Министерство науки и высшего образования Российской Федерации
НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ТОМСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ (НИ ТГУ)

Институт прикладной математики и компьютерных наук

УТВЕРЖДЕНО:
Директор
А. В. Замятин

Оценочные материалы по дисциплине

Прикладные аспекты машинного обучения

по направлению подготовки

09.03.03 Прикладная информатика

Направленность (профиль) подготовки:
Искусственный интеллект и большие данные

Форма обучения
Очная

Квалификация
Бакалавр

Год приема
2024

СОГЛАСОВАНО:
Руководитель ОП
С.П.Сущенко

Председатель УМК
С.П.Сущенко

Томск – 2024

1. Компетенции и индикаторы их достижения, проверяемые данными оценочными материалами

Целью освоения дисциплины является формирование следующих компетенций:

ПК-6 Способен разрабатывать и применять методы машинного обучения для решения задач.

Результатами освоения дисциплины являются следующие индикаторы достижения компетенций:

ИПК-6.1 Проводит анализ требований и определяет необходимые классы задач машинного обучения

ИПК-6.2 Принимает участие в оценке и выборе используемых методов машинного обучения

2. Оценочные материалы текущего контроля и критерии оценивания

Элементы текущего контроля:

– тесты;

– контрольная работа;

Тест (ИПК-6.1, ИПК-6.2)

1. Чем характеризуется задача классификации?
    а) Результирующий сигнал является непрерывным
    б) Присутствует система вознаграждений
    в) Используется информация о метках
    г) Наличием обучающей выборки
2. Какая мера не используется при построении дерева решений-классификатора?
    а) энтропия
    б) R2
    в) ошибка классификации
    г) неоднородности Джини
3. Какой из классификаторов считается наихудшим для практического использования?
    а) с AUC=0.75
    б) с AUC=0.29
    в) с AUC=0.44
    г) с AUC=0.11
4. Что не относится к ленивому обучению?
    а) Низкие затраты на обучение
    б) Возможность обработки линейно-неразделимых классов
    в) Обязательное требование стандартизации данных
    г) Компактное представление модели
5. Для какого из инструментов обучения этап приведения данных к одной шкале можно опустить?
    а) Деревья решения
    б) К-ближайших соседей
    в) Логистическая регрессия
    г) Метод опорных векторов
6. Какое действие в процессе обучения приводит в общем случае к снижению вероятности получения случайных зависимостей, отсутствующих в генеральной совокупности?
    а) Увеличение числа уровней для деревьев решений
    б) Уменьшение параметра C для SVM
    в) Уменьшение числа деревьев в случайном лесе
    г) Увеличение параметра регуляризации для логистической регрессии

7. Какой из методов машинного обучения использует мажоритарное голосование?
    а) К-ближайших соседей
    б) Персептрон
    в) Метод опорных векторов
    г) Логистическая регрессия

Ключи: 1 в), 2 б), 3 в), 4 г), 5 а), 6 б), 7 а).

Критерии оценивания: тест считается пройденным, если обучающий ответил правильно как минимум на половину вопросов.

Контрольная работа (ИПК-6.1, ИПК-6.2)

Контрольная работа состоит из 2 теоретических вопросов и 3 задач.

Перечень теоретических вопросов:

1. Для фиксированного числа наблюдений в наборе данных введение большего количества переменных обычно генерирует модель, которая лучше подходит для данных. В чем может заключаться недостаток такой стратегии подгонки модели?

2. Для фиксированного числа наблюдений в наборе данных введение большего количества переменных обычно генерирует модель, которая лучше подходит для данных. В чем может заключаться недостаток

3. Опишите, как в модели логистической регрессии применяется преобразование к переменной отклика для получения распределения вероятностей. Почему это считается более информативным представлением отклика?

4. Запишите формулы для дивергенции Кульбака-Лейблера между двумя дискретными функциями плотности вероятности P и Q.

5. Существует несколько мер, с помощью которых можно определить, как оптимально разделить атрибуты в дереве решений. Перечислите три наиболее часто используемые меры и напишите их формулы.

6. Как определяется число кластеров в алгоритме К-средних?

7. Как проверить стационарность временной последовательности?

8. Объясните преимущества и недостатки методов бустинга

9. Какие метрики используются при обучении регрессионных моделей?

10. Как работает алгоритм оценки важности на основе перестановок?

11. Как вычисляются значения Шепли для задачи, связанной с табличными наборами данных?

12. Что такое непараметрический алгоритм?

