Машинное обучение: оценка качества

Н. Поваров, И. Куралёнок

СПб,
2018

Задача

Хотим понять хорошо ли будет работать решающая функция на практике.

“If you can’t measure it, you can’t improve it”

Lord Kelvin

“Гораздо легче что-то измерить, чем понять, что именно вы измеряете.”

Джон Уильям Салливан

График целевой метрики

График целевой метрики

Вспомним о чём ML

- $$\underset{\mathrm{F,B}:\,\mathrm{F}_0 = \mathrm{B}(\mathrm{F})}{\operatorname{argmax}} \; \mathrm{A}(\Gamma, \mathrm{F}_0)$$

A — цели эксплуатации (например деньги) на всей области работы

B^* — способ оптимизации, который реализуем

* — здесь прячутся D и T.

Грубая классфикация способов оценки

- Black Box методы
  - Online
  - Offline
- Glass Box методы
  - VC-оценки
  - PAC-Bayes bounds
  - Оценки по Воронцову

Грубая классфикация способов оценки

- Black Box методы
  - Online
  - Offline
- Glass Box методы
  - VC-оценки
  - PAC-Bayes bounds
  - Оценки по Воронцову

Online

- Наблюдение
- Эксперимент

Online

- Наблюдение
- Эксперимент

Online

- ~~Наблюдение~~
- Эксперимент

Online

+ В условиях эксплуатации
+ В положительные результаты эксперимента обычно верят
+ Легко хвастаться результатом

- Вряд ли цель эксплуатации
- Можно навредить пользователям

Offline

+ Нельзя навредить пользователям
+ Обычно можно проводить сильно больше экспериментов

- Обычно нужны данные (примеры)
- Сложно «хвастаться» результатом

Оценка качества как система принятия решений

		Nature	
		P	N
СПР	P	TP	FP
	N	FN	TN

Offline

Offline на данных

- Hold Out
- Cross-fold Validation
- Bootstrap

Hold-out

Можно поделить множество на две части и обучить на одной, а оценить на другой:

$$DS = L \cup T, \qquad L \cap T = \emptyset$$

+ Расскажет о качестве предсказания
+ Можно посмотреть на качество на L и на T

- Использует меньше данных в обучении
- Если исходное множество непоказательно, то всё плохо

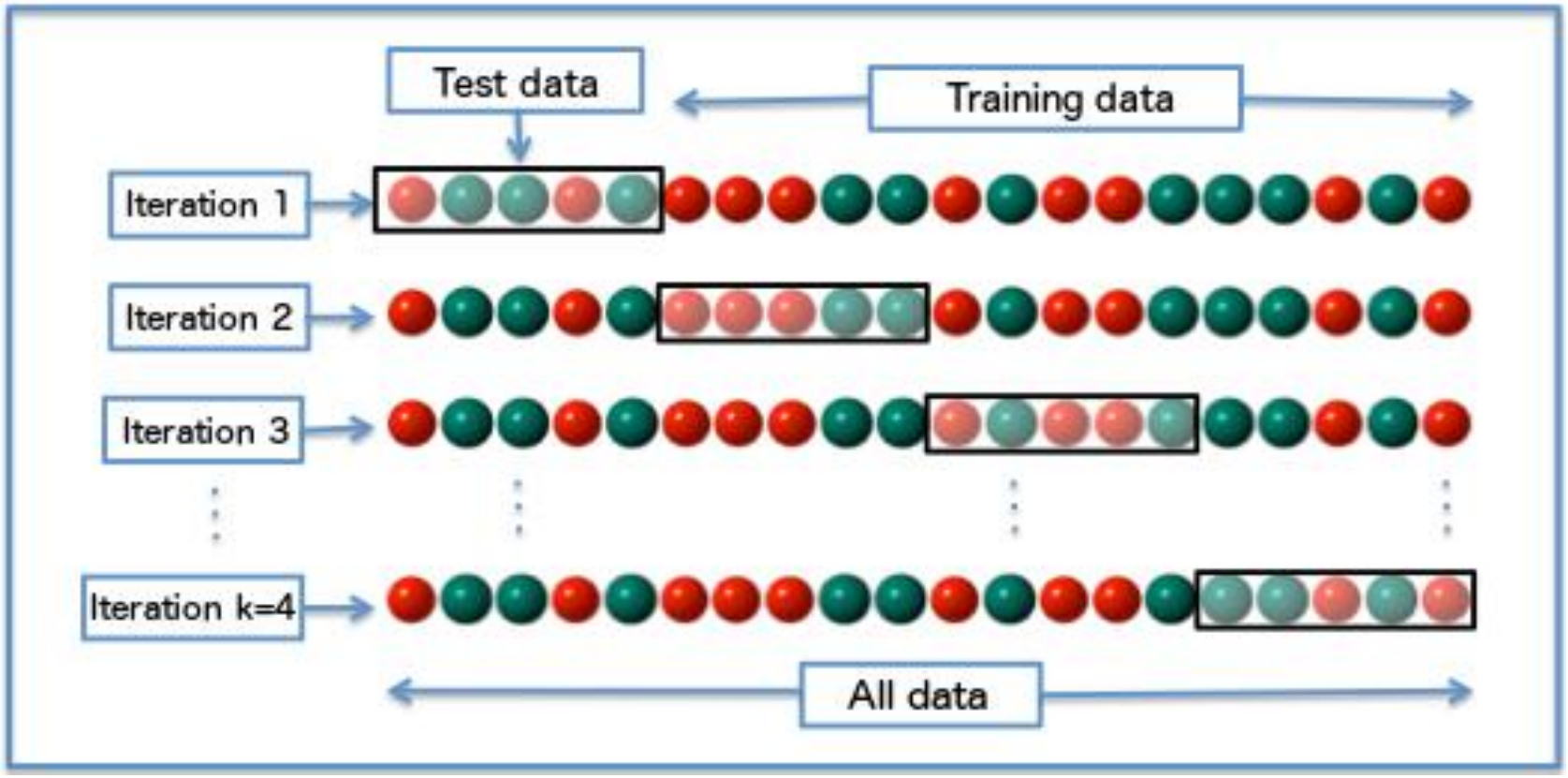
Cross-fold Validation

- Можно поделить множество на k частей и обучить на k-1, а оценить на оставшейся:

$$DS = L \cup T, \qquad L \cap T = \emptyset$$

Проделать так k раз.

