2021

Лекция 7. Повторные выборки и выбор модели

МЕТОДЫ МАШИННОГО ОБУЧЕНИЯ

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

Дополнительные материалы доступны по ссылке: https://github.com/mlmethods

Содержание

1. Повторные выборки

Ранее мы рассматривали кривую обучения и определили два вида ошибок: при обучении и тестировании.

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

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

Методы предсказания должные не столько давать хорошее качество предсказания на обучающем множестве, сколько на тех данных, которое модель ещё не видела. То, как ведет себя модель на новых данные, представляет наибольший интерес. Каким образом это можно измерить, если мы ограничены в количестве наблюдений?

Прежде всего определимся с возможными задачами. Можно выделить две:

- выбор модели/гиперпараметров модели (сравнение и выбор)
- оценка качества предсказания модели (оценка действительной ошибки с использованием количественных интерпретируемых метрик)

Существует несколько техник для решения обозначенных задач:

- Оценка ошибки тестирования за счет корректировки ошибки обучения (C_p, AIC, BIC, скорректированные R^2). Данный подход может быть использован для выбора модели с разным количеством параметров;
- Оценка ошибки тестирования за счет исключения части данных из обучающего множества и последующего использования его для тестирования. Подходит как для выбора модели, так и для оценки качества. Вариантами данного подхода являются:
    - Отложенная выборка
    - Повторные выборки

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)
  Сравнение различных моделей для выбора наилучшей
- Оценки качества модели
  Оценка ошибки предсказания на новых данных

Бутстреп предназначен для измерения точности оценки параметров или модели предсказания

Когда у нас 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, например, степени полинома или коэффициента регуляризации)
- Тестовое множество (для определения ошибки предсказания выбранной модели)

2. Отложенное множество/выборка

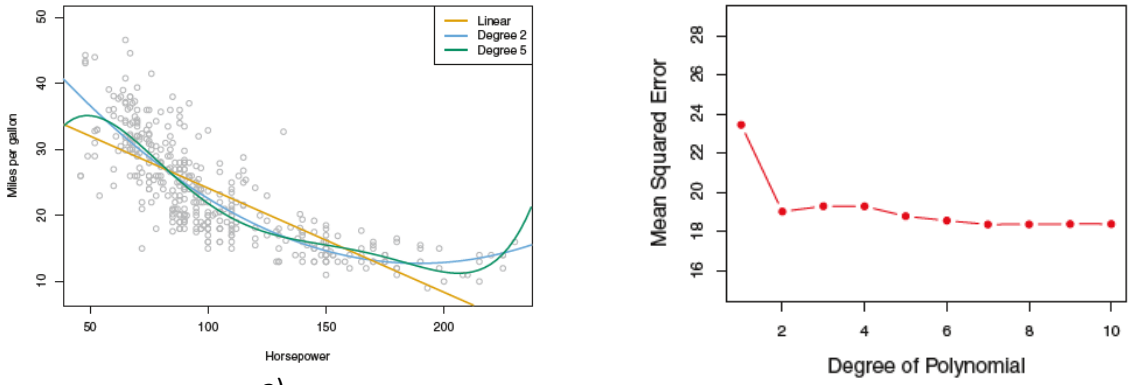
Подход с отложенным множеством заключается в случайном разделении доступного множества наблюдений на две части: обучающее множество и проверочное (или тестовое) множество. Эти части сопоставимы по размеру (по количеству входящих в них элементов). Как правило, это 80 на 20, 70 на 30, 60 на 40.

Следует отметить, что в случае с определением качества модели отложенное множество будем называть тестовым (test set), а при выборе гиперпараметров (или модели) – проверочным (validation set).

Для настройки модели используется обучающее множество, после чего обученная модель применяется для предсказания отклика на наблюдениях из отложенной выборки.

Итоговая ошибка на отложенном множестве дает оценку ошибки тестирования.

