Вопросы занятия

1. Обучающая и тестовая выборка, кросс-валидация;

2. Метрики качества: *accuracy, precision, recall*;

3. Смещение и разброс (*bias-variance tradeoff*);

4. Признаки переобучения и регуляризация.

В конце занятия научимся:

- проводить кросс-валидацию модели;
- оценивать качество разных версий модели по AUC;
- подбирать параметры модели, чтобы бороться с переобучением.

Обучающая и тестовая выборка

ОБУЧАЮЩАЯ ВЫБОРКА

Содержит значения признаков и целевой переменной.

На обучающей выборке строим модель.

Обучающая и тестовая выборка

ТЕСТОВАЯ ВЫБОРКА

Содержит значения признаков, по которым необходимо предсказать значение целевой переменной.

Оцениваем качество различных вариантов модели.

Обучающая и тестовая выборка

ПРОБЛЕМЫ

Модель может хорошо работать на обучающей выборке, однако сильно терять в качестве на тестовой (один из вариантов - переобучение).

Преобразования данных на обучающей выборке должны быть повторены и иметь смысл для тестовой.

Обучающая и тестовая выборка

РАЗБИВАЕМ ОБУЧАЮЩУЮ ВЫБОРКУ

Разбиваем обучающую выборку на 2 части. На одной будем тренировать модель, на другой – проверять (т. е. использовать в качестве тестовой, только с известной целевой переменной)

```
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split( X, y, test_size = 0.3, random_state = 0 )
```

PRECISION RECALL
ТОЧНОСТЬ И ПОЛНОТА

Оценка качества модели

ПОРОГ ДЛЯ ТЕСТОВОЙ ВЫБОРКИ

```
model = LogisticRegression()
```

```
model.fit(X_train, y_train)
predictions = model.predict_proba(X_test)
```

```
zip(predictions[:, 1], y_test)
```

```
[(0.64583193796528038, 0),
 (0.075906148028446599, 0),
 (0.2704606033743272, 0),
 (0.26938542699540474, 0),
 (0.26433391263337475, 1),
 (0.1443590034736055, 0),
 (0.17840859560894495, 0),
 (0.21871761029690232, 0),
 (0.75293068528621931, 1),
 (0.2694630112685994, 0),
 (0.11209927315788928, 0),
 (0.18717054508217956, 0),
 (0.081787486664569364, 0),
```

Выберем порог, выше которого будем считать полученное значение принадлежащим 1. А ниже – нулю.

Это определит долю угаданных моделью значений

Оценка качества модели

МАТРИЦА ОШИБОК ДЛЯ ПОРОГА

	Actual positive	Actual negative
Predicted positive	True positive	False positive
Predicted negative	False negative	True negative

False positive — ошибка I рода (ложная тревога).

False negative — ошибка II рода (пропуск цели).

Оценка качества модели

ТОЧНОСТЬ

	Actual positive	Actual negative
Predicted positive	True positive	False positive
Predicted negative	False negative	True negative

Accuracy – доля правильно предсказанных от всех вариантов.

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN}$$

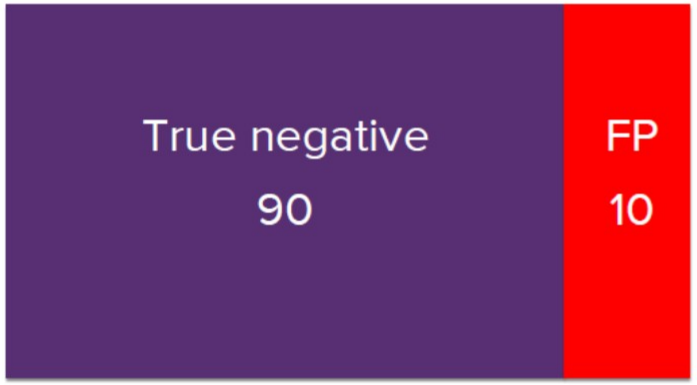
ПРАКТИКА

Logres_affair.IPYNB

Оценка качества модели

ПОЧЕМУ ТОЧНОСТИ НЕДОСТАТОЧНО

100 обычных писем



10 спам-писем

На почту пришло 100 обычных писем. И 10 писем спама.

Наша модель из 100 обычных 10 классифицировала как спам. Из 10 спам-писем – 5 как спам

Оценка качества модели

ПОЧЕМУ ТОЧНОСТИ НЕДОСТАТОЧНО

	Actual positive	Actual negative
Predicted positive	5	5
Predicted negative	10	90

Accuracy – доля правильно предсказанных от всех вариантов.

$$Accuracy = \frac{5 + 90}{5 + 90 + 10 + 5} = 86\%$$

Оценка качества модели

ПОЧЕМУ ТОЧНОСТИ НЕДОСТАТОЧНО

100 обычных писем

True negative
100

10 спам-писем

False negative
10

Возьмём модель, которая считает все письма обычными.

