Лекция 6. Повторные выборки. Выбор модели

Папулин С.Ю. (papulin.study@yandex.ru)

Содержание

Повторные выборки

Когда достаточно много исходных данных/наблюдений, то наилучшим подходом будет разделение их на три части:

- Обучающее множество (для обучения модели, например, оценки коэффициентов линейной регрессии)
- Проверочное множество (для оценки ошибки предсказания при выборе модели)
- Тестовое множество (для определения ошибки предсказания выбранной модели)

Ошибка при тестировании (test error) есть средняя ошибка, которая вычисляется по результатам предсказания обученной модели на новых данных, которые не использовались при обучении.

Ошибка при обучении (training error) вычисляется по предсказаниям модели на обучающем множестве, т.е. по тем данными, которые использовались при обучении модели.

При отсутствии большого набора данных для тестирования, которые могут быть использованы для оценки ошибки, применяются следующие техники с использованием обучающих данных:

- Регулировка ошибки обучения для оценки ошибки тестирования
- Класс методов, которые оценивают ошибку тестирования за счет исключения части данных из обучающего множества (отложенной выборке) и последующего использования его для тестирования

Методы повторной выборки заключаются в повторяющемся извлечении экземпляров из обучающего множества и повторном обучении модели для каждой новой выборки для получения дополнительной информации о модели предсказания.

Основные методы повторной выборки:

- Кросс-валидация (Cross-validation)
- Бутстреп (Bootstrap)

Кросс-валидация используется для:

- Выбора модели (Model selection)
  Оценка производительности различных моделей для выбора наилучшей
- Оценки качества модели
  Оценка ошибки предсказания на новых данных

Бутстреп предназначен для измерения точности оценки параметров или модели предсказания

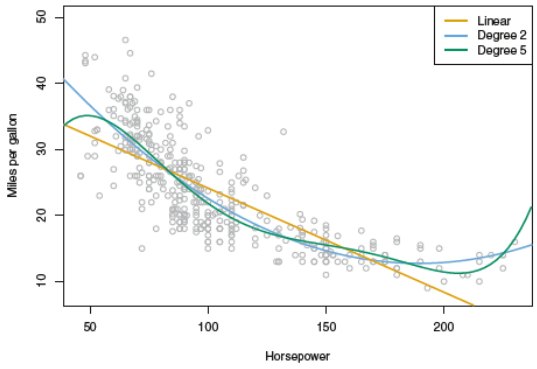
Отложенное множество/выборка

Подход с проверочным множеством (validation set) заключается в случайном разделении доступного множества наблюдений на две части: обучающее множество и проверочное множество, или отложенное множество. Данные части сопоставимы по размеру (по количеству входящих в них элементов).

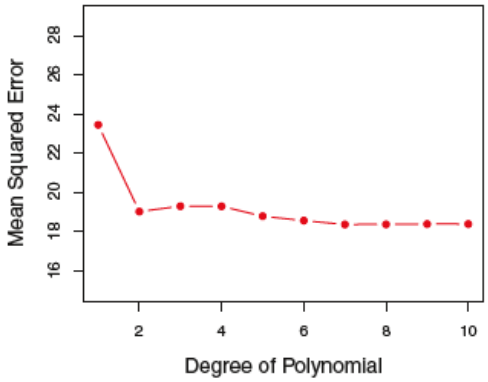
Для настройки модели используется обучающее множество, после чего обученная модель используется для предсказания ответа на наблюдениях из проверочного множества.

Итоговая ошибка на проверочном множестве дает оценку ошибки тестирования.

Пример

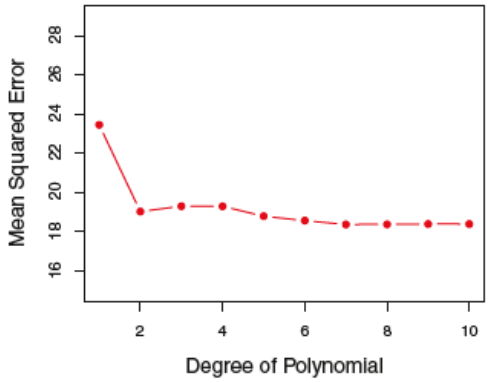


Зависимость потребления потлива (miles per gallon) от количества лошадиных сил (horsepower)