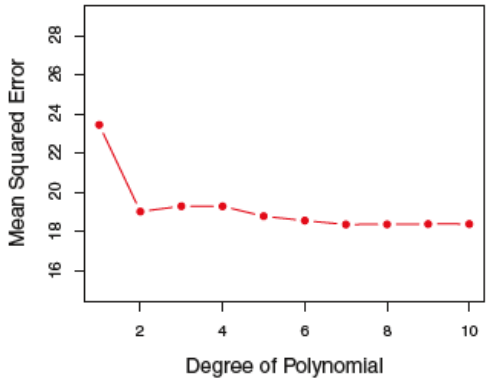
Пример



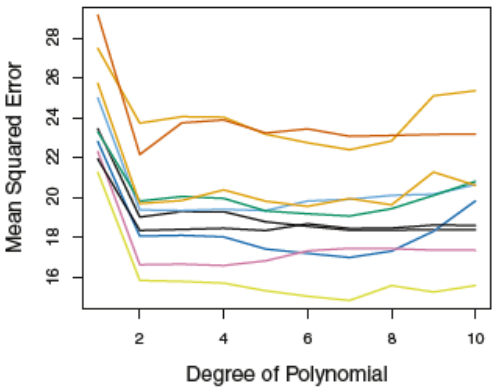
а) б)

Рисунок – Пример оценки ошибки предсказания посредством отложенной выборки [1]: а) зависимость потребления потлива (miles per gallon) от количества лошадиных сил (horsepower); б) кривая оценки тестовой ошибки, полученная посредством отложенной выборки

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



а)



б)

Рисунок – Сравнение оценки тестовой ошибки при использовании техники отложенной выборки при различных разбиения исходные данных на обучающее и проверочное [1]: а) одна отложенная выборка; б) 10 отложенных выборок (разбиение каждый раз случайным образом)

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 на проверочном 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.

3. Кросс-валидация с leave-one-out

Кросс-валидация с leave-one-out (LOOCV) подобна ранее рассмотренному подходу с отложенной выборкой, когда множество наблюдений разделяется 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

количеству наблюдений для каждого элемента из S. Таким образом, обучаются n моделей и соответственно получаем 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), так как основана на единственном наблюдении (x_i, y_i)

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

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

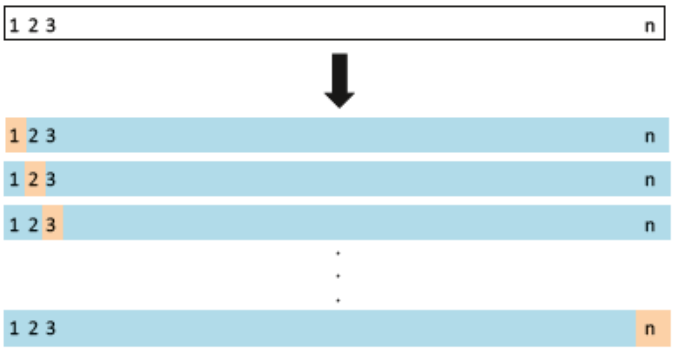


Рисунок – Пример разделения данных при LOOCV [1]

LOOCV имеет несколько важных преимуществ над обычным подходом 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, то есть нет случайности в делении множества наблюдений на обучающее и проверочное.

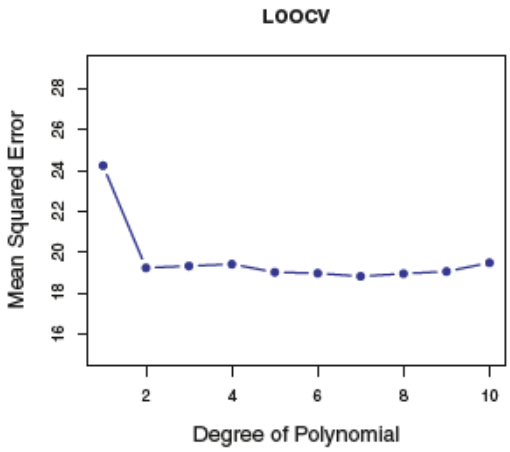


Рисунок – Оценка текстовой ошибки посредством LOOCV [1]

Из-за того, что в LOOCV необходимо обучить n моделей (то есть 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:

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

где

$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{j=1}^{n}\left(x_j - \bar{x}\right)^2}.$$

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

4. Кросс-валидация с k-Folds

4.1. Общие сведения

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 $k = 10$.

