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 или результатов лечения по совокупности клинических и лабораторных показателей;
- 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 возникновения некоторого критического события до различных моментов времени.
- Задачи, в которых прогнозируемая величина принимает значения из множества, содержащего несколько элементов принято называть задачей распознавания, Например, к задачам распознавания относятся задачи прогнозирования категориальных переменных.

Генеральная совокупность

Множество объектов, которые могут возникать в рамках рассматриваемой задачи, называется генеральной совокупностью Ω

Обычно генеральная совокупность рассматривается как множество элементарных событий, на котором заданы σ - алгебра событий Σ и вероятностная мера $\mathbf{P}$. То есть генеральная совокупность рассматривается как вероятностное пространство $<\Omega, \Sigma, \mathbf{P}>$

Значение Y для некоторых объектов Ω может оказаться неизвестным. Однако данное значение может быть восстановлено по вектору известных значений переменных $X_1, \ldots, X_n$

.

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 объекта s_j ;

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

$j = 1, \ldots, m$;

m - число объектов в $\tilde{S}_t$;

Обучающая выборка

Обычно предполагается, что обучающая выборка может рассматриваться как независимая, выборка объектов из Ω . Иными словами предполагается, что $\tilde{S}_t$ является элементом декартова произведения $\Omega_m = \Omega \times \ldots \times \Omega$. При этом предполагается, что на Ω_m задана σ - алгебра Σ_m , содержащая всевозможные декартовы произведения вида $\mathbf{a}_1 \times \ldots \times \mathbf{a}_m$, где $\mathbf{a}_i \in \Sigma, i = 1, \ldots, m$, и вероятностная мера $\mathbf{P}^m$, удовлетворяющая условию

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

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)

Способы поиска закономерностей

Основным способом поиска закономерностей является поиск некотором априори заданном семействе алгоритмов прогнозирования

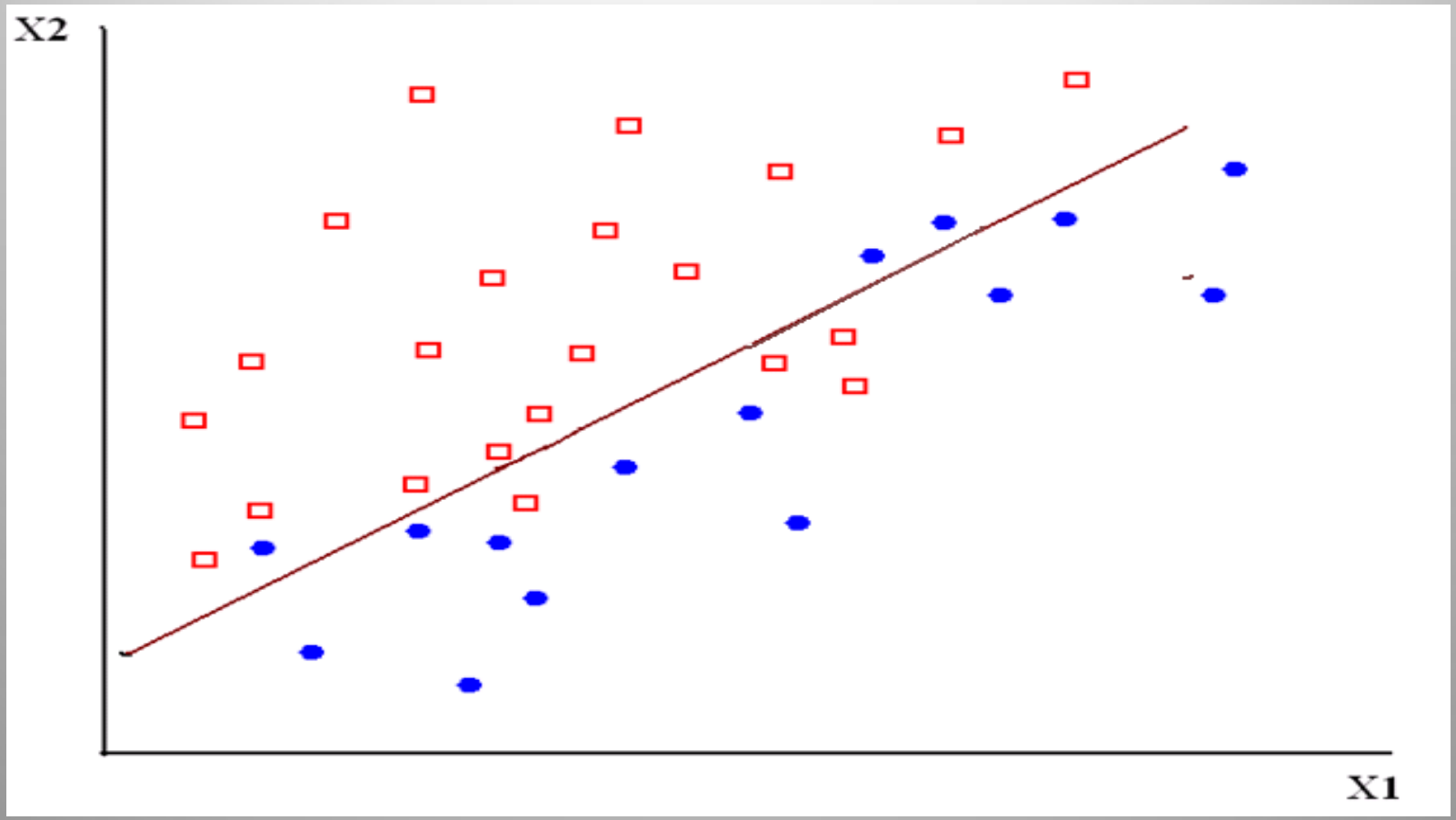
$\tilde{M} = \{A : \tilde{X} \to \tilde{Y}\}$ алгоритма, наилучшим образом аппроксимирующего связь переменных из набора $X_1, \ldots, X_n$ с переменной Y на обучающей выборке, где $\tilde{X}$ - область возможных значений векторов переменных $X_1, \ldots, X_n$

