МАТЕМАТИЧЕСКИЕ ОСНОВЫ ТЕОРИИ ПРОГНОЗИРОВАНИЯ

Лектор

Сенько Олег Валентинович

Лекция I

Задачи диагностики и прогнозирования некоторой величины Y по доступным значениям переменных $X_1, \ldots, X_n$ часто возникают в различных областях человеческой деятельности:

- постановка медицинского диагноза по результатам анализов;
- прогноз результатов лечения;
- прогноз свойств ещё не синтезированного химического соединения по его молекулярной формул;
- диагностика хода технологического процесса;
- Диагностика состояния технического оборудования;
- прогноз финансовых индикаторов;
- и многие другие задачи

Типы прогнозируемых величин

Прогнозируемая величина Y может иметь различную природу:

- принимать значения из отрезка непрерывной оси;
- принимать значения из конечного множества;
- являться кривой, описывающей вероятность дожития до какого-то момента времени.

В случаях когда прогнозируемая величина является категориальной и принимает значения из множества, содержащего несколько элементов, задачу прогнозирования принято называть задачей распознавания.

Методы, основанные на обучении по прецедентам

В случаях, когда существует выборка прецедентов, для которых известны значения прогнозируемой величины Y и переменных $X_1, \ldots, X_n$ для решения задач прогнозирования могут быть использованы методы, основанные на обучении по прецедентам. Выборку прецедентов принято называть

Обучающей выборкой

Обучающая выборка имеет вид $\tilde{S}_t = \{(y_1, \mathbf{x}_1), \ldots, (y_m, \mathbf{x}_m)\}$,

где y_j - значение переменной Y для j-го объекта;

$\mathbf{x}_j$ - значение вектора переменных $X_1, \ldots, X_n$ для j-го объекта;

$j = 1, \ldots, m$;

m - число объектов в $\tilde{S}_t$;