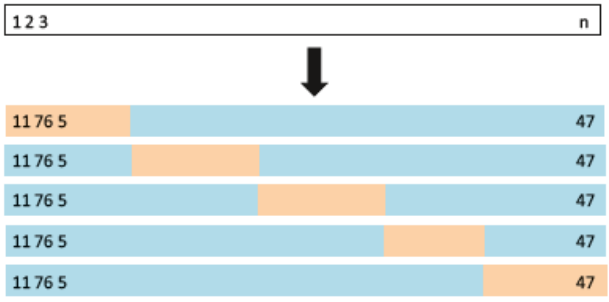


Рисунок – Пример разделения данных при k-folds [1]

Преимущества по сравнению с LOOCV:

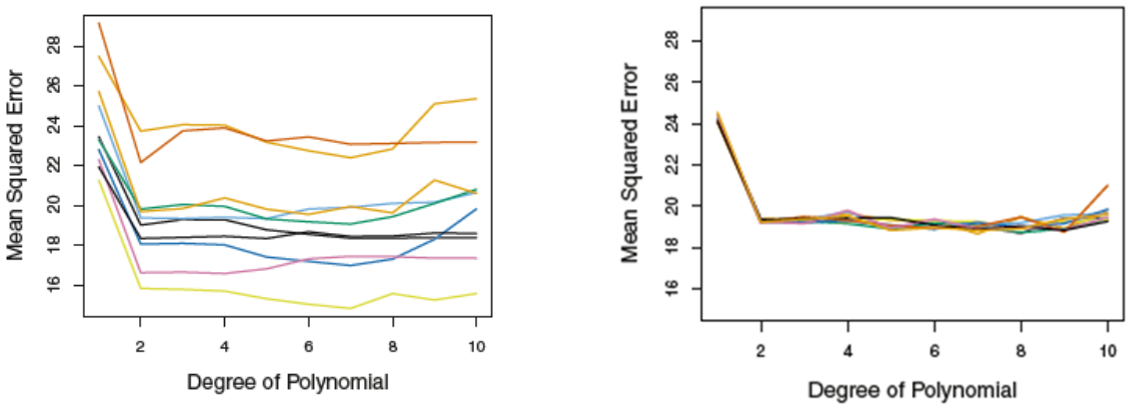
- Вычислительные, так как меньше моделей для обучения.
- Чаще дает более точную оценку ошибки тестирования. Оценка ошибки тестирования при LOOCV имеет склонность к более высокой дисперсии, чем оценка при использовании k-Folds.

Пример

Разделяем исходные наблюдения на 10 частей случайным образом. Каждая часть используется как проверочное множество, 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, то есть 9 раз 10-folds



а) б)

Рисунок – Оценки тестовых ошибок [1]: а) 10 раз запускаем обучение с отложенной выборкой; б) 10 раз запускаем обучение с кросс-валидацией

