Лекция 1

Задачи машинного обучения, обобщающая способность, метод минимизации эмпирического риска, проблема переобучения кросс-валидация

Лектор – *Сенько Олег Валентинович*

Курс «Методы МО и ИАД»

Содержание лекции

Задачи прогнозирования

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

Модели прогнозирования

Для решения подобных задач могут быть использованы методы, основанные на использовании точных знаний. Например, могут использоваться методы математического моделирования, основанные на использовании физических законов. Однако сложность точных математических моделей нередко оказывается слишком высокой. Кроме того при использовании физических моделей часто требуется знание различных параметров, характеризующих рассматриваемое явление или процесс. Значения некоторых из таких параметров часто известны только приблизительно или неизвестны вообще. Все эти обстоятельства ограничивают возможности эффективного использования физических моделей.

Прогнозирования по прецендентам

В прикладных исследованиях нередко возникают ситуации, когда математическое моделирование, основанное на использовании точных законов оказывается затруднительны, но в распоряжении исследователей оказывается выборка прецедентов - результатов наблюдений исследуемого процесса или явления, включающих значения прогнозируемой величины Y и переменных $X_1, \ldots, X_n$. В этих случаях для решения задач диагностики и прогнозирования могут быть использованы методы, основанные на обучении по прецедентам.

Прогнозирование по прецедентам. Генеральная совокупность

Предположим, что задача прогнозирования решается для некоторого процесса или явления $\mathbf{F}$. Множество объектов, которые потенциально могут возникать в рамках $\mathbf{F}$, называется генеральной совокупностью, далее обозначаемой Ω . Предполагается, что прогнозируемая величина Y и переменные $X_1, \ldots, X_n$ заданы на Ω. Однако значение Y для некоторых объектов Ω может по разным причинам оказаться недоступным исследователю. При этом значения по крайней мере части переменных $X_1, \ldots, X_n$ известны. Целью математических методов прогнозирования, рассматриваемых в курсе, является построение алгоритма, вычисляющего недоступные значений Y по известным значениям переменных $X_1, \ldots, X_n$. Обычно генеральная совокупность рассматривается как множество элементарных событий, на котором заданы - алгебра событий Σ и вероятностная мера $\mathbf{P}$. То есть генеральная совокупность рассматривается как вероятностное пространство $\langle \Omega, \Sigma, P \rangle$.

Методы, основанные на обучении по прецедентам

Поиск алгоритма, вычисляющего осуществляется по выборке прецедентов, которая обычно является случайной выборкой объектов из Ω с известными значениями $Y, X_1, \ldots, X_n$, Выборку прецедентов также принято называть обучающей выборкой.
Обучающая выборка имеет вид $\tilde{S}_t = \{(y_1, \boldsymbol{x}_1), \ldots, (y_m, \boldsymbol{x}_m)\}$, где
y_j – значение переменной Y для объекта s_j, $j = 1, \ldots, m$;
$\boldsymbol{x}_j$ – значение вектора переменных $X_1, \ldots, X_n$ для объекта s_j;
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).

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.

Примеры задач машинного обучения

Задача распознавания (классификации) ириса на три класса. Здесь целевая переменная $Y \in \{setosa, versicolor, virginica\}$, признаки $X_1, \ldots, X_4 \in \mathbb{R}$.

Классы:

Setosa

Versicolor

Virginica

Признаки:

