

Лекция 1

Задачи прогнозирования,
обобщающая способность, байесовский классификатор,
скользящий контроль

Лектор – Сенько Олег Валентинович

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

- 1 Основные понятия теории прогнозирования по прецедентам
- 2 Обобщающая способность и эффект переобучения
- 3 Байесовский классификатор
- 4 Поиск оптимальных алгоритмов прогнозирования
- 5 Методы оценки обобщающей способности и скользящий контроль

Задачи диагностики и прогнозирования некоторой величины Y по доступным значениям переменных X_1, \dots, 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

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.

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

В задачах и методах, рассматриваемых в настоящем курсе, переменные X_1, \dots, X_n являются непрерывными или дискретными скалярными величинами. При этом переменная X_i считается непрерывной, если она принимает значения из некоторого подмножества \mathbb{R} , имеющего мощность континуума. Например, непрерывная переменная может принимать значения из некоторого интервала числовой оси. Непрерывные переменные также называют вещественными. Переменная X_i считается дискретной, если она принимает значения из некоторого конечного или счётного множества \tilde{d}_i . Дискретная переменная X_i называется порядковой, если существует нумерация элементов \tilde{d}_i , соответствующая общепринятым представлениям об их взаимной близости. В противном случае X_i называется номинальной. Следует выделить также бинарные переменные, принимающие значения из множества с двумя элементами. Обычно такие переменные указывают на наличие или отсутствие какого-либо свойства у описываемых объектов.

Прогнозируемая величина Y также может иметь различную природу. Методы прогнозирования разделяются в зависимости от типа прогнозируемой величины.

- Задачи прогнозирования непрерывной Y обычно решаются с помощью методов **регрессионного анализа**.
- Задачи, в которых прогнозируемая величина принимает значения из множества, содержащего относительно небольшое число элементов, принято называть **задачей распознавания**. Например, к задачам распознавания относятся задачи прогнозирования номинальных переменных.
- Следует выделить также задачи, в которых Y является кривой, описывающей вероятность возникновения некоторого критического события до различных моментов времени. Подобные задачи решаются с помощью методов анализа выживаемости (надёжности).

Отметим, что всегда можно построить взаимно-однозначное соответствие между областью значения дискретной переменной и некоторым множеством чисел. Поэтому переменные X_1, \dots, X_n практически всегда можно считать числовыми, а множество значений X_1, \dots, X_n можно считать вектором x в пространстве \mathbb{R}_n . Встречаются задачи прогнозирования, в которых переменные X_1, \dots, X_n являются только вещественными или только дискретными. Однако во многих задачах для прогнозирования используются одновременно непрерывные и дискретные переменные.

Предположим, что задача прогнозирования решается для некоторого процесса или явления F . Множество объектов, которые потенциально могут возникать в рамках F , называется генеральной совокупностью, далее обозначаемой Ω . Предполагается, что прогнозируемая величина Y и переменные X_1, \dots, 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, \dots, X_n известны. Целью математических методов прогнозирования, рассматриваемых в курсе, является построение алгоритма, вычисляющего недоступные значения Y по известным значениям переменных X_1, \dots, X_n . Обычно генеральная совокупность рассматривается как множество элементарных событий, на котором заданы - алгебра событий Σ и вероятностная мера P . То есть генеральная совокупность рассматривается как вероятностное пространство $\langle \Omega, \Sigma, P \rangle$.

Поиск алгоритма, вычисляющего осуществляется по выборке прецедентов, которая обычно является случайной выборкой объектов из Ω с известными значениями Y, X_1, \dots, X_n , Выборку прецедентов также принято называть **обучающей выборкой**.

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

y_j – значение переменной Y для объекта s_j , $j = 1, \dots, m$;

\mathbf{x}_j – значение вектора переменных X_1, \dots, X_n для объекта s_j ;

m – число объектов в \tilde{S}_t .