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

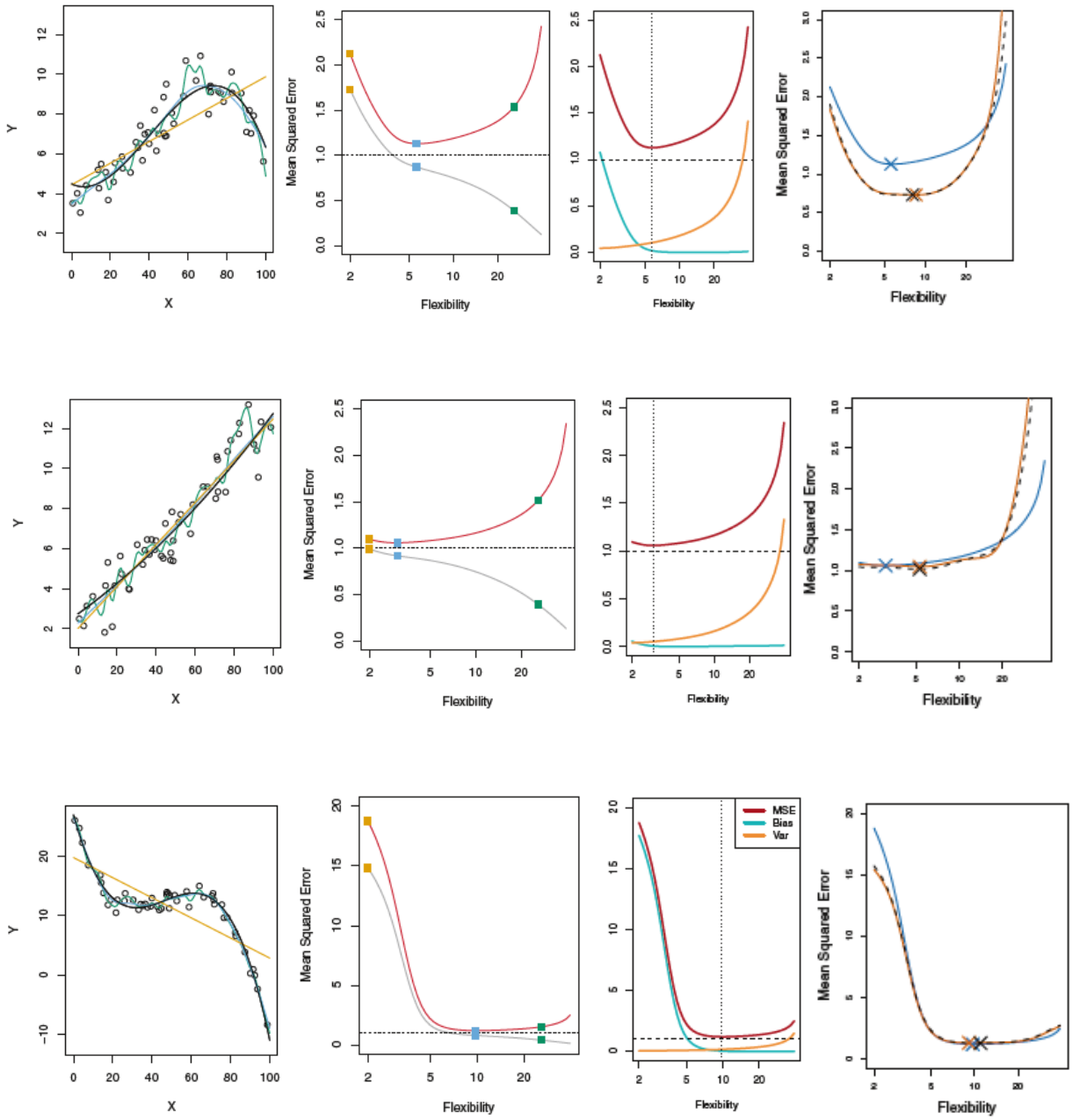


Рисунок – Действительная тестовая ошибка и её оценка для разных наборов данных [1]

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.

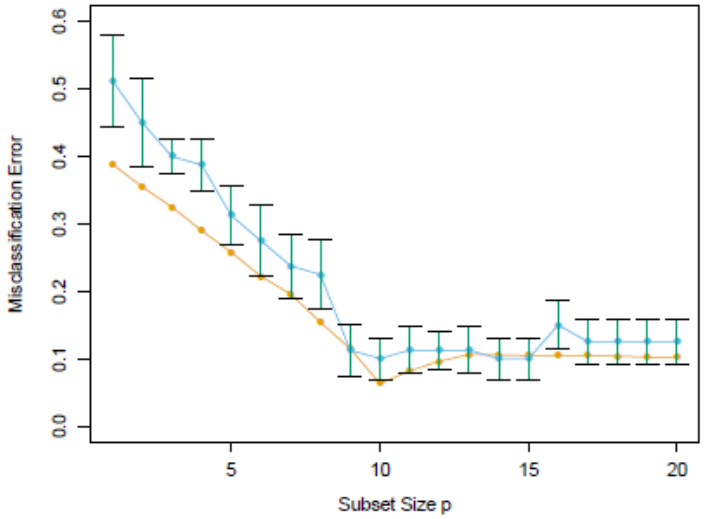


Рисунок – Применение правила одной сигмы [2]

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

4.2. Кросс-валидация для задачи классификации

Вычисляется аналогично LOOCV и k-Folds CV.

