

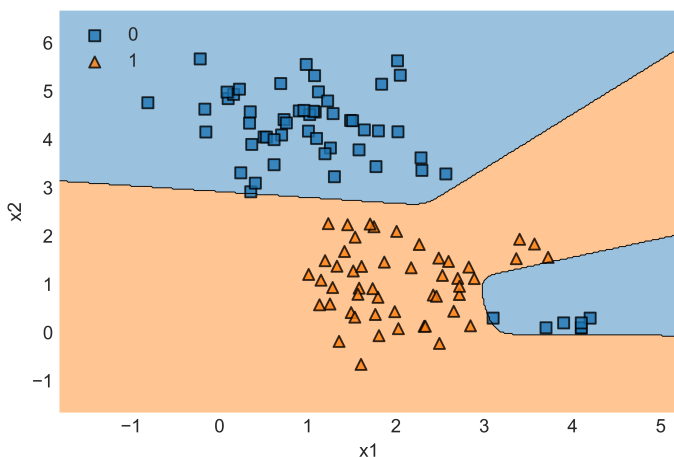
Переобучение и регуляризация

Переобучение — состояние, когда модель показывает высокое качество на тренировочных данных, но низкое на валидационных или тестовых.

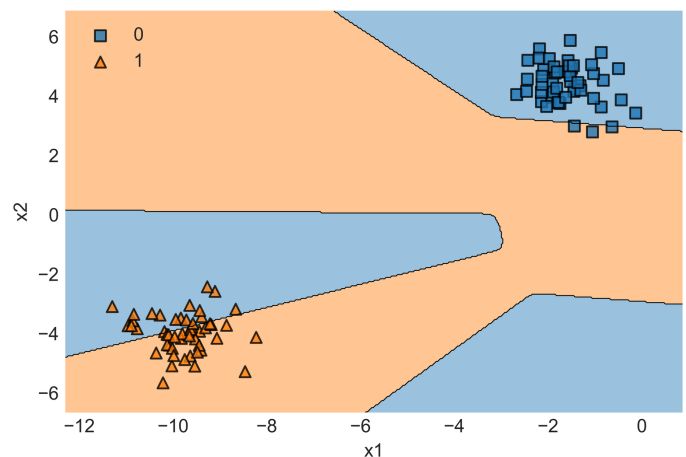
Недообучение — состояние, когда модель слишком простая и не уловила закономерности в данных.

Обнаружение проблемы переобучения

Переобученная модель реагирует на выбросы и описывает каждое наблюдение. Чрезмерная подгонка модели под тренировочные данные ведёт к уменьшению предсказательной силы: модель будет хуже классифицировать наблюдения из тестовой выборки:



Предсказания переобученной модели на тренировочной выборке



Предсказания переобученной модели на тестовой выборке

Существует дилемма между точностью и гибкостью модели. Нужно искать баланс — подбирать модель, которая хорошо описывает данные, но при этом не слишком сильно под них подстраивается.

Регуляризация

Регуляризация — это метод борьбы с переобучением, который добавляет дополнительные ограничения на модель за то, что она подстраивается под тренировочные данные.

- **L2-регуляризация**, или ридж-регуляризация. Накладывает на модель штраф за сумму квадратов коэффициентов. Чем выше λ , тем больше второе слагаемое в формуле.

$$(f(w, x_i) - y)^2 + \lambda ||w||^2 \rightarrow \min$$

- **L1-регуляризация**, или лассо-регуляризация. Накладывает штраф на основе абсолютных значений коэффициентов модели.

$$(f(w, x_i) - y)^2 + \lambda ||w|| \rightarrow \min$$

Чтобы регуляризовать логистическую или линейную регрессию, нужно во время инициализации модели указать в настройке **penalty** подходящий вам тип: 'l1', 'l2', 'elasticnet', а также указать силу регуляризации **C**.

Для регуляризации моделей SVM нужно при инициализации выбрать значения коэффициентов **C** и **gamma**.