Во втором примере несимметричная функция ошибки, являющаяся кусочно-линейной функцией, имеет вид

$$S(\mathbf{w}) = \sum_{i\in\mathcal{I}}\sum_{s}\begin{cases} a_s + b_s(f(\mathbf{w},\mathbf{x}_i) - y_i), \text{ при } f \in (z_{s-1}, z_s); \\ 0, \text{ в противном случае.} \end{cases} \tag{29}$$

Параметры a_s, b_s и концы отрезков z_s выбираются согласно расчетам убытков при совершении торговых операций при условии непрерывности функции ошибки и её первой производной. В обоих случаях происходит отказ от принятия гипотезы порождения данных, и функция ошибки $S(\mathbf{w})$ оптимизируется, исходя из условий поставленной задачи. Например, при восстановлении регрессии измерений некоторых физических величин используется метод наименьших модулей [?, ?, ?], согласно которому функция ошибки задана как сумма модулей регрессионных остатков. Задача нахождения минимального значения функций вида (??) или (??) решается методами линейного программирования. В таблице ?? приведен набор функций ошибок, часто используемых при решении задач прогнозирования.

Таблица 3. Функции ошибок регрессионных моделей.

Среднее арифметическое модулей остатков	$\text{MAE} = \frac{1}{m}\sum\limits_{i=1}^{m}\lvert\varepsilon_i\rvert$
Среднее арифметическое модулей относительных остатков	$\text{MAPE} = \frac{1}{m}\sum\limits_{i=1}^{m}\lvert\frac{\varepsilon_i}{y_i}\rvert$
Среднее отклонение модулей остатков	$\text{PMAD} = \sum\limits_{i=1}^{m}\lvert\varepsilon_i\rvert\,(\sum_{i=1}^{m}\lvert y_i\rvert)^{-1}$
Среднеквадратичная ошибка	$\text{MSE} = \frac{1}{m}\sum\limits_{i=1}^{m}\varepsilon_i^2$
Корень среднеквадратичной ошибки	$\text{RMSE} = \frac{1}{\sqrt{m}}\sqrt{\sum_{i=1}^{m}\varepsilon_i^2}$
Сила прогноза	$\text{SS} = 1 - \frac{\text{MSE}_{\text{forecast}}}{\text{MSE}_{\text{history}}}$

Функция ошибки и разбиение выборки. В данной работе не предполагается, что для оценки наиболее вероятных параметров модели, либо для выбора наиболее правдоподобной модели из некоторого множества требуется разбиение множества индексов $\mathcal{I}$ элементов выборки $\mathfrak{D} = \{(\mathbf{x}_i, y_i)\}, i \in \mathcal{I}$ на обучающую и контрольную: $\mathcal{I} = \mathcal{L} \sqcup \mathcal{C}$. Тем не менее, следует отметить, что при выборе моделей такое разбиение является одним из наиболее эффективных способов избежать переобучения, см. [?, ?, ?]. Поэтому ниже приведен ряд примеров эвристических функций ошибок, предложенных авторами метода группового учета аргументов. Данные функции называются критериями. Значительная их часть опубликована на сайте [?].

Используемые в этом подразделе обозначения $\mathbf{X}_\mathcal{C}, \mathbf{y}_\mathcal{C}, \mathbf{w}_\mathcal{L}$ означают, что значения переменных $\mathbf{X}, \mathbf{y}, \mathbf{w}$ фиксированы, в выборку $(\mathbf{X}_\mathcal{C}, \mathbf{y}_\mathcal{C})$ вошли только объекты с индексами из множества $\mathcal{C} \in \mathcal{I} \neq \emptyset$, а оцека вектора параметров $\mathbf{w}_\mathcal{L}$ получена с использованием выборки, состоящей из элементов с индексами из множества $\mathcal{L} \subset \mathcal{I} \neq \emptyset$:

$$\mathbf{w}_\mathcal{L} = \arg\min_{i\in\mathcal{L}\subset\mathcal{I}} S(\mathbf{w}|\mathbf{X}_\mathcal{L}, \mathbf{y}_\mathcal{L}, f).$$

При этом считается, что множество индексов $\mathcal{I}$ элементов выборки разбито на подмножества

$$\mathcal{I} = \mathcal{L} \sqcup \mathcal{C} \sqcup \mathcal{V},$$

в котором $\mathcal{L}$ — обучающая выборка, $\mathcal{C}$ — котрольная выборка, $\mathcal{V}$ — валидационная выборка. Последняя в ряде задач может быть пустой.

Метод группового учета аргументов [?, ?, ?] использует внутренний и внешний критерий, так как при оценке параметров моделей и при выборе моделей используются разные элементы выборки. *Внутренний критерий* используется для оценки параметров: их значения оцениваются на подвыборке элементов с индексами из $\mathcal{L}$. Выбор моделей производится с помощью *внешнего критерия*, значение которого вычисляется на множестве $\mathcal{C}$. При выборе минимум внешнего критерия означает, что модель, доставляющая такой минимум, является искомой.

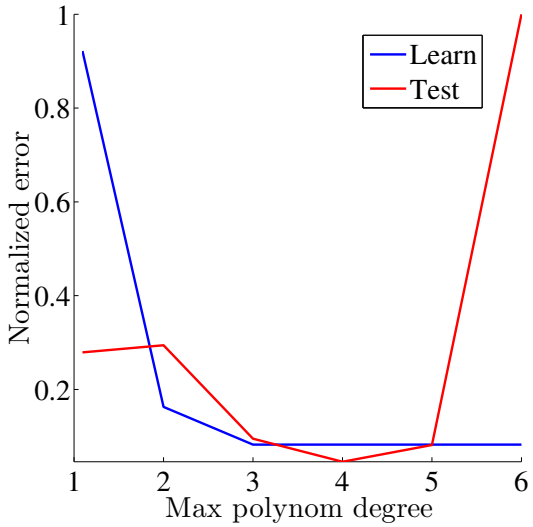

Рис. 5. Внешний и внутренний критерии при различных значениях структурной сложности.