13. В чём разница между ошибками первого и второго рода?

14. Что такое опорные вектора в SVM?

15. Методы поиска аномалий

16. Можно ли использовать PCA для уменьшения размерности нелинейного набора данных со многими переменными?

17. Есть ли какие-либо вероятностные результаты от SVM?

18. В каких многочисленных случаях модели машинного обучения вызывают недообучение?

19. В каких случаях модели машинного обучения могут оказаться переобученными?

20. Можно ли использовать логистическую регрессию для более чем двух классов?

21. Каково определение принципа Парето?

22. Что именно означает ядерный трюк?

23. Улучшит ли переклассификация категориальных переменных в непрерывные переменные прогностическую модель?

24. Какие стратегии выбора признаков доступны для выбора подходящих переменных для создания эффективных моделей прогнозирования?

25. Что делать, если набор данных содержит переменные с более чем 30% пропущенных значений?

26. Что отличает задачи анализа временных рядов от других задач регрессии?

27. Что такое перекрестная проверка и как она работает?

28. В чем разница между собственными векторами и собственными значениями?

29. Перечислите критерии переобучения и недообучения.

30. Какие стратегии используются для формирования выборок? В чем главное преимущество каждой стратегии?

31. Что означает наличие высоких и низких p-значений?

32. Какие характеристики модели показывает ROC-кривая?

33. Укажите разницу между перекрестной проверкой K-fold и стратифицированной перекрестной проверкой.

34. Если у нас есть ансамбль из K обученных моделей и ошибки некоррелированы, то путем их усреднения средняя ошибка модели уменьшается в K раз или нет?

35. Верно ли высказывание: Точность обобщения ансамбля увеличивается с количеством хорошо обученных моделей, из которых он состоит?

36. Верно ли высказывание: Обучение ансамбля единой монолитной архитектуры приводит к снижению разнообразия моделей и, возможно, снижению точности прогнозирования модели?

37. Что такое граница принятия решения?

38. Истинно ли высказывание: “Идеальный ансамбль состоит из максимально корректных классификаторов, различающихся между собой”?

39. При бэггинге мы повторно выбираем обучающий набор с повторами, и поэтому это может привести к тому, что некоторые экземпляры будут представлены много раз, а другие — не будут представлены вообще. Истинно ли это высказывание?

40. Как выбрать k в алгоритме KNN?

Примеры задач:

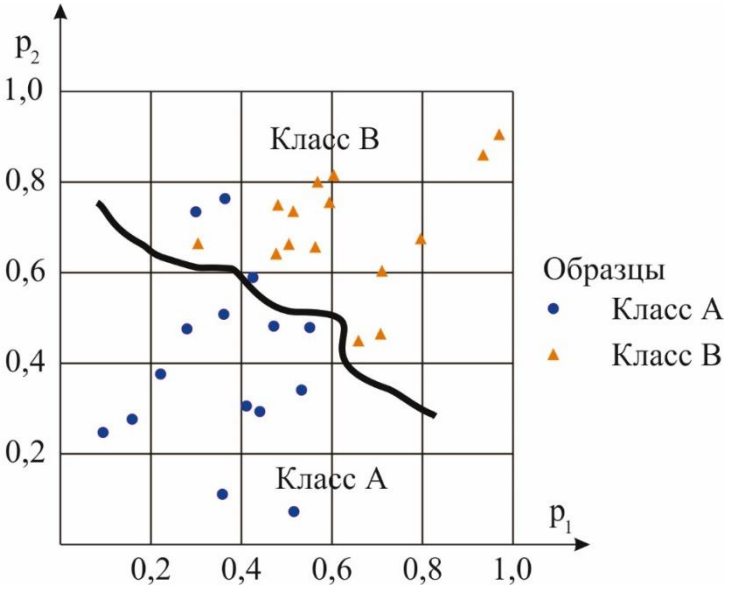
Задача 1 (ИПК-6.1, ИПК-6.2)

Для класса B найти значение специфичности

		Предсказанная метка		
		A	B	C
Истинная метка	A	8	2	1
	B	1	7	1
	C	2	0	9

Задача 2 (ИПК-6.1, ИПК-6.2)

Записать значения матрицы неточностей для классификатора, приведенного на рисунке



Задача 3 (ИПК-6.1, ИПК-6.2)

По приведенным ниже данным рассчитать индекс Данна

$$D = min_{i,j\in\{1\dots c\},i\neq j}\left\{\frac{d(c_i,c_j)}{\max_{k\in\{1\cdots c\}} diam(c_k)}\right\} \qquad (10)$$

d — межкластерное расстояние: $d(c_i,c_j) = min_{x\in c_i, y\in c_j}\|x-y\|$

$diam(c_i)$ — диаметр кластера: $diam(c_i) = \max_{x,y\in c_i}\|x-y\|$.

