Методы борьбы с переобучением

1. Проблема переобучения

1.1. Понятия недообучения и переобучения

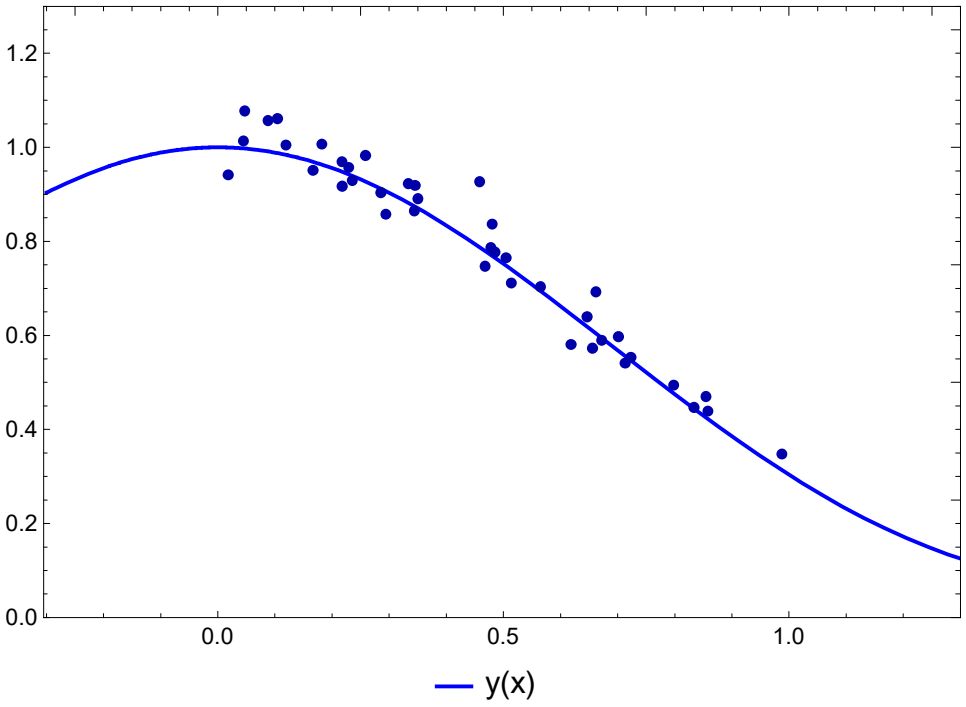
Недообучением называют нежелательное явление, когда алгоритм обучения не обеспечивает достаточно малой величины средней ошибки на обучающей выборке. Недообучение возникает при использовании недостаточно сложных моделей.

Переобучение, переподгонка (overtraining, overfitting) – нежелательное явление, когда вероятность ошибки обученного алгоритма на объектах тестовой выборки оказывается существенно выше, чем средняя ошибка на обучающей выборке. Переобучение возникает при использовании избыточно сложных моделей.

Таким образом, важно следить, чтобы модель хорошо описывала данные, но при этом не слишком настраивалась под них.
Говорят, что алгоритм обучения обладает **обобщающей способностью**, если вероятность ошибки на тестовой выборке достаточно мала или хотя бы предсказуема, то есть не сильно отличается от ошибки на обучающей выборке.