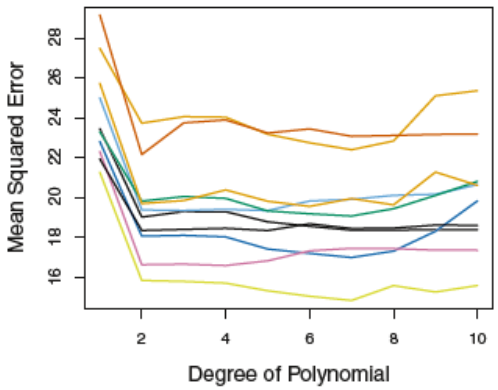


Оценка на проверочном множестве для одного разделения наблюдений на обучающее и проверочное множества

Если повторять процесс разделения на два множества случайным образом несколько раз, то получим разные оценки для тестового MSE



Оценка на проверочном множестве для одного разделения наблюдений на обучающее и проверочное множества



Оценка на проверочном множестве для 10 разделений наблюдений на обучающее и проверочное множества

В результате можно выделить следующие особенности:

- Все 10 кривых показывают, что модель со степенью 2 имеет меньшую ошибку MSE на проверочном множестве, чем для линейного признака (степень=1)
- Все 10 кривых показывают, что включение 3 степени и более высокого порядка признаков не ведет к существенному улучшению
- Каждая из 10 кривых дает в результате различную оценку ошибки тестирования для каждой из 10 рассмотренных моделей регрессии.

На основе вариативности кривых единственное, что можно сказать с некоторой долей уверенности, это то, что линейная модель (со степенью 1) не подходит к имеющимся данными/наблюдениями.

Подход с отложенной выборкой прост и легко применим.

Можно выделить следующие недостатки:

- Оценка ошибки тестирования на проверочном множестве может иметь высокую вариативность, т.е. может сильно зависеть от наблюдений, которые включены в обучающее и проверочное множества.
- При оценке ошибки тестирования по проверочному множестве она может быть завышена по сравнению с ошибкой тестирования при обучении на всем наборе данных. Это связанно с тем, что в общем случае модель хуже обучается на меньшем количестве наблюдений. В данным случае уменьшается количество наблюдения для обучения из-за отложенной выборки для проверочного множества.

Кросс-валидация – усовершенствованный подход с отложенным множеством, который используется для решения вышеуказанных проблем.

Кросс-валидация с leave-one-out

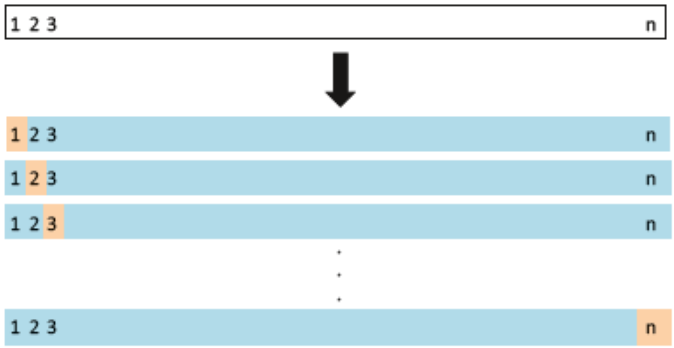
Кросс-валидация с leave-one-out (LOOCV) подобна ранее рассмотренному подходу с отложенной выборкой и также разделяется множество наблюдений на две части. Однако вместо создания двух соразмерных подмножеств, только один элемент (x_i, y_i) используется для проверочной части, а все остальные $\{(x_j, y_j)|(x_j, y_j) \in S, j \neq i\}$ для обучения. Более того производится n разделений по

количестве наблюдений для каждого элемента из S. Таким образом, обучаются n моделей и соответственно n значений ошибок, $MSE_1, MSE_2, \dots, MSE_n$, на проверочном множестве.