Cross-fold Validation



Cross-fold Validation

Можно поделить множество на k частей и обучить на k-1, а оценить на оставшейся:

$$DS = L \cup T, \qquad L \cap T = \emptyset$$

Проделать так k раз.

+ Расскажет о качестве предсказания
+ Можно посмотреть на качество на L и на T

- Использует меньше данных в обучении
- Если исходное множество непоказательно, то всё плохо
- Надо обучающим показать всё множество
- Живёт такая система не очень долго

Повторные выборки

Можно посэмплить.

- $|\mathrm{L}| = |\mathrm{T}| = |\mathrm{D}|$
- Random sampling with replacement

+ Объём в обучении максимальный
+ T можно делать много раз

- Между L и T возникли зависимости

Где в offline система принятия решений

		Верная гипотеза	
		H_0	H_1
Результат применения критерия	H_0	H_0 верно принята	H_0 неверно принята (Ошибка *второго* рода)
	H_1	H_0 неверно отвергнута (Ошибка *первого* рода)	H_0 верно отвергнута

Где в offline система принятия решений

- В случае hold-out — «много» наблюдений в T.
- В случае cross-fold — большое k.
- В случае повторных выборок — можно сделать много раз.

		Верная гипотеза	
		H_0	H_1
Результат применения критерия	H_0	H_0 верно принята	H_0 неверно принята (Ошибка *второго* рода)
	H_1	H_0 неверно отвергнута (Ошибка *первого* рода)	H_0 верно отвергнута

Как выбрать статтест?

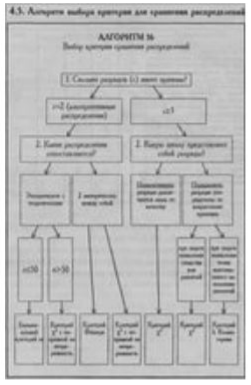
Яндекс

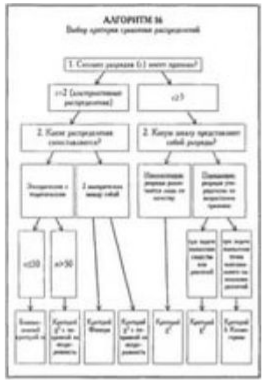
схема выбора статистического критерия

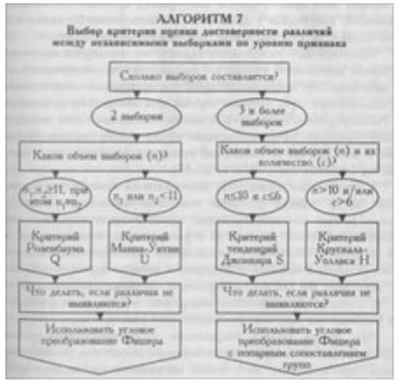
Найти

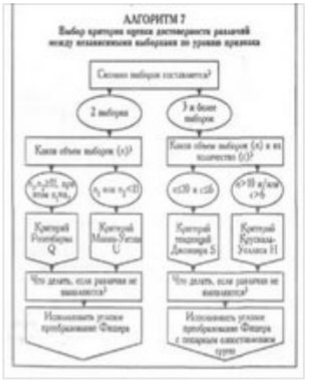
ПОИСК | КАРТИНКИ | ВИДЕО | КАРТЫ | МАРКЕТ | НОВОСТИ | ПЕРЕВОДЧИК | ЕЩЁ

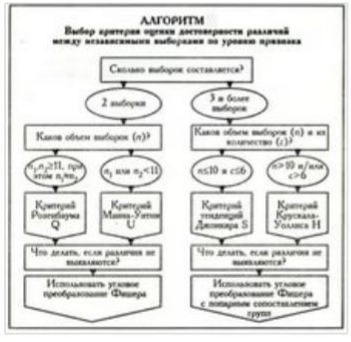
РАЗМЕР | ОРИЕНТАЦИЯ | ТИП | ЦВЕТ | ФАЙЛ | ТОВАРЫ | СВЕЖИЕ | ОБОИ 2560×1600 | НА САЙТЕ