Для LOOCV:

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

где

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

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

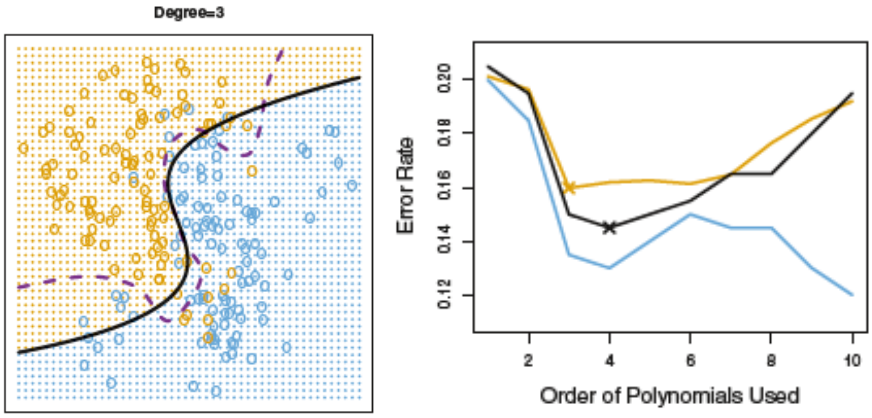


Рисунок – Выбор степени полинома для задачи классификации [1]

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 4ая степень, что достаточно близко к действительному значению 3.

4.3. Виды кросс-валидации с k-folds

Виды кросс-валидации с k-folds:

- С последовательным разделением

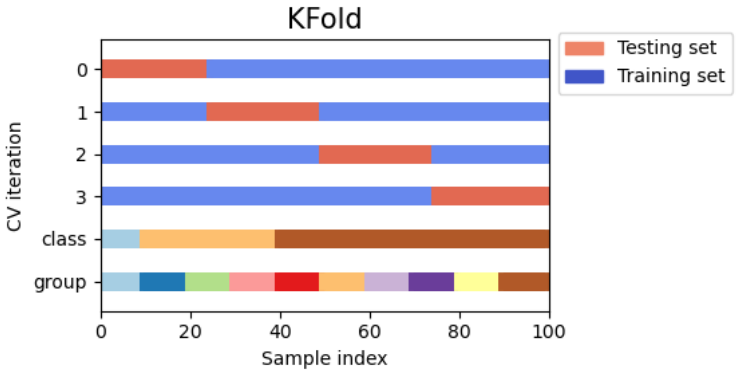


Рисунок x – K-fold кросс-валидация [3]

- С тасовкой
- Стратифицированная с последовательным разделением

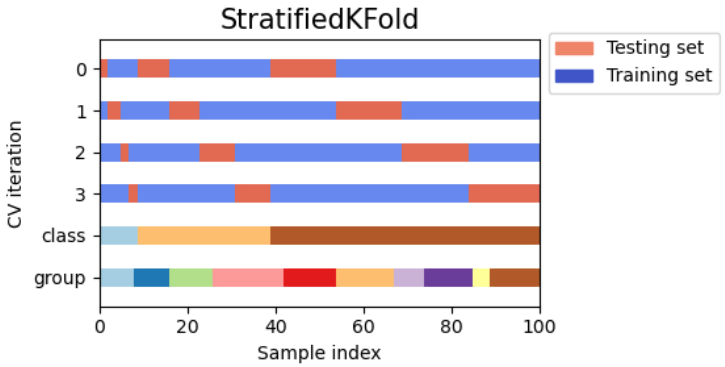
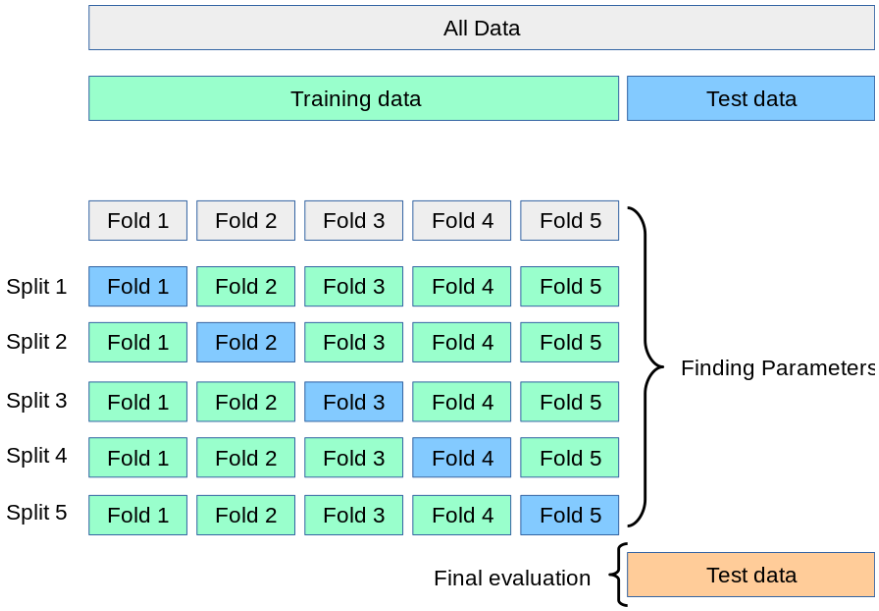