Методы, основанные на обучении по прецедентам

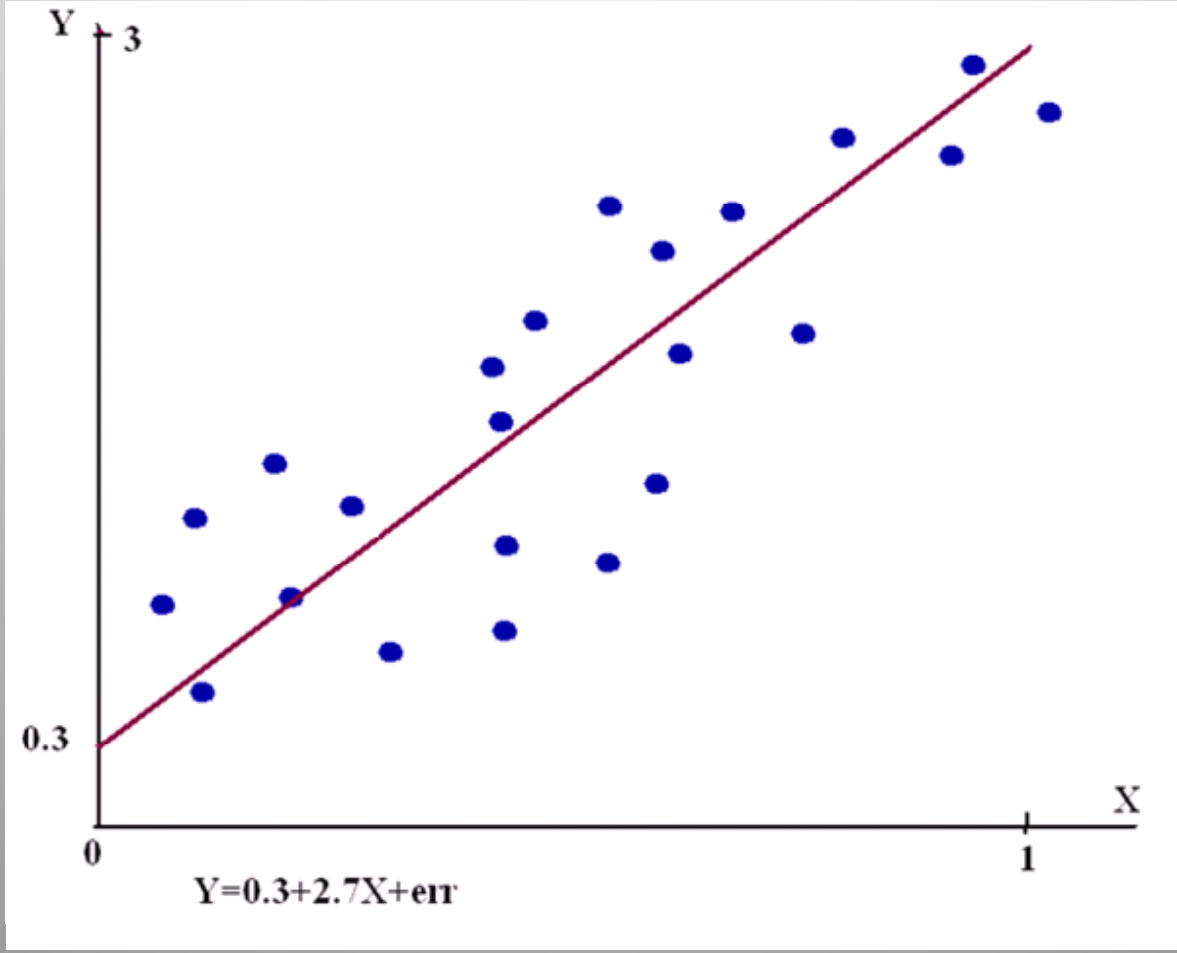
В процессе обучения производится поиск эмпирических закономерностей, связывающих прогнозируемую переменную Y с переменными $X_1, \ldots, X_n$.

Данные закономерности далее используются при прогнозировании.