$\tilde{Y}$ - 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

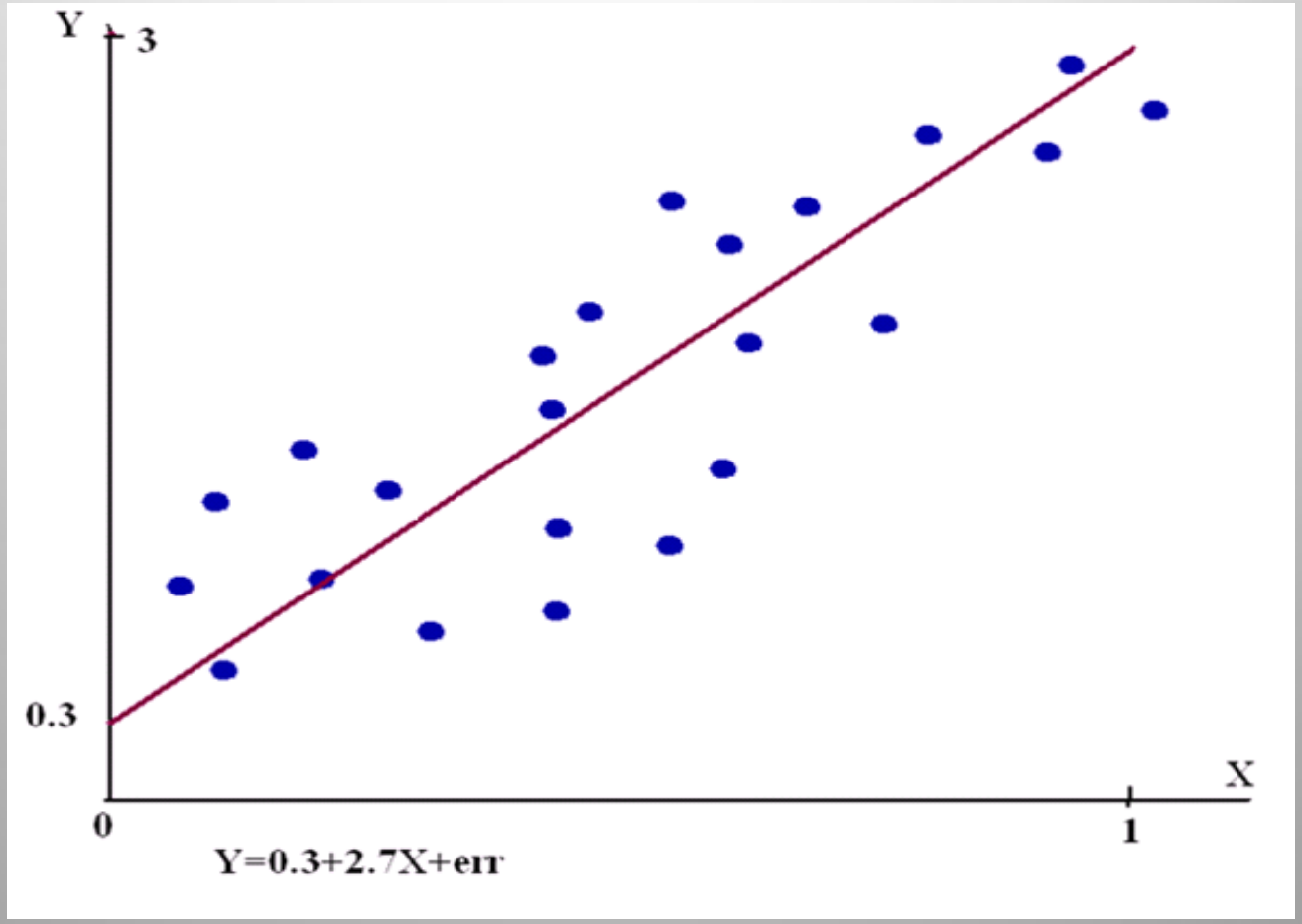
$$Q(\tilde{S}_t, A) = \frac{1}{m} \sum_{j=1}^{m} \lambda[y_j, A(\mathbf{x}_j)]$$