Рисунок x – Стратифицированная k-fold кросс-валидация [3]

- Стратифицированная с тасовкой

4.4. Выбор модели и оценка качества предсказания

Схема разбиения данных с k-fold кросс-валидацией для выбора модели и отложенной выборкой для оценки качества [3]:



4.5. Вложенная кросс-валидация

Мы можем воспользоваться кросс-валидацией для оценки качества модели. В этом случае у нас нет отдельной тестовой части и все исходные данные будут разбиты на k частей. В итоге получим k сплитов, каждый из которых будет формировать обучающее и тестовое множества. При этом полагаем, что модель уже известна, то есть нет необходимости выбирать модель. Для оценки качества предсказания мы для каждого сплита обучаем нашу модель на обучающей части и выполняем оценку на тестовой. В результате получаем k оценок. Их среднее значение и будет итоговая оценка качества модели.

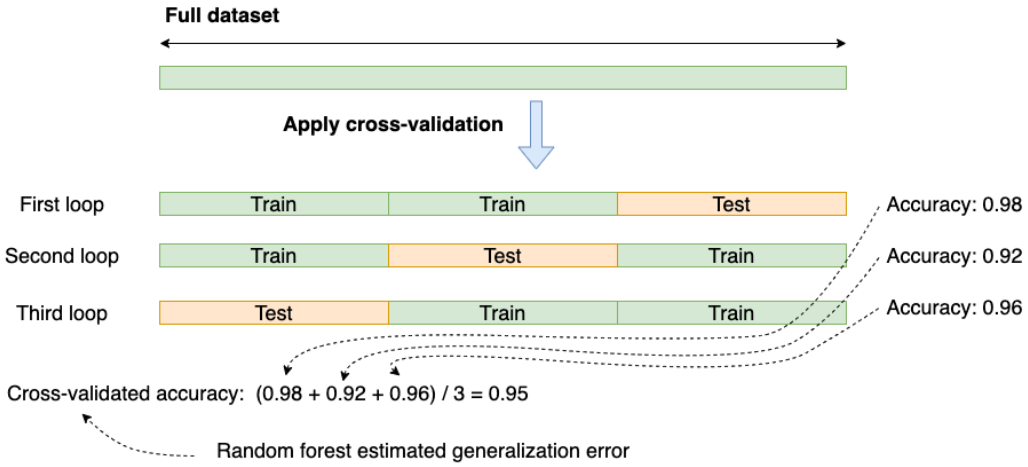


Рисунок – Кросс-валидация для оценки тестовой ошибки [4]

Если ставится задача выбора модели, то мы можем повторить ранее приведенную процедуру для каждой модели и выбрать ту, которая покажет лучший результат. Однако в этом случае, как правило, оценка качества модели будет завышена.

Для получения более адекватной оценки используется вложенная кросс-валидация. Данная техника предполагает наличие внешней и внутренней кросс-валидации. При этом количество частей для внешней и внутренней кросс-валидации, на которые разбиваются данные, могут не совпадать. Обучающая часть каждого внешнего сплита подвергается внутренней кросс-валидации. Внутренняя кросс-валидация применяется для выбора гиперпараметров. Другими словами, мы оцениваем качество предсказания на проверочном множестве для каждого сплита внутренней кросс-валидации и усредняем их. Таким же образом поступаем для каждого набора гиперпараметров. Выбираем тот набор, который показывает лучшее качество предсказывания. После этого заново обучаем модель с выбранными гиперпараметрами на всей обучающей части внешнего сплита и оцениваем качество предсказания на тестовой части этого же сплита. Процедура повторяется для каждого внешнего сплита. В результате получаем k_{outer} оценок, которые усредняем и получает итоговое значения оценки качества предсказания нашей модели.

После это для выбора итоговых гиперпараметров модели, запускаем обычную кросс-валидацию на всем наборе данных. При этом оценка качества будет та, которую мы получили посредством вложенной кросс-валидации.