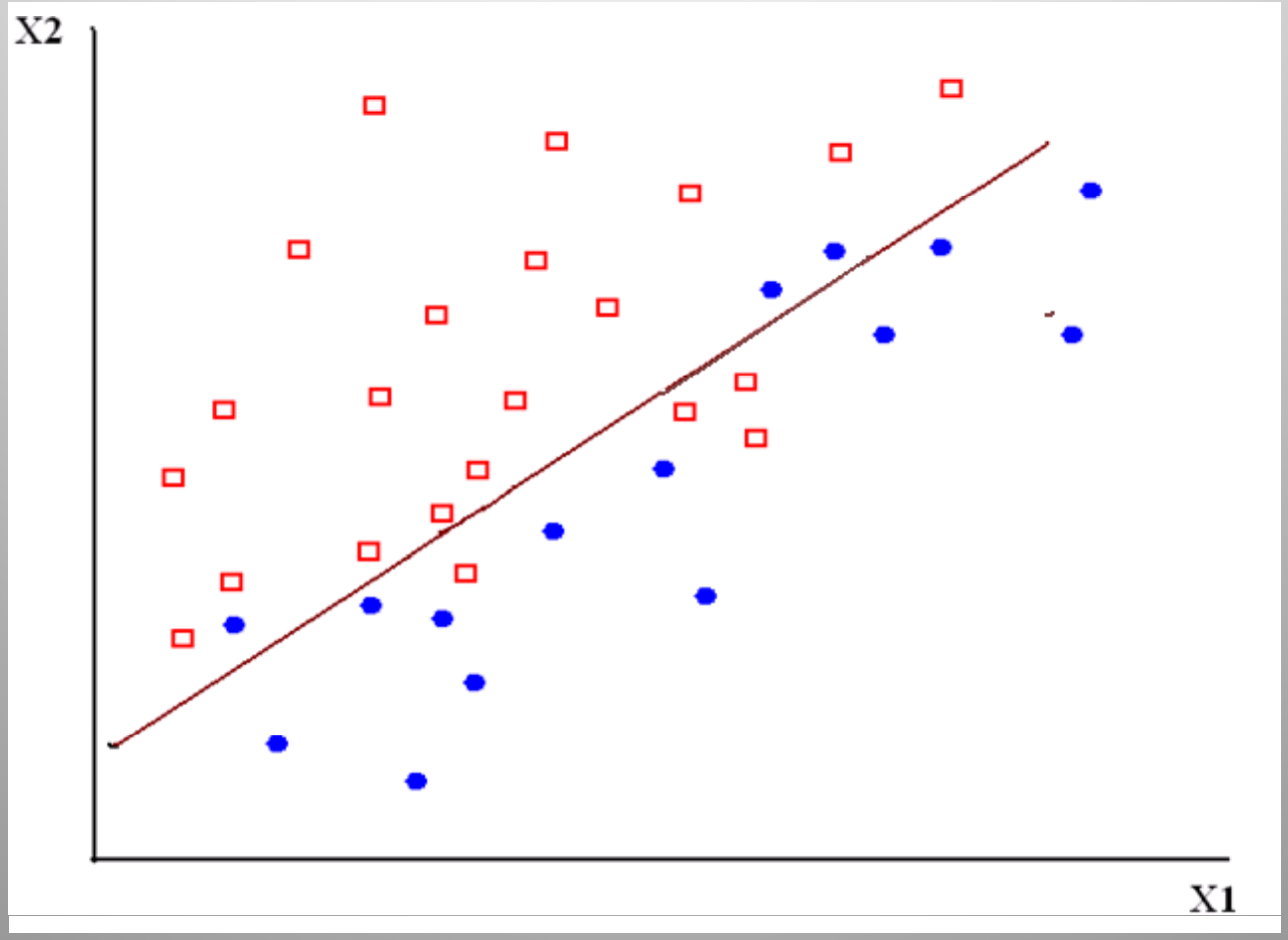
Методы, основанные на обучении по прецедентам, также принято называть

Методами машинного обучения (Machine learning)

Примеры



Примеры



Способы поиска закономерностей

Основным способом поиска закономерностей является поиск в некотором априори заданном семействе алгоритмов $\tilde{M} = \{A : \tilde{X} \to \tilde{Y}\}$ прогнозирования алгоритма, наилучшим образом аппроксимирующей связь подмножества переменных набора $X_1, \ldots, X_n$ с переменной Y на обучающей выборке,

где $\tilde{X}$ - область возможных значений векторов переменных $X_1, \ldots, X_n$, $\tilde{Y}$ - область возможных значений переменной Y.

Пусть $\lambda[y_j, A(\mathbf{x}_j)]$ - величина "потерь", произошедших в результате использования $A(\mathbf{x}_j)$ в качестве прогноза значения Y. Тогда одним из способов обучения является минимизация функционала эмпирического риска на обучающей выборке

$$Q(\tilde{S}_t, A) = \frac{1}{m} \sum_{j=1}^{m} \lambda[y_j, A(\mathbf{x}_j)]$$

Способы поиска закономерностей

Частные случаи функции потерь

$$\lambda[y_j, A(\mathbf{x}_j)] = [y_j - A(\mathbf{x}_j)]^2$$ квадрат ошибки

$$\lambda[y_j, A(\mathbf{x}_j)] = | y_j - A(\mathbf{x}_j) |$$ модуль ошибки

В случае задачи распознавания функция потерь может быть равной 0 при правильной классификации и 1 в при ошибочном. При этом функционал эмпирического риска равен числу ошибочных классификаций.