MSE_i является несмещенной (unbiased) оценкой ошибки тестирования с высокой вариабельностью (разбросом/колебанием), т.к. основана на единственном наблюдении (x_i, y_i)

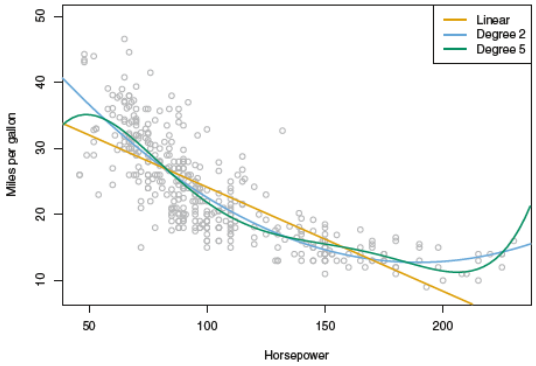
Оценка MSE тестирования при LOOCV вычисляется как среднее значение индивидуальных оценок $\{MSE_i\}$:

$$\mathrm{CV}_{(n)} = \frac{1}{n}\sum_{i=1}^{n}\mathrm{MSE}_i.$$

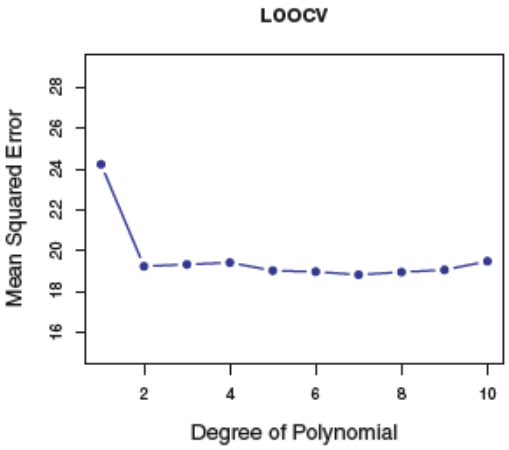


LOOCV имеет несколько важных преимуществ над обычным походом с отложенной выборкой

- Имеет меньшее смещение (bias) из-за большего количества наблюдений, участвующих в обучении. В результате данный подход менее склонен к завышению ошибки тестирования (test error), чем подход с отложенной выборкой.
- Выполнение LOOCV множество раз дает один и тот же результат, т.е. нет случайности в делении множества наблюдений на обучающее и проверочное.



Зависимость потребления потлива (miles per gallon) от количества лошадиных сил (horsepower)



Оценка при LOOCV

Из-за того, что в LOOCV необходимо обучить n моделей (т.е. по количеству наблюдений), вся эта процедура может быть сложно реализуемой или потребует значительных вычислительных ресурсов.

Для линейной регрессии есть более простой вариант, чтобы избежать обучение n моделей. В этом случае модель просто обучается на всем множестве наблюдений, а потом используется следующая формула вычисления ошибки:

$$\mathrm{CV}_{(n)} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \hat{y}_i}{1 - h_i}\right)^2,$$

$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{i'=1}^{n}(x_{i'} - \bar{x})^2}.$$

LOOCV является общим подходом и может быть использован для любого рода моделей предсказания.

Кросс-валидация с k-Folds

Альтернативой для LOOCV является k-folds кросс-валидация.

Данный подход заключается в случайном разделении множества наблюдений на k групп, непересекающихся частей (folds), одинакового размера.

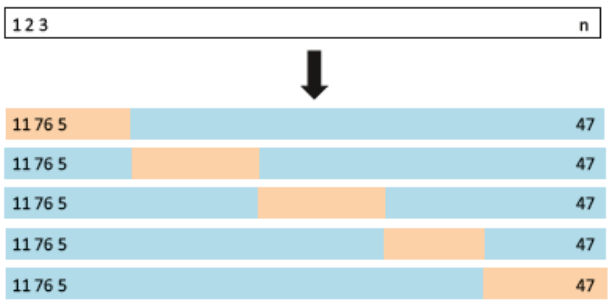
Первая часть выполняет роль проверочного множества, остальные $k-1$ используются для обучения. Таким образом, MSE_1 вычисляется на отложенной выборке.

Данный процесс повторяется k раз. При этом каждый раз используется одна отличная часть как проверочное множество, а остальные как одно обучающее множество.

В результате получается k оценок ошибки тестирования, $MSE_1, MSE_2, \ldots, MSE_k$. Общая оценка с k-folds кросс-валидацией вычисляется следующим образом:

$$\mathrm{CV}_{(k)} = \frac{1}{k}\sum_{i=1}^{k}\mathrm{MSE}_i.$$

Как правило, используется $k = 5$ или $k = 10$.



