Модель линейной регрессии

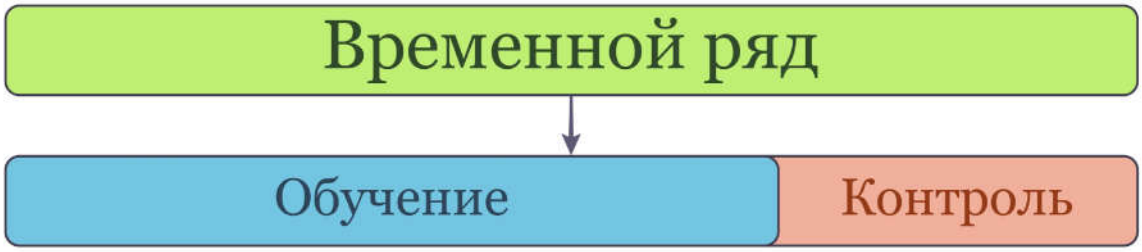
1. Методы оценки качества моделей прогнозирования

Чтобы оценить, насколько хорошо модель справляется с задачей прогнозирования, сравнить между собой различные модели или выбрать набор признаков, позволяющий наилучшим образом описать данные, используются различные методы оценки и сравнения качества.

1.1. Методы разбиения выборки

1.1.1. Отложенная выборка

Для того чтобы оценка была достоверной, неправильно сравнивать качество модели на тех же данных, на которых она обучалась, то есть на которых были оценены параметры модели. Очевидным способом оценки качества является разбиение всей выборки на две части: обучающую длины l и контрольную длины k.

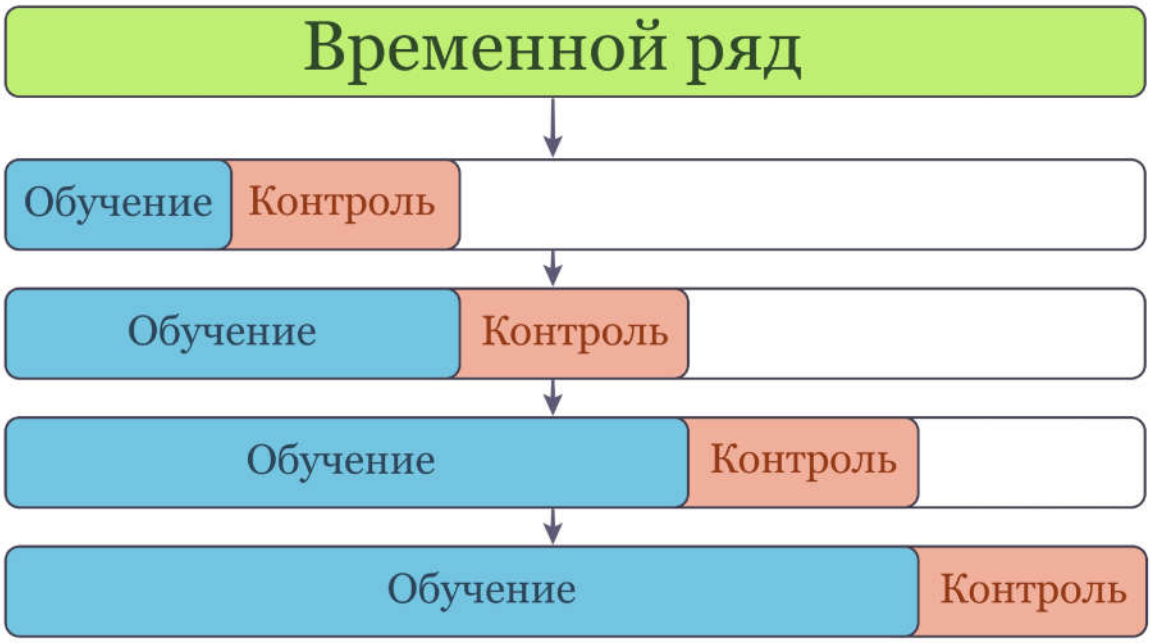


Упражнение

Какими недостатками обладает такой способ разбиения выборки?