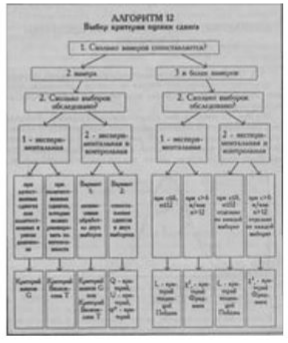


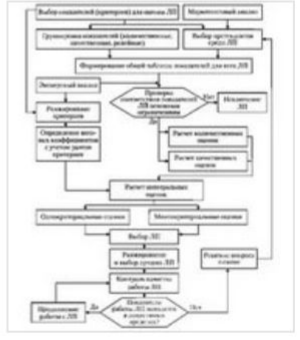


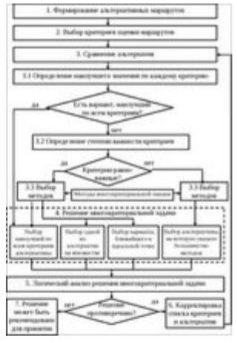


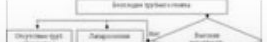


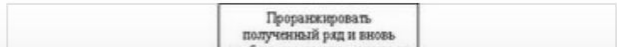


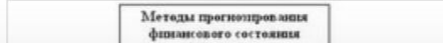


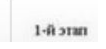












Классический подход что надо помнить

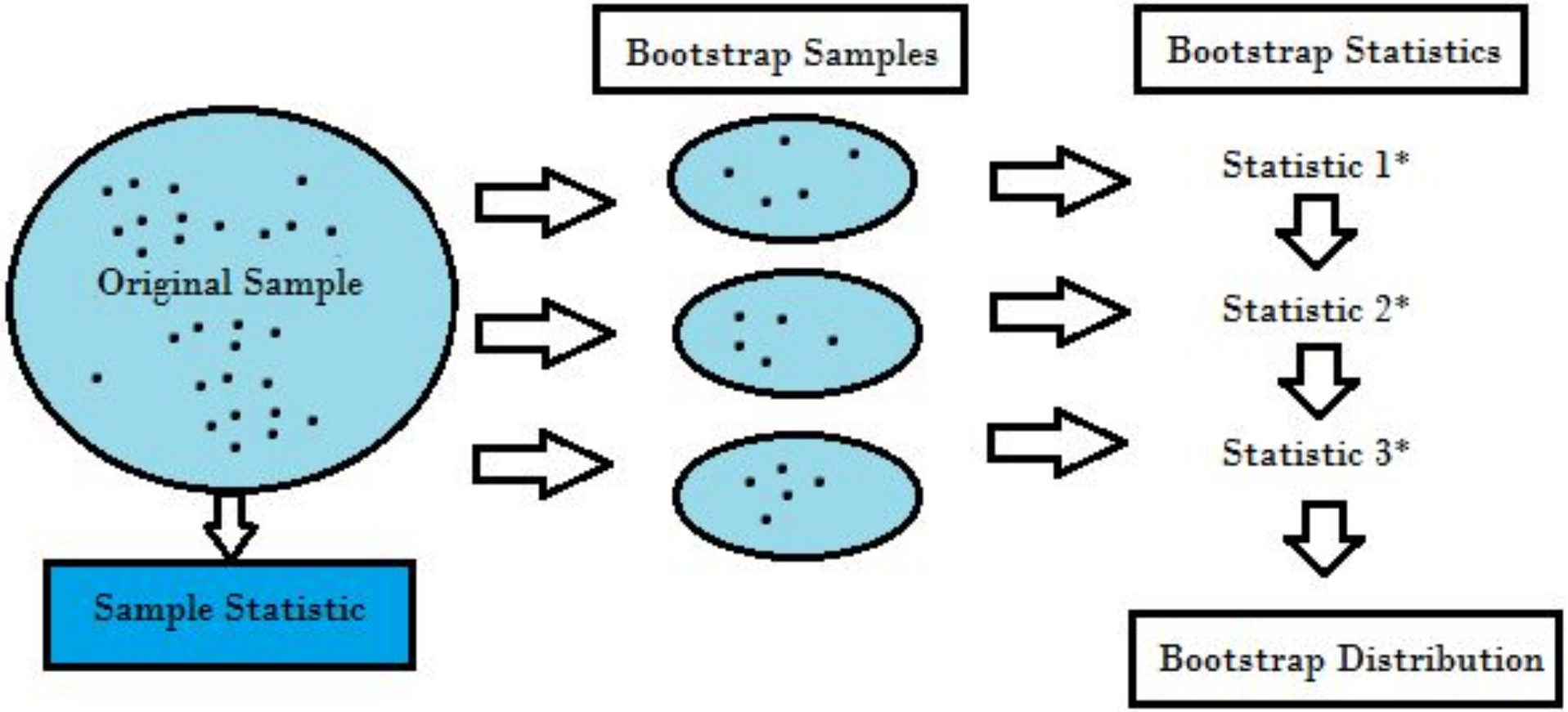
- Направленность
- Шкала измерений (отношений, порядка, номинальная, бинарная)
- Связанные/независимые наблюдения
- Параметрические/непараметрические
- ...

Есть метод проще!

Правильный подход на примере hold-out

- A bootstrap sample is a random sample taken with replacement from the original sample, of the same size as the original sample.
- A bootstrap statistic is the statistic computed on bootstrap sample.
- A bootstrap distribution is the distribution of many bootstrap statistic.