Преимущества по сравнению с LOOCV:

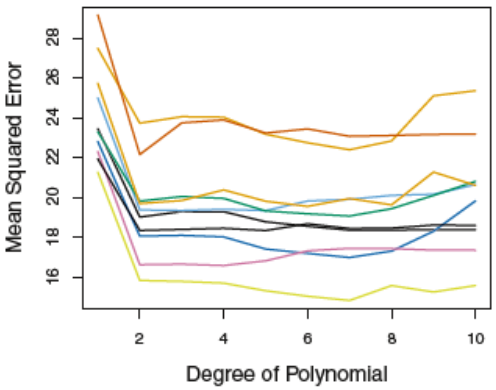
- Вычислительные, т.к. меньше моделей для обучения.
- Чаще дает более точную оценку ошибки тестирования. Оценка ошибки тестирования при LOOCV имеет склонность к более высокой дисперсии, чем оценка при использовании k-Folds.

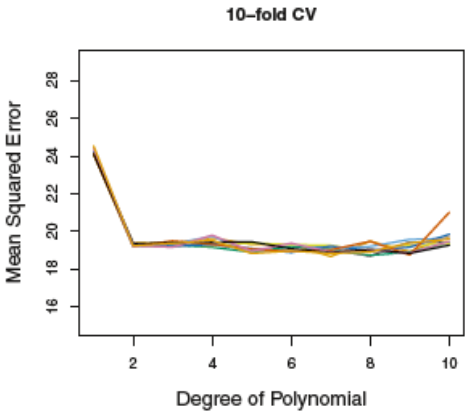
Пример

Разделяем исходные наблюдения на 10 частей случайным образом. Каждая часть используется как проверочное множества, а на оставшихся обучается модель.

В итоге получаем 10 моделей и 10 MSE для каждой. Усредняем и получаем значение CV. Так повторяем для каждой степени полинома с использованием исходных 10 частей.

На рисунке справа показаны оценки ошибки тестирования для 9 случайных разбиений на 10 частей, т.е. 9 раз 10-folds