Обычно предполагается, что объекты обучающей выборки \tilde{S}_t случайно и независимо друг от друга извлекаются из генеральной совокупности Ω . Иными словами предполагается, что \tilde{S}_t является элементом декартова произведения $\Omega_m = \Omega \times \dots \times \Omega$. При этом считается, что на Ω_m задана σ -алгебра Σ_m , содержащая всевозможные декартовы произведения вида $\mathbf{a}_1 \times \dots \times \mathbf{a}_m$, где $\mathbf{a}_i \in \Sigma$, $i = 1, \dots, m$, и вероятностная мера P^m , удовлетворяющая условию

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

Пусть x_1, \dots, x_m значения переменной X на объектах обучающей выборки \tilde{S}_t . Тогда x_1, \dots, x_m могут интерпретироваться как независимые одинаково распределённые случайные величины (н.о.р.с.в.).

В процессе обучения производится поиск эмпирических закономерностей, связывающих прогнозируемую переменную Y с переменными X_1, \dots, X_n . Данные закономерности далее используются при прогнозировании. Методы, основанные на обучении по прецедентам, также принято называть **методами машинного обучения (machine learning)**.

Примеры задач машинного обучения

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

Классы:



Setosa



Versicolor



Virginica

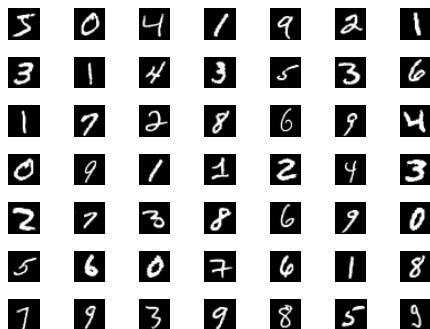
Признаки:

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

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

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

Примеры объектов:



Данные: <http://yann.lecun.com/exdb/mnist/>



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>

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

\tilde{X} – область возможных значений векторов переменных X_1, \dots, 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:**

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

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.

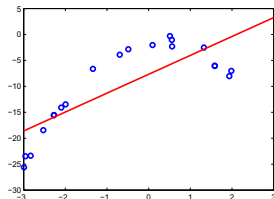
В случае задачи распознавания функция потерь может быть равной 0 при правильной классификации и 1 при ошибочной. При этом функционал эмпирического риска равен числу ошибочных классификаций.

Примеры поиска закономерностей

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**).

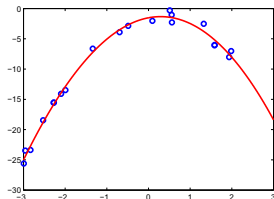
Поиск зависимости в семействе линейных функций

$$\tilde{M} = \{y = ax + b, a, b \in \mathbb{R}\}:$$



Поиск зависимости в семействе кубических функций

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

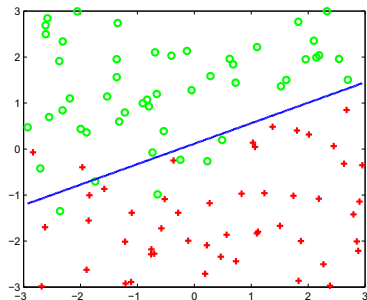


Примеры поиска закономерностей

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:

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



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 может быть записана в виде

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

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

Интегрирование ведётся по области M , принадлежащей пространству \mathbb{R}^n вещественных векторов размерности n , из которой принимают значения X_1, \dots, X_n .

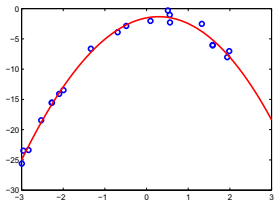
При решении задач прогнозирования основной целью является достижение **максимальной обобщающей способности**.

Расширение модели $\tilde{M} = \{A : \tilde{X} \rightarrow \tilde{Y}\}$, 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**.