Примеры



Показана закономерность: объекты класса K1 и класса K2 из обучающей выборке St находятся по разные стороны прямой.

Примеры



Показана закономерность: наличие линейной зависимости между Y и X

Способы поиска закономерностей

Частные случаи функции потерь.

При прогнозировании непрерывных величин могут использоваться

$$\lambda[y_j, A(\mathbf{x}_j)] = [y_j - A(\mathbf{x}_j)]^2$$ - 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 при ошибочной. 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 математическое ожидание потерь по генеральной совокупности - $E_{\Omega}\{\lambda[Y, A(\mathbf{x})]\}$

Обобщающая способность

Обобщающая способность может быть записана в виде

$$E_{\Omega}\{\lambda[Y, A(\mathbf{x})]\} = \int_{\mathbf{M}} E\{\lambda[Y, A(\mathbf{x})] \mid \mathbf{x}\}\, p(\mathbf{x}) dx_1 \ldots dx_n,$$

где $p(\mathbf{x})$ - плотность вероятности в точке $\mathbf{x}$

Интегрирование ведётся по области M , принадлежащей пространству $\mathbf{R}^n$ вещественных векторов размерности ***n,*** из которой принимают значения $X_1, \ldots, X_n$

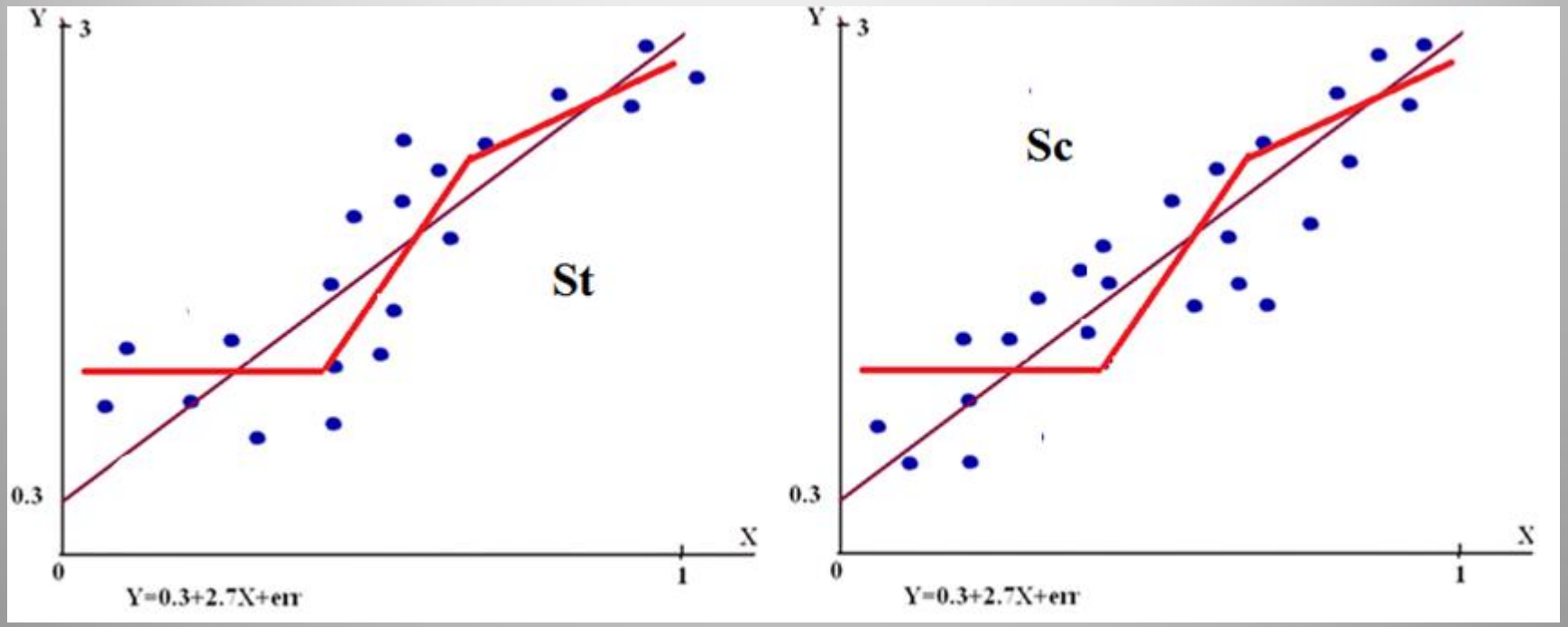
Обобщающая способность

При решении задач прогнозирования основной целью является достижение максимальной обобщающей способности

