GeekBrains

Логистическая регрессия. Разделение данных и валидация модели

Что будет на уроке?

1. Bias-variance tradeoff.
2. Тренировочная, тестовая и валидационная выборки.
3. Гиперпараметры и их подбор.
4. Проблема переобучения и недообучения.
5. Логистическая регрессия.

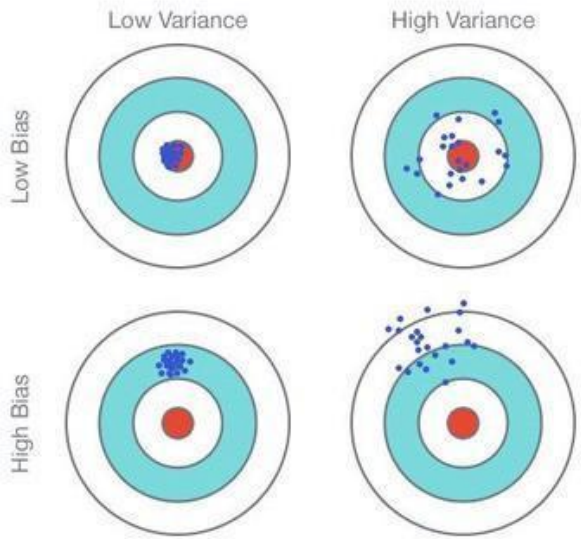


Fig. 1: Graphical Illustration of bias-variance trade-off , Source: Scott Fortmann-Roe., Understanding Bias-Variance Trade-off

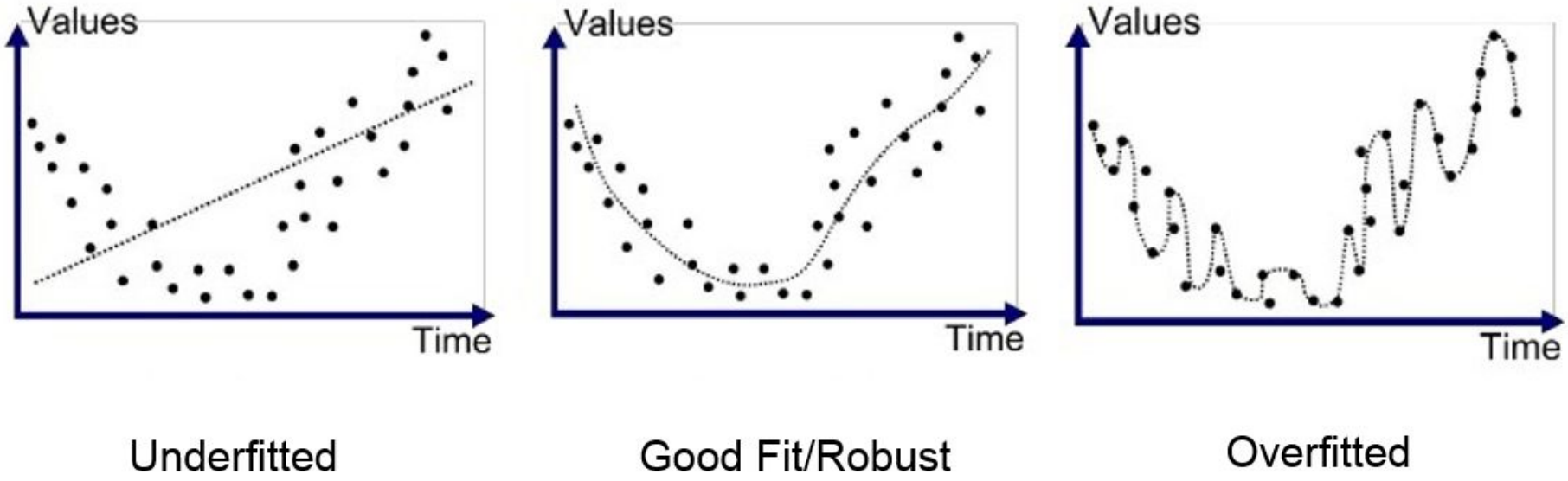
Bias-Variance Tradeoff

- **Bias (смещение)** — ошибка, возникающая в результате того, что алгоритм не научился находить связь между признаками и целевой переменной.

