Введение в искусственный интеллект. Машинное обучение

Лекция 1. Вводная

Бабин Д.Н., Иванов И.Е., Петюшко А.А.

МаТИС

15 февраля 2019г.

План лекции

1. Организационные вопросы
2. Постановка основных задач машинного обучения
3. Тестирование моделей, выбор лучшей
4. Декомпозиция ошибки, недообучение и переобучение

Авторы курса

Руководитель курса	Лектор	Лектор
д.ф.-м.н. Бабин Дмитрий Николаевич	к.ф.-м.н. Иванов Илья Евгеньевич	к.ф.-м.н. Петюшко Александр Александрович
		

- Авторы имеют более 10 лет опыта участия в проектах, связанных с машинным обучением и компьютерным зрением
- Являются постоянными участниками группы распознавания образов кафедры МаТИС
- В качестве научных консультантов работали с такими российскими и международными компаниями как Нейроком, LSI Research, Fotonation, Huawei и др.

Зачем посещать этот курс

1. Это возможность получить знания, которые пригодятся в работе

Зачем посещать этот курс

1. Это возможность получить знания, которые пригодятся в работе
2. Специалисты по анализу данных сейчас очень востребованы

Зачем посещать этот курс

1. Это возможность получить знания, которые пригодятся в работе
2. Специалисты по анализу данных сейчас очень востребованы
3. Это шанс максимально использовать своё образование

Зачем посещать этот курс

1. Это возможность получить знания, которые пригодятся в работе
2. Специалисты по анализу данных сейчас очень востребованы
3. Это шанс максимально использовать своё образование
4. Это просто очень интересно и затягивает

Что же такое искуссвенный интеллект?

Естественный интеллект (человек)

- Может мыслить, принимать решения, анализировать информацию

Что же такое искуссвенный интеллект?

Естественный интеллект (человек)

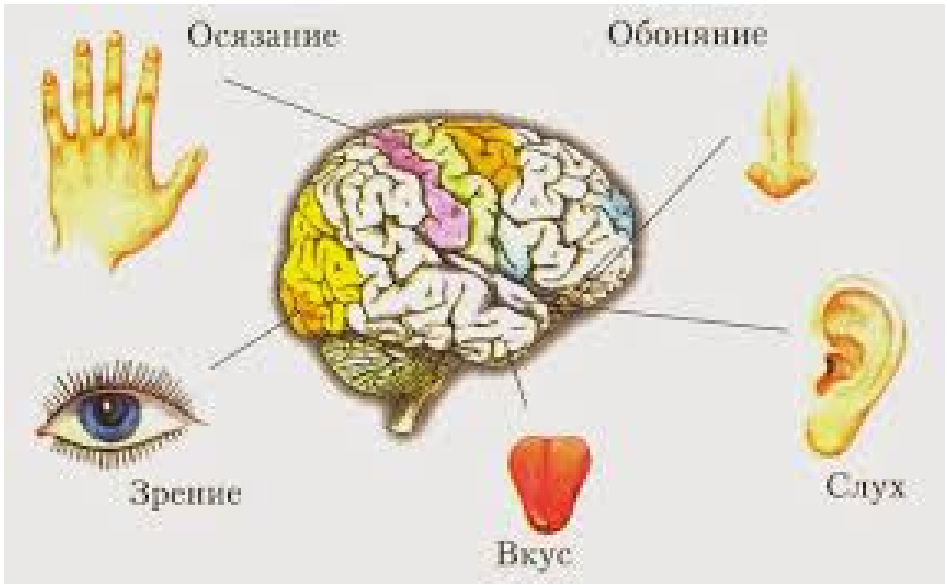
- Может мыслить, принимать решения, анализировать информацию

Искусственный интеллект

- (в широком смысле) то же самое, что и естественный, только с использованием компьютера вместо человека
- (в узком смысле) алгоритмы способные сами обучаться, чтобы выполнять задачи вместо человека