Оценка качества модели

ПОЧЕМУ ТОЧНОСТИ НЕДОСТАТОЧНО

	Actual positive	Actual negative
Predicted positive	0	10
Predicted negative	0	100

Возьмем модель, которая считает все письма обычными

$$Accuracy = \frac{0 + 100}{0 + 100 + 0 + 10} = 91\%$$

Оценка качества модели

PRECISION

	Actual positive	Actual negative
Predicted positive	True positive	False positive
Predicted negative	False negative	True negative

Precision – доля правильно предсказанных среди причисленных моделью к категории 1.

$$Precision = \frac{TP}{TP + FP}$$

Способность алгоритма отличать данный класс от других.

Оценка качества модели

RECALL

	Actual positive	Actual negative
Predicted positive	True positive	False positive
Predicted negative	False negative	True negative

Recall — доля правильно предсказанных среди категории 1.

$$Recall = \frac{TP}{TP + FN}$$

Синоним — True Positive Rate (sensivity)

Способность алгоритма обнаруживать данный класс вообще

Оценка качества модели

PRECISION И RECALL ДЛЯ СПАМА

100 обычных писем

True negative
100

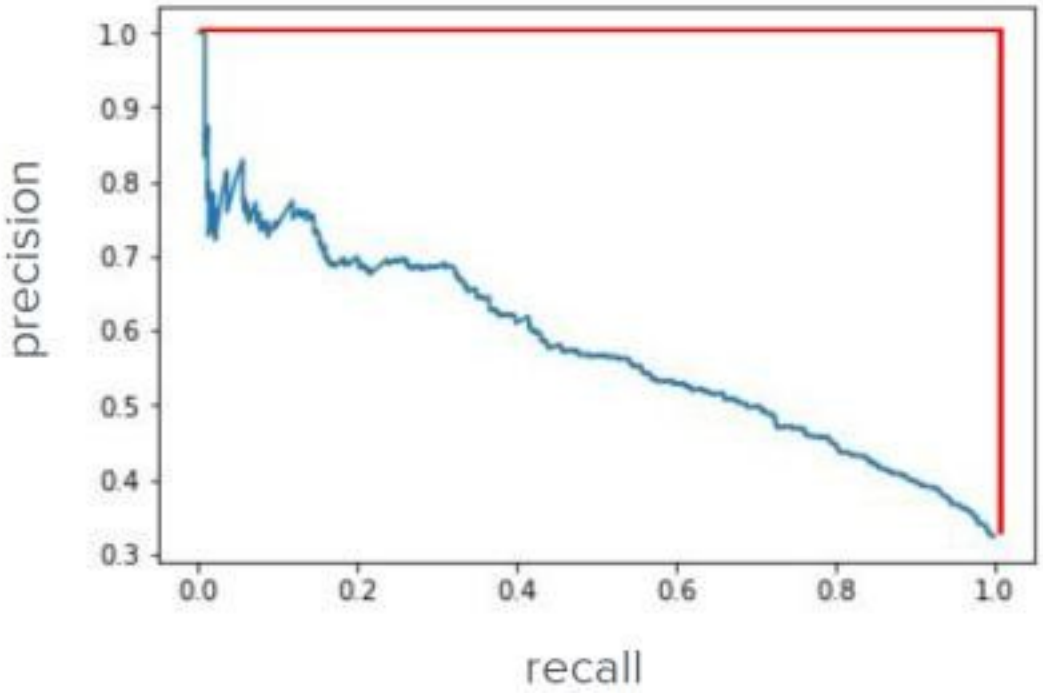
10 спам-писем

False negative
10

	Actual positive	Actual negative
Predicted positive	0	10
Predicted negative	0	100

Оценка качества модели

КРИВАЯ PRECISION-RECALL



Модель тем лучше, чем выше площадь под кривой.

AREA UNDER CURVE

Оценка качества модели

TRUE POSITIVE RATE

	Actual positive	Actual negative
Predicted positive	True positive	False positive
Predicted negative	False negative	True negative