- **Variance (дисперсия)** — ошибка под влиянием отклонений в данных. Модель видит связи там, где их нет.

src: https://medium.com/@mp32445/understanding-bias-variance-tradeoff-ca59a22e2a83

Переобучение и недообучение



src: https://medium.com/greyatom/what-is-underfitting-and-overfitting-in-machine-learning-and-how-to-deal-with-it-6803a989c76

Регуляризация

Регуляризация - процесс искусственного ограничения обучения модели.

В основном различают 2 вида регуляризации:

- L1 регуляризацию (Lasso)
- L2 регуляризацию (Ridge)

L2 регуляризация

$$\sum_{i=1}^{n}(y_i - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

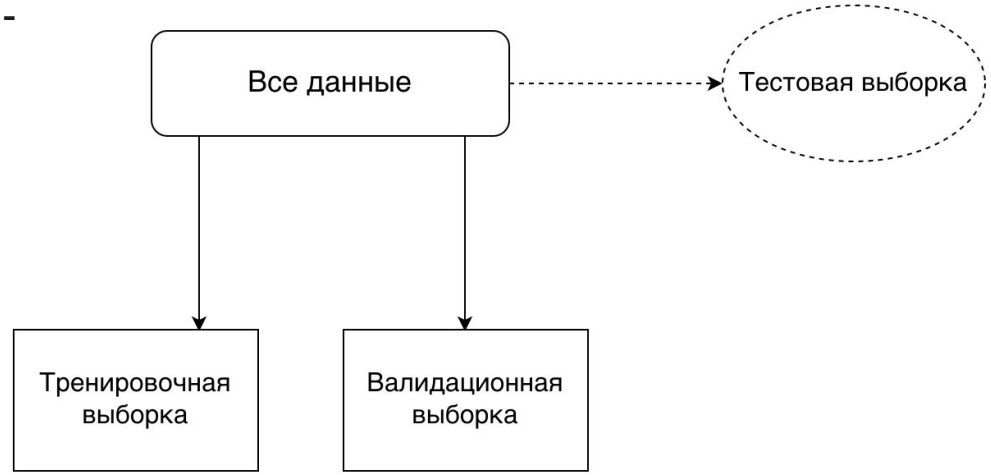
L1 регуляризация

$$\sum_{i=1}^{n}(Y_i - \sum_{j=1}^{p} X_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

Разделение данных

В некоторых случаях, например, на DS-контестах, тестовые данные не предоставляются.

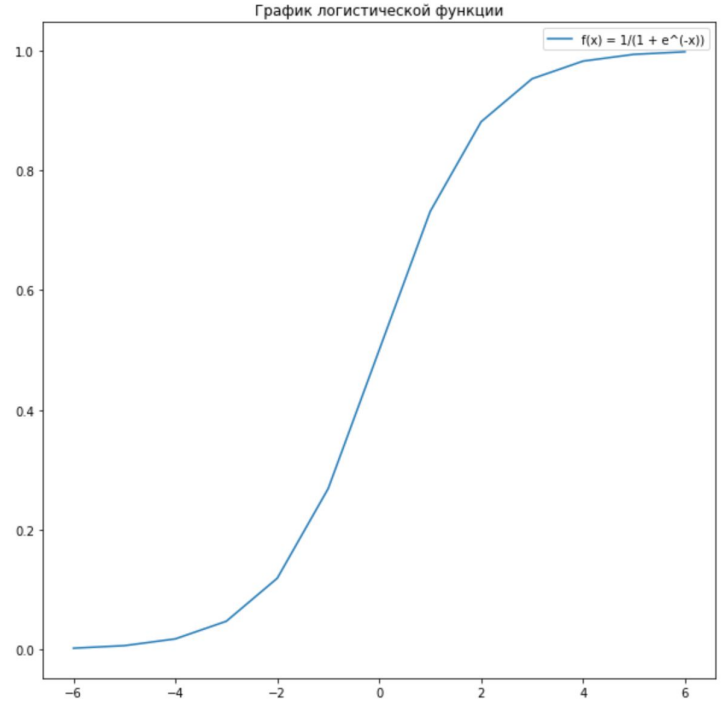
Тогда целесообразно выделить их самостоятельно из тренировочных данных.



