Машинное обучение в экономике.
Домашнее задание.
2024-2025

Дедлайн: домашнее задание загружается в **SmartLMS** в формате **pdf** и **обязательно дублируется** вместе с файлом, содержащим **код**, на почту:

studypotanin@gmail.com

до конца дня **16.03.2025** включительно (по московскому времени). Тема письма должна иметь следующий формат: "МО Фамилия Имя Группа", например, "МО Потанин Богдан 123".
Оформление: первый лист задания должен быть титульным и содержать лишь информацию об именах, фамилиях и группах студентов, а также фамилиях своих семинаристов. Если файл содержит фотографии, то они должны быть разборчивыми и повернуты правильной стороной. Все графики и таблицы в тексте должны быть пронумерованы и подписаны, а также на них должна быть дана ссылка в тексте.
В тексте работы **обязательно** должен быть продублирован текст каждого задания и сразу следом за ним дан развернутый комментарий по поводу проделанной работы, например:

Задание 3.1

Отберите признаки, которые могут быть полезны при прогнозировании переменной воздействия и кратко обоснуйте выбор каждой из них. Не включайте в число этих признаков целевую переменную.

Ответ

Развернутый ответ, объясняющий причины выбора соответствующих признаков.
Санкции: домашние задания, не удовлетворяющие требованиям к оформлению, выполненные несамостоятельно или сданные позже срока получают 0 баллов.
Самостоятельность: задания выполняются в группах до 2-х человек включительно. С целью проверки самостоятельности выполнения домашнего задания любой студент может быть вызван на устное собеседование, по результатам которого оценка может быть либо сохранена, либо снижена вплоть до обнуления. На собеседовании студент должен продемонстрировать уверенное владение материалами, касающимися работы всех участников группы.
Распределение оценки: при выполнении работы в группах все участники получают одинаковую оценку. **Рекомендация:** для написания текста в группах рекомендуется использовать **overleaf**.
Важно: принимаются лишь аккуратно оформленные работы в формате pdf. Например, работы, сданные в формате doc или в форме блокнота (ноутбука) python, оцениваются в 0 баллов. Также, работа **не** должна содержать код. Содержащие код задания оцениваются в 0 баллов.
Цитирование: все цитируемые в работе исследования должны входить в международную систему цитирования Scopus или Web of Science. Ссылки приводятся в алфавитном порядке (сперва идут русскоязычные источники) и должны быть оформлены в едином формате, например, с использованием одного из распространенных стилей (APA, MLA, Chicago и т.д.).

Стилистика: при написании текста следует придерживаться академической стилистики.
Оформление кода: код должен быть аккуратно оформлен в рамках одного файла, воспроизводиться без ошибок, а также содержать указания на то, к какому из заданий он относится. Допускается использование R, python, Julia и matlab, однако, все задания должны быть выполнены на одном из этих языков программирования.
Оценивание: при выставлении оценки за работу учитывается как качество выполнения отдельных пунктов, так и то, насколько она в целом интересна, оригинальна и грамотно выполнена как исследование.

Задание

1 Обоснование темы

1. Придумайте **непрерывную** зависимую (целевую) переменную (например, заработная плата или прибыль) и **бинарную** переменную воздействия (например, образование или факт занятий спортом).

2. Опишите, для чего может быть полезно изучение влияния переменной воздействия на зависимую переменную. В частности, укажите, как эта информация может быть использована бизнесом или государственными органами.

3. Обоснуйте наличие причинно-следственной связи между зависимой переменной и переменной воздействия. Приведите не менее 2-х источников из научной литературы, подкрепляющих ваши предположения.

4. Кратко опишите результаты предшествовавших исследований по схожей тематике и критически оцените методологию этих работ с точки зрения гибкости (жесткости предпосылок) использовавшихся методов эконометрического анализа. Объясните, в чем заключается преимущество и недостатки применяемых вами методов в сравнении с теми, что ранее использовались в литературе.