Если необходимо выбрать из нескольких моделей, каждая из которых имеет свой набор гиперпараметров, то для каждой модели в отдельности применяется аналогическая процедур. Выбирается та модель, которая дает лучшую оценку предсказания. В данном случае нас интересует только оценка качества предсказания, а не конкретные гиперпараметры. Гиперпараметры для выбранной модели подбираются на следующем этапе посредством обычной кросс-валидации на всем наборе данных.

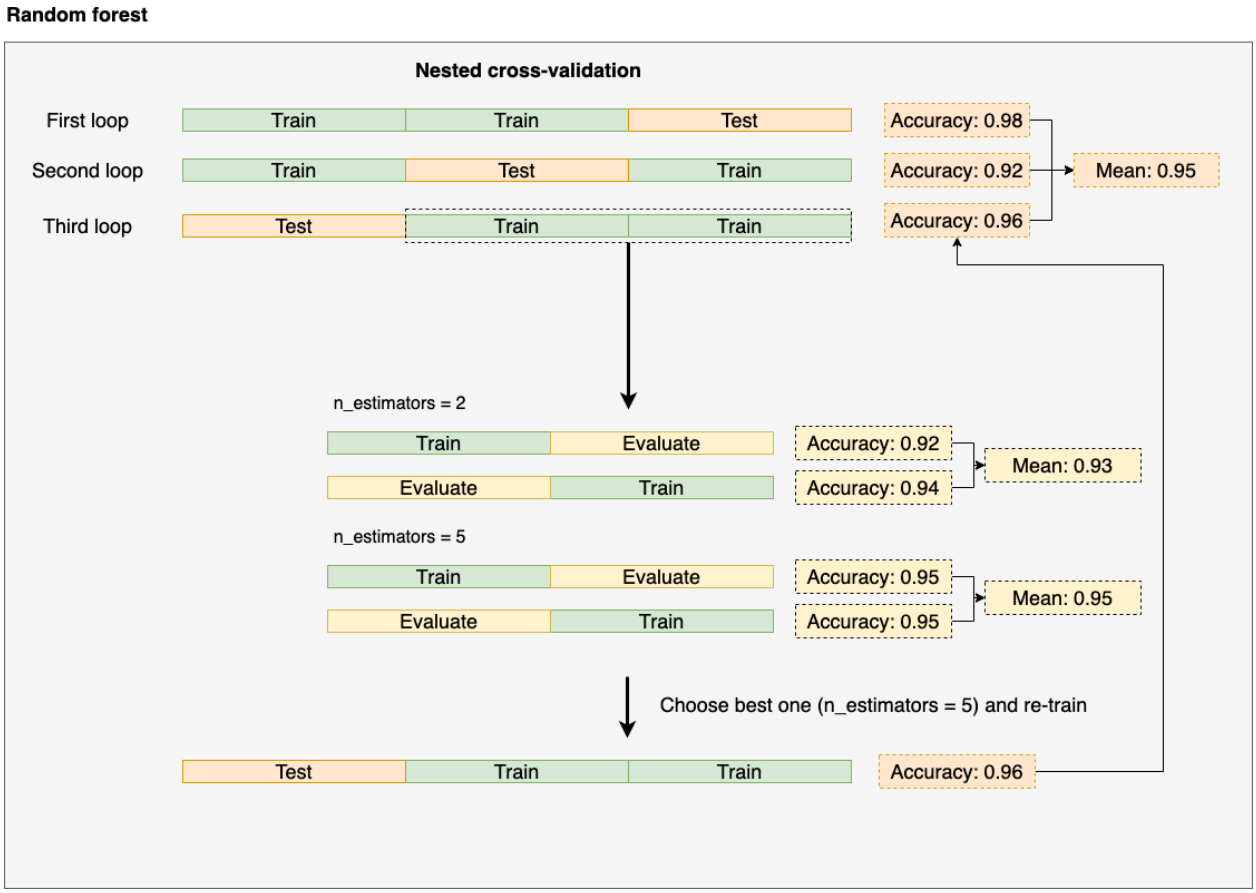
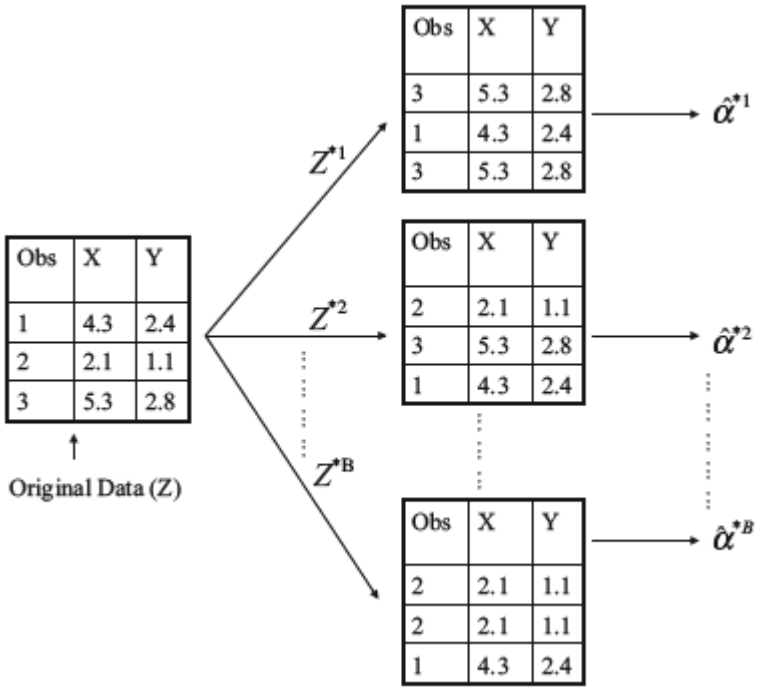