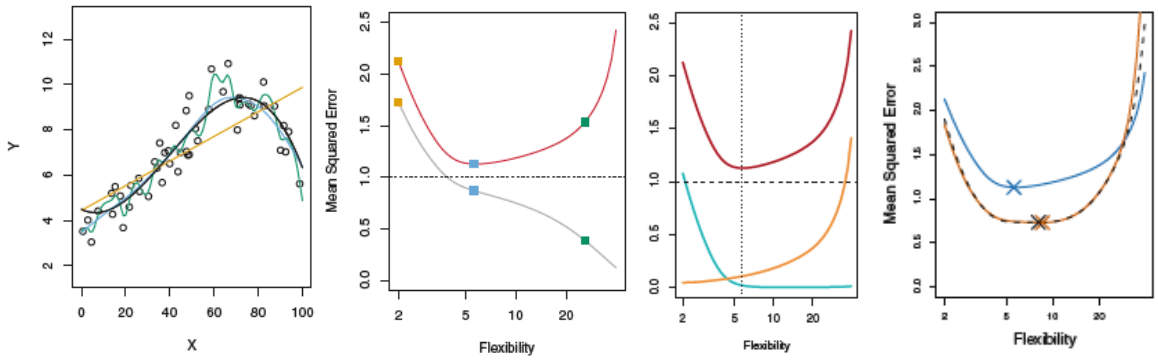


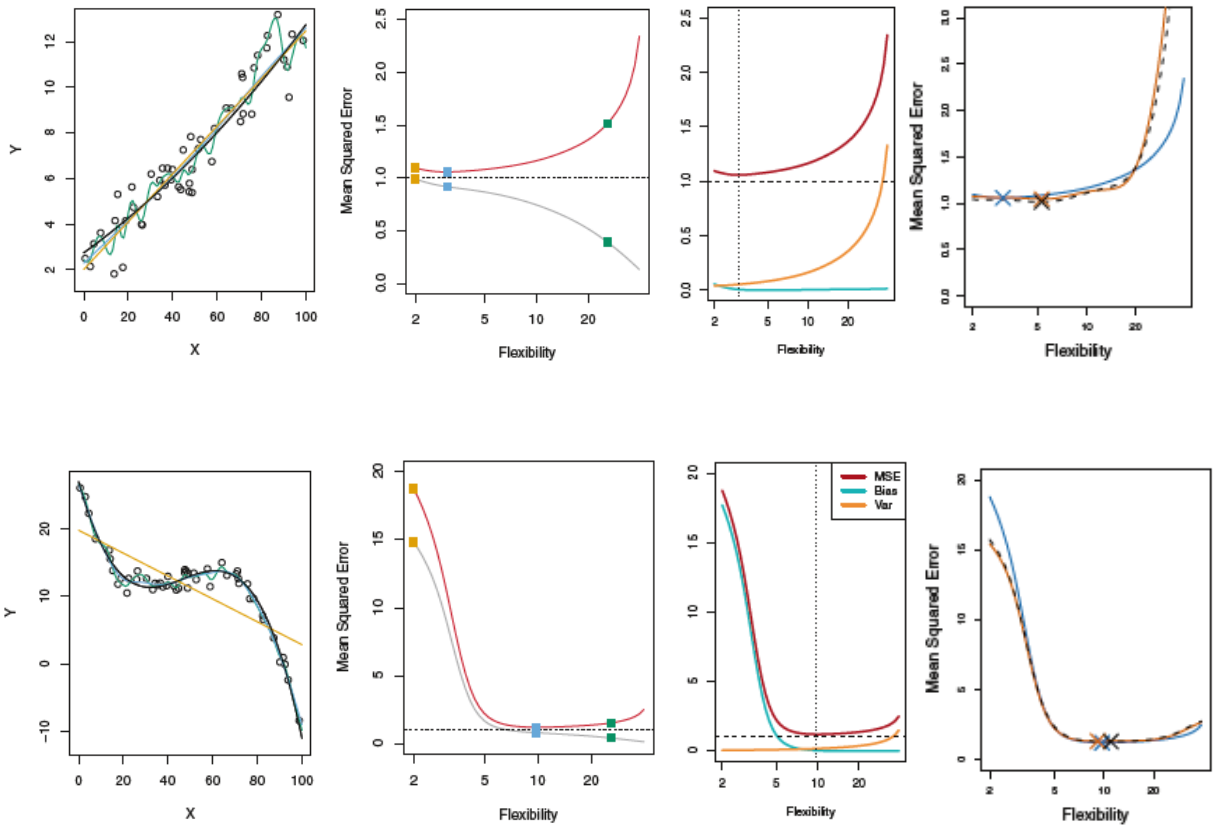
Оценка на проверочном множестве для 10 разделений наблюдений на обучающее и проверочное множества случайным образом

Примеры

Крайний правый график:

- синим – действительная ошибка тестирования по MSE (true test MSE) (это сгенерированные данные, поэтому известно действительная функция линейной регрессии, и поэтому можно определить true test MSE). Когда мы исследуем реальные данные, мы не знаем true test MSE и поэтому сложно определить точность оценки с кросс-валидацией
- черным пунктиром – LOOCV
- оранжевым – 10-fold CV





Комментарии к примеру:

- На всех трех графиках (трех справа) кросс-валидационные оценки (LOOCV и k-Folds CV) очень близки
- На верхнем графике CV имеет правильную форму, но занижает действительную ошибку тестирования по MSE (true test MSE)
- Несмотря на то, что все графики CV иногда занижают true test MSE, все они достаточно близко определяют корректный уровень гибкости (здесь используется сглаживающая регрессия с разным количеством узловых точек, которые определяют гибкость. Аналогия со степенями полиномиальной регрессии)

Особенности использования CV:

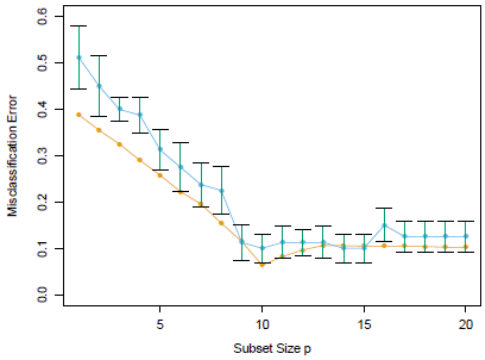
- *Оценка качества модели*

  При использовании кросс-валидации основной целью может быть определение того, как хорошо модель может работать на новых данных, которые не использовались при обучении. В этом случае интерес заключается в оценке действительной ошибки тестирования MSE.