1.1.2. Кросс-валидация на временных рядах

Для сглаживания недостатков предыдущего метода применяется так называемая кросс-валидация (или перекрестная проверка) на временных рядах. Особенность задачи прогнозирования временных рядов состоит в том, что оценка качества модели может производиться только последовательно.



Частным случаем метода кросс-валидации является контрольной по отдельным объектам (leave-one-out).

1.2. Метрики качества

Для получения числовой оценки качества прогнозов под конкретную задачу выбирается метрика качества. Рассмотрим некоторые из них.

1.2.1. Коэффициент детерминации

В эконометрике часто рассчитывают коэффициент детерминации R^2 для оценки качества аппроксимации модели:

$$R^2 = 1 - \frac{\sum_{t=k}^{T}(y_t - \hat{y}_t)^2}{\sum_{t=k}^{T}(y_t - \overline{y}_t)^2}.$$

Здесь числитель показывает дисперсию ошибки модели, а знаменатель – дисперсию рассматриваемого ряда. Таким образом, второе слагаемое показывает долю необъясненной моделью дисперсии ряда.

Коэффициент детерминации принимает значения от 0 до 1. Чем ближе значение к 1, тем сильнее зависимость. Коэффициент детерминации может принимать и отрицательные значения.

Серьезным недостатком R^2 является то, что его значение увеличивается с увеличением числа параметров модели, что не всегда является показателем того, что модель стала предсказывать лучше и обладает хорошей обобщающей способностью. Чтобы сгладить данный недостаток можно использовать скорректированный коэффициент детерминации (R^2-adjusted):

$$R^2_{\text{adj}} = 1 - \frac{K-1}{K-N} R^2,$$

где $K = T - k + 1$ – число наблюдений (в соответствие с обозначениями выше), N – число параметров модели.

1.2.2. Среднеквадратичная ошибка

Часто в качестве метрики качества берут непосредственно квадрат отклонения фактических значений от прогнозных (mean squared error), которая выступает в качестве оптимизируемой функции при оценке параметров модели:

$$\text{MSE} = \frac{1}{T-k+1} \sum_{t=k}^{T} (y_t - \hat{y}_t)^2.$$

Недостатки: квадрат сильно завышает ошибку, если прогнозируемые значения большие. Или занижает, если $-1 \le y \le 1$.

1.2.3. Средняя абсолютная ошибка

Наиболее интерпретируемой метрикой выступает средняя абсолютная ошибка (mean absolute error), которая показывает среднее отклонение прогнозных значений от фактических:

$$\text{MAE} = \frac{1}{T-k+1} \sum_{t=k}^{T} |y_t - \hat{y}_t|.$$

Недостатки: отсутствие возможности сравнивать качество, если решается задача прогнозирования нескольких временных рядов. Таким примером может выступать прогнозирование объемов продаж сети гипермаркетов, где разные продукты имеют разные единицы измерения (шт., кг).

1.2.4. Средняя абсолютная процентная ошибка

Для сравнения качества для нескольких рядов можно сравнивать не абсолютную ошибку, а процентную (mean absolute percentage error):

$$\text{MAPE} = \frac{1}{T-k+1} \sum_{t=k}^{T} \left| \frac{y_t - \hat{y}_t}{y_t} \right|.$$

Упражнение

1. Какие очевидные недостатки у данной метрики качества?
2. Предположим, фактическое значение $y_t = 20$, а оценка модели $\hat{y}_t$=100. Чему равна абсолютная процентная ошибка в данном случае? А если, наоборот, $y_t = 100$, а $\hat{y}_t$=20?