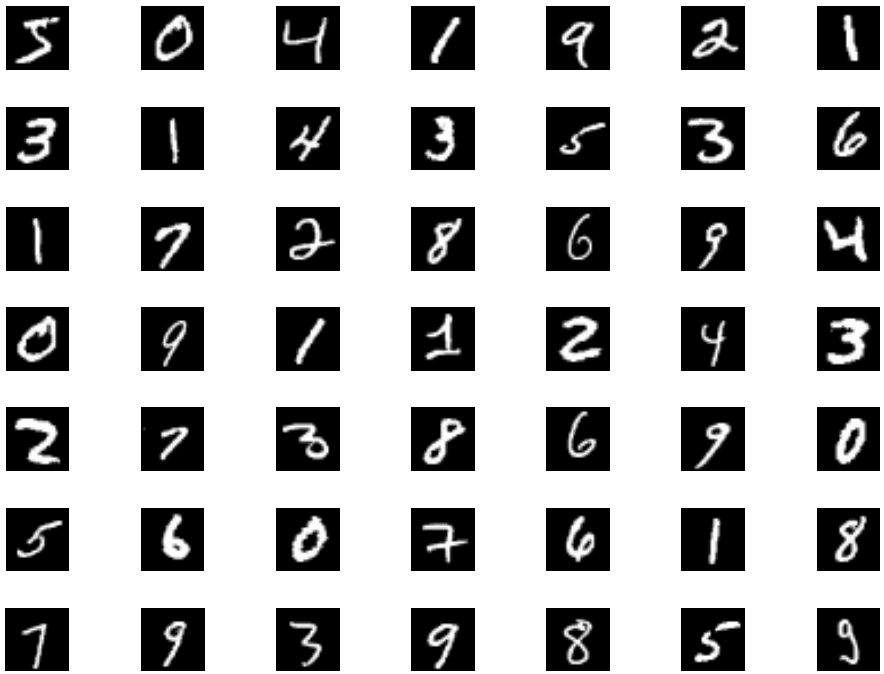
- длина чашелистика (см)
- ширина чашелистика (см)
- длина лепестка (см)
- ширина лепестка (см)

Данные: `http://archive.ics.uci.edu/ml/datasets/Iris`

Примеры задач машинного обучения

Задача распознавания рукописных цифр. Целевая переменная $Y \in \{0, 1, \ldots, 9\}$, признаки $X_1, X_2, \ldots, X_{784} \in [0, 255]$ – пикселы изображения размера 28×28.

Примеры объектов:



Данные: `http://yann.lecun.com/exdb/mnist/`

Примеры задач машинного обучения

Задача прогноза стоимости жилья в различных пригородах Бостона (задача восстановления регрессии).

Целевая переменная Y – цена жилья. Признаки:

- уровень криминала в районе
- концентрация окисей азота
- доля жилья, построенного до 1940 года
- среднее расстояние до основных районов концентрации рабочих мест
- уровень налогообложения
- отношение числа учителей к числу учеников в школах
- и другие

Данные: `http://archive.ics.uci.edu/ml/datasets/Housing`

Способы обучения

Основным способом обучения является поиск в 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 $\lambda[y_j, A(\boldsymbol{x}_j)]$ – величина «потерь», произошедших в результате использования $A(\boldsymbol{x}_j)$ в качестве прогноза значения Y. Тогда одним из способов обучения является минимизация функционала эмпирического риска на обучающей выборке:

$$Q(\tilde{S}_t, A) = \frac{1}{m} \sum_{j=1}^{m} \lambda[y_j, A(\boldsymbol{x}_j)] \to \min_{A \in \tilde{M}}.$$

Частные случаи функции потерь

При прогнозировании непрерывных величин могут использоваться

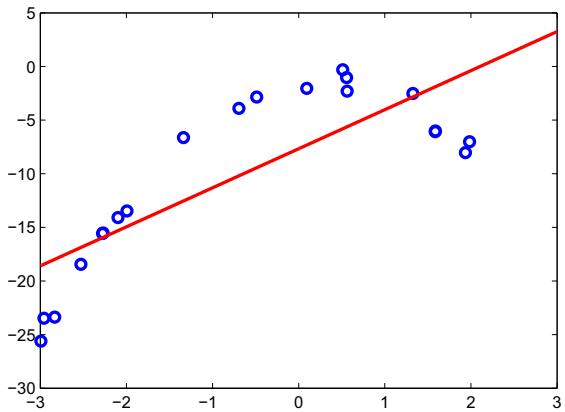
$\lambda[y_j, A(\boldsymbol{x}_j)] = (y_j - A(\boldsymbol{x}_j))^2$ – квадрат ошибки,
$\lambda[y_j, A(\boldsymbol{x}_j)] = |y_j - A(\boldsymbol{x}_j)|$ – модуль ошибки.