Правильный подход на примере hold-out



Common knowledge

Model Selection

Обобщающая способность алгоритма

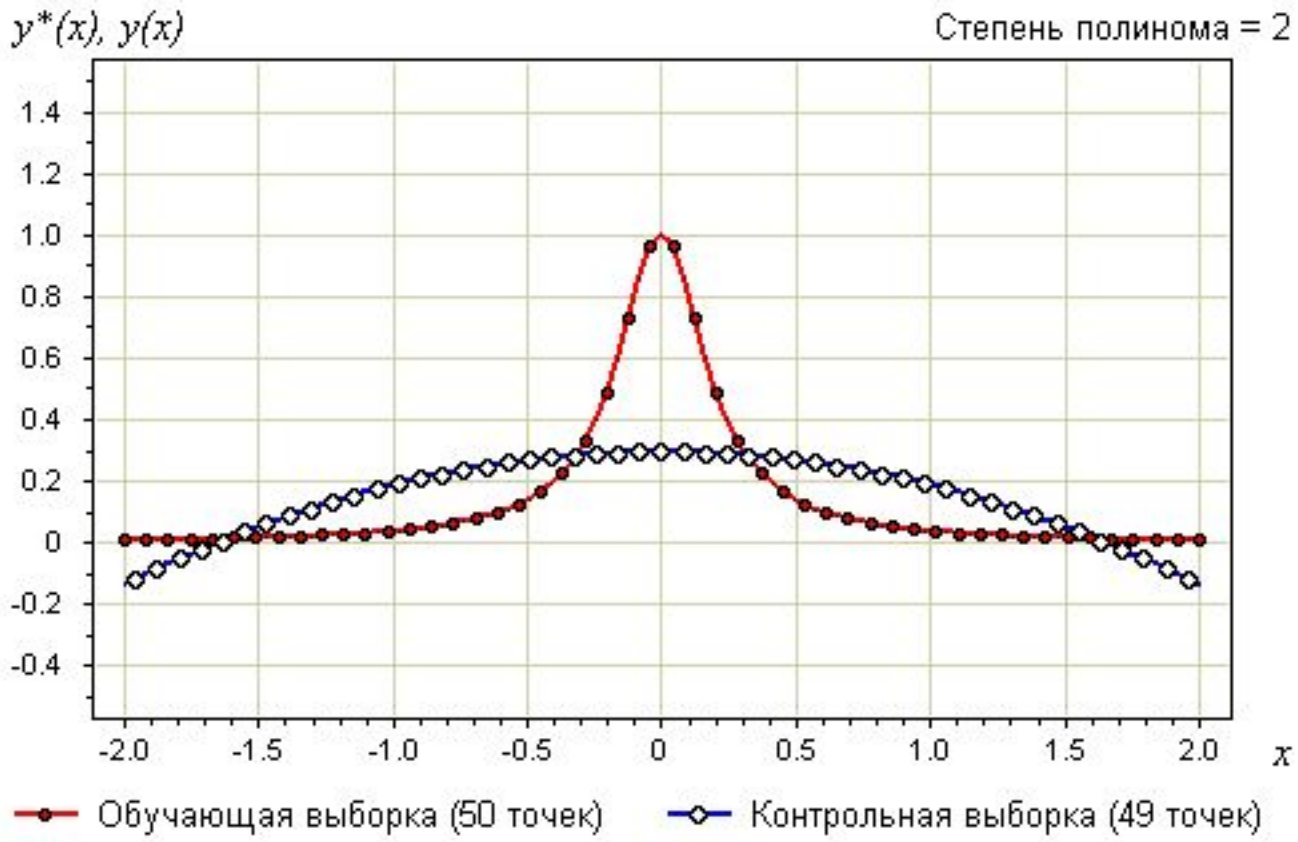
Flexibility

Bias-Variance trade-off

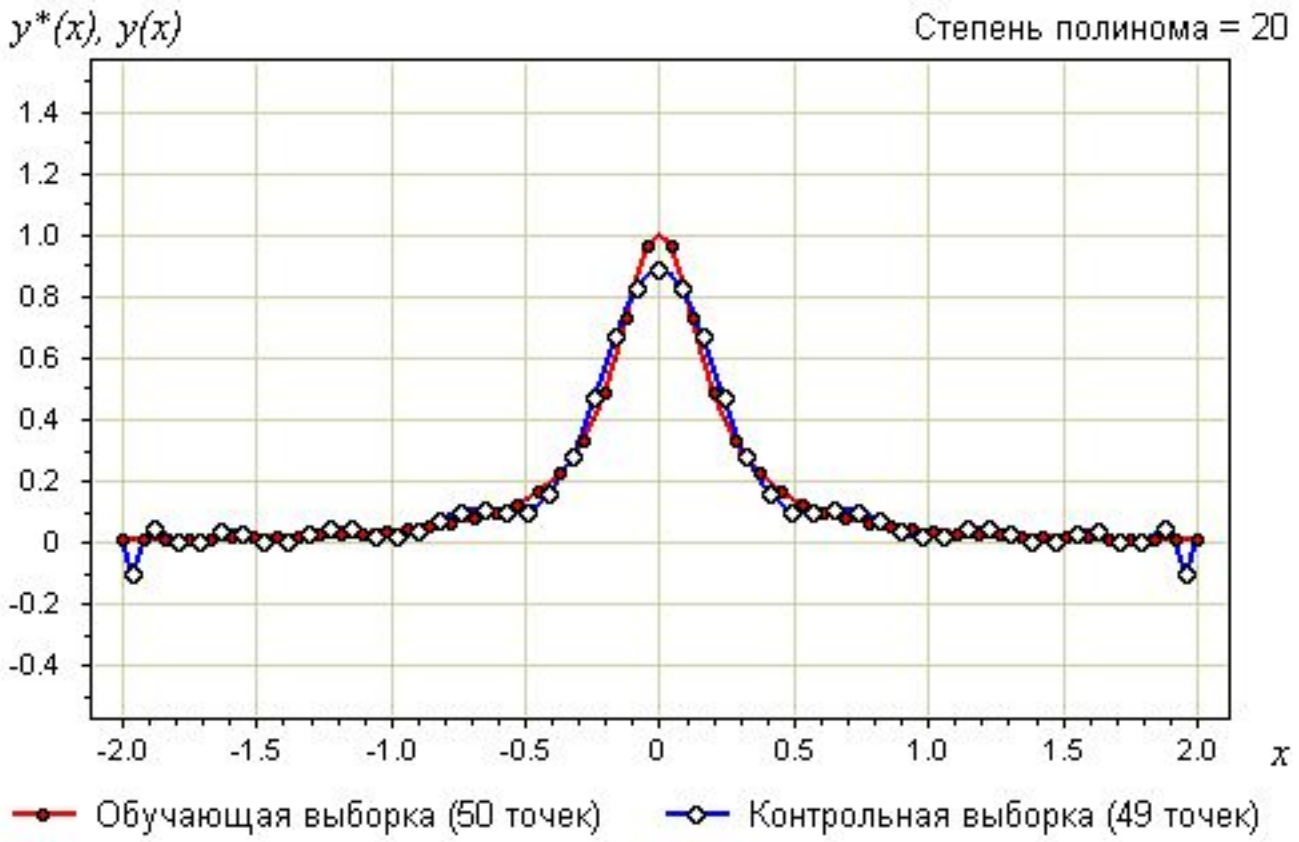
Выбор сложности модели