Данные кластеризации: Кластер 1: точки A, B, C, кластер 2: точки D, E, F. Расстояния между точками: AB=2, AC=3, AD=5, AE=6, AF=5, BC=2, BD=6, BE=9, BF=4, CD=7, CE=8, CF=7, DE=2, DF=1, EF=4

Критерии оценивания:

Результаты контрольной работы определяются оценками «отлично», «хорошо», «удовлетворительно», «неудовлетворительно».

Оценка «отлично» выставляется, если даны правильные ответы на все теоретические вопросы и все задачи решены без ошибок.

Оценка «хорошо» выставляется, если корректные ответы даны на большую часть вопросов, но были отмечены неуверенность в ответе и информация представлена фрагментарно, также задачи были решены правильно с небольшими замечаниями.

Оценка «удовлетворительно» выставляется, если даны правильные ответы на половину теоретические вопросы и одна из задач решена с негрубой ошибкой.

Оценка «неудовлетворительно» выставляется, если корректные ответов меньше половины и задачи были решены неправильно.

3. Оценочные материалы итогового контроля (промежуточной аттестации) и критерии оценивания

Экзаменационный билет состоит из трех частей.

Первая часть представляет собой тест из 5 вопросов, проверяющих ИПК-6.1, ИПК-6.2. Ответы на вопросы первой части даются путем выбора из списка предложенных.

Вторая часть содержит один вопрос, проверяющий ИПК-6.1, ИПК-6.2. Ответ на вопрос второй части дается в развернутой форме.

Третья часть содержит 2 вопроса, проверяющих ИПК-6.1, ИПК-6.2 и оформленные в виде практических задач. Ответы на вопросы третьей части предполагают решение задач и краткую интерпретацию полученных результатов.

Перечень теоретических вопросов:

1. Для фиксированного числа наблюдений в наборе данных введение большего количества переменных обычно генерирует модель, которая лучше подходит для данных. В чем может заключаться недостаток такой стратегии подгонки модели?
2. Для фиксированного числа наблюдений в наборе данных введение большего количества переменных обычно генерирует модель, которая лучше подходит для данных. В чем может заключаться недостаток
3. Опишите, как в модели логистической регрессии применяется преобразование к переменной отклика для получения распределения вероятностей. Почему это считается более информативным представлением отклика?
4. Запишите формулы для дивергенции Кульбака-Лейблера между двумя дискретными функциями плотности вероятности P и Q.
5. Существует несколько мер, с помощью которых можно определить, как оптимально разделить атрибуты в дереве решений. Перечислите три наиболее часто используемые меры и напишите их формулы.
6. Как определяется число кластеров в алгоритме К-средних?
7. Как проверить стационарность временной последовательности?
8. Объясните преимущества и недостатки методов бустинга
9. Какие метрики используются при обучении регрессионных моделей?
10. Как работает алгоритм оценки важности на основе перестановок?
11. Как вычисляются значения Шепли для задачи, связанной с табличными наборами данных?
12. Что такое непараметрический алгоритм?
13. В чём разница между ошибками первого и второго рода?
14. Что такое опорные вектора в SVM?
15. Методы поиска аномалий
16. Можно ли использовать PCA для уменьшения размерности нелинейного набора данных со многими переменными?
17. Есть ли какие-либо вероятностные результаты от SVM?
18. В каких многочисленных случаях модели машинного обучения вызывают недообучение?
19. В каких случаях модели машинного обучения могут оказаться переобученными?
20. Можно ли использовать логистическую регрессию для более чем двух классов?
21. Каково определение принципа Парето?
22. Что именно означает ядерный трюк?
23. Улучшит ли переклассификация категориальных переменных в непрерывные переменные прогностическую модель?
24. Какие стратегии выбора признаков доступны для выбора подходящих переменных для создания эффективных моделей прогнозирования?
25. Что делать, если набор данных содержит переменные с более чем 30% пропущенных значений?
26. Что отличает задачи анализа временных рядов от других задач регрессии?
27. Что такое перекрестная проверка и как она работает?
28. В чем разница между собственными векторами и собственными значениями?
29. Перечислите критерии переобучения и недообучения.
30. Какие стратегии используются для формирования выборок? В чем главное преимущество каждой стратегии?