Взаимодействие со средой

- 83 % информации поступает через зрение
- 10 % информации поступает через слух



Вывод

Чтобы построить интеллектуальную систему, её необходимо научить взаимодействовать со средой

Общая структура курса «Введение в компьютерный интеллект»

1. **Машинное обучение**
   - Необходимые основы для всего курса

Общая структура курса «Введение в компьютерный интеллект»

1. **Машинное обучение**
   - Необходимые основы для всего курса
2. Компьютерное зрение
   - Извлечение информации из визуальных образов (изображений и видео)

Общая структура курса «Введение в компьютерный интеллект»

1. **Машинное обучение**
   - Необходимые основы для всего курса
2. Компьютерное зрение
   - Извлечение информации из визуальных образов (изображений и видео)
3. Обработка естественного языка
   - Извлечение информации из речи и текста

Общая структура курса «Введение в компьютерный интеллект»

1. **Машинное обучение**
   - Необходимые основы для всего курса
2. Компьютерное зрение
   - Извлечение информации из визуальных образов (изображений и видео)
3. Обработка естественного языка
   - Извлечение информации из речи и текста
4. Обучение с подкреплением
   - Интерактивное взаимодействие со средой

Машинное обучение в жизни

- Предсказание стоимости недвижимости
- Предсказание платёжеспособности клиента
- Предсказание оттока клиентов
- Классификация заболевания
- Предсказание клика пользователя по рекламному баннеру
- И многие другие задачи. . .

Что будет в этом курсе

Теоретическая часть

- Постановка задач машинного обучения. Тестирование моделей
  - Precision / Recall, TPR / FPR, ROC, AUC, Cross-Validation, Bootstrap, ...
- Методы классификации
  - SVM, Random Forest, Decision Tree, Stochastic Gradient Descent, ...
- Методы восстановления регрессии
  - Linear Regression / Least Squares, Ridge Regression, LASSO, ...
- Композиции алгоритмов
  - Bagging, Boosting, AdaBoost, AnyBoost, ...

Практическая часть

- Обработка и анализ данных на python
  - SciPy, Scikit-Learn, XGBoost, Pandas, ...
- Соревнования по машинному обучению

Чего не будет в этом курсе

- Глубокое обучение

Чего не будет в этом курсе

- Глубокое обучение
- Обучение без учителя

Чего не будет в этом курсе

- Глубокое обучение
- Обучение без учителя
- Частичное обучение

Чего не будет в этом курсе

- Глубокое обучение
- Обучение без учителя
- Частичное обучение
- Методы ранжирования

Чего не будет в этом курсе

- Глубокое обучение
- Обучение без учителя
- Частичное обучение
- Методы ранжирования
- Прогнозирование временных рядов

Чего не будет в этом курсе

- Глубокое обучение
- Обучение без учителя
- Частичное обучение
- Методы ранжирования
- Прогнозирование временных рядов
- Рекомендательные системы

Чего не будет в этом курсе

- Глубокое обучение
- Обучение без учителя
- Частичное обучение
- Методы ранжирования
- Прогнозирование временных рядов
- Рекомендательные системы
- Цифровая обработка сигналов

Оценки за курс

- Оценки за курс будут выставляться в соответствии с набранными баллами за выполнение домашних заданий.
- По курсу будут предложены домашние задания трёх видов:
  - теоретические
  - практические
  - соревнования
- В конце семестра состоится экзамен, на котором при желании можно будет повысить свою оценку
- Предварительная шкала оценок:

Оценка	Процент выполненых заданий
Отлично	80 %
Хорошо	60 %
Зачет	40 %

Полезные ресурсы

- Страница курса:
  `https://github.com/mlcoursemm/mlcoursemm2019spring`
- Телеграмм-канал:
  `https://t.me/joinchat/AAAAAEUmx5cJLOdLXsOt8g`
- Группа обсуждения:
  `https://t.me/joinchat/B2iAGkx8IrWpQdM3LL-H-w`
- Почта курса: `mlcoursemm@gmail.com`