Обобщающая способность

Точность алгоритма прогнозирования на всевозможных новых не использованных для обучения объектах, которые возникают в результате процесса, соответствующего рассматриваемой задаче прогнозирования принято называть

Обобщающей способностью

Иными словами обобщающую способность алгоритма прогнозирования можно определить как точность на всей генеральной совокупности. Мерой обобщающей способности служит

$$E_{\Omega}\{\lambda[Y, A(\mathbf{x})]\} = \int_{\Omega} \lambda[Y, A(\mathbf{x})] P(d\omega)$$

Обобщающая способность

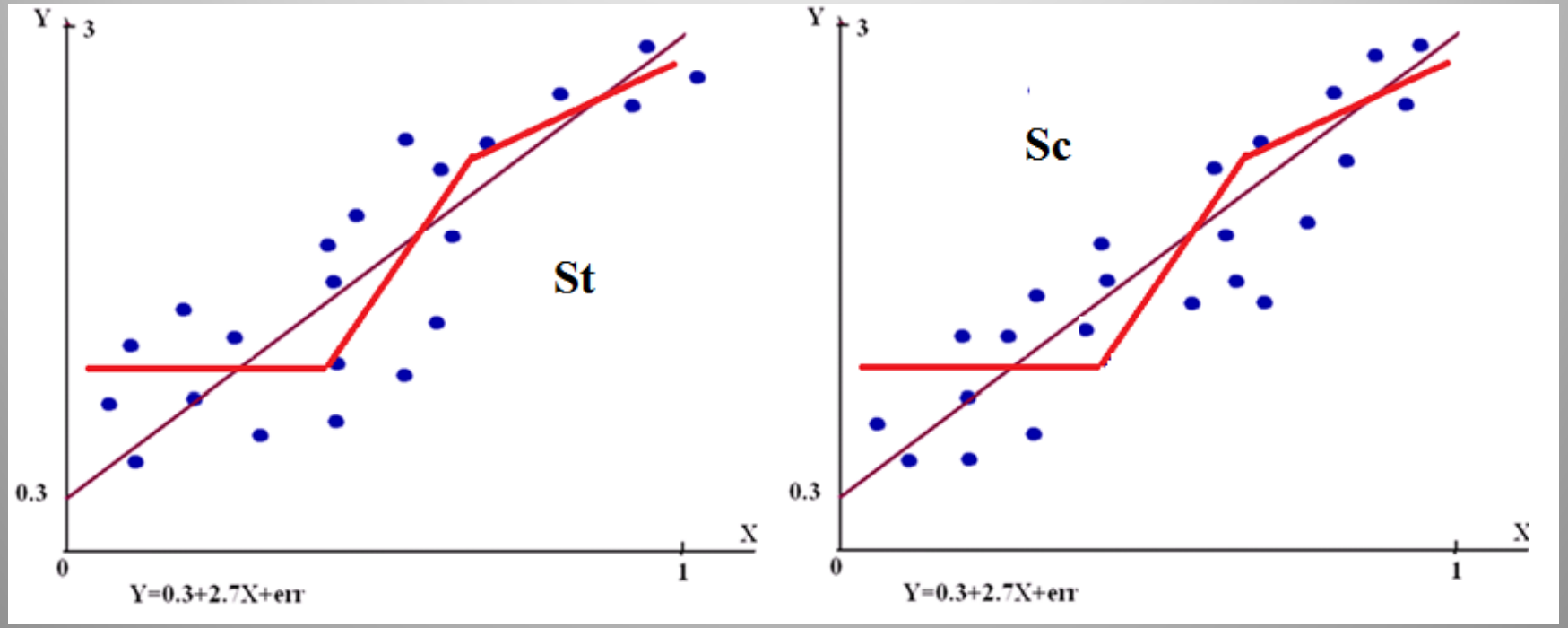
Математическое ожидание берётся по вероятностному пространству Ω содержащему всевозможные объекты процесса, соответствующего решаемой задаче

При решении задач прогнозирования основной целью является достижение максимальной обучающей способности