31. Что означает наличие высоких и низких p-значений?

32. Какие характеристики модели показывает ROC-кривая?

33. Укажите разницу между перекрестной проверкой K-fold и стратифицированной перекрестной проверкой.

34. Если у нас есть ансамбль из K обученных моделей и ошибки некоррелированы, то путем их усреднения средняя ошибка модели уменьшается в K раз или нет?

35. Верно ли высказывание: Точность обобщения ансамбля увеличивается с количеством хорошо обученных моделей, из которых он состоит?

36. Верно ли высказывание: Обучение ансамбля единой монолитной архитектуры приводит к снижению разнообразия моделей и, возможно, снижению точности прогнозирования модели?

37. Что такое граница принятия решения?

38. Истинно ли высказывание: “Идеальный ансамбль состоит из максимально корректных классификаторов, различающихся между собой”?

39. При бэггинге мы повторно выбираем обучающий набор с повторами, и поэтому это может привести к тому, что некоторые экземпляры будут представлены много раз, а другие — не будут представлены вообще. Истинно ли это высказывание?

40. Как выбрать k в алгоритме KNN?

Примеры задач:

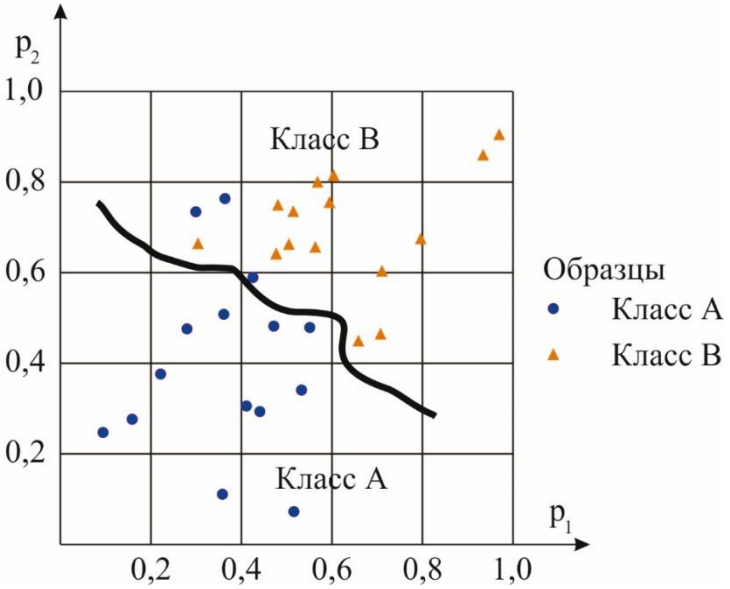
Задача 1 (ИПК-6.1, ИПК-6.2)

Для класса B найти значение специфичности

		Предсказанная метка		
		A	B	C
Истинная метка	A	8	2	1
	B	1	7	1
	C	2	0	9

Задача 2 (ИПК-6.1, ИПК-6.2)

Записать значения матрицы неточностей для классификатора, приведенного на рисунке



Задача 3 (ИПК-6.1, ИПК-6.2)

По приведенным ниже данным рассчитать индекс Данна

$$D = min_{i,j\in\{1\dots c\},i\neq j}\left\{\frac{d(c_i,c_j)}{\max_{k\in\{1\dots c\}} diam(c_k)}\right\} \qquad (10)$$

d — межкластерное расстояние: $d(c_i,c_j) = min_{x\in c_i, y\in c_j}\|x-y\|$

$diam(c_i)$ — диаметр кластера: $diam(c_i) = \max_{x,y\in c_i}\|x-y\|$.

Данные кластеризации: Кластер 1: точки A, B, C, кластер 2: точки D, E, F. Расстояния между точками: AB=2, AC=3, AD=5, AE=6, AF=5, BC=2, BD=6, BE=9, BF=4, CD=7, CE=8, CF=7, DE=2, DF=1, EF=4

Критерии оценивания:

Результаты экзамена определяются оценками «отлично», «хорошо», «удовлетворительно», «неудовлетворительно».

Оценка «отлично» выставляется, если даны правильные ответы на все теоретические вопросы и все задачи решены без ошибок.

Оценка «хорошо» выставляется, если корректные ответы даны на большую часть вопросов, но были отмечены неуверенность в ответе и информация представлена фрагментарно, также задачи были решены правильно с небольшими замечаниями.