Способы машинного обучения

Определения

- X — множество объектов
- Y — множество ответов
- $y : X \to Y$ — неизвестная зависимость

Основные способы машинного обучения

- **С учителем**
  - Достаточное количество обучающего материала, то есть пар (x_i, y_i)
- Частичное обучение
  - Малое количество размеченных данных и много неразмеченных примеров x_i
- Без учителя
  - Нет размеченных пар, только примеры x_i
- С подкреплением
  - Формирование отклика на основе взаимодействия со средой

Постановка задачи обучения с учителем

- Дано:
  - $\{(x_1, y_1), ..., (x_n, y_n)\} \subset X \times Y$ — обучающая выборка
- Найти
  - Решающую функцию $a : X \to Y$, которая приближает целевую зависимость y.
- Необходимо детализировать:
  - Как определяются объекты
  - Как задаются ответы
  - Что значит, что одна зависимость приближает другую

Как определяются объекты

Определение

Объект = совокупность признаков

Типы признаков

- Бинарный признак
- Категориальный признак
- Порядковый признак
- Количественный признак

Как задаются ответы

Задачи классификации

- Бинарная классификация $Y = \{-1, 1\}$ или $Y = \{0, 1\}$
- Многоклассовая классификация $Y = \{0, 1, ..., M - 1\}$
- Многозначная бинарная классификация $Y = \{0, 1\}^M$

Задачи восстановления регрессии

$Y = \mathbb{R}$ или $Y = \mathbb{R}^n$

Функция потерь

Опредедение

Функция потерь (loss function) $\mathcal{L}(a, x)$ — величина ошибка алгоритма a на объекте x

Функции потерь для задачи классификации

$\mathcal{L}(a, x) = [a(x) \neq y]$ — индикатор ошибки

Функции потерь для задач регрессии

$\mathcal{L}(a, x) = (a(x) - y)^2$ — квадратичная ошибка

Сравнение моделей машинного обучения

Как понять, что одна модель лучше другой?

Для этого используют независимое от **обучающего** множества множество, которое называется **тестовым**

Зачем вообще это понимать?

- Существует множество алгоритмов машинного обучение и важно понимать, какой из них более применим в конкретной задаче
- Даже в рамках одной модели есть много параметров

Как выбирать лучшую модель

Наивный подход

Обучить модели с различными параметрами и посмотреть, что будет на тесте

Минусы наивного подхода

- Так как тест состоит из случайной выборки теста, то результат на теста тоже является некоторым приближение случайной величины
- Если все модели тестировать на тестовом датасете и выбирать лучшую таким образом, то будет происходить неявное обучение на тесте и на другом независимом тесте возможны сюрпризы

Что же делать?

Чтобы неявно не обучиться на тестовых данных – надо использовать кросс-валидацию

Кросс-валидация

Общая идея

Основная идея кросс-валидации состоит в разбиении обучающего множества на два непересекающихся множества (возможно многократном):

$$X^{learn} = X^{train} \sqcup X^{val}$$

На одном из них происходит обучение, а на другом происходит валидация модели.

