\frac{k'}{m} - P_{err}(A)| < \varepsilon} C_m^{k'} [p_{err}(A)]^{k'} [1 - p_{err}(A)]^{m-k'} \qquad (1)$$

Теоретические подходы к исследованию обобщающей способности

Теория Вапника-Червоненкиса

В силу интегральной теоремы Муавра–Лапласа сумма (1) при больших может быть оценена сверху с помощью выражения:

$$\frac{2\sigma}{\sqrt{2\pi m\varepsilon}} e^{\frac{-\varepsilon^2 m}{2\sigma^2}}$$

где $\sigma^2 = [1 - p_{err}(A)]p_{err}(A) \leq \frac{1}{2}$

Таким образом

$$\Pr\{|\nu_{err}(A) - p_{err}(A)| > \varepsilon\} \leq \frac{2\sigma}{\sqrt{2\pi m\varepsilon}} e^{\frac{-\varepsilon^2 m}{2\sigma^2}} \leq \frac{1}{\sqrt{2\pi m\varepsilon}} e^{-2\varepsilon^2 m} \quad (2)$$

Теоретические подходы к исследованию обобщающей способности

Теория Вапника-Червоненкиса

На самом деле в процессе обучения оценивается большое число всевозможных алгоритмов модели . 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:

$$P\{\max_{A\in\tilde{M}} |\nu_{err}(A) - p_{err}(A)| > \varepsilon\} \to 0 \quad \text{при} \quad m \to \infty$$

Теоретические подходы к исследованию обобщающей способности

Теория Вапника-Червоненкиса

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

$$P\{\max_{A\in\tilde{M}} |\nu_{err}(A) - p_{err}(A)| > \varepsilon\} = P\{\bigcup_{A\in\tilde{M}} \tilde{A}_{\varepsilon m}\} \le \sum_{A\in\tilde{M}} P\{\tilde{A}_{\varepsilon m}\}$$

Принимая во внимание неравенство (2) получаем

$$P^{l}\{\max_{A\in\tilde{A}} |\nu_{err}(A) - p_{err}(A)| > \varepsilon\} \le \sum_{A\in\tilde{M}} \frac{1}{\sqrt{2\pi m\varepsilon}} e^{-2\varepsilon^2 m} \qquad (3)$$

Теоретические подходы к исследованию обобщающей способности

Теория Вапника-Червоненкиса

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

$$P\{\max_{A\in\tilde{M}} | \nu_{err}(A) - p_{err}(A) | > \varepsilon\} \le \frac{N}{\sqrt{2\pi m\varepsilon}} e^{-2\varepsilon^2 m}$$

В теории Вапника-Червоненкиса предлагается использовать для оценки разнообразия модели $\tilde{M}$ число входящих в него алгоритмов, делающих ошибки на одних и тех же объектах обучающей выборки $\tilde{S}_t$

Теоретические подходы к исследованию обобщающей способности

Теория Вапника-Червоненкиса

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

$$\boldsymbol{\mu}(\tilde{A},m)=\max_{\tilde{S}_t\in\Omega_m}\Delta(\tilde{M},\tilde{S}_t)$$

Теоретические подходы к исследованию обобщающей способности

Теория Вапника-Червоненкиса

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

$$P\{\max_{A \in \tilde{M}} |\nu_{err}(A) - p_{err}(A)| > \varepsilon\} \leq \frac{\boldsymbol{\mu}(\tilde{M}, m)}{\sqrt{2\pi m\varepsilon}} e^{-2\varepsilon^2 m} \qquad (4)$$

Теоретические подходы к исследованию обобщающей способности

Теория Вапника-Червоненкиса

Свойства функции роста

Существует два типа моделей

Для первого типа при любом объёме m существует выборка произвольное разбиение которой на два подмножества может быть реализовано алгоритмами из $\tilde{M}$.

Иными словами $\mu(\tilde{M},m)=2^m \quad \forall m$

Для второго типа существует такой объём m^*, для которого отсутствуют выборки, разделимые на два протзвольных подмножества алгоритмами из $\tilde{M}$,

Иными словами $\exists m^* \quad \mu(\tilde{A},m^*)<2^{m^*}$

Теоретические подходы к исследованию обобщающей способности

Теория Вапника-Червоненкиса

Во втором случае говорится, что ёмкость модели $\tilde{M}$ конечна и равна m^*. В случаях, когда отсутствует такой объём выборки, при котором $\tilde{M}$ позволяет реализовать произвольное разбиение обучающей выборки на две подвыборки при любом объёме последней, считается, что ёмкость $\tilde{M}$ бесконечна ,

Было показано, что $m > m^*$ случае для функции роста справедливо ограничение сверху

$$\mu(\tilde{A},m) \le 1{,}5\frac{m^{(m^*-1)}}{(m^*-1)!}$$

Поскольку $\mu(\tilde{A},m)$ ограничено сверху полиномом конечной степени , то

$$\forall \varepsilon \quad \lim_{m\to\infty} \frac{e^{-2m\varepsilon^2}\mu(\tilde{M},m)}{\sqrt{2\pi m}} = 0$$

Теоретические подходы к исследованию обобщающей способности

Теория Вапника-Червоненкиса

Из стремления к 0 правой части неравенства (4) следует

$$P\{\max_{A\in\widetilde{M}} |\nu_{err}(A) - p_{err}(A)| > \varepsilon\} \to 0 \quad \text{при} \quad m \to \infty,$$

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 даже при очень больших объёмах