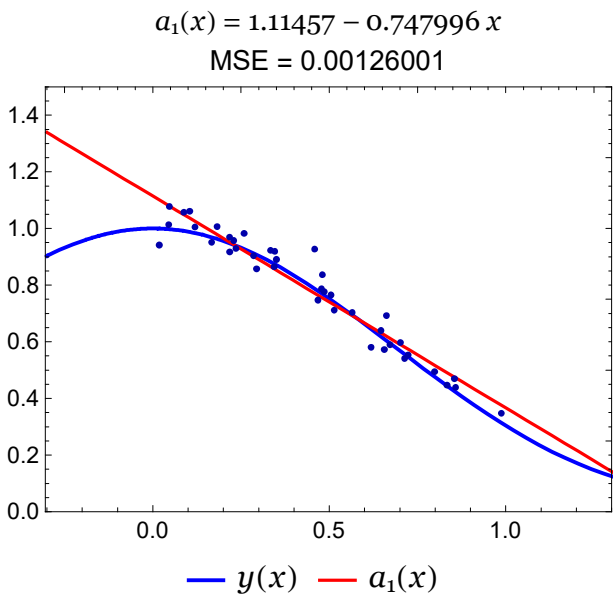
Пример

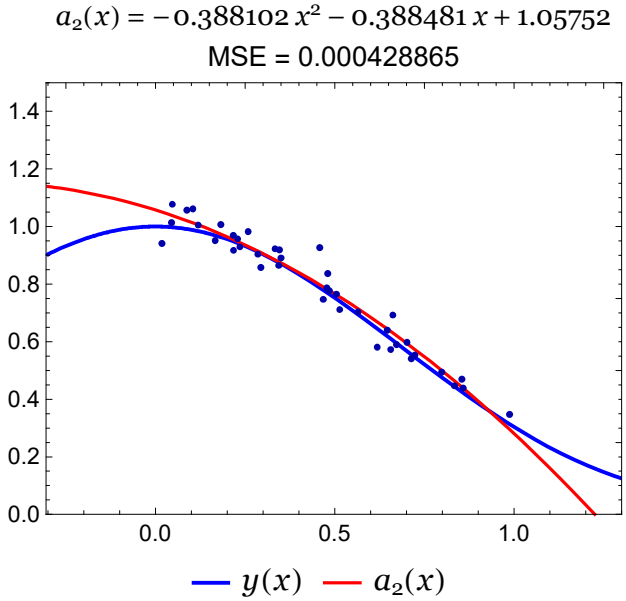
Рассмотрим функцию $y(x) = \cos\left(\frac{3}{2}\sin(x)\right)$. В качестве обучающей выборки возьмем 30 точек, сгенерированных по данному правилу с добавлением некоторого шума.



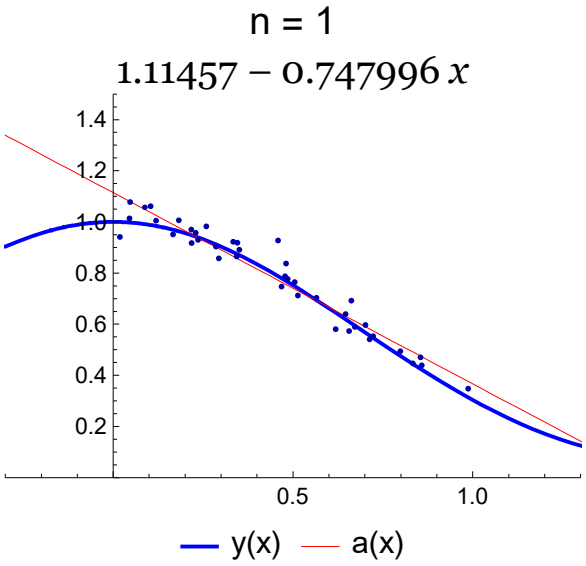
Для аппроксимации данной зависимости возьмем полиномы первой и второй степеней:

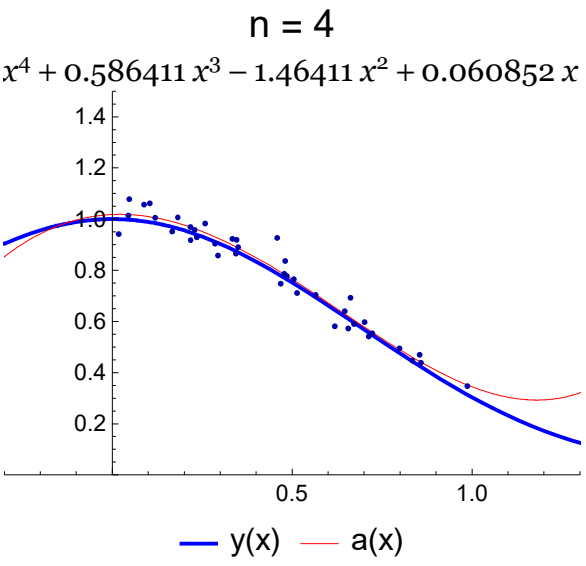
$$a_1(x) = \theta_0 + \theta_1 x,$$
$$a_2(x) = \theta_0 + \theta_1 x + \theta_2 x^2.$$