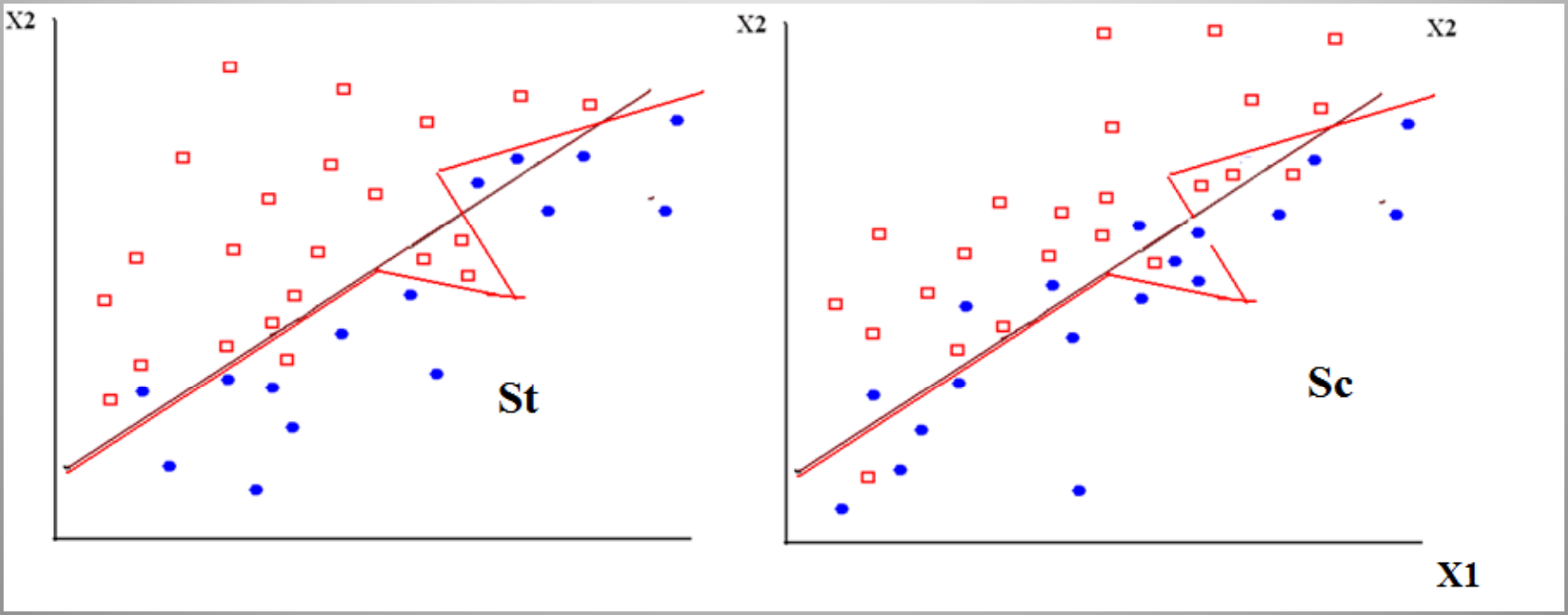
Эффект переобучения

- Расширение модели $\tilde{M} = \{A : \tilde{X} \to \tilde{Y}\}$, увеличение её сложности всегда приводит к повышению точности аппроксимации на обучающей выборке. Однако повышение точности на обучающей выборке, связанное с увеличением сложности модели, часто не ведёт к увеличению обобщающей способности. Более того, обобщающая способность может даже снижаться. Различие между точностью на обучающей выборке и обобщающей способностью при этом возрастает. Данный эффект называется эффектом переобучения.

Эффект переобучения



Эффект переобучения



Для какого алгоритма достигается максимальная обобщающая способность?

