

Цель работы

Проверить освоение студентами методы оценки и диагностики моделей машинного обучения.

Методические указания

Контрольная работа выполняется в аудитории. В начале занятия преподаватель распределяет по студентам датасеты (задания). Распределение должно проходить случайным образом.

После получения задания студенты выполняют работу на лабораторных компьютерах. Работа должна быть выполнена студентом самостоятельно. Работа выполняется в виде ноутбука формата и названия файла: Фамилия_группа.ipynb (Иванов_ПМ22-1.ipynb)

После завершения выполнения работы студент должен сдать работу преподавателю. Работы сдаются в электронном виде с корпоративной почты студента на корпоративную почту преподавателя. Работа должна быть сдана не позже окончания аудиторного занятия.

Контрольная работа выполняется на лабораторный компьютеров.

Во время выполнения работы можно пользоваться всем (документацией, прошлыми работами, Интернетом), кроме помощи других студентов, работа выполняется самостоятельно.

Критерии оценки

1. Код должен выполняться без ошибок. Программные ошибки - сразу 0 баллов.
2. Код должен содержать пояснения в виде ячеек с текстом. В пояснениях нужно писать значимые выводы по ходу выполнения работы.
3. Все графики должны быть правильно подписаны и понятны.
4. Текстовый вывод должен быть нагляден и отформатирован. Численные значения - округлены до разумного количества значащих цифр.

Содержание работы

1. (1 балл) Загрузите предложенный вам датасет с помощью функции ``sklearn.datasets.fetch_openml``. Выведите текстовое описание загруженного датасета. Определите какая бизнес-задача ставится. Переведите бизнес-задачу в задачу машинного обучения.
2. (1 балл) Проведите предварительный анализ датасета. Выведите основную статистическую информацию о данных. Проверьте дисбаланс целевого признака.

3. (1 балл) Убедитесь, что данные пригодны для моделирования. В данных не должно быть пропущенных значений, все признаки должны быть численными. Если эти условия нарушаются, исправьте это.
4. (1 балл) Обучите любую линейную модель, которая решает поставленную задачу. Выведите с помощью метода ``score()`` результат метрики оценки модели. Опишите какая метрика подсчитывается в методе ``score()``.
5. (1 балл) К поставленной задаче машинного обучения подберите еще 3 метрики. С помощью их оцените обученную линейную модель.
6. (1 балл) Для задач классификации: Постройте для этой модели кривые ROC. Вычислите для этой модели метрики площади под кривыми ROC-AUC. Найдите оптимальные значения порога модели.
6. (1 балл) Для задач регрессии: Постройте совместное распределение теоретических и эмпирических значений целевой переменной. Сделайте вывод о наличии либо отсутствии систематической компоненты в отклонениях, их общей величине.
7. (1 балл) Для задач классификации: Постройте для этой модели кривые PR. Вычислите для этой модели метрики площади под кривыми PR-AUC. Найдите оптимальные значения порога модели.
7. (1 балл) Для задач регрессии: Постройте ту же модель, но с предварительной нормализацией значений признаков. Сделайте вывод о влиянии нормализации на качество работы модели.
8. (1 балл) Обоснуйте выбор типа разбиения данных при кросс-валидации для вашей задачи машинного обучения. Обучите любую другую модель используя кросс-валидацию данных.
9. (2 балл) Проведите диагностику любой модели. Постройте кривые обучения. Сделайте вывод о переобученности или недообученности модели.
10. (2 балл) Исследуйте как любой из гиперпараметров модели влияет на эффективность модели.

Варианты заданий

1. students_scores
2. wholesale-customers
3. pc1_req
4. vertebra-column
5. prnn_crabs
6. cloud
7. fertility

8. ricci_vs_destefano
9. machine_cpu
- 10.lupus
- 11.energy-efficiency
- 12.LED-display-domain-7digit
- 13.diggle_table_a2
- 14.ecoli
- 15.breast-tissue
- 16.prnn_fglass
- 17.confidence
- 18.glass
- 19.user-knowledge
- 20.olindda_outliers