5. Придумайте хотя бы 3 контрольные переменные, по крайней мере одна из которых должна быть бинарной и хотя бы одна – непрерывной. Кратко обоснуйте выбор каждой из них.

6. Придумайте **бинарную** инструментальную переменную и обоснуйте, почему она удовлетворяет необходимым условиям.

7. В случае необходимости приведите дополнительные содержательные комментарии о целях, задачах, методологии и вкладе вашего исследования.

2 Генерация и предварительная обработка данных

1. Опишите математически предполагаемый вами процесс генерации данных.
   Примечание: оценивается в том числе оригинальность предложенного вами процесса, поэтому, в частности, не рекомендуется использовать совсем простые линейные модели.

2. Обоснуйте предполагаемые направления связей зависимой переменной и переменной воздействия с контрольными переменными.

3. Симулируйте данные в соответствии с предполагаемым вами процессом и приведите корреляционную матрицу, а также таблицу со следующими описательными статистиками:

   - Для непрерывных переменных: выборочное среднее, выборочное стандартное отклонение, медиана, минимум и максимум.
   - Для бинарных переменных: доля и количество единиц.

Указания:

- Необходимо сгенерировать не менее 1000 наблюдений.
- Доля единиц не должна быть меньше 0.1 или больше 0.9 ни для одной из бинарных переменных.

4. Разделите выборку на обучающую и тестовую. Тестовая выборка должна включать от 20% до 30% наблюдений.

5. В случае необходимости проведите дополнительный анализ и приведине дополнительные комментарии о процессе генерации данных, описательных статистиках и т.д.

3 Классификация

В каждом из заданий, если не сказано иного, необходимо использовать **хотя бы** 3 (на ваш выбор) из следующих методов: наивный Байесовский классификатор, метод ближайших соседей, случайный лес, градиентный бустинг и логистическая регрессия.

1. Отберите признаки, которые могут быть полезны при прогнозировании переменной воздействия и обоснуйте выбор каждой из них. Не включайте в число этих признаков целевую переменную.

2. Выберите произвольные значения гиперпараметров, а затем оцените и сравните (между методами) точность прогнозов:

   - на обучающей выборке.
   - на тестовой выборке.
   - с помощью кросс-валидации (используйте только обучающую выборку).

   Проинтерпретируйте полученные результаты.

3. Для каждого метода с помощью кросс-валидации на обучающей выборке подберите оптимальные значения гиперпараметров (тюнинг). В качестве критерия качества используйте точность ACC. Результат представьте в форме таблицы, в которой для каждого метода должны быть указаны:

   - изначальные и подобранные значения гиперпараметров.
   - кросс-валидационная точность на обучающей выборке с исходными и подобранными значениями гиперпараметров.
   - точность на тестовой выборке с исходными и подобранными значениями гиперпараметров.

   Проинтерпретируйте полученные результаты и далее используйте методы с подобранными значениями гиперпараметров.
   Повышенная сложность: подберите на обучающей выборке оптимальные значения гиперпараметров случайного леса ориентируясь на значение OOB

(out-of-bag) ошибки. Сопоставьте гиперпараметры и точность на тестовой выборке для случайного леса в зависимости от того, используется кросс-валидация или OOB ошибка. Объясните преимущества и недостатки OOB ошибки по сравнению с кросс-валидацей.

4. Повторите предыдущий пункт, используя любой альтернативный критерий качества модели. Обоснуйте возможные преимущества и недостатки этого альтернативного критерия.
   Повышенная сложность: дополнительно самостоятельно запрограммируйте не представленный в стандартных библиотеках критерий качества и используйте его для тюнинга гиперпараметров. Сравните результат стандартного и вашего критериев, а также опишите его преимущества и недостатки: в каких случаях его разумно применять, а в каких случаях он может оказаться не очень полезен.

5. Постройте ROC-кривую для ваших моделей и сравните их по AUC на тестовой выборке.
   Повышенная сложность: дополнительно выполните это задание для Байесовской сети и сравните ее ROC-кривую и AUC с теми, что были получены для иных методов.