Evaluation

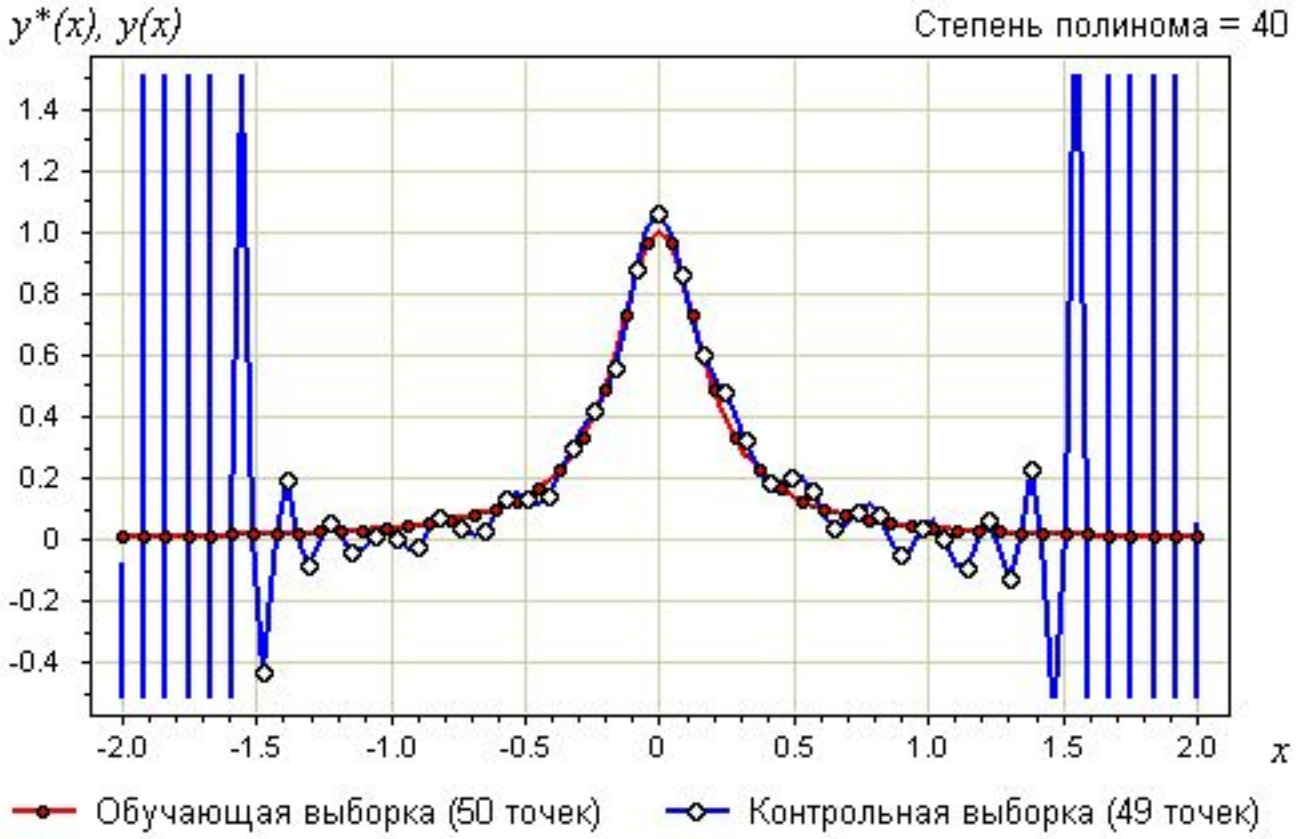
На примере


y*(x), y(x)
Степень полинома = 2
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-2.0
-1.5
-1.0
-0.5
0
0.5
1.0
1.5
2.0
x
Обучающая выборка (50 точек)
Контрольная выборка (49 точек)