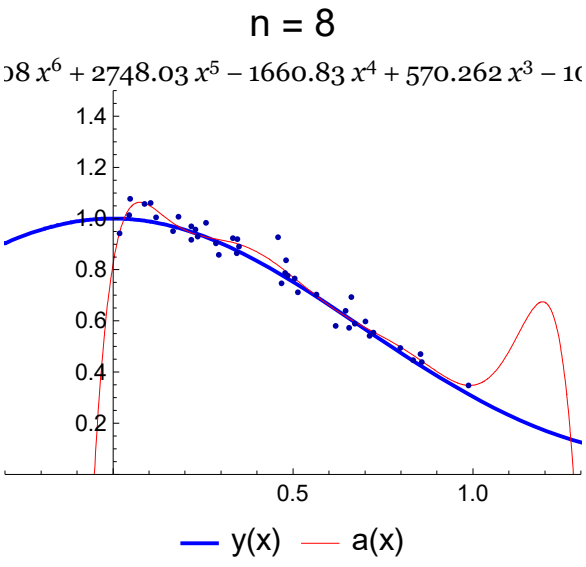


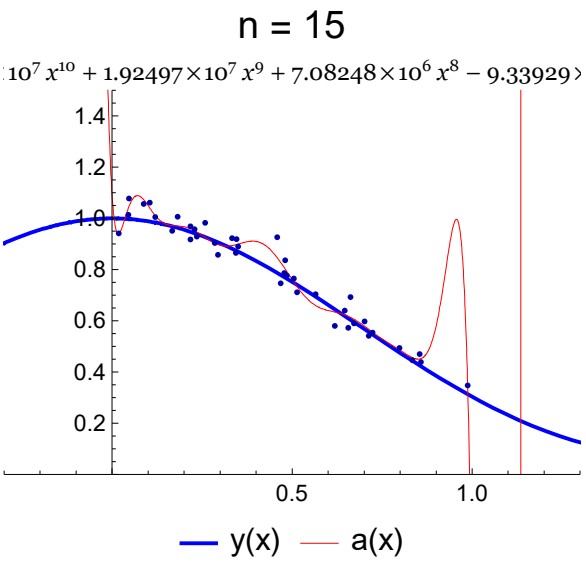


Очевидно, сложность модели недостаточная для описания исходной зависимости, что говорит о недообучении. Попробуем взять полиномы более высоких степеней. Заметим, что полином третьей степени неплохо описывает данные, однако с увеличением степени полинома, модель начинает «подстраиваться» под обучающую выборку.









Overfitting Thriller: https://www.youtube.com/watch?v=DQWI1kvmwRg

2. Регуляризация

Если используется слишком сложная модель, а данных недостаточно, чтобы точно определить ее параметры, эта модель легко может получиться переобученной. Бороться с этим можно различными способами:

- взять больше данных. Такой подход редко бывает доступен, поскольку дополнительные данные стоят дополнительных денег, а также иногда недоступны совсем. Например, в задачах веб-поиска, несмотря на наличие терабайтов данных, эффективный объем выборки, описывающей персонализированные данные, существенно ограничен: в этом случае можно использовать только историю посещений данного пользователя;
- выбрать более простую модель или упростить модель, например исключив из рассмотрения некоторые признаки. Процесс отбора признаков представляет собой нетривиальную задачу. В частности, не понятно, какой из двух похожих признаков следует оставлять, если признаки сильно зашумлены;
- использовать регуляризацию. Ранее было показано, что у переобученной линейной модели значения параметров в модели становятся огромными и разными по знаку. Если ограничить значения весов модели, то с переобучением можно до какой-то степени бороться.

Рассмотрим подробнее методы регуляризации. Основными способами являются: добавление L_2-регуляризатора (ridge-регрессия или гребневая регрессия) и добавление L_1-регуляризатора (lasso-регрессия или лассо регрессия, least absolute shrinkage and selection operator).

2.1. Гребневая регрессия

Метод наименьших квадратов состоит в минимизации функционала качества:

$$Q(a,\, X^l) = \sum_{i=1}^{l}\left(\sum_{j=0}^{m}\theta_j\, x_{ij} - y_i\right)^2 \to \min_{\theta}.$$

Из соображения, что большие значения параметров θ приводят к переобучению, добавим в функционал качества «штраф» на слишком большие значения θ:

$$Q(a,\, X^l) = \sum_{i=1}^{l}\left(\sum_{j=0}^{m}\theta_j\, x_{ij} - y_i\right)^2 + \lambda\sum_{j=1}^{m}\theta_j^2 \to \min_{\theta}.$$

Такая модель называется гребневой регрессией.