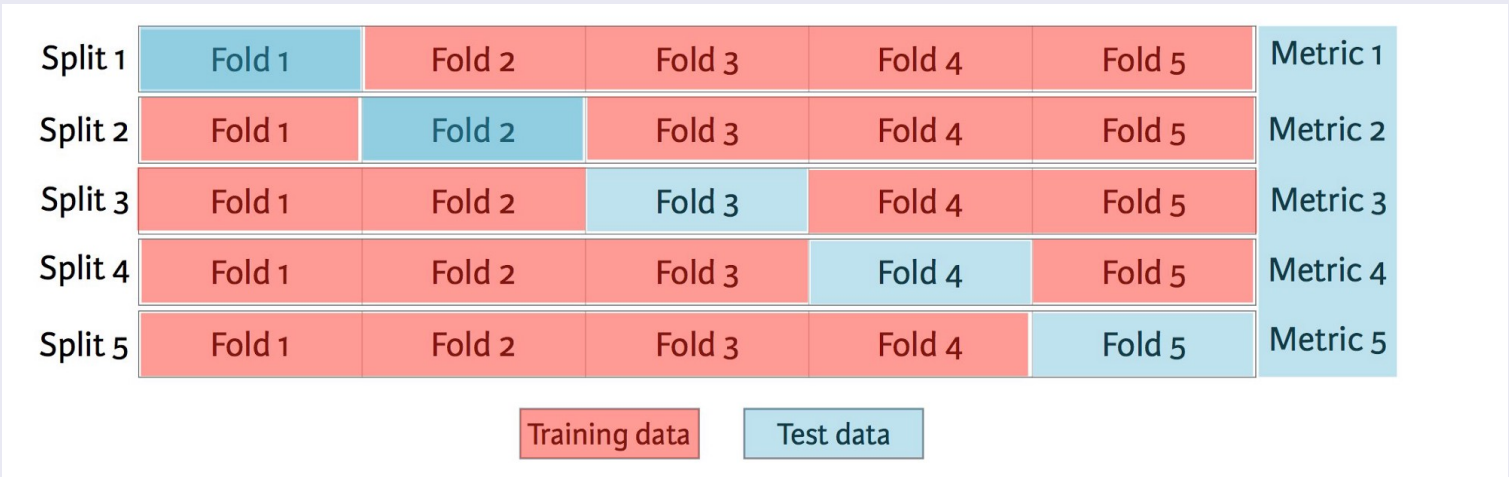
Частные случаи

1. Простейшая кросс-валидация (Holdout) — однократное разделение множества

Кросс-валидация

Частные случаи

2. k-fold валидация[a]:



3. Leave one out (LOO) валидация — частный случай k-fold валидации если k равно мощности обучающего множества
4. Многократная k-fold валидация — повторение k-fold валидации несколько раз с разными разбиениями.

[a] https://towardsdatascience.com

Переобучение

Определение

Переобучение (overfitting) — нежелательное явление, возникающее при решение задач обучения по прецедентам, когда вероятность ошибки обученного алгоритма на объектах тестовой выборки оказывается существенно выше, чем средняя ошибка на обучающей выборке. Переобучение возникает при использовании избыточно сложной модели

Причины возникновения

Одной из причин переобучения избыточная сложность пространства параметров модели, "лишние"степени свободы используются для точной подгонки на обучающую выборку

Методы обнаружения

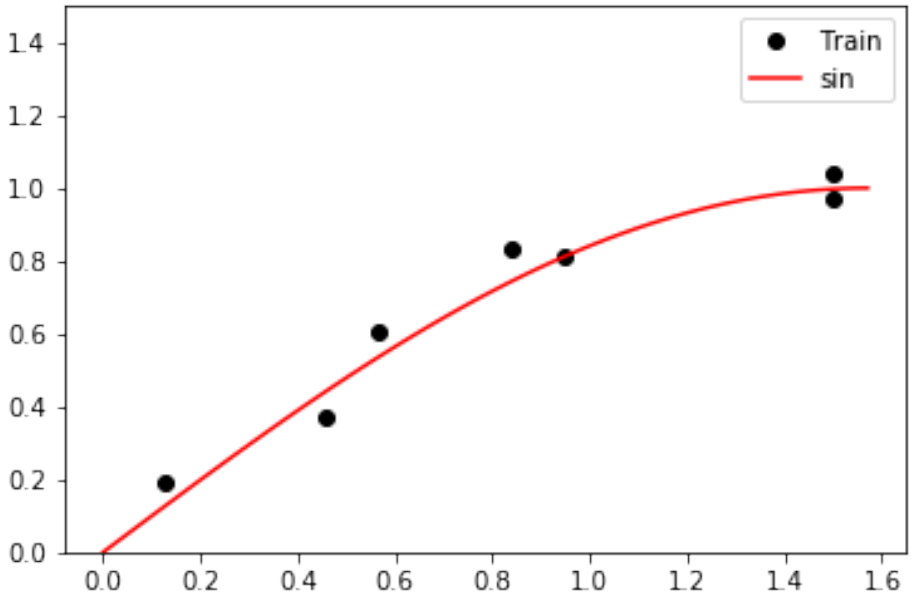
Основным методом обнаружения является использование кросс-валидации