1.2.5. Средняя симметричная абсолютная ошибка

Идея симметричной абсолютной процентной ошибки: в знаменателе указать среднее между фактическим и прогнозным значениями:

$$\text{SMAPE} = \frac{1}{T-k+1}\sum_{t=k}^{T}\frac{2\,|y_t - \hat{y}_t|}{|y_t| + |\hat{y}_t|}.$$

Упражнение

Если фактическое значение $y_t = 100$, а оценка модели $\hat{y}_t$=110. Чему равна симметричная абсолютная процентная ошибка в данном случае? А если $y_t = 100$, а $\hat{y}_t$=90?

2. Модель линейной регрессии

1.1. Основные понятия

Обучающая выборка – набор пар $X^l = (x_i, y_i)_{i=1}^l$, в котором для каждого объекта $x_i \in X^l$, характеризующегося рядом признаков $(x_{i1}, x_{i2}, \ldots, x_{im})$, известно значение целевой переменной y_i.

Признаки – числовые характеристики рассматриваемых объектов $X \to D_j,\ j = 1, \ldots, M$.

В зависимости от принимаемых значений можно выделить следующие типы признаков:

- $D_j \in \{0, 1\}$ – бинарный;
- $|D_j| < \infty$ – номинальный;
- $|D_j| < \infty$, D_j упорядочено – порядковый;
- $D_j \in \mathbb{R}$ – количественный (вещественный).

Задача состоит в том, чтобы по обучающей выборке определить неизвестную зависимость $a : X \to Y$.

Предсказательная модель – параметрическое семейство функций

$$A = \{a(x) = g(x, \theta) \mid \theta \in \Theta\},$$

где $g : X \times \Theta \to Y$ – фиксированная функция, Θ – множество параметров модели.

Обучение модели – процесс настройки параметров модели $\theta \in \Theta$ путем решения задачи оптимизации:

$$Q(a, X^l) \to \min_{a \in A},$$

где $Q(a, X^l)$ – функционал качества алгоритма a на обучающей выборке X^l. Величина ошибки на каждом объекте $x \in X^l$ называется **функцией потерь** $L(a, x)$, таким образом,

$$Q(a, X^l) = \frac{1}{l}\sum_{i=1}^{l} L(a, x_i).$$

2.1. Общий вид модели

В соответствии с введенными ранее обозначениями, модель линейной регрессии может быть представлена в виде:

$$a(x) = g(x,\, \theta) = \theta_0 + \sum_{j=1}^{m} \theta_j\, x_j,$$

где θ_0 – свободный член. Если добавить фиктивный признак $x_0 = 1$ для каждого объекта, то модель линейной регрессии можно переписать несколько проще:

$$a(x) = g(x,\, \theta) = \sum_{j=0}^{m} \theta_j\, x_j = \left\langle \vec{\theta},\, \vec{x} \right\rangle = \vec{\theta}^{\,T}\, \vec{x}.$$

Таким образом, наблюдаемые значения y описываются следующим выражением:

$$\vec{y} = X\vec{\theta} + \epsilon,$$

где X – матрица «объекты-признаки», ϵ – случайная непрогнозируемая ошибка модели.

2.2. Метод наименьших квадратов

Для оценки неизвестных параметров $\theta \in \Theta$ в модели линейной регрессии применяется метод наименьших квадратов, который состоит в минимизации следующего функционала:

$$Q\left(a,\, X^l\right) = \frac{1}{2\,l} \sum_{i=1}^{l} (y_i - g(x_i,\, \theta))^2 \to \min_{\theta}.$$

Оценка наименьших квадратов оптимальна в классе линейных несмещенных моделей в соответствии с теоремой Гаусса-Маркова. При этом должны выполняться следующие требования:

- ошибки не носят систематического характера: $\forall\, i : \mathbb{E}(\epsilon_i) = 0$;
- дисперсия ошибок одинакова и конечна (ошибки гомоскедастичны): $\forall\, i :\, \mathrm{Var}(\epsilon_i) = \sigma^2 < \infty$;
- случайные ошибки некоррелированы: $\forall\, i \neq j :\, \mathrm{Cov}(\epsilon_i,\, \epsilon_j) = 0$.