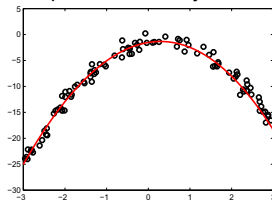
Примеры эффекта переобучения

Задача восстановления регрессии по одному признаку.

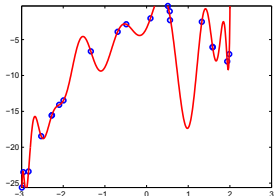
Восстановление кубической зависимости по обучающим данным:



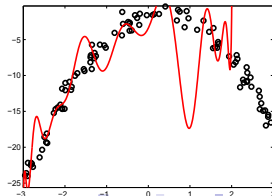
Применение восстановленной зависимости к тестовым данным из той же генеральной совокупности:



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



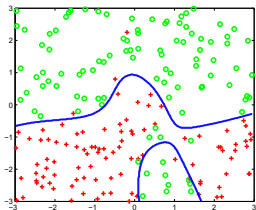
Применение сложной зависимости к тестовым данным обнаруживает переобучение:



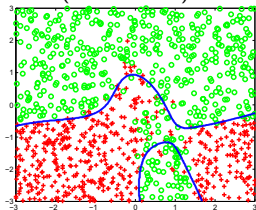
Примеры эффекта переобучения

Задача классификации на два класса по двум признакам.

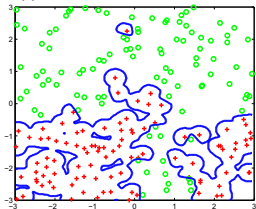
Поиск простой разделяющей кривой по обучающим данным (ошибка 5%):



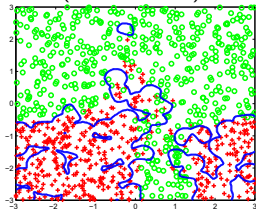
Применение разделяющей кривой к тестовым данным из той же генеральной совокупности (ошибка 6%):



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 \mathbf{x} используется величина $A(\mathbf{x})$, а величиной потерь является квадрат ошибки (т.е.

$\lambda[y_j, A(\mathbf{x}_j)] = (y_j - A(\mathbf{x}_j))^2$), справедливо разложение:

$$\begin{aligned} E\{\lambda[Y, A(\mathbf{x})]|\mathbf{x}\} &= E\{[Y - A(\mathbf{x})]^2|\mathbf{x}\} = \\ &E\{[Y - E(Y|\mathbf{x}) + E(Y|\mathbf{x}) - A(\mathbf{x})]^2|\mathbf{x}\} = \\ &E\{[Y - E(Y|\mathbf{x})]^2|\mathbf{x}\} + E\{[A(\mathbf{x}) - E(Y|\mathbf{x})]^2|\mathbf{x}\} + \\ &2[A - E(Y|\mathbf{x})]E\{[Y - E(Y|\mathbf{x})]|\mathbf{x}\}. \end{aligned}$$

Для какого алгоритма прогнозирования достигается максимальная обобщающая способность?

Далее мы воспользуемся простейшими свойствами условных математических ожиданий. Для произвольных случайных функций ζ_1 и ζ_2

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

Для произвольной константы C и произвольной случайной функции ζ

$$E[C\zeta|\mathbf{x}] = CE[\zeta|\mathbf{x}].$$

Также $E[1|\mathbf{x}] = 1$.

Для какого алгоритма прогнозирования достигается максимальная обобщающая способность?

Однако

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

Отсюда следует, что

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

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

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

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

Иными словами распознаваемый объект должен быть отнесён к одному из классов, вероятность принадлежности которому в точке \mathbf{x} максимальна. В случае, если максимальная условная вероятность достигается только для одного из классов K_1, \dots, K_L , распознаваемый объект должен быть однозначно отнесён только к этому классу.

Такое решающее правило получило название **байесовского классификатора**.