Недообучение

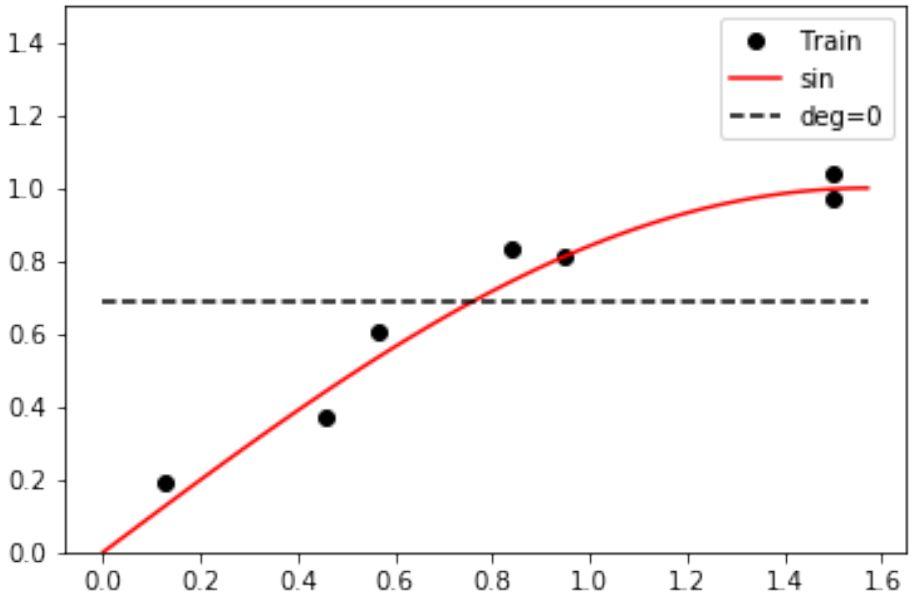
Определение

Недообучение (underfitting) – нежелательное явление, возникающее при решении задач обучения по прецендентам, когда алгоритм обучения не обеспечивает достаточно малой величины средней ошибки на обучающей выборке. Недообучение возникает при использовании недостаточно сложных моделей