Оценка «удовлетворительно» выставляется, если даны правильные ответы на половину теоретические вопросы и одна из задач решена с негрубой ошибкой.

Оценка «неудовлетворительно» выставляется, если корректные ответов меньше половины и задачи были решены неправильно.

4. Оценочные материалы для проверки остаточных знаний (сформированности компетенций)

Тест

1. Какой узел дерева является наиболее однородным?
   а) С энтропией =0.4
   б) С энтропией =0.1
   в) С энтропией =0.6
   г) С энтропией =0.8
2. Какой из методов использует оптимизационную задачу максимизации расстояния между гиперплоскостью и самыми близкими к ней тренировочными образцами?

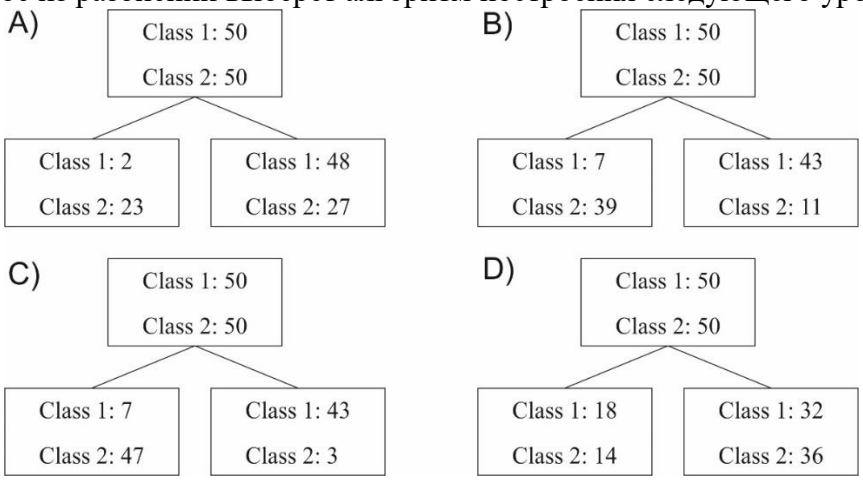
а) Логистическая регрессия
б) Случайный лес
в) SVM
г) К-ближайших соседей

3. Какой из классификаторов считается наилучшим для дальнейшего практического использования?
   а) с AUC=0.07
   б) с AUC=0.37
   в) с AUC=0.64
   г) с AUC=0.76
4. Что характеризует качественное обучение?
   а) Высокое смещение и высокая дисперсия
   б) Высокое смещение и низкая дисперсия
   в) Низкое смещение и высокая дисперсия
   г) Низкое смещение и низкая дисперсия
5. Чем характеризуется задача регрессии?
   а) Выполняется без учителя
   б) Обработка только для числовых признаков
   в) Требует предварительной кластеризации данных
   г) Результирующий сигнал является непрерывным
6. Учитывая значения оценки Силуэта, какое разбиение предпочтительнее (в скобках даны значения Силуэта для объекта каждого кластера):
   а) Кластер А (0.8, 0.7, 0.7, 0.4), Кластер В (0.9, 0.7, 0.2), Кластер С (0.9, 0.3, 0.4)
   б) Кластер А (0.8, 0.6, 0.5, 0.5), Кластер В (0.8, 0.6, 0.1), Кластер С (0.8, 0.7, 0.1)
   в) Кластер А (0.8, 0.7, 0.6, 0.5), Кластер В (0.8, 0.6, 0.6, 0.5, 0.4, 0.3)
   г) Кластер А (0.8, 0.6, 0.6, 0.6, 0.5), Кластер В (0.7, 0.5, 0.5, 0.4, 0.3)
7. В каком случае крайне рекомендуется использовать К-блочную перекрестную проверку?
   а) Несбалансированность выборки
   б) Использование модели с большим числом параметров
   в) Малый объем выборки
   г) Большое число классов в выборке
8. Каким наиболее существенным недостатком обладает алгоритм DBSCAN по сравнению с другими алгоритмами кластеризации
   а) Требование нормализации данных перед процедурой обучения
   б) Необходимость четкого задания желаемого числа кластеров
   в) Длительная процедура
   г) Невозможность получения кластера произвольной формы
9. Какое утверждение ложно?
   а) Критерий SSE оценивает компактность
   б) Критерий SSE зависит числа кластеров
   в) Критерий SSE не подходит для алгоритма плотностной кластеризации
   г) Критерий SSE учитывает отделимость
10. Какая метрика используется при решении задачи регрессии?
    а) TPR
    б) Чувствительность
    в) Коэффициент детерминации
    г) Верность
11. Каким наиболее существенным недостатком обладает алгоритм К-средних по сравнению с другими алгоритмами кластеризации?
    а) Требование нормализации данных перед процедурой обучения