Эффект переобучения

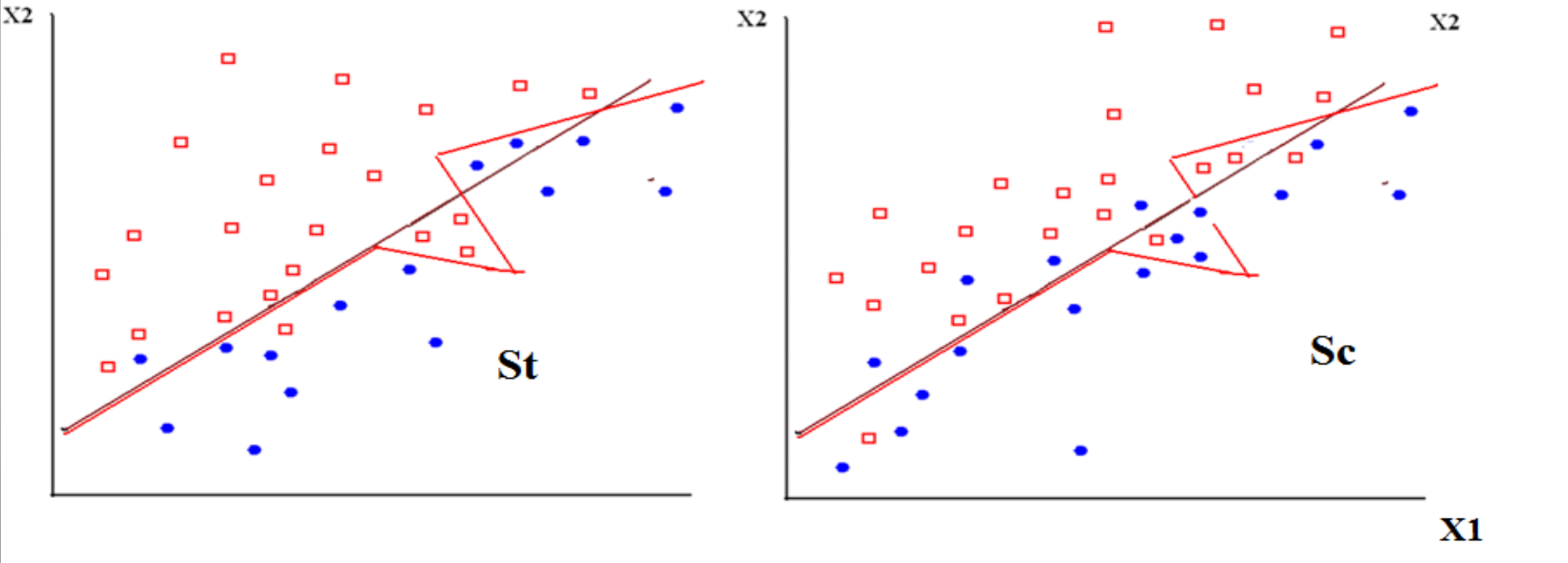
- Расширение модели $\tilde{M} = \{A : \tilde{X} \to \tilde{Y}\}$, увеличение её сложности всегда приводит к повышению точности аппроксимации на обучающей выборке. Однако повышение точности на обучающей выборке, связанное с увеличением сложности модели, часто не ведёт к увеличению обобщающей способности. 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



На левом слайде показано, что использование кусочно-линейной модели (красная линия)позволяет значительно лучше аппроксимировать зависимость на обучающей выборке St, чем простая линейная регрессия (тёмно-синяя прямая). Однако оказывается (правый слайд), что точность аппроксимации новой контрольной выборки Sc, взятой из той же самой генеральной совокупности, для простой линейной регрессии значительно лучше, чем для кусочно-линейной.

Эффект переобучения



На левом слайде показано, что использование кусочно-линейной границы (красная линия)позволяет значительно лучше разделить объекты класса K1 и класса K2 обучающей выборке St, чем простая линейная граница (тёмно-синяя прямая). Однако оказывается (правый слайд), что точность на новой контрольной выборки Sc, взятой из той же самой генеральной совокупности, для простой линейной границы значительно лучше, чем для кусочно-линейной

Для какого алгоритма прогнозирования достигается максимальная обобщающая способность?

В случае, если при прогнозе Y в точке $\mathbf{x}$ используется Величина $A(\mathbf{x})$, а величиной потерь является квадрат ошибки

(т.е. $\lambda[Y, A(\mathbf{x})] = [Y - A(\mathbf{x})]^2$) , справедливо разложение:

$$E\{\lambda[Y, A(\mathbf{x})] | \mathbf{x}\} = E\{[Y - A(\mathbf{x})]^2 | \mathbf{x}\} =$$

$$= E\{[Y - E(Y|\mathbf{x}) + E(Y|\mathbf{x}) - A(\mathbf{x})]^2 | \mathbf{x}\} = E\{[Y - E(Y|\mathbf{x})]^2 | \mathbf{x}\} +$$

$$+ E\{[A(\mathbf{x}) - E(Y|\mathbf{x})]^2 | \mathbf{x}\} + 2[A(\mathbf{x}) - E(Y|\mathbf{x})]E\{[Y - E(Y|\mathbf{x})] | \mathbf{x}\}$$

Для какого алгоритма прогнозирования достигается максимальная обобщающая способность?