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.

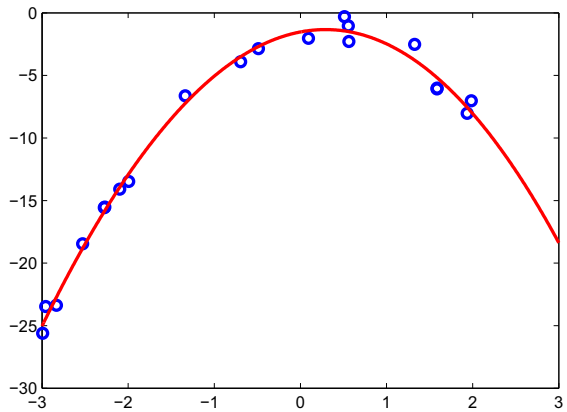
Примеры применения методов МО

Рассмотрим задачу восстановления регрессии по одному признаку. Здесь $\tilde{Y} = \mathbb{R}$, $\tilde{X} = \mathbb{R}$. Поиск зависимости между регрессионной переменной Y и признаком X в рамках семейства отображений $\tilde{M}$ осуществляется с помощью минимизации функционала эмпирического риска с функцией потерь $\lambda[y, A(x)] = (y - A(x))^2$ (т.н. метод наименьших квадратов).

Поиск зависимости в семействе линейных функций $\tilde{M} = \{y = ax + b, a, b \in \mathbb{R}\}$:



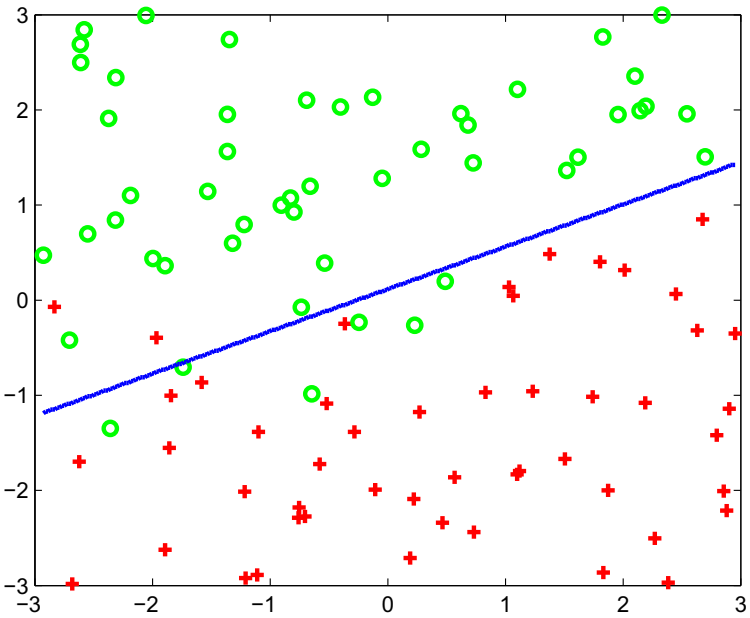
Поиск зависимости в семействе кубических функций $\tilde{M} = \{y = ax^3 + bx^2 + cx + d, a, b, c, d \in \mathbb{R}\}$:



Примеры поиска закономерностей

Рассмотрим задачу классификации на два класса по двум признакам.
Здесь $\tilde{Y} = \{1, 2\}$, $\tilde{X} = \mathbb{R}^2$.
Поиск зависимости в семействе линейных разделителей:

$$y = \begin{cases} 1, & \text{если } ax_1 + bx_2 + c \geq 0, \\ 2, & \text{иначе.} \end{cases}$$



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 $E_{\Omega}\{\lambda[Y, A(\boldsymbol{x})]\}$.

Обобщающая способность

Обобщающая способность может быть записана в виде

$$E_\Omega\{\lambda[Y, A(\boldsymbol{x})]\} = \int_M E\{\lambda[Y, A(\boldsymbol{x})]|\boldsymbol{x}\}p(\boldsymbol{x})dx_1 \dots dx_n,$$

где $p(\boldsymbol{x})$ – плотность вероятности в точке $\boldsymbol{x}$.
Интегрирование ведётся по области M, принадлежащей пространству $\mathbb{R}^n$ вещественных векторов размерности n, из которой принимают значения $X_1, \dots, X_n$.