True Positive Rate — доля правильно предсказанных среди категории 1

$$TPR = \frac{TP}{TP + FN}$$

Оценка качества модели

FALSE POSITIVE RATE

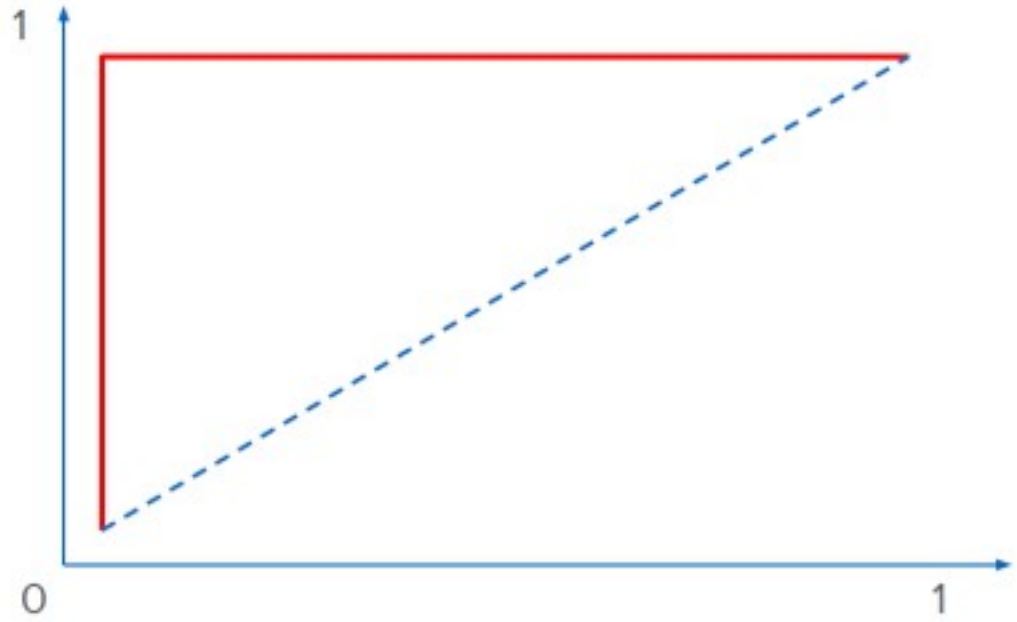
	Actual positive	Actual negative
Predicted positive	True positive	False positive
Predicted negative	False negative	True negative

False Positive Rate – доля неправильно предсказанных среди относящихся к категории 0.

$$FPR = \frac{FP}{FP + TN}$$

Оценка качества модели

ИДЕАЛЬНЫЙ СЛУЧАЙ



Модель предсказывает абсолютно верно

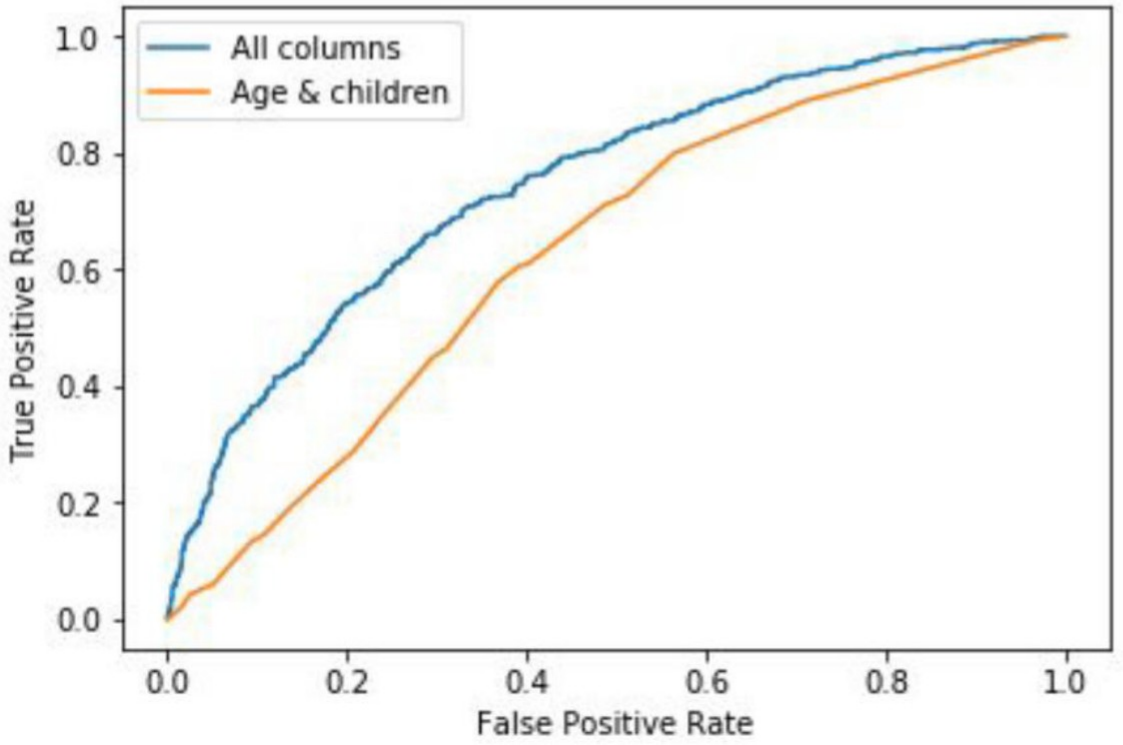
TPR = 1

FPR = 0

----- случайные предсказания

Оценка качества модели

СРАВНЕНИЕ ДВУХ МОДЕЛЕЙ



ПРАКТИКА

ATHLETES_CLASSIFIER.IPYNB

Дана статистика спортсменов ОИ 2016. Необходимо построить модель, предсказывающая пол спортсмена по имеющимся признакам (кроме столбца sex).