- *Выбор модели*

  Иногда необходимо знать только положение точки с минимальной оценкой ошибки, пример, при выборе степени полинома. В этом случае особую роль играет форма кривой оценки ошибки и положении её минимальной оценки. Точность самой оценки не имеет значения.

Для уменьшения количества параметров применяется правило одной сигмы. Рассчитывается стандартное отклонение в точке с минимальным значением ошибки. На рисунке ниже это точка 10. Если значение ошибок слева от рассматриваемой точки укладывается в вычисленный диапазон, то выбирается модель с меньшим количеством параметров/признаков. В данном случае выбирается модель с 9 признаками.



k-Folds CV является общим подходом и может быть использован для любого рода моделей предсказания.

Кросс-валидация для задачи классификации

Вычисляется аналогично LOOCV и k-Folds CV.

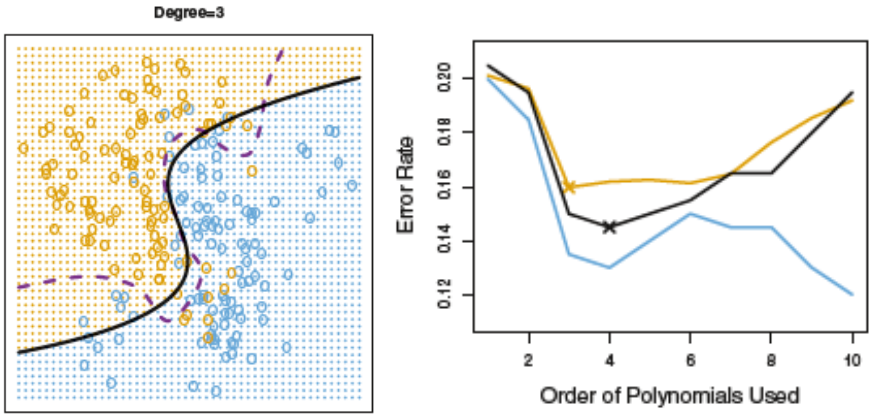
Для LOOCV:

$$\mathrm{CV}_{(n)} = \frac{1}{n}\sum_{i=1}^{n}\mathrm{Err}_i,$$

$$\mathrm{Err}_i = I(y_i \neq \hat{y}_i)$$

Пример:

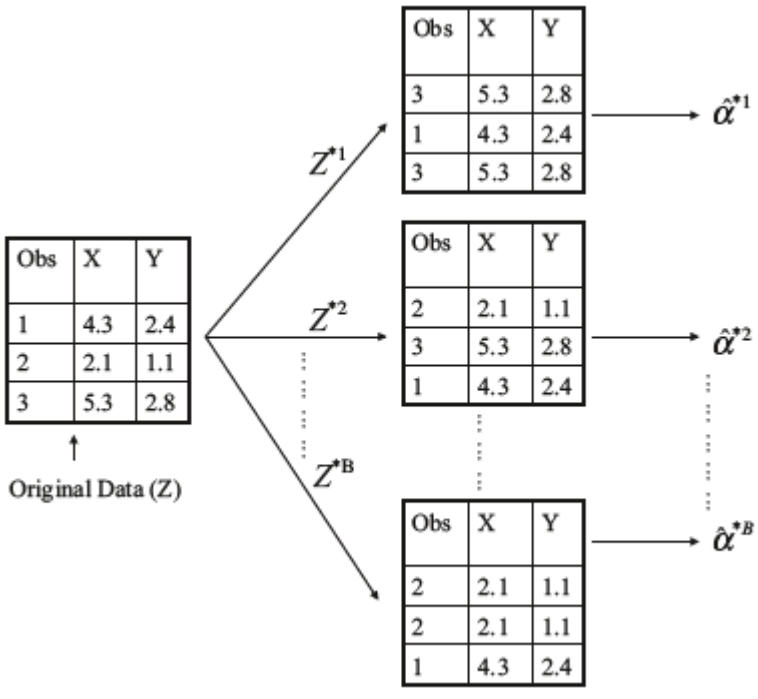
- Черным – CV 10-Folds оценка ошибки тестирования при использовании логистической регрессии для различных степеней полинома
- Синим – CV 10-Folds ошибка обучения
- Оранжевым – действительная ошибка тестирования



Ошибка обучения уменьшается при увеличении гибкости модели. Поэтому её нельзя использовать для выбора модели. Хотя ошибка тестирования с кросс-валидацией немного занижает действительную ошибку, она дает хорошее приближение относительно того, какую модель необходимо выбрать. В данном случает выбирается 4ую степень, что достаточно близко к действительному значению 3.