Рисунок – Вложенная кросс-валидация для оценки тестовой ошибки [4]

5. Бутстреп (Bootstrap)

Бутстреп может быть использован в условиях небольшого количества исходных наблюдений. Данный подход эмулирует процесс получения новых выборочных множеств таким образом, что можно оценить вариабельность некоторого параметра без новых наблюдение, то есть 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.

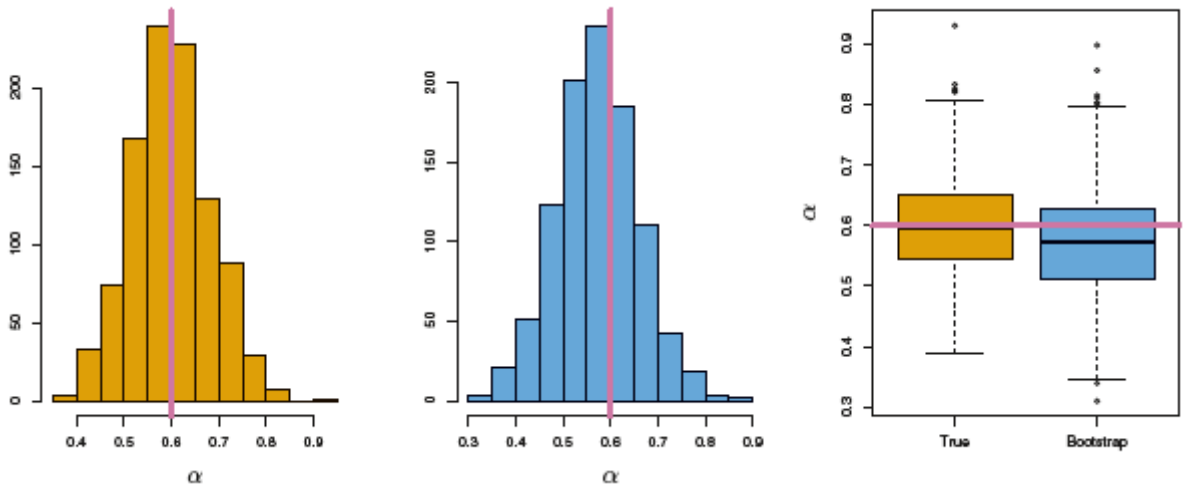


Примеры использования:

- 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, то есть из одного исходного ограниченного набора наблюдений.
- Справа приведена диаграмма размаха для оценок полученных двумя способами, которая показывает, что оценки бутстрепом 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/
3. Cross-validation: evaluating estimator performance // Sklearn: User Guide. URL: https://scikit-learn.org/stable/modules/cross_validation.html
4. Model selection done right: A gentle introduction to nested cross-validation. URL: https://ploomber.io/blog/nested-cv/