Построить графики Precision-Recall и FPR-TPR, посчитать AUC.

ПРАКТИКА

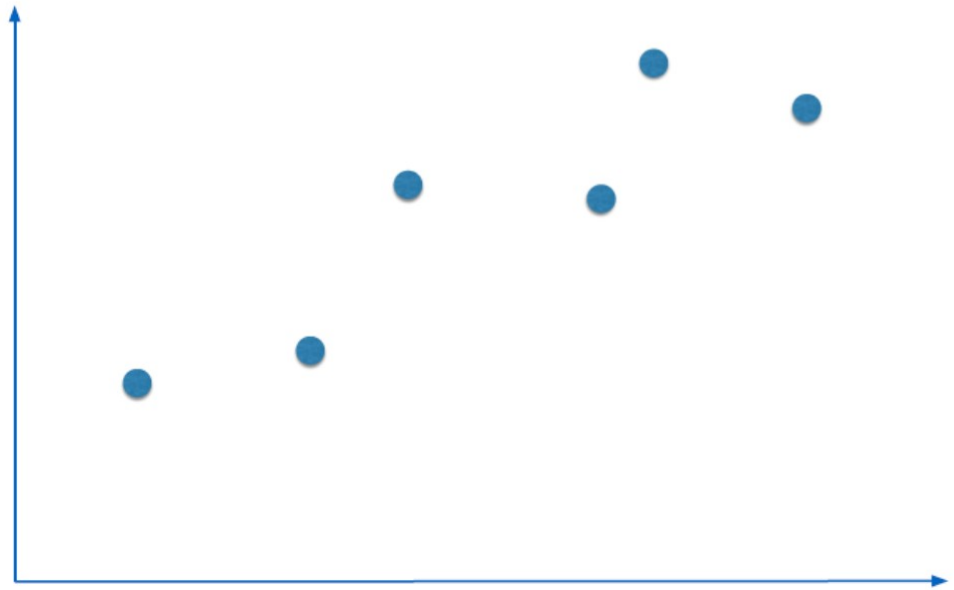
SHELTER.IPYNB

Оценка многоклассовой классификации

Борьба с переобучением

ПРИМЕР ПЕРЕОБУЧЕНИЯ

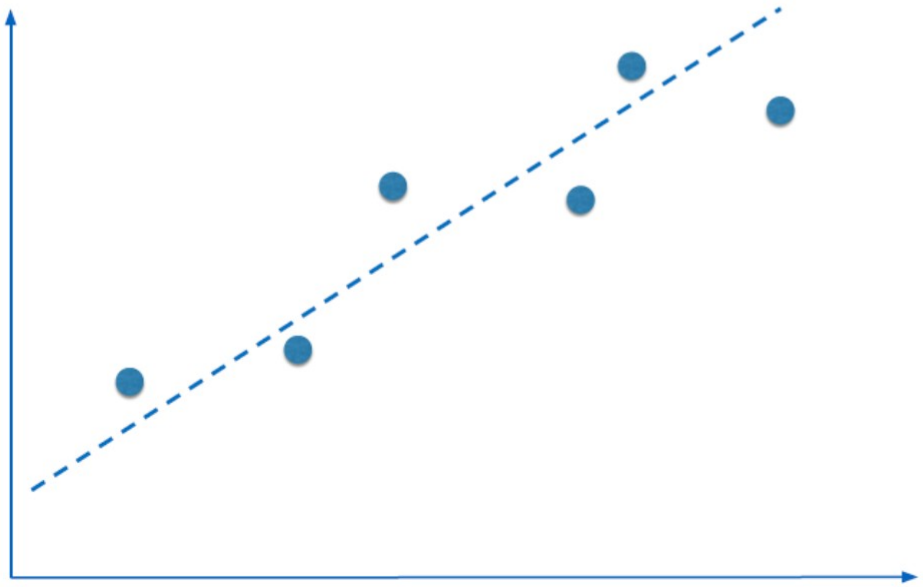
Имеются данные из 6 точек



Борьба с переобучением

ПРИМЕР ПЕРЕОБУЧЕНИЯ

Имеются данные из 6 точек

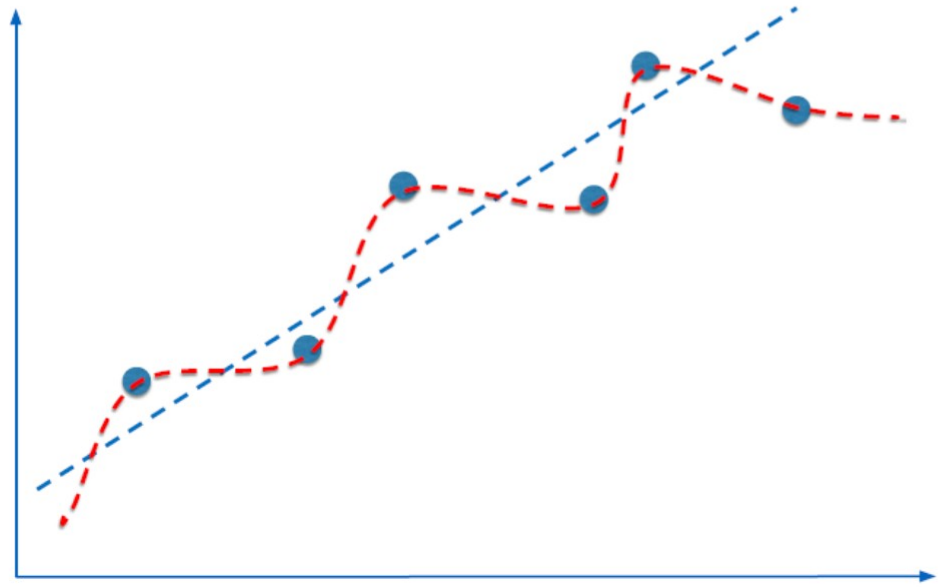


----- y = kx + b; есть ошибка > 0

Борьба с переобучением

ПРИМЕР ПЕРЕОБУЧЕНИЯ

Имеются данные из 6 точек