Покажем, что при справедливости предположения о том, что всю доступную информацию о распределении объектов по классам содержат переменные X_1, \dots, X_n , байесовский классификатор обеспечивает наименьшую ошибку распознавания. Пусть используется классификатор, относящий в некоторой точке \mathbf{x} в классы K_1, \dots, K_L доли объектов $\nu_1(\mathbf{x}), \dots, \nu_L(\mathbf{x})$, соответственно. Из предположения о содержании всей информации о классах переменными X_1, \dots, X_n следует, что внутри множества объектов с фиксированным \mathbf{x} истинный номер класса не зависит от вычисленного прогноза. Откуда следует, что вероятность отнесения в класс K_i объекта, который классу K_i в действительности принадлежит составляет $\nu_i(\mathbf{x})P(K_i|\mathbf{x})$. Вероятность ошибочного отнесения в классы отличные от K_i объекта, в действительности принадлежащего K_i , составляет $[1 - \nu_i(\mathbf{x})]P(K_i|\mathbf{x})$.

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

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

Задача поиска минимума ошибки сводится к задаче линейного программирования вида

$$\begin{aligned} \sum_{i=1}^L \nu_i P(K_i | \mathbf{x}) &\rightarrow \max_{\nu_1, \dots, \nu_L}, \\ \sum_{i=1}^L \nu_i &= 1, \\ \nu_i &\geq 0, i = 1, \dots, L. \end{aligned}$$

Одна из вершин симплекса, задаваемого ограничениями задачи линейного программирования, является её решением. Данное решение представляет собой бинарный вектор размерности L , имеющий вид $(0, \dots, 1, \dots, 0)$. При этом значение 1 находится в позиции i' , для которой выполняется набор неравенств $P(K_{i'}|\mathbf{x}) \geq P(K_i|\mathbf{x})$, $i = 1, \dots, L$. В случае, если максимальная условная вероятность достигается только для одного класса $K_{i'}$, решение задачи линейного программирования достигается в единственной точке, задаваемой бинарным вектором с единственной 1, находящейся в позиции i' .

Предположим, что максимальная условная вероятность достигается для нескольких классов $K_{j(1)}, \dots, K_{j(l')}$. Тогда решением задачи является вектор ν_1, \dots, ν_L , компоненты которого удовлетворяют условиям:

$$\nu_i = 0, i \neq j(r), r = 1, \dots, l'.$$

Из этого следует, что любая стратегия, при которой объекты относятся в один из классов $K_{j(1)}, \dots, K_{j(l')}$, является оптимальной.

Однако для вычисления условных математических ожиданий $E(Y|\mathbf{x})$ или условных вероятностей $P(K_i|\mathbf{x})$, $i = 1, \dots, L$, 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 (ММП) используется в математической статистике для аппроксимации вероятностных распределений по выборкам данных. В общем случае ММП требует априорных предположений о типе распределений. Значения параметров $(\theta_1, \dots, \theta_r)$, задающих конкретный вид распределений, ищутся путём максимизации функционала правдоподобия. Функционал правдоподобия представляет собой произведение плотностей вероятностей на объектах обучающей выборки.

Функционал правдоподобия имеет вид

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

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.

Пример

Пусть x_1, x_2, \dots, x_m – независимые одинаково распределённые случайные величины (н.о.р.с.в), причём

$$x_i \sim \mathcal{N}(x_i|\theta, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x_i - \theta)^2}{2\sigma^2}\right).$$

Требуется с помощью метода максимального правдоподобия оценить значение θ .

Пример

Пусть x_1, x_2, \dots, x_m – независимые одинаково распределённые случайные величины (н.о.р.с.в), причём

$$x_i \sim \mathcal{N}(x_i|\theta, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x_i - \theta)^2}{2\sigma^2}\right).$$

Требуется с помощью метода максимального правдоподобия оценить значение θ .