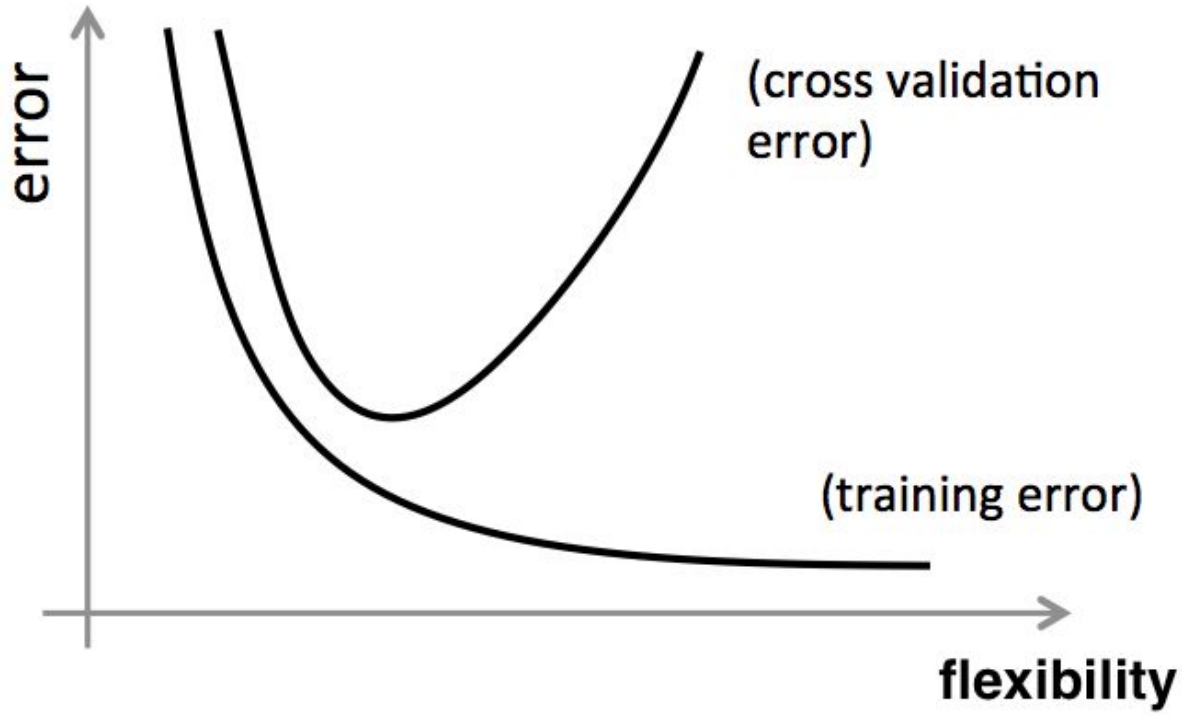
На примере


y*(x), y(x)
Степень полинома = 20
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-2.0
-1.5
-1.0
-0.5
0
0.5
1.0
1.5
2.0
x
Обучающая выборка (50 точек)
Контрольная выборка (49 точек)

На примере


y*(x), y(x)
Степень полинома = 40
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-2.0
-1.5
-1.0
-0.5
0
0.5
1.0
1.5
2.0
x
Обучающая выборка (50 точек)
Контрольная выборка (49 точек)

Формальная картинка же



Сложность (гибкость) модели

Чем больше в модели параметров, тем больше информации она может нести.

— Ваш К. О.

Какая бывает информация в параметрах:

- про генеральную совокупность;
- про выборку;
- про random seed.

Определение I

Переобучение, переподгонка (overfitting, high variance) — нежелательное явление, возникающее при решении задач обучения по прецедентам, когда вероятность ошибки обученного алгоритма на объектах тестовой выборки оказывается существенно выше, чем средняя ошибка на обучающей выборке.