----- y = kx + b; есть ошибка > 0

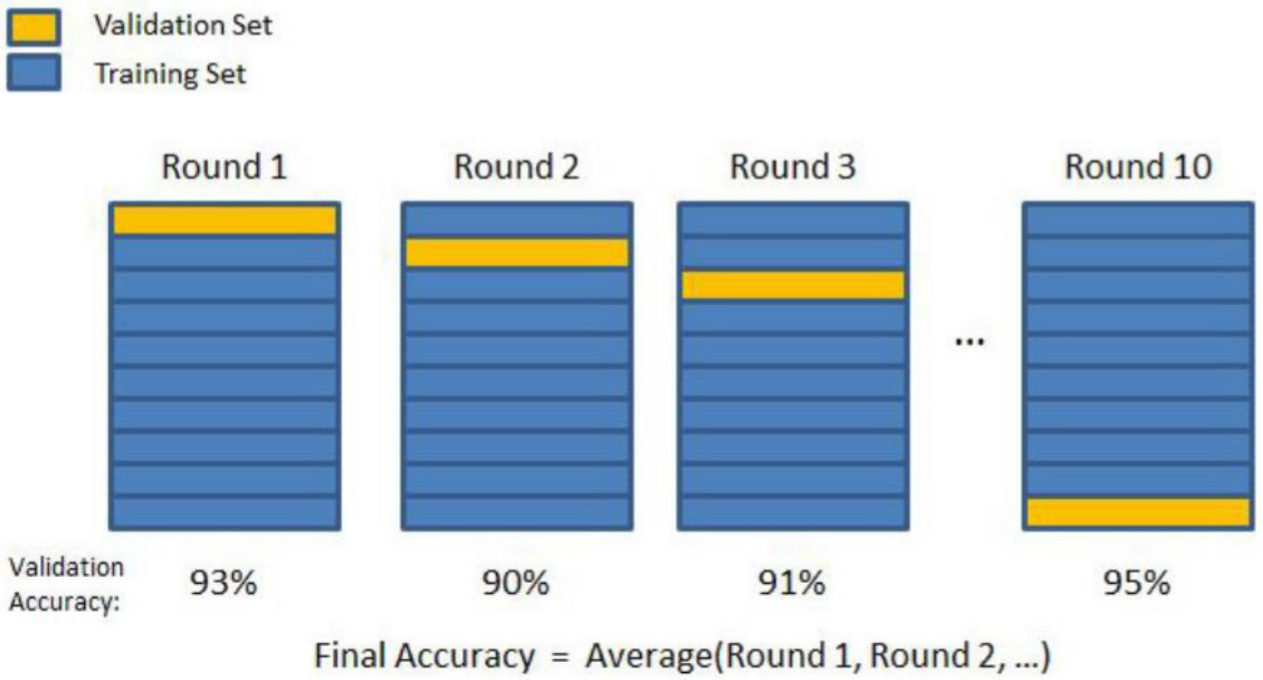
----- ошибка = 0. Круто?

$$y = a_0 + a_1x + a_2x^2 + a_3x^3 + a_4x^4 + a_5x^5$$

Борьба с переобучением

КРОСС-ВАЛИДАЦИЯ

k-fold cross validation



Лучше, чем случайная выборка

ПРАКТИКА

CROSS_VAL_SCORE.IPYNB

Дана статистика картинок цифр, каждая из которых описывается набором из 64 признаков.

Используя модель DecisionTreeClassifier, подберите значение параметра модели max_depth, при котором точность модели (accuracy) максимальна.