Примеры недообучения и переобучения

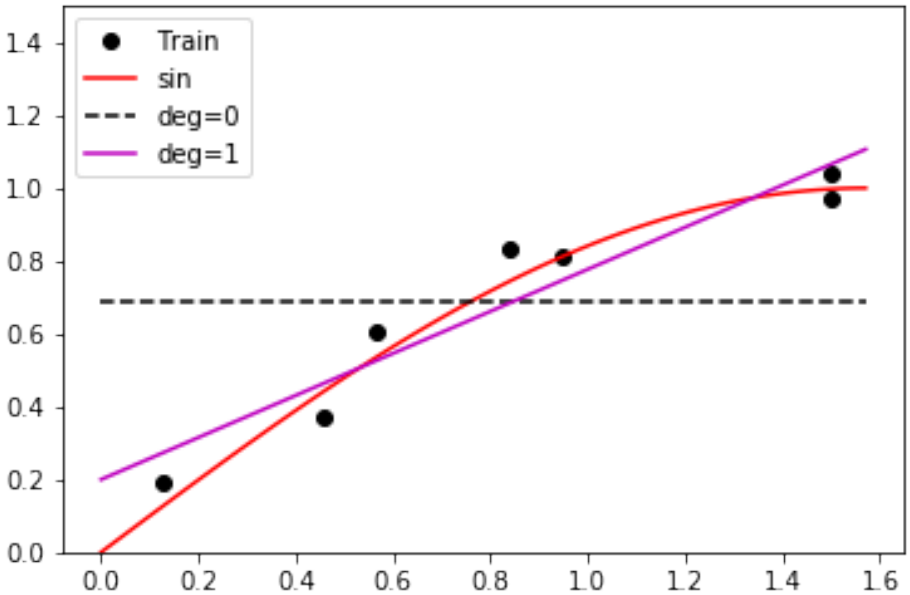


Примеры недообучения и переобучения



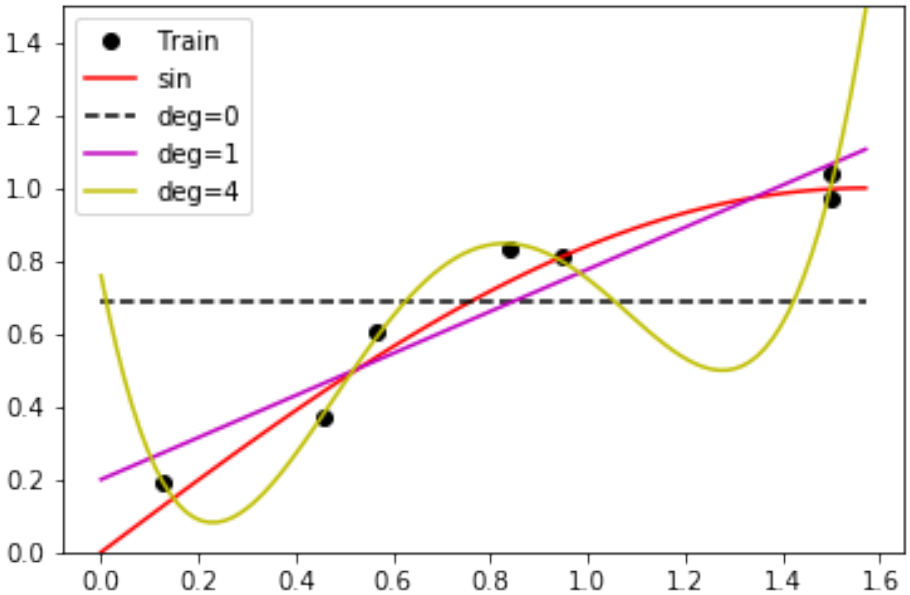
- Полином нулевой степени не может хорошо приближать зависимость в силу ограниченности параметров модели

Примеры недообучения и переобучения



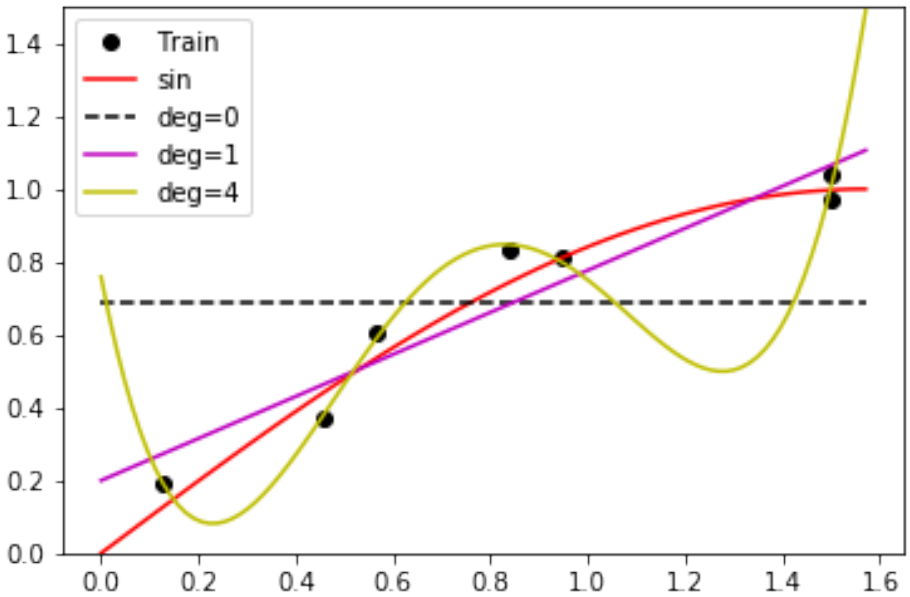
- Полином нулевой степени не может хорошо приближать зависимость в силу ограниченности параметров модели
- Линейная и квадратичная модели адекватно описывают закономерность