2.2. Лассо регрессия

Отличие лассо регрессии лишь в том, что штрафующим слагаемым выступает модуль:

$$Q(a,\, X^l) = \sum_{i=1}^{l}\left(\sum_{j=1}^{m}\theta_j\, x_{ij} - y_i\right)^2 + \lambda\sum_{j=0}^{m}\left|\theta_j\right| \to \min_{\theta}.$$

2.3. Особенности регуляризаторов

Рассмотрим особенности каждого из регуляризаторов.

Пусть матрица «объекты-признаки» X является единичной матрицей размера $l \times l$:

$$X = \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 1 \end{pmatrix}.$$

Тогда при решении задачи линейной регрессии использование метода наименьших квадратов без регуляризации:

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

дает следующий вектор θ:

$$\theta_i^* = y_i.$$

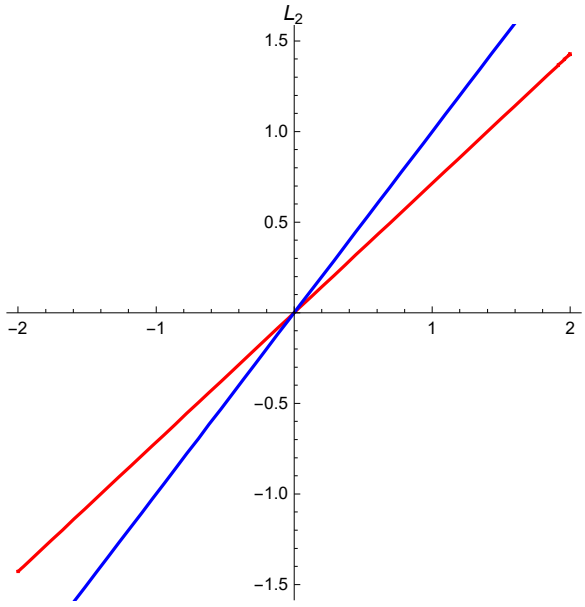
При использовании гребневой регуляризации компоненты вектора θ имеют вид:

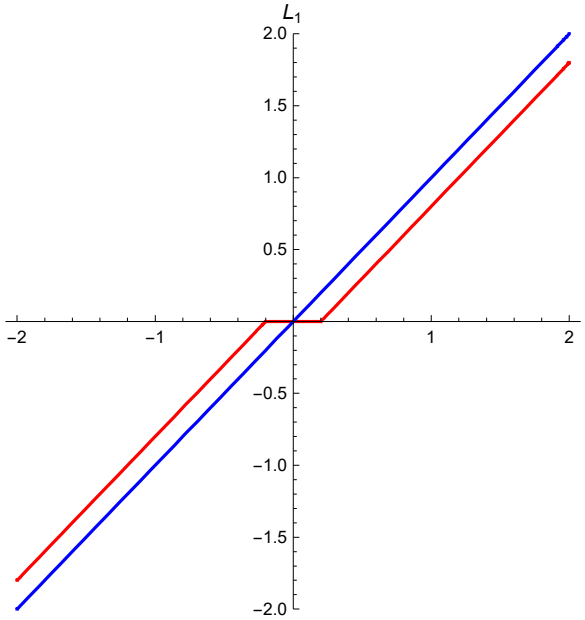
$$\theta_i^* = \frac{y_i}{1+\lambda},$$

а при использовании лассо регуляризации:

$$\theta_i^* = \begin{cases} y_i - \frac{\lambda}{2}, & \theta_i > 0, \\ y_i + \frac{\lambda}{2}, & \theta_i < 0, \\ 0, & \theta_i = 0. \end{cases}$$

При использовании только МНК без регуляризации $\theta_i^* = y_i$. Соответствующая линия изображена синей линией на обоих графиках. При использовании L_2-регуляризации зависимость θ_i^* от y_i все еще линейная, компоненты вектора весов ближе расположены к нулю.





В случае L_1-регуляризации график выглядит несколько иначе: существует область (размера λ) значений y_i, для которых $\theta_i = 0$. То есть lasso, или L_1-регуляризация, позволяет отбирать признаки, а именно: параметры (веса) признаков, обладающих низкой предсказательной способностью, оказываются равными нулю.