Принимая во внимание условие независимости наблюдений запишем совместное распределение величин x_1, \dots, x_m :

$$\begin{aligned} p(\mathbf{x}|\theta, \sigma^2) &= p(x_1, x_2, \dots, x_m|\theta, \sigma^2) = \\ &= \prod_{j=1}^m p(x_j|\theta, \sigma^2) = \prod_{j=1}^m \mathcal{N}(x_j|\theta, \sigma^2). \end{aligned}$$

Функция $p(\mathbf{x}|\theta, \sigma^2)$:

- как функция от \mathbf{x} – **условная плотность**, в частности, $\int p(\mathbf{x}|\theta, \sigma^2) d\mathbf{x} = 1$;
- как функция от θ, σ^2 – **функция правдоподобия**.

Оценка θ с помощью максимизации правдоподобия соответствует задаче оптимизации:

$$L(\theta, \sigma^2) = p(\mathbf{x}|\theta, \sigma^2) \rightarrow \max_{\theta}$$

При работе с функционалами в форме произведений удобно переходить к логарифму:

$$\begin{aligned}\theta_{ML} &= \arg \max_{\theta} L(\theta, \sigma^2) = \arg \max_{\theta} \log L(\theta, \sigma^2) = \\ &= \arg \max_{\theta} \left[-\frac{1}{2\sigma^2} \sum_{j=1}^m (x_j - \theta)^2 - \underbrace{\frac{m}{2} \log 2\pi - m \log \sigma}_{const} \right]\end{aligned}$$

Дифференцируя $\log L(\theta, \sigma^2)$ по θ и приравнивая производную к нулю, получаем:

$$\frac{\partial}{\partial \theta} \log L(\theta, \sigma^2) = -\frac{1}{2\sigma^2} \left(2m\theta - 2 \sum_{j=1}^m x_j \right) = 0.$$

Отсюда

$$\theta_{ML} = \frac{1}{m} \sum_{j=1}^m x_j.$$

Заметим, что оценка θ_{ML} не зависит от дисперсии σ^2 .

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.

- 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

Обобщающая способность алгоритма прогнозирования на генеральной совокупности Ω может оцениваться по случайной выборке объектов из Ω , которую обычно называют контрольной выборкой. При этом контрольная выборка не должна содержать объектов из обучающей выборки. В противном случае оценка величины потерь обычно оказывается заниженной. Контрольная выборка имеет вид

$\tilde{S}_c = \{(y_1, \mathbf{x}_1), \dots, (y_{m_c}, \mathbf{x}_{m_c})\}$, где

y_j – значение переменной Y для j -го объекта;

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

m_c – число объектов в \tilde{S}_c .

Обобщающая способность A может оцениваться с помощью функционала риска

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

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

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

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.

- 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 как

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

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 случайной независимой выборкой объектов из генеральной совокупности Ω .

Напомним, что в этом случае \tilde{S}_{in} является элементом вероятностного пространства $\langle \Omega_m, \Sigma_m, P_m \rangle$. Подвыборка \tilde{S}_{in} размером $m' < m$ с произвольным порядком объектов является элементом вероятностного пространства $\langle \Omega_{m'}, \Sigma_{m'}, P_{m'} \rangle$, которое строится также, как и вероятностное пространство $\langle \Omega_m, \Sigma_m, P_m \rangle$.

Математическое ожидание оценки потерь по методу скользящего контроля может быть представлено в виде средней величины математических ожиданий оценок потерь на каждом шаге:

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

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 .

Из равенства, приведённого на предыдущем слайде, а также из известной теоремы Фубини, обосновывающей правомерность изменения порядка интегрирования, следует равенство

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

Таким образом,

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

В результате приходим к заключению о справедливости равенства

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

что по определению соответствует несмещённости оценки потерь по методу скользящего контроля. Отметим, что скользящий контроль является достаточно трудоёмкой процедурой. Снижение трудоёмкости может быть достигнуто при формировании контрольных выборок не из одного, а из нескольких объектов. При этом контрольная выборка, формируемая на каждом шаге не должна пересекаться с контрольными выбоками, формируемыми на предыдущих шагах.