Логистическая регрессия

Логистическая регрессия (англ. logit model) — статистическая модель для задачи классификации.

1. Может найти нелинейные зависимости.
2. Более устойчива к выбросам.



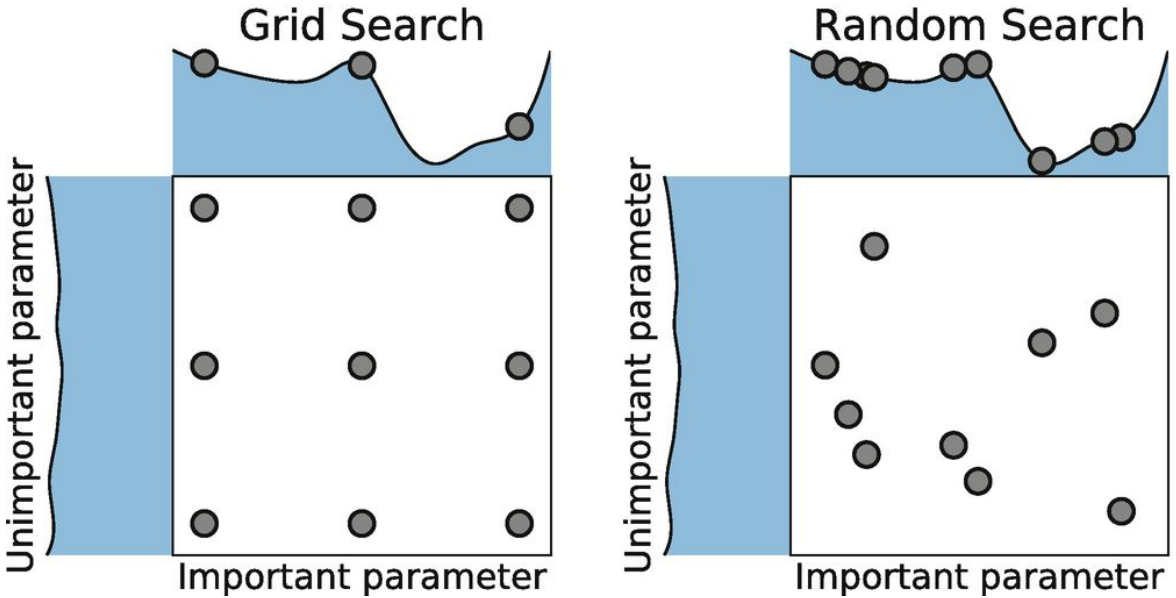
Гиперпараметры и их подбор

В машинном обучении **гиперпараметры** — параметры алгоритмов, которые задаются до начала процесса обучения. Они используются для гибкой настройки алгоритма под конкретную задачу.

Примеры гиперпараметров:

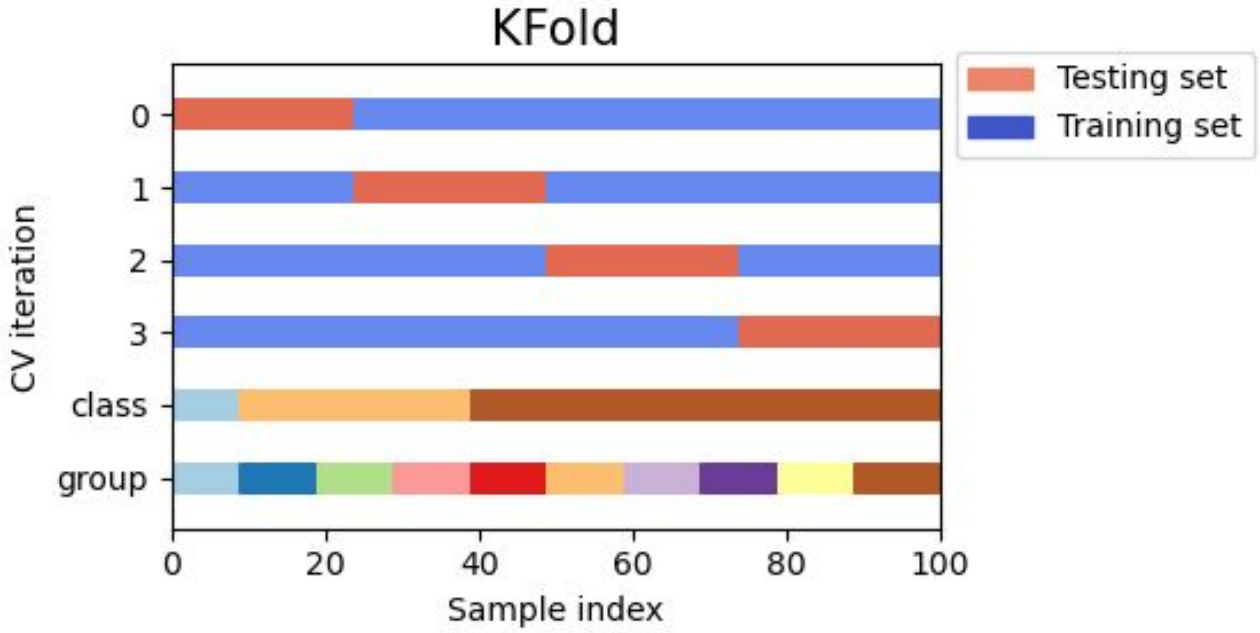
1. Количество эстиматоров (estimators).
2. Скорость обучения (learning rate).
3. Значения дискриминационного порога.
4. Вероятность отключения нейрона.

Случайный поиск и поиск по сетке



src: https://medium.com/@cjl2fv/an-intro-to-hyper-parameter-optimization-using-grid-search-and-random-search-d73b9834ca0a

KFold и его разновидности



https://scikit-learn.org/stable/modules/cross_validation.html#k-fold

Модели машинного обучения. Линейные модели

Плюсы

+ Быстрые
+ Легко интерпретируются
+ Подходят для работы с по-настоящему «большими» данными

Минусы

– Обычно уступают в точности более сложным моделям

Всегда используйте линейные модели как отправную точку.

Модели машинного обучения. Ансамбли

Плюсы

\+ Высокая точность

Минусы

– Тяжелы в вычислениях

– Плохо интерпретируемые

– Тяжело с большими данными

Стоит использовать, когда данных относительно немного и важны метрики, а не интерпретация.

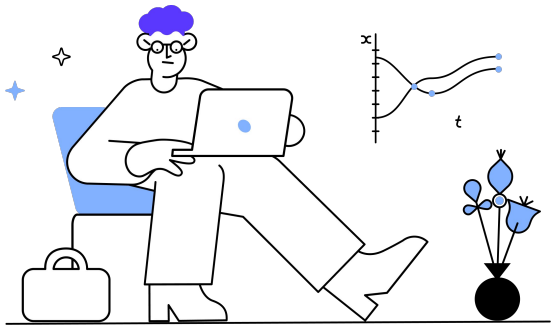
Модели машинного обучения. Нейросети

Плюсы

+ Хорошо работают на нетабличных данных и нестандартных задачах (NLP, CV и т. д.)

Минусы

– ОЧЕНЬ тяжёлые вычисления
– Плохо интерпретируемые
– Дорого обучать



Стоит использовать только для специфичных задач и только специалистам, иначе вы, скорее всего, зря потратите время и деньги.

Практическое задание

1. Изучить методические материалы к занятию.
2. Пройти тест с выбором варианта ответа.