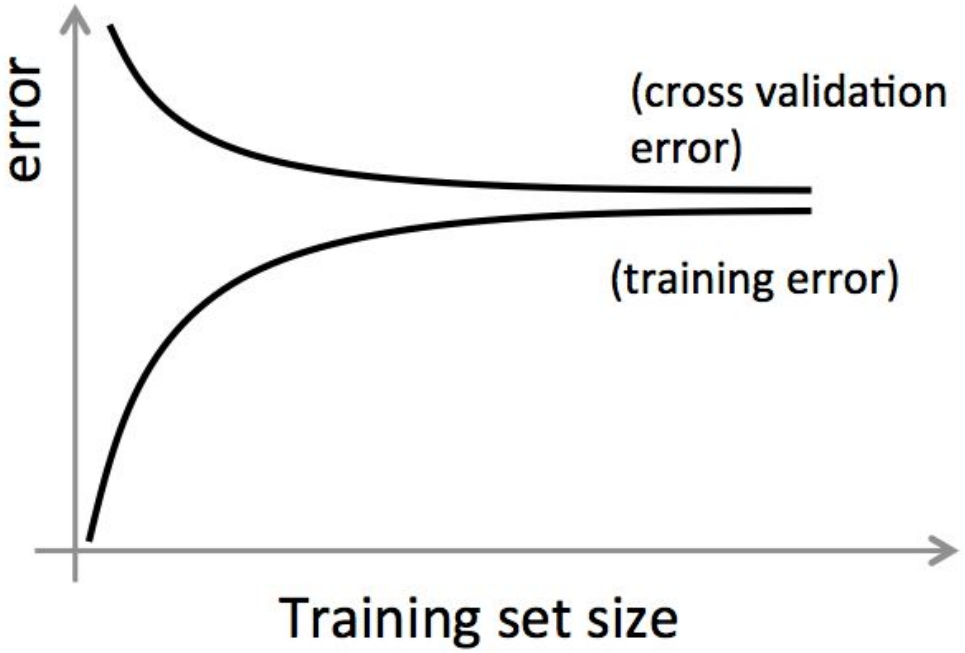
Определение II

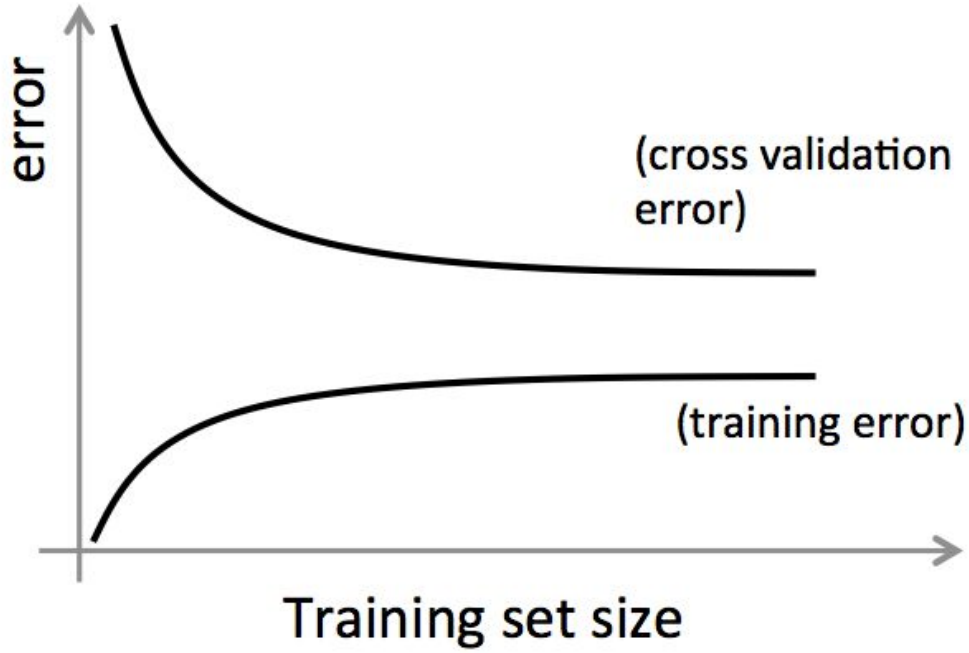
Недообучение (underfitting, high bias) — нежелательное явление, возникающее при решении задач обучения по прецедентам, когда алгоритм обучения не обеспечивает достаточно малой величины средней ошибки на обучающей выборке.

Зачем знать

- Увеличение числа примеров для обучения исправляет high variance
- Меньшее число факторов исправляет high variance
- Уменьшение сложности модели исправляет high variance
- Увеличение числа факторов исправляет high bias
- Увеличение сложности модели исправляет high bias