6. Постройте матрицу ошибок и предположите цены различных видов прогнозов. Исходя из критерия максимизации прибыли на обучающей выборке подберите оптимальный порог прогнозирования для каждого из методов и сравните прибыли на тестовой выборке при соответствующих порогах. Результат представьте в форме таблицы, в которой должны быть указаны как AUC, так и прибыли (на тестовой выборке). Проинтерпретируйте полученный результат.
   Повышенная сложность: предложите, содержательно обоснуйте и примените собственную, отличную от линейной функцию прибыли от прогнозов.

7. Опишите предполагаемые связи между переменными в форме ориентированного ациклического графа (DAG). Обучите структуру Байесовской сети на обучающей выборке и сравните точность прогнозов вашего и обученного DAG на тестовой выборке.

8. На основании проделанного анализа выберите **лучший** и **худший** из обученных классификаторов. Обоснуйте сделанный выбор.
   Примечание: необхоидмо самостоятельно сформулировать разумный критерий, в соответствии с которым будут определяться лучшая и худшая модели.

9. **Повышенная сложность**: включите в анализ дополнительный метод классификации, не рассматривавшийся в курсе и не представленный в библиотеке scikit-learn. Опишите данный метод (принцип работы, преимущества и недостатки) и осуществите тюнинг гиперпараметров. Сопоставьте его точность на тестовой выборке с точностью лучшего из обученных вами ранее методов. Добавление обычной регуляризации к классическим методам не считается отдельным методом.

10. В случае необходимости проведите дополнительный анализ.

4 Регрессия

В каждом из заданий, если не сказано иного, необходимо использовать **хотя бы** 3 (на ваш выбор) из следующих методов: случайный лес, метод наименьших квадратов, метод ближайших соседей и градиентный бустинг.

1. Отберите признаки, которые могут быть полезны при прогнозировании целевой (зависимой) переменной. Не включайте в число этих признаков переменную воздействия. Содержательно обоснуйте выбор признаков.

2. Выберите произвольные значения гиперпараметров, а затем оцените и сравните (между методами) точность прогнозов с помощью RMSE и MAPE:

   - на обучающей выборке.
   - на тестовой выборке.
   - с помощью кросс-валидации (используйте только обучающую выборку).

   Проинтерпретируйте полученные результаты.

3. Для каждого метода с помощью кросс-валидации на обучающей выборке подберите оптимальные значения гиперпараметров (тюнинг). В качестве критерия качества используйте RMSE. Результат представьте в форме таблицы, в которой для каждого метода должны быть указаны:

   - изначальные и подобранные значения гиперпараметров.
   - кросс-валидационное значение RMSE на обучающей выборке с исходными и подобранными значениями гиперпараметров.
   - значение RMSE на тестовой выборке с исходными и подобранными значениями гиперпараметров.

   Проинтерпретируйте полученные результаты.
   Повышенная сложность: подберите на обучающей выборке оптимальные значения гиперпараметров градиентного бустинга ориентируясь на значение OOB (out-of-bag) ошибки. Сопоставьте гиперпараметры и точность на тестовой выборке для градиентного бустинга в зависимости от того, используется кросс-валидация или OOB ошибка.

4. На основании проделанного анализа выберите **лучший** и **худший** из обученных классификаторов. Обоснуйте сделанный выбор.

5. **Повышенная сложность**: включите в анализ дополнительный метод регрессии, не рассматривавшийся в курсе и не представленный в библиотеке scikit-learn. Опишите данный метод (принцип работы, преимущества и недостатки) и осуществите тюнинг гиперпараметров. Сопоставьте его точность на тестовой выборке с точностью лучшего из обученных вами ранее методов.

6. В случае необходимости проведите дополнительный анализ.

5 Эффекты воздействия

При выполнении данного задания необходимо объединить обучающую и тестовую выборки в одну.