Критерий регулярности S_{Δ^2} равен норме разности вектора значений зависимой переменной и вектора значений функции регрессии на тестовой подвыборке $\mathcal{C}$ при параметрах, оцененных на обучающей подвыборке $\mathcal{L}$.

$$S_{\Delta^2_{\mathcal{C}}} = \|\mathbf{y}_{\mathcal{C}} - \mathbf{X}_{\mathcal{C}}\mathbf{w}_{\mathcal{L}}\|^2,$$

где

$$\mathbf{w}_{\mathcal{L}} = (\mathbf{X}^{\mathsf{T}}_{\mathcal{L}}\mathbf{X}_{\mathcal{L}})^{-1}(\mathbf{X}^{\mathsf{T}}_{\mathcal{L}}\mathbf{y}_{\mathcal{L}}).$$

Этот критерий может быть нормирован выражениями $\|\mathbf{y}_{\mathcal{L}}\|^2$ или $\|\mathbf{y}_{\mathcal{L}} - \mathrm{mean}(\mathbf{y}_{\mathcal{L}})\|^2$.

Критерий предсказательной способности — модификация критерия регулярности для задач прогнозирования. Этот критерий включает среднеквадратичную ошибку для валидационной выборки $\mathcal{V}$, которая не используется ни при оценке параметров, ни при выборе модели. В этом случае выборка делится на три части. Критерий предсказательной способности имеет вид

$$S_{\Delta^2_{\mathcal{V}}} = \frac{\|\mathbf{y}_{\mathcal{I}} - \mathbf{X}_{\mathcal{I}}\mathbf{w}_{\mathcal{V}}\|^2}{\|\mathbf{y}_{\mathcal{I}} - \mathrm{mean}(\mathbf{y}_{\mathcal{I}})\|^2}.$$

Критерий минимального смещения или *критерий непротиворечивости*: модель, которая имеет на обучающей и на контрольной выборках различные векторы невязок, называется противоречивой. Критерий задан разностью между значениями функции регрессии, вычисленными на двух различных выборках, заданных множествами $\mathcal{L}$ и $\mathcal{C}$ и требует, чтобы оценки параметров, вычисленные на этих выборках, различались минимально. Он имеет вид:

$$S_{\eta^2_{\text{bs}}} = \|\mathbf{X}_{\mathcal{I}}\mathbf{w}_{\mathcal{C}} - \mathbf{X}_{\mathcal{I}}\mathbf{w}_{\mathcal{L}}\|^2,$$

модификация:

$$S_{\eta^2_{\text{a}}} = \|\mathbf{w}_{\mathcal{C}} - \mathbf{w}_{\mathcal{L}}\|^2,$$

где $\mathbf{w}_{\mathcal{C}}$ и $\mathbf{w}_{\mathcal{L}}$ — векторы параметров, полученные с использованием подвыборок $\mathcal{C}$ и $\mathcal{L}$.

Критерий иммунитета к шуму имеет вид

$$\begin{aligned} S_{V^2} =& (\mathbf{X}_{\mathcal{I}}\mathbf{w}_{\mathcal{C}} - \mathbf{X}_{\mathcal{I}}\mathbf{w}_{\mathcal{I}})^{\intercal}(\mathbf{X}_{\mathcal{I}}\mathbf{w}_{\mathcal{I}} - \mathbf{X}_{\mathcal{I}}\mathbf{w}_{\mathcal{L}}) = \\ & (\mathbf{w}_{\mathcal{C}} - \mathbf{w}_{\mathcal{I}})^{\intercal}\mathbf{X}_{\mathcal{I}}^{\intercal}\mathbf{X}_{\mathcal{I}}(\mathbf{w}_{\mathcal{I}} - \mathbf{w}_{\mathcal{L}}), \end{aligned}$$

где $\mathbf{w}_{\mathcal{I}}$ — вектор параметров, полученный с использованием полной выборке $\mathcal{I}$. Утверждается [**?**], что с помощью этого критерия в сильно зашумленных данных можно найти скрытые закономерности.

Комбинированный критерий позволяет использовать при выборе моделей линейную комбинацию нескольких критериев. Комбинированный критерий

$$S_{\kappa^2} = \sum_{k=1}^{K} v_k S_k, \quad \text{при условии} \quad \sum_{k=1}^{K} v_k = 1.$$

Здесь S_k — принятые на рассмотрение критерии, а v_k — веса этих критериев, назначенные в начале вычислительного эксперимента.

Используются также нормализованные значения критериев. При этом предыдущая формула имеет вид

$$S_{\kappa^2} = \sum_{i=1}^{K} v_k \frac{S_k}{\max\limits_{f \in \mathfrak{F}}(S_k)}.$$

Максимальное значение критерия $\max(S_k)$ берется по вычисленным значениям критериев $S_k(f)$ для всех порожденных моделей $f \in \mathfrak{F}$.

1.3. Задачи регрессионного анализа

Задача восстановления регрессии (**??**) имеет несколько разных постановок, каждую их которых можно условно отнести к одному из следующих типов:

1) задачи оценки параметров модели,

2) задачи выбора признаков или объектов регрессионной выборки,

3) задачи выбора регрессионных моделей,

4) задачи проверки гипотезы порождения данных.