СМЕЩЕНИЕ И РАЗБРОС

Смещение и разброс

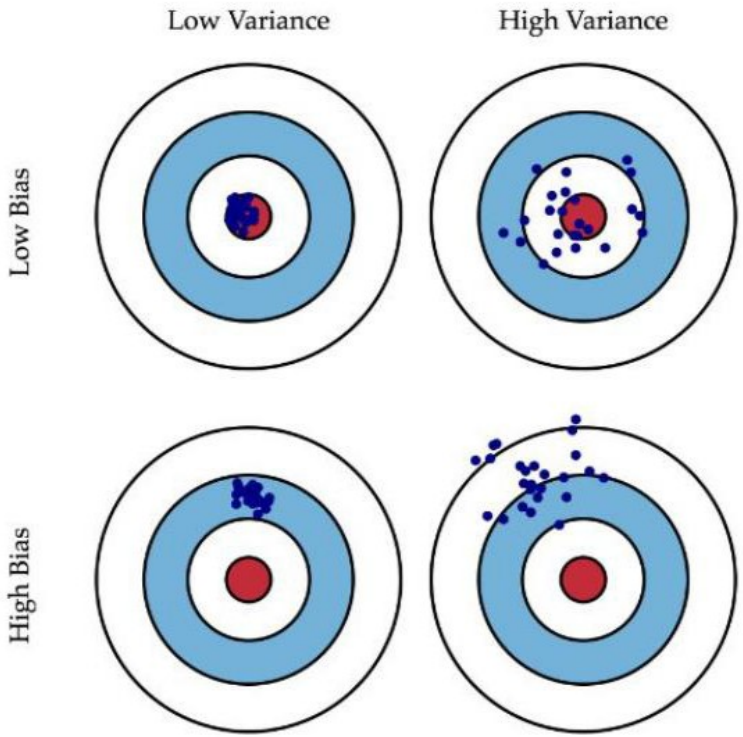
Ошибка прогноза

Можем разложить на слагаемые:

- Bias — средняя ошибка прогноза
- Variance — изменение ошибки при обучении на разных наборах данных
- Неустранимая ошибка

Смещение и разброс

ОШИБКА ПРОГНОЗА

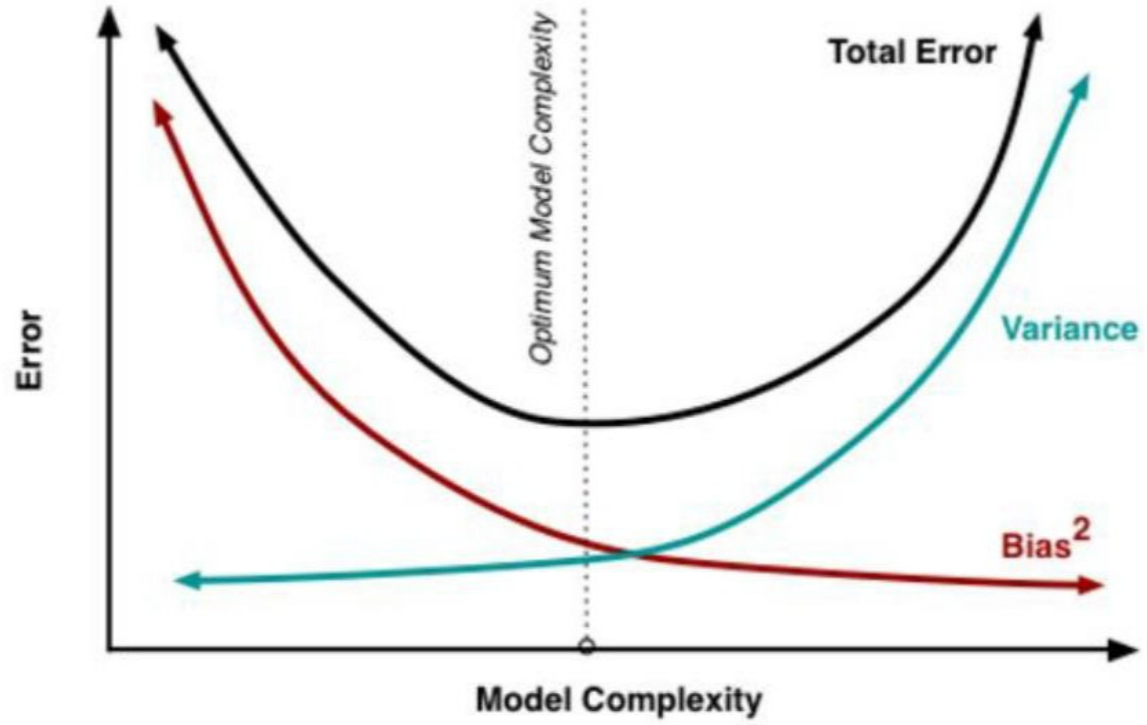


Сложная модель (учитывает много признаков) увеличивает разброс ошибки

Слишком простая модель (мало признаков) вызывает смещение в пользу одного признака.