Для произвольного прогнозирующего алгоритма величина м математическое ожидание риска в точке $\mathbf{x}$ записывается как $\int_{\Omega(\mathbf{x})} \lambda[Y, A(\mathbf{x})]\mathbf{P}_{\mathbf{x}}(d\omega)$, где $\mathbf{P}_{\mathbf{x}}(a) = \mathbf{P}(a)/\mathbf{P}[\Omega(\mathbf{x})]$

$\Omega(\mathbf{x})$ - множество объектов для которых вектор $X_1, \ldots, X_n$ равен $\mathbf{x}$

В случае, если $\lambda[Y, A(\mathbf{x})] = [Y - A(\mathbf{x})]^2$

$$\int_{\Omega(\mathbf{x})} [Y - A(\mathbf{x})]^2 \mathbf{P}_{\mathbf{x}}(d\omega) = \int_{\Omega(\mathbf{x})} \lambda[Y, A(\mathbf{x})]\mathbf{P}_{\mathbf{x}}(d\omega) =$$

$$\int_{\Omega(\mathbf{x})} [Y - E_{\Omega(\mathbf{x})}Y + E_{\Omega(\mathbf{x})}Y - A(\mathbf{x})]^2 \mathbf{P}_{\mathbf{x}}(d\omega) =$$

$$\int_{\Omega(\mathbf{x})} \{[Y - E_{\Omega(\mathbf{x})}Y]^2 + [E_{\Omega(\mathbf{x})}Y - A(\mathbf{x})]^2 - 2[E_{\Omega(\mathbf{x})}Y - A(\mathbf{x})][Y - E_{\Omega(\mathbf{x})}Y]\}\mathbf{P}_{\mathbf{x}}(d\omega)$$

Для какого алгоритма достигается максимальная обобщающая способность

Однако

$$\int_{\Omega(\mathbf{x})} \{2[E_{\Omega(\mathbf{x})}Y - A(\mathbf{x})][Y - E_{\Omega(\mathbf{x})}Y]\}\mathbf{P}(d\omega) =$$

$$2[E_{\Omega(\mathbf{x})}Y - A(\mathbf{x})] \int_{\Omega(\mathbf{x})} \{[Y - E_{\Omega(\mathbf{x})}Y]\}\mathbf{P}(d\omega) = 0$$

$$\mathbf{x}$$

Откуда следует, что

$$\int_{\Omega(\mathbf{x})} [Y - A(\mathbf{x})]^2 \mathbf{P}_{\mathbf{x}}(d\omega) = [E_{\Omega(\mathbf{x})}Y - A(\mathbf{x})]^2 + \int_{\Omega(\mathbf{x})} [Y - E_{\Omega(\mathbf{x})}Y]^2 \mathbf{P}_{\mathbf{x}}(d\omega) \quad (1)$$

Из формулы (1) хорошо видно, что наилучший прогноз должен обеспечивать алгоритм вычисляющий прогноз равный

$$E_{\Omega(\mathbf{x})}Y = E(Y \mid \mathbf{x})$$

Для какого алгоритма достигается максимальная обобщающая способность

Байесовский классификатор

Пусть в точке $\mathbf{x} \in \mathbf{R}^n$ объекты из классов $K_1, \ldots, K_L$ встречаются с вероятностями $\mathbf{P}(K_1 \mid \mathbf{x}), \ldots, \mathbf{P}(K_L \mid \mathbf{x})$. Тогда распознаваемый объект со значением вектора прогностических переменных $\mathbf{x}$ должен быть отнесён в класс K_* с максимальным значением $\mathbf{P}(K_* \mid \mathbf{x})$

Байесовский классификатор

Покажем , что при справедливости предположения о том, что всю доступную информацию о распределении объектов по классам содержат переменные $X_1, \ldots, X_n$, байесовский классификатор обеспечивает наименьшую ошибку распознавания.

Пусть используется классификатор, относящий классам $K_1, \ldots, K_L$ доли объектов

$\nu_1, \ldots, \nu_L$, соответственно.