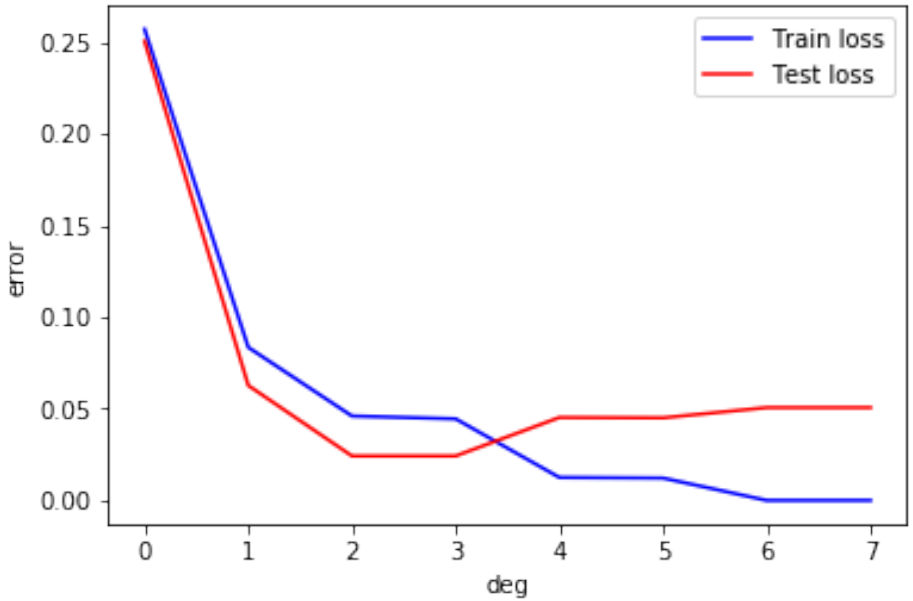
Примеры недообучения и переобучения



- Полином нулевой степени не может хорошо приближать зависимость в силу ограниченности параметров модели
- Линейная и квадратичная модели адекватно описывают закономерность
- Полиномы высоких степеней могут в точности пройти через точки обучающей выборки

Примеры недообучения и переобучения





- Полином нулевой степени не может хорошо приближать зависимость в силу ограниченности параметров модели
- Линейная и квадратичная модели адекватно описывают закономерность
- Полиномы высоких степеней могут в точности пройти через точки обучающей выборки

Немного математики

Определения

Пусть $y = y(x) = f(x) + \varepsilon,\ \varepsilon \sim N(0, \sigma^2)$ — целевая зависимость, и $a(x)$ — алгоритм машинного обучения.

Разложение квадрата ошибки

$$E(y-a)^2 = E(y^2 + a^2 - 2ya) = Ey^2 + Ea^2 - 2Eya =$$
$$= Ey^2 + Ea^2 - 2E(f+\varepsilon))a = Ey^2 + Ea^2 - 2Efa =$$
$$= Ey^2 - (Ey)^2 + (Ey)^2 + Ea^2 - (Ea)^2 + (Ea)^2 - 2fEa =$$
$$= Dy + Da + (Ey)^2 + (Ea)^2 - 2fEa =$$
$$= Dy + Da + (E(f-a))^2 = \sigma^2 + variance(a) + bias^2(f,a)$$