Бутстреп (Bootstrap)

Бутстреп может быть использован в условия небольшого количества исходных наблюдений. Данный подход эмулирует процесс получения новых выборочных множеств таким образом, что можно оценить вариабельность некоторого параметра без новых наблюдение, т.е. основываясь только на имеющихся данных. Таким образом, вместо получения новых наблюдений из генеральной совокупности, генерируем различные наборы данных посредством многократных выборок с возвратом из исходного набора наблюдений.

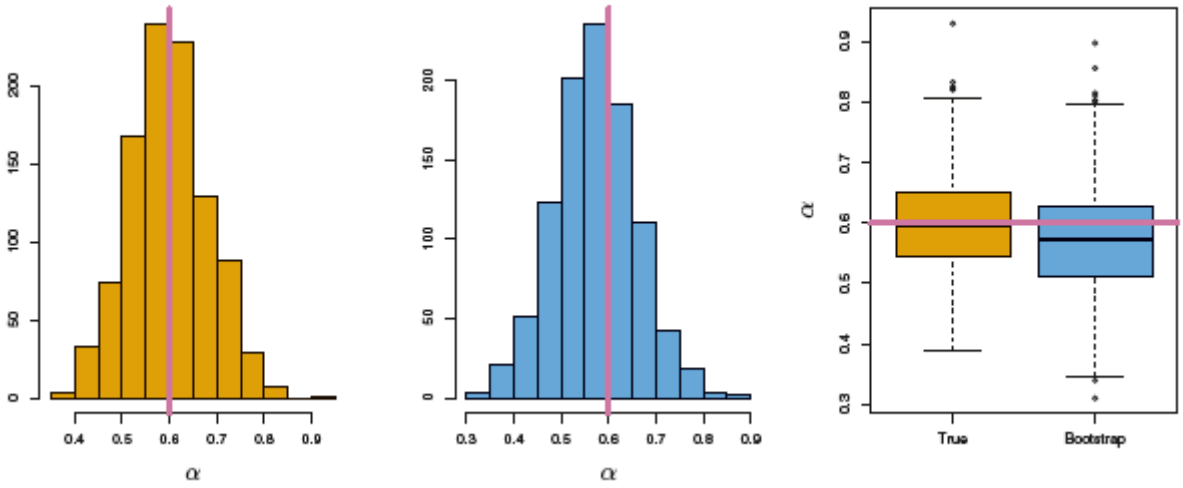


Примеры использования:

- Оценка точности отдельных статистик (математическое ожидание, дисперсия, среднее, стандартное отклонение)
- Оценка точности моделей предсказаний: оценка стандартных ошибок коэффициентов линейной регрессии

Пример

Оценка некоторого параметра α (подробности в [1])



- Слева изображена гистограмма оценок некоторого параметра α, полученные посредством симуляции 1000 выборок из генеральной совокупности (мы можем получать новые наблюдения в неограниченном количестве)
- Посередине гистограмма 1000 оценок параметра α, полученных посредством бутстрепа, т.е. из одного исходного ограниченного набора наблюдений.
- Справа приведена диаграмма размаха для оценок полученных двумя способами, которая показывает, что оценки бутстрепам схожи с оценками при неограниченном наборе данных.
- Розовая линия – реальное значение параметра α

Список литературы

1. Chapter 5. Resampling Methods // An Introduction to Statistical Learning by Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshir. pp. 175–190. URL: http://faculty.marshall.usc.edu/gareth-james/ISL/
2. Chapter 7. Model Assessment and Selection // The Elements of Statistical Learning by Trevor Hastie, Robert Tibshirani, Jerome Friedman. pp. 219–257. URL: https://web.stanford.edu/~hastie/ElemStatLearn/