Байесовский классификатор

Поэтому объекты, отнесённые в класс K_i по прежнему распределены с вероятностями $\mathbf{P}(K_1 \mid \mathbf{x}), \ldots, \mathbf{P}(K_L \mid \mathbf{x})$, а вероятность ошибочных решений среди объектов отнесённых в класс K_i составляет $1 - \mathbf{P}(K_i \mid \mathbf{x})$

Общая вероятность ошибочных классификаций в точке $\mathbf{x}$ составляет

$$\sum_{i=1}^{L} \nu_i [1 - \mathbf{P}(K_i \mid \mathbf{x})] = 1 - \sum_{i=1}^{L} \nu_i \mathbf{P}(K_i \mid \mathbf{x}) \qquad (1)$$

Байесовский классификатор

Задача поиска минимума ошибки (2) сводится к задаче линейного программирования

$$\sum_{i=1}^{L} \nu_i \mathbf{P}(K_i \mid \mathbf{x}) \to \max$$

при ограничениях

$$\sum_{i=1}^{L} \nu_i = 1$$

$$\nu_i \geq 0 \qquad при \quad i = 1, \ldots, L$$

Байесовский классификатор

Решение задачи линейного программирования находится

в вершине симплекса задаваемого ограничения и является бинарным вектором размерности *L*

(0, ..., 1, ..., 0). При этом 1 находится в позиции, соответствующей максимальной условной вероятности

$$\mathbf{P}(K_i \mid \mathbf{x})$$

МЕТОДЫ ПРОГНОЗИРОВАНИЯ

- Однако для вычисления условных математических ожиданий $E(Y \mid \mathbf{x})$ или условных вероятностей
- $\mathbf{P}(K_i \mid \mathbf{x})$ необходимы знания вероятностных распределений, присущих решаемой задаче. Для подавляющего числа приложений ни общий вид распределений, ни значения конкретных их параметров неизвестны.
- В связи с эти возникло большое число разнообразных подходов к решению задач прогнозирования, использование которых позволяло добиваться определённых успехов при решении конкретных задач.

МЕТОДЫ ПРОГНОЗИРОВАНИЯ

- Статистические методы
- Линейные модели регрессионного анализа
- Различные методы, основанные на линейной разделимости
- Методы, основанные на ядерных оценках
- Нейросетевые методы
- Комбинаторно-логические методы и алгоритмы вычисления оценок
- Алгебраические методы
- Решающие или регрессионные деревья и леса
- Методы, основанные на опорных векторах

Эмпирические методы оценки обобщающей способности

Обобщающая способность может оценивать по случайной выборке объектов из одной и той же генеральной совокупности, соответствующей исследуемому процессу, которую принято называть контрольной выборкой. Контрольная выборка не должна содержать из обучающей выборки.

- Контрольная выборка имеет вид $\tilde{S}_c = \{(y_1, \mathbf{x}_1), \ldots, (y_{m_c}, \mathbf{x}_{m_c})\}$

где y_j - значение переменной Y для j-го объекта;

$\mathbf{x}_j$ - значение вектора переменных $X_1, \ldots, X_n$ для j-го объекта;

m_c - число объектов в $\tilde{S}_c$;

Эмпирические методы оценки обобщающей способности

- Обобщающая способность A может оцениваться с помощью функционала риска

$$Q(\tilde{S}_c, A) = \frac{1}{m} \sum_{j=1}^{m_c} \lambda[y_j, A(\mathbf{x}_j)]$$

При $m_c \to \infty$ согласно закону больших чисел $Q(\tilde{S}_c, A) \to E_\Omega\{\lambda[Y, A(\mathbf{x})]\}$

Эмпирические методы оценки обобщающей способности

Обычно при решении задачи прогнозирования по прецедентам в распоряжении исследователей сразу оказывается весь массив существующих эмпирических данных $\tilde{S}_{in}$. Для оценки точности прогнозирования могут быть использованы следующие стратегии.