Дополнительные определения

Определение

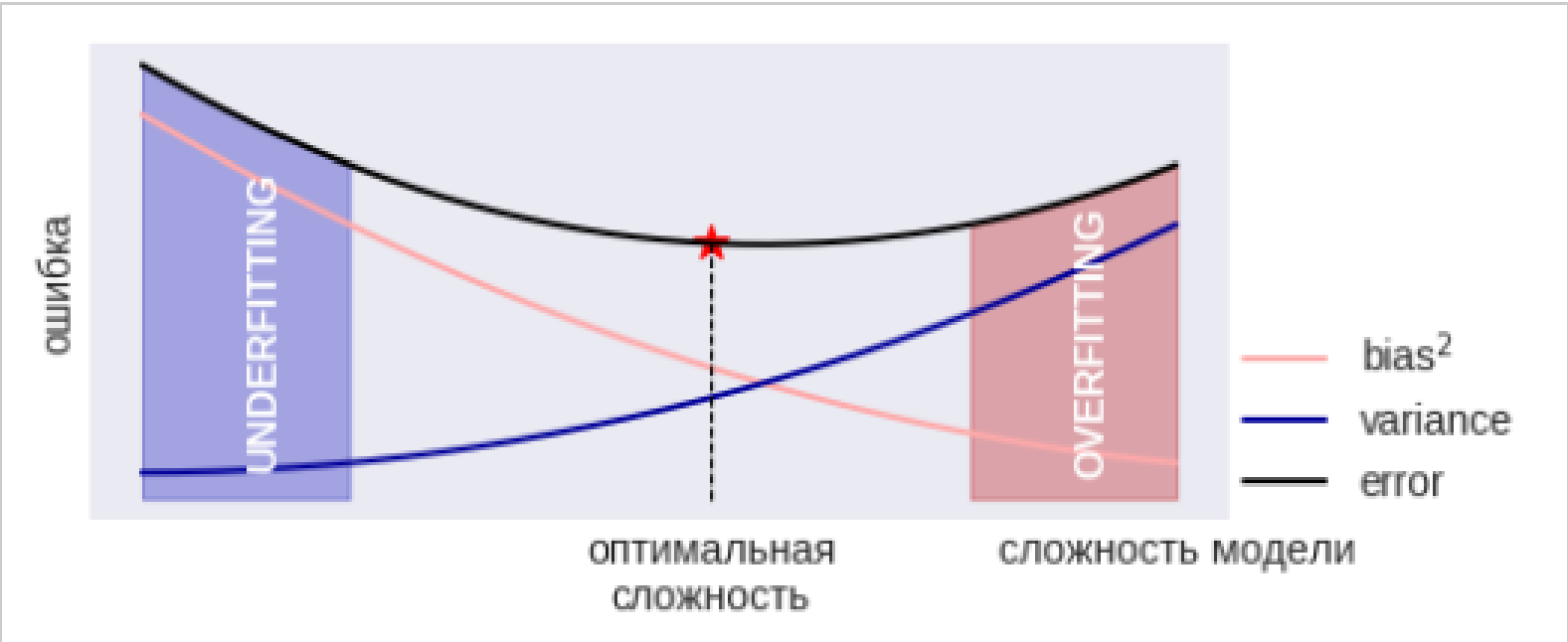
Разброс (variance) – дисперсия ответов алгоритмов a(x). Характеризует разнообразие алгоритмов (из-за случайности обучающей выборки, шума, стохастичности обучения и т.д.)

Определение

Смещение (bias) – матожидание разности между истинным ответом и выбанным алгоритмом. Характеризует способность модели настраивться на целевую зависимость

Модель оптимальной сложности

- Для простых моделей характерно недообучение
- Для сложных моделей характерно переобучение
- Оптимальная сложность модели где-то между[1]



[1] https://dyakonov.org

Спасибо за внимание!