Обобщающая способность

При решении задач прогнозирования основной целью является достижение максимальной обобщающей способности.

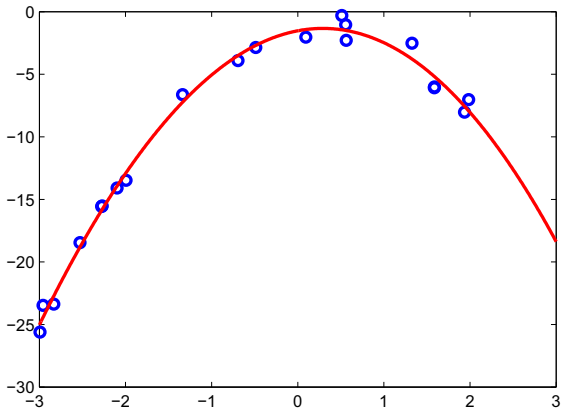
Эффект переобучения

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.

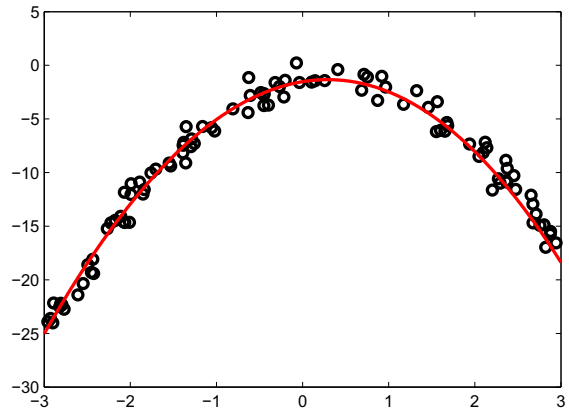
Примеры эффекта переобучения

Задача восстановления регрессии по одному признаку.

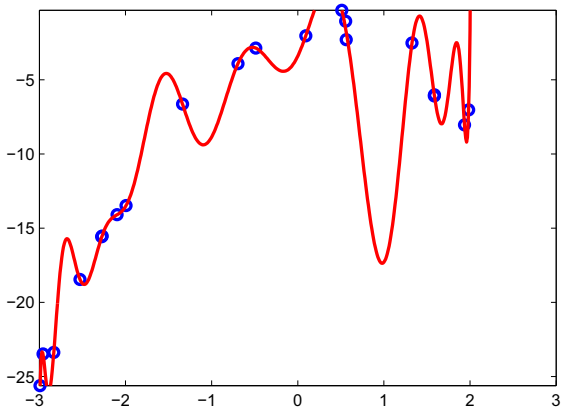
Восстановление кубической зависимости по обучающим данным:



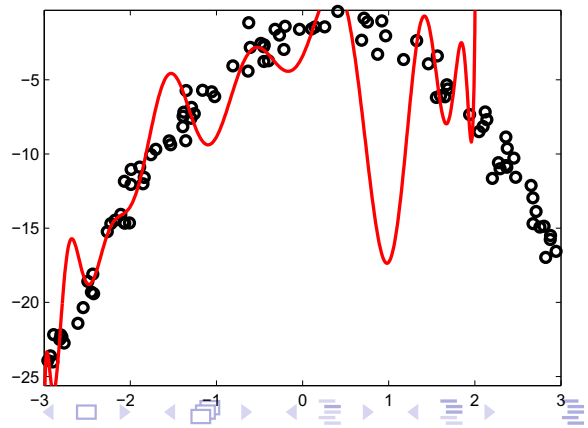
Применение восстановленной зависимости к тестовым данным из той же генеральной совокупности:



Увеличение сложности восстанавливаемой зависимости (степень полинома = 20) приводит к повышению точности на обучающих данных:



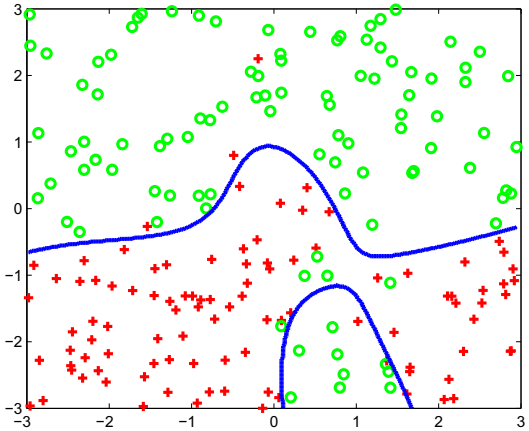
Применение сложной зависимости к тестовым данным обнаруживает переобучение:



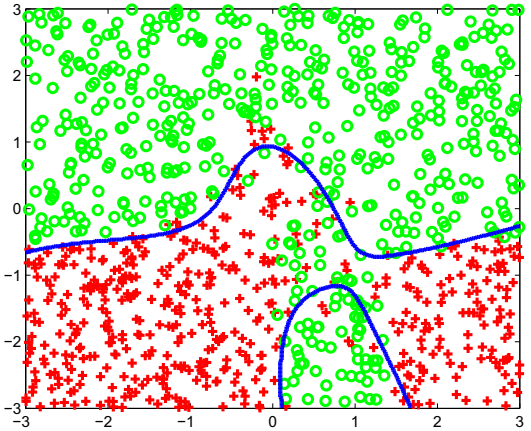
Примеры эффекта переобучения

Задача классификации на два класса по двум признакам.

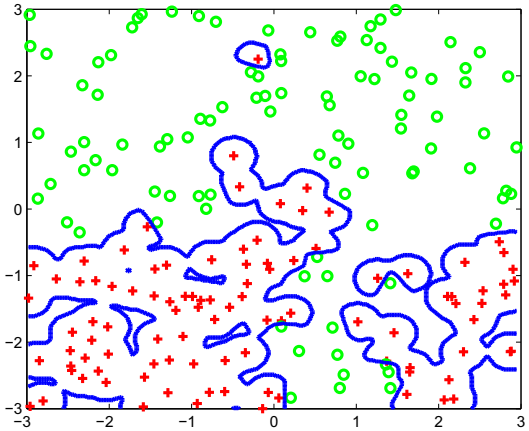
Поиск простой разделяющей кривой по обучающим данным (ошибка 5%):



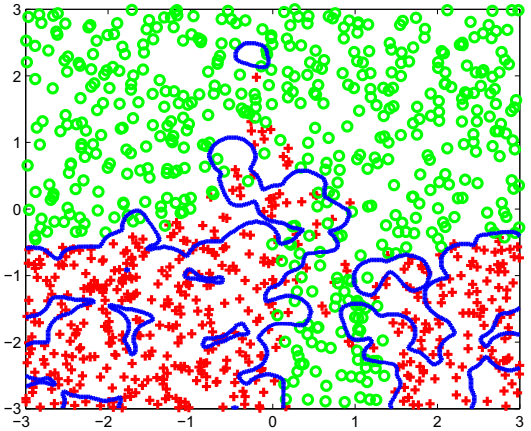
Применение разделяющей кривой к тестовым данным из той же генеральной совокупности (ошибка 6%):



Увеличение сложности разделяющей кривой приводит к 100-процентной точности на обучающих данных:



Применение сложной разделяющей кривой к тестовым данным обнаруживает переобучение (ошибка 14%):



Для какого алгоритма прогнозирования достигается максимальная обобщающая способность?

В случае, если при прогнозе Y в точке $\boldsymbol{x}$ используется величина $A(\boldsymbol{x})$, а величиной потерь является квадрат ошибки (т.е. $\lambda[y_j, A(\boldsymbol{x}_j)] = (y_j - A(\boldsymbol{x}_j))^2$), справедливо разложение:

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

Для какого алгоритма прогнозирования достигается максимальная обобщающая способность?