1) Выборка $\tilde{S}_{in}$ случайным образом расщепляется на выборку $\tilde{S}_t$ для обучения алгоритма прогнозирования и выборку $\tilde{S}_c$ для оценки точности

2) Процедура кросс-проверки. Выборка $\tilde{S}_{in}$ случайным образом расщепляется на выборки $\tilde{S}_A$ и $\tilde{S}_B$. На первом шаге $\tilde{S}_A$ используется для обучения и $\tilde{S}_B$ для контроля. На следующем шаге $\tilde{S}_A$ и $\tilde{S}_B$ меняются местами

Эмпирические методы оценки обобщающей способности

- 3) Процедура скользящего контроля выполняется по полной выборке $\tilde{S}_{in}$ за $m = |\tilde{S}_{in}|$ шагов .

на j -ом шаге формируется обучающая выборка $\tilde{S}_t^j = \tilde{S}_{in} \setminus s_j$, где $s_j = (y_j, \mathbf{x}_j)$ j- ый объект $\tilde{S}_{in}$,

и контрольная выборка $\tilde{S}_c$, состоящая из единственного объекта s_j.

Процедура скользящего контроля вычисляет оценку обобщающей способности

$$Q_{sc}(\tilde{S}_{in}, A) = \frac{1}{m}\sum_{j=1}^{m} \lambda[y_j, A(\mathbf{x}_j, \tilde{S}_t^j)]$$

Несмещённость оценки скользящего контроля

Пусть Ω_m вероятностное пространство, элементами которого являются выборки по m объектов из генеральной совокупности, соответствующей рассматриваемому процессу

Под несмещённостью оценки скользящего контроля понимается выполнение следующего равенства

$$E_{\Omega_m}\{Q_{sc}[\tilde{S}_m, A]\} = E_{\Omega_{m-1}} E_{\Omega}\{\lambda[Y, A(\mathbf{x}, \tilde{S}_{m-1})]\}$$

Несмещённость оценки скользящего контроля

Покажем, что несмещённость имеет место, если выборка $\tilde{S}_{in}$ является случайной выборкой из одной и той же генеральной совокупности.

В этом случае пространство Ω_m является декартовым произведением m пространств Ω ($\Omega_m = \Omega \times \ldots \times \Omega$) с вероятностной мерой $\mathbf{P}^m$, удовлетворяющей условию

$$\mathbf{P}^m(\mathbf{a}_1 \times \ldots \times \mathbf{a}_m) = \prod_{i=1}^{m} \mathbf{P}(\mathbf{a}_i)$$

Несмещённость оценки скользящего контроля

$$E_{\Omega_m}\{Q_{sc}[\tilde{S}_m, A] = E_{\Omega_m}\{\frac{1}{m}\sum_{j=1}^{m}\lambda[y_j, A(\mathbf{x}_j, \tilde{S}_t^j)]\} =$$

$$= \frac{1}{m}\sum_{j=1}^{m} E_{\Omega_m}\{\lambda[y_j, A(\mathbf{x}_j, \tilde{S}_t^j)]\}$$

Однако выборка $\tilde{S}_t^j$ является элементом пространства Ω_{m-1}.

Объект $(y_j, \mathbf{x}_j)$ является элементом Ω. Откуда

$$E_{\Omega_m}\{\lambda[y_j, A(\mathbf{x}_j, \tilde{S}_t^j)]\} = E_{\Omega_{m-1}} E_{\Omega}\{\lambda[Y, A(\mathbf{x}, \tilde{S}_{m-1})]\}$$

и

$$E_{\Omega_m}\{Q_{sc}[\tilde{S}_m, A]\} = E_{\Omega_{m-1}} E_{\Omega}\{\lambda[Y, A(\mathbf{x}, \tilde{S}_{m-1})]\}$$