Смещение и разброс

ОПТИМАЛЬНЫЙ ВАРИАНТ

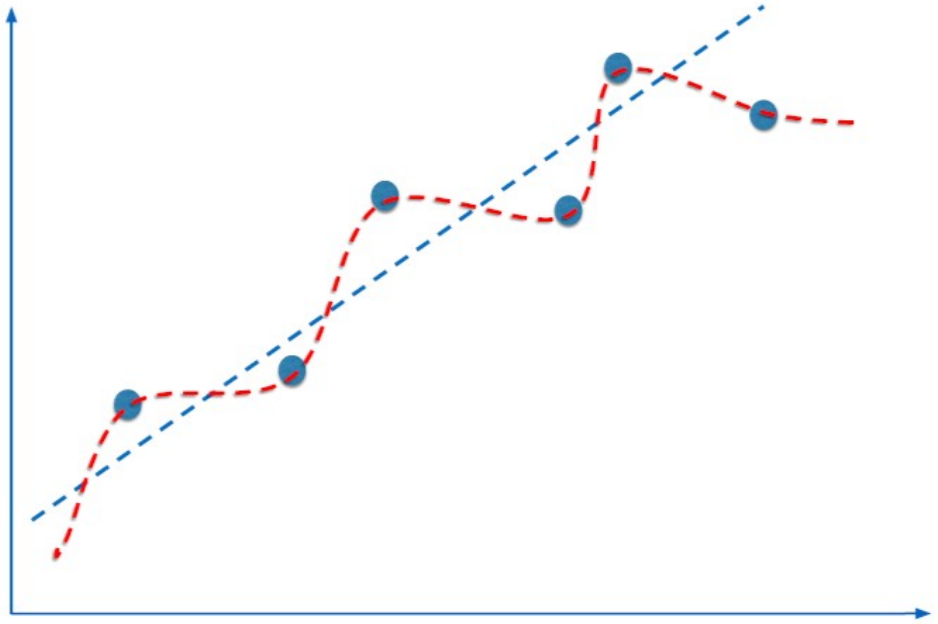


Можно ли повлиять на стабильность модели, то есть уменьшить Variance?

L1 И L2 РЕГУЛЯРИЗАЦИЯ

Регуляризация

ПРОШЛЫЙ ПРИМЕР ПЕРЕОБУЧЕНИЯ



Переберем модели, увеличивая степень функции

$$y = a_0 + a_1 x$$

$$y = a_0 + a_1 x + a_2 x^2$$

$$y = a_0 + a_1 x + a_2 x^2 + a_3 x^3$$

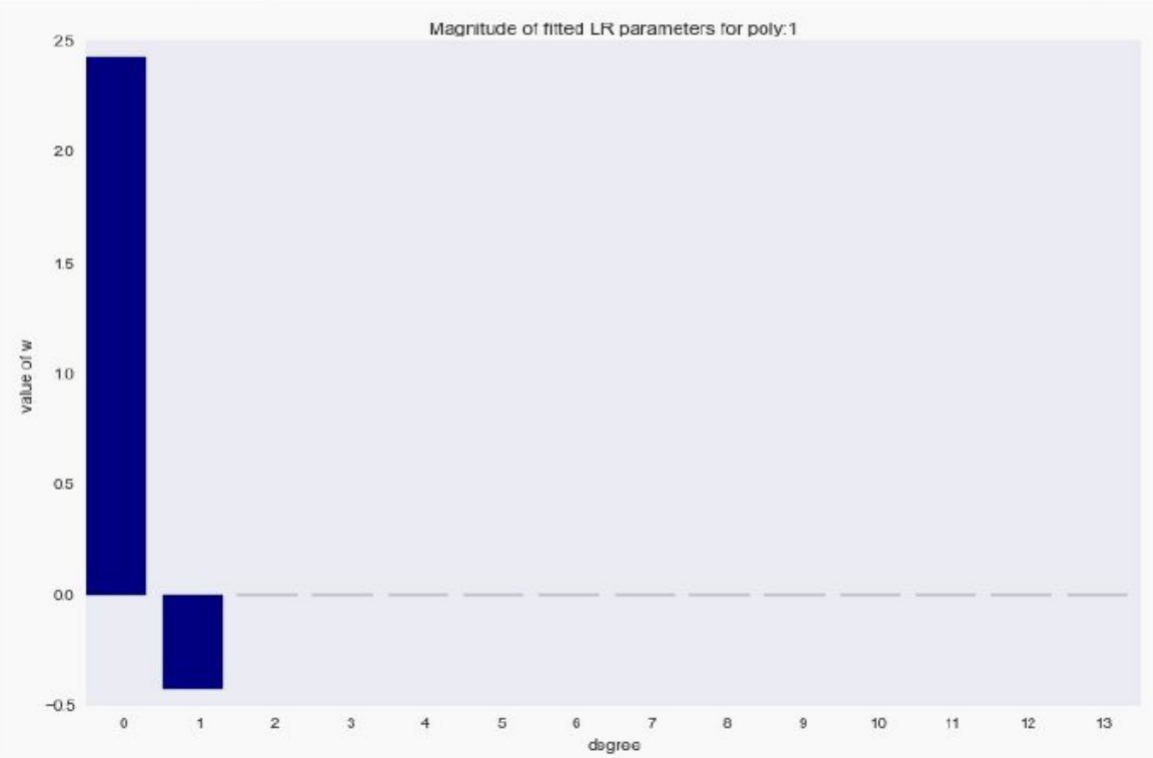
...