1. Математически запишите и содержательно проинтерпретируйте потенциальные исходы целевой переменной. Объясните, как они связаны с наблюдаемыми значениями целевой переменной.

2. Используя симулированные вами, но недоступные в реальных данных потенциальные исходы (гипотетические значения), получите оценки среднего эффекта воздействия, условных средних эффектов воздействия и локального среднего эффекта воздействия. Для ATE и LATE результаты представьте в форме таблицы, а для CATE постройте гистограмму или ядерную оценку функции плотности. Проинтерпретируйте полученные значения.
   Примечание: для получения очень точных оценок эффектов воздействия с помощью потенциальных исходов (гипотетических переменных) можно сперва симулировать очень большого число наблюдений, например, несколько миллионов. Затем, для ускорения вычислений, для оценивания эффектов воздействия с помощью наблюдемых значений можно использовать часть выборки, например, десять тысяч наблюдений.

3. Оцените средний эффект воздействия как разницу в средних по выборкам тех, кто получил и не получил воздействие. Опишите недостатки соответствующего подхода с учетом специфики рассматриваемой вами экономической проблемы.
   Примечание: в этом пункте и далее, если не сказано иное, используются лишь наблюдаемые значения целевой переменной.

4. Используя оценки, полученные лучшими из обученных ранее классификационных и регрессионных моделей, оцените средний эффект воздействия с помощью:

   - метода наименьших квадратов.
   - условных математических ожиданий.
   - взвешивания на обратные вероятности (в случае возникновения ошибок убедитесь в отсутствии оценок вероятностей, равных 0 или 1 и при необходимости измените метод оценивания).
   - метода, обладающего двойной устойчивостью.
   - двойного машинного обучения.

   Сравните результаты и назовите ключевую предпосылку этих методов. Содержательно обсудите причины, по которым она может соблюдаться или нарушаться в вашем случае. Приведите содержательную экономическую интерпретацию оценки среднего эффекта воздействия.
   Повышенная сложность: включите дополнительный метод, не рассматривавшийся в курсе, и опишите его принцип работы, а также преимущества и недостатки по сравнению с другими методами.

5. Оцените локальный условный эффект воздействия с помощью:

- двойного машинного обучения без инструментальной переменной.
- двойного машинного обучения с инструментальной переменной.

Сопоставьте результаты и объясните, в чем в вашем случае будет заключаться различие между средним эффектом воздействия и локальным средним эффектом воздействия. Приведите содержательную экономическую интерпретацию оценки локального среднего эффекта воздействия.
Повышенная сложность: воспользуйтесь также параметрической моделью, например, с помощью пакета switchSelection. Обсудите преимущества и недостатки такого подхода по сравненинию с двойным машинным обучением. Обычный метод инструментальных переменных параметрическим подходом не считается.

6. Оцените условные средние эффекты воздействия с помощью:

   - метода наименьших квадратов.
   - S-learner.
   - T-learner.
   - метода трансформации классов.
   - X-learner.

   Сравните результаты и обсудите, насколько в вашем случае мотивированы применение метода X-learner. Опишите, как можно было бы использовать полученные вами оценки в бизнесе или при реализации государственных программ.
   Повышенная сложность: включите дополнительный метод, не рассматривавшийся в курсе и опишите его принцип работы, а также преимущества и недостатки по сравнению с другими методами.

7. Выберите лучшую модель оценивания условных средних эффектов воздействия, используя:

   - истинные значения условных средних эффектов воздействия.
   - прогнозную точность моделей.
   - псевдоисходы.

   Проинтерпретируйте различия в результатах различных подходов.

8. Оцените средние эффекты воздействия и локальные средние эффекты воздействия используя **худшие** из обученных классификационных и регрессионных моделей. Сопоставьте результаты с теми, что были получены с помощью **лучших** моделей. Сдеайте вывод об устойчивости результатов к качеству используемых методов машинного обучения.

9. Резюмируйте ключевые выводы проведенного в данном разделе анализа.

10. В случае необходимости проведите дополнительный анализ.