Здесь мы воспользовались простейшими свойствами условных математических ожиданий. Для произвольных случайных функций ζ_1 и ζ_2

$$E[(\zeta_1+\zeta_2)|\mathbf{x}]=E[\zeta_1|\mathbf{x}]+E[\zeta_2|\mathbf{x}] \quad ;$$

для произвольной константы C и произвольной ζ случайной функции $E[C\zeta|\mathbf{x}]=CE[\zeta|\mathbf{x}]$;

$$E(1|\mathbf{x})=1$$

Для какого алгоритма прогнозирования достигается максимальная обобщающая способность

Однако

$$2[E(Y \mid \mathbf{x}) - A(\mathbf{x})]E\{[Y - E(Y \mid \mathbf{x})] \mid \mathbf{x}\} = 0$$

x

Откуда следует, что

$$E\{[Y - A(\mathbf{x})]^2 \mid \mathbf{x}\} = [E(Y \mid \mathbf{x}) - A(\mathbf{x})]^2 + E\{[Y - E(Y \mid \mathbf{x})]^2 \mid \mathbf{x}\} \quad (1)$$

Из формулы (1) хорошо видно, что наилучший прогноз должен обеспечивать алгоритм вычисляющий прогноз равный $A(\mathbf{x}) = E(Y \mid \mathbf{x})$

Для какого алгоритма распознавания 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_* 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 в некоторой точке $\mathbf{x}$ в классы $K_1, \ldots, K_L$ доли объектов $\nu_1(\mathbf{x}), \ldots, \nu_L(\mathbf{x})$, соответственно.

Байесовский классификатор

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

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

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

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

При ограничениях

$$\sum_{i=1}^{L}\nu_i = 1 \qquad \nu_i \geq 0 \qquad при \quad i = 1,\ldots,L$$

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

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

Поиск оптимальных алгоритмов прогнозирования и распознавания

Однако для вычисления условных математических ожиданий $E(Y \mid \mathbf{x})$ или условных вероятностей $\mathbf{P}(K_i \mid \mathbf{x})$ $i = 1, \ldots, L$ необходимы знания конкретного вида вероятностных распределений, присущих решаемой задаче. Такие знания в принципе могут быть получены с использованием известного метода максимального правдоподобия.

Метод максимального правдоподобия

Метод максимального правдоподобия используется в математической статистике для аппроксимации вероятностных распределений по выборкам данных. В общем случае ММП требует априорных предположений о типе распределений. Значения параметров $(\theta_1,\ldots,\theta_r)$, задающих конкретный вид распределений, ищутся путём максимизации функционала правдоподобия. Функционал правдоподобия представляет собой произведение плотностей вероятностей на объектах обучающей выборки.

Метод максимального правдоподобия

Функционал правдоподобия имеет вид

$$L(\tilde{S}_t, \theta_1, \ldots, \theta_r) = \prod_{j=1}^{m} p(y_j, \mathbf{x}_j \mid \theta_1, \ldots, \theta_r)$$

Наряду с методом минимизации эмпирического риска метод ММП является одним из важнейших инструментов настройки алгоритмов распознавания или регрессионных моделей. Следует отметить тесную связь между обоими подходами.

Поиск оптимальных алгоритмов прогнозирования и распознавания

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 оцениваться по случайной выборке объектов из одной и той же генеральной совокупности, соответствующей исследуемому процессу, которую принято называть контрольной выборкой. Контрольная выборка не должна содержать объекты 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

Под несмещённостью оценки скользящего контроля понимается выполнение следующего равенства

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

Покажем, что несмещённость имеет место, если выборка $\tilde{S}_{in}$ является независимой выборкой объектов из генеральной совокупности Ω

Несмещённость оценки скользящего контроля

Напомним, что в этом случае $\tilde{S}_{in}$ является элементом вероятностного пространства $(\Omega_m, \Sigma_m, \mathbf{P}_m)$. Произвольная подвыборка $\tilde{S}_{in}$ размером $m' < m$ с произвольным порядком объектов является элементом вероятностного пространства $(\Omega_{m'}, \Sigma_{m'}, \mathbf{P}_{m'})$, которое строится также как и вероятностное пространство $(\Omega_m, \Sigma_m, \mathbf{P}_m)$

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 из ранее сказанного следует, что $\forall j$ выборка $\tilde{S}_t^j$ является элементом пространства Ω_{m-1}.

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

Несмещённость оценки скользящего контроля

Из упомянутых свойств, а также из теоремы Фубини следует

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

Таким образом

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

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