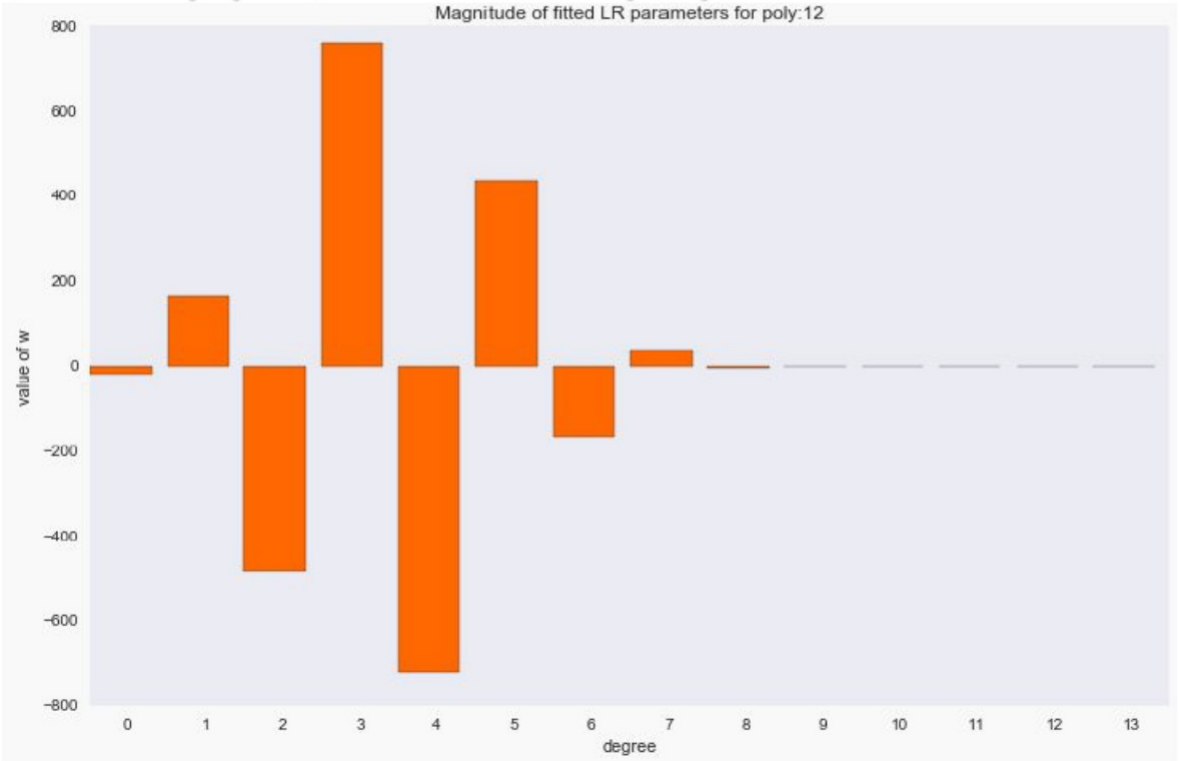
$$y = a_0 + a_1 x + a_2 x^2 + \cdots + a_5 x^5$$

Регуляризация

КАК БУДУТ ВАРЬИРОВАТЬСЯ α?

При увеличении степени полинома вариация коэффициентов быстро растет





Регуляризация

НАДО УМЕНЬШИТЬ РАЗБРОС КОЭФФИЦИЕНТОВ

Имеем модель целевой переменной *y* и коэффициентами *a*

$$\text{Целевая функция} = \sum_i (y_{\text{факт}} - Xa)^2$$

Регуляризация

ШТРАФ ЗА СЛОЖНОСТЬ

Основные варианты регуляризации

$$L_1 = \sum_i (y_{\text{факт}} - Xa)^2 + \lambda \sum_i |a_i|$$

$$L_2 = \sum_i (y_{\text{факт}} - Xa)^2 + \lambda \sum_i a_i{}^2$$

ПРАКТИКА

Дана статистика пользователей adult.csv.

Получите значения AUC для различных моделей и их параметров.

ЧТО МЫ СЕГОДНЯ УЗНАЛИ

1. Изучили метрики оценки качества моделей.

2. На практике потренировались в проведении кросс-валидации моделей.

3. Изучили признаки и способы борьбы с переобучением на примере L1 и L2 регуляризации.