Однако

$$2[E(Y|\boldsymbol{x}) - A(\boldsymbol{x})]E\{[Y - E(Y|\boldsymbol{x})]|\boldsymbol{x}\} = 0.$$

Отсюда следует, что

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

Из этой формулы хорошо видно, что наилучший прогноз должен обеспечивать алгоритм, вычисляющий прогноз как $A(\boldsymbol{x}) = E(Y|\boldsymbol{x})$.

Классификатор с максимальной обобщающей способностью

Пусть в точке $\boldsymbol{x} \in \mathbb{R}^n$ объекты из классов $K_1, \ldots, K_L$ встречаются с вероятностями $\mathbf{P}(K_1|\boldsymbol{x}), \ldots, P(K_L|\boldsymbol{x})$. Тогда распознаваемый объект со значением вектора прогностических переменных $\boldsymbol{x}$ должен быть отнесён в класс $K_{i'}$, для которого выполнены все неравенства

$$P(K_{i'}|\boldsymbol{x}) \geq \mathbf{P}(K_i|\boldsymbol{x}), i \in \{1, \ldots, L\}.$$

Иными словами распознаваемый объект должен быть отнесён к одному из классов, вероятность принадлежности которому в точке $\boldsymbol{x}$ максимальна . В случае, если максимальная условная вероятность достигается только для одного из классов $K_1, \ldots, K_L$, распознаваемый объект должен быть днозначно отнесён только к этому классу. Такое решающее правило получило название байесовского классификатора.

Однако для вычисления условных математических ожиданий $E(Y|\boldsymbol{x})$ или условных вероятностей $P(K_i|\boldsymbol{x})$, $i = 1, \ldots, L$, 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 (ММП) 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

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

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

Для подавляющего числа приложений ни общий вид распределений, ни значения конкретных их параметров неизвестны. В связи с этим 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 $\tilde{S}_c = \{(y_1, \boldsymbol{x}_1), \ldots, (y_{m_c}, \boldsymbol{x}_{m_c})\}$, где

y_j – значение переменной Y для j-го объекта;

$\boldsymbol{x}_j$ – 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

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

При $m_c \to \infty$ согласно закону больших чисел

$$Q(\tilde{S}_c, A) \to E_\Omega\{\lambda[Y, A(\boldsymbol{x})]\}.$$

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;

3. Процедура k-фолдовая кросс-валидации (скользящего контроля) выполняется по полной выборке $\tilde{S}_{in}$ за k шагов. Выборка $\tilde{S}_{in}$ случайным образом расщепляется k по-возможности равных подвыборок $\widetilde{S}_1, \ldots, \widetilde{S}_k$. На первом шаге в качестве обучающей выборки используется объединение всех выборок кроме $\widetilde{S}_1$, которая считается контрольной

$$\tilde{S}_t^1 = \cup_{i=2}^k \widetilde{S}_i$$

$$\tilde{S}_c^1 = \widetilde{S}_1$$

На шаге j

$$\tilde{S}_t^j = (\cup_{i=1}^k \widetilde{S}_i) \setminus \widetilde{S}_j$$

$$\tilde{S}_c^j = \widetilde{S}_j$$

Общая величина потерь при использовании кросс-валидацции с k фолдами может быть рассчитана по формуле

$$Q_{sc}(\tilde{S}_{in}, A) = \frac{1}{k} \sum_{j=1}^{k} \frac{1}{|\widetilde{S}_c^j|} \sum_{(y_i, \boldsymbol{x}_i) \in \widetilde{S}_c^j} \lambda[y_i, A(\boldsymbol{x}_i, \tilde{S}_t^j)].$$

В случае, если число фолдов равно числу объектов в выборке, процедуру кросс-валидации принято называть валидацией Leave One Out. Величина потерь при этом может быть вычислена по формуле

$$Q_{sc}(\tilde{S}_{in}, A) = \frac{1}{m} \sum_{i=1}^{m} \lambda[y_i, A(\boldsymbol{x}_i, \tilde{S}_t^j)].$$

Несмещённость оценки скользящего контроля

Под несмещённостью оценки скользящего контроля понимается выполнение следующего равенства

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