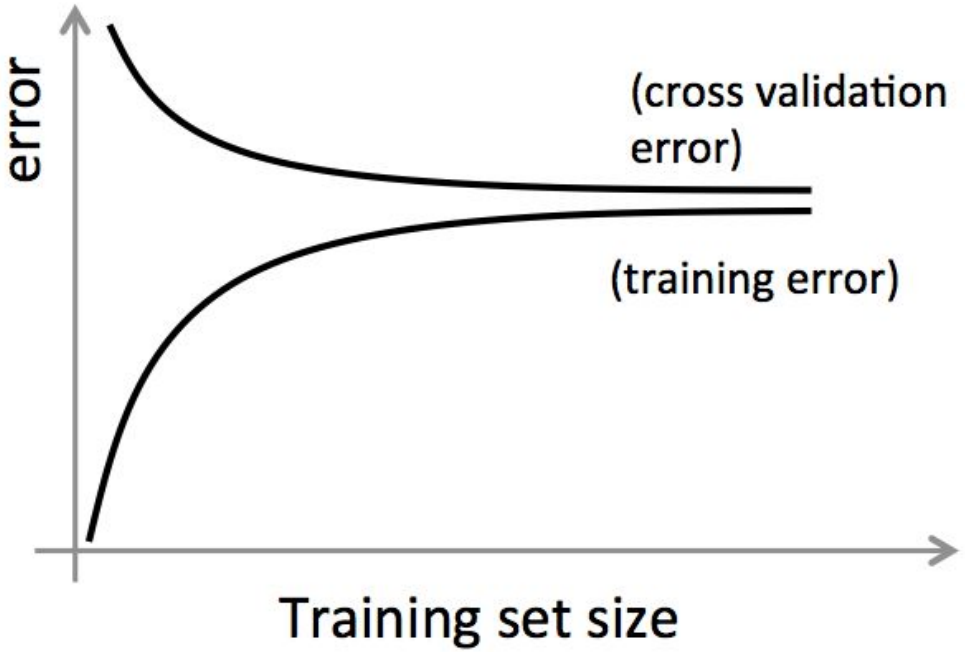
Как понять где мы находимся



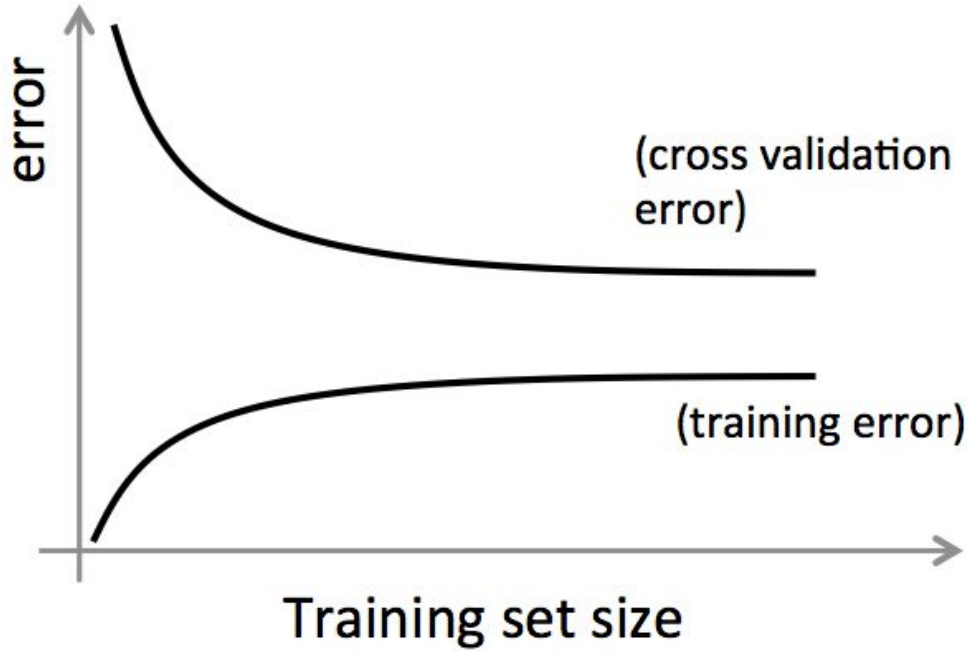


Как понять где мы находимся

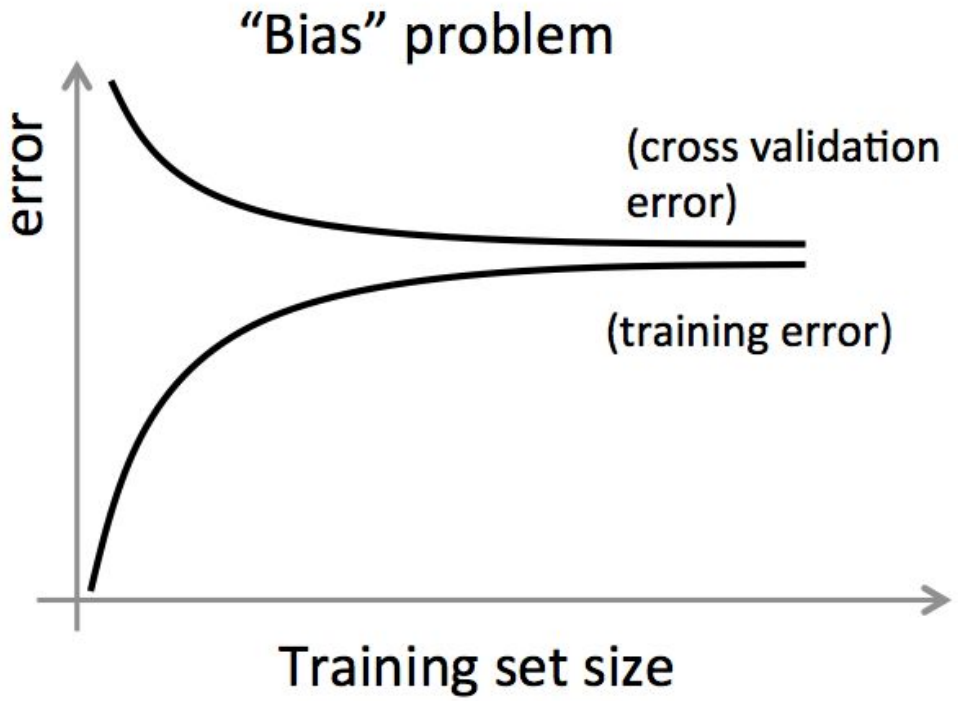
Underfit

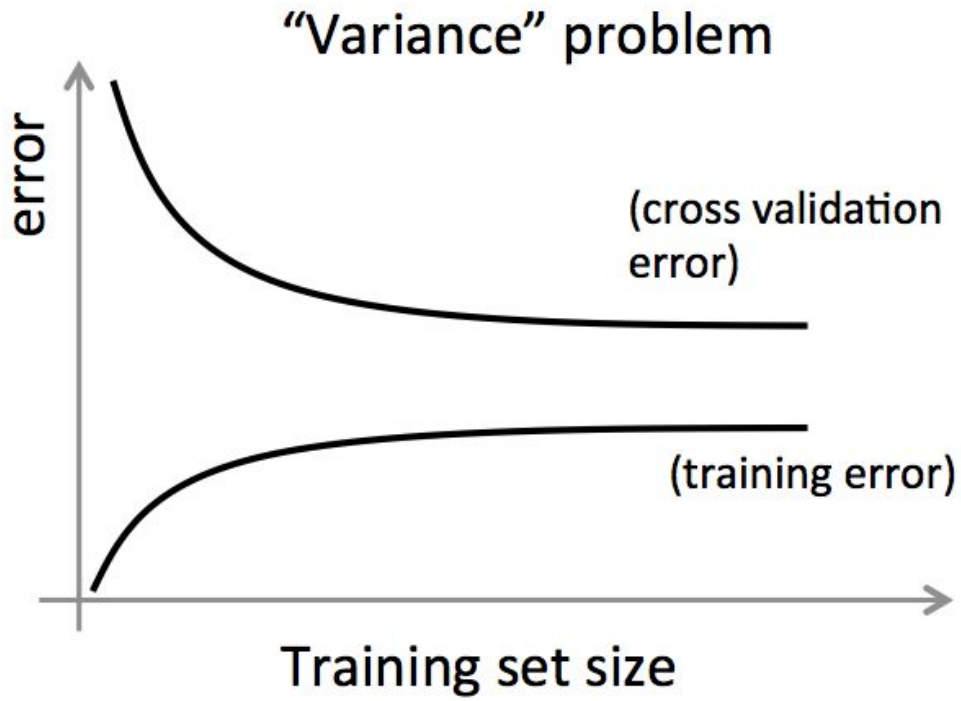


Overfit



Как понять где мы находимся





Как можно переобучиться

- Линейные модели: степень полинома
- Деревья решений: глубина дерева
- Нейронные сети: ширина и глубина
- SVM: Kernel trick
- …

Итого offline

- В академической среде чаще всего это про сложность модели
- В промышленности это про выбор лучшей решающей функции
- Смотреть на это надо как на СПР
- У нас есть проверка гипотез

Вопросы?