б) Необходимость неоднократного перезапуска для получения оптимального решения
в) Вычислительная сложность
г) Невозможность получения кластера произвольной формы

12. Почему кластеризация относится у к задаче обучения без учителя
а) не используются метки класса в процессе обучения
б) Процедура обучения контролируется настройками пользователя
в) Не используется оценка изменения параметров модели в процессе обучения
г) Пользователь выбирает желаемое число кластеров

13. Какое из разбиений выберет алгоритм построения следующего уровня дерева?



Ключи: 1 б), 2 в), 3 а), 4 г), 5 г), 6 а), 7 в), 8 в), 9 г), 10 в), 11 г), 12 а), 13 а).

Пример задач

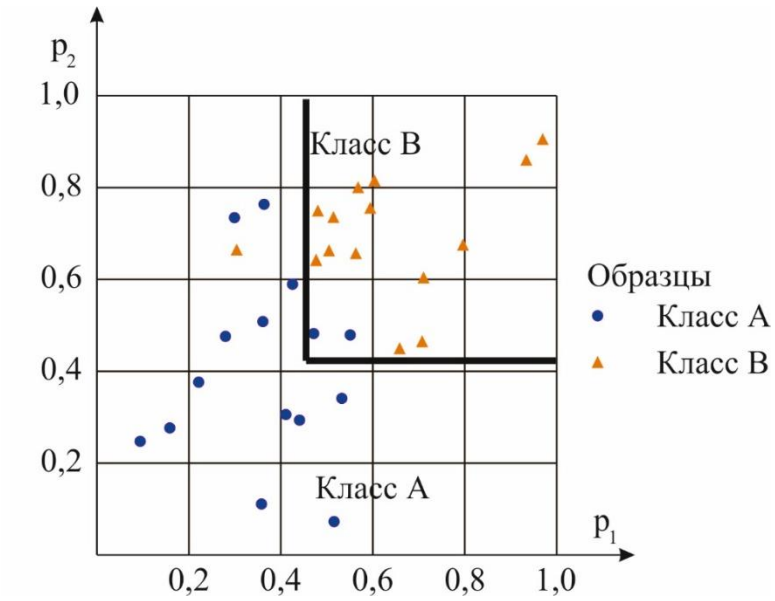
Задача 1 (ИПК-6.1, ИПК-6.2)

Для класса С найти значение TNR

		Предсказанная метка		
		A	B	C
Истинная метка	A	8	2	1
	B	1	7	1
	C	2	0	9

Задача 2 (ИПК-6.1, ИПК-6.2)

Записать значения матрицы неточностей для классификатора, приведенного на рисунке



Задача 3 (ИПК-3.3)

По приведенным ниже данным рассчитать индекс Данна

$$D = min_{i,j\in\{1\dots c\},i\neq j}\left\{\frac{d(c_i,c_j)}{\max_{k\in\{1\dots c\}} diam(c_k)}\right\} \qquad (10)$$

d — межкластерное расстояние: $d(c_i,c_j) = min_{x\in c_i, y\in c_j}\|x-y\|$

$diam(c_i)$ — диаметр кластера: $diam(c_i) = \max_{x,y\in c_i}\|x-y\|$.

Данные кластеризации: Кластер 1: точки A, B, C, кластер 2: точки D, E, F. Расстояния между точками: AB=1, AC=5, AD=6, AE=8, AF=5, BC=4, BD=3, BE=7, BF=2, CD=4, CE=8, CF=6, DE=3, DF=4, EF=3

Теоретические вопросы:

1. Для фиксированного числа наблюдений в наборе данных введение большего количества переменных обычно генерирует модель, которая лучше подходит для данных. В чем может заключаться недостаток такой стратегии подгонки модели?
2. Для фиксированного числа наблюдений в наборе данных введение большего количества переменных обычно генерирует модель, которая лучше подходит для данных. В чем может заключаться недостаток
3. Опишите, как в модели логистической регрессии применяется преобразование к переменной отклика для получения распределения вероятностей. Почему это считается более информативным представлением отклика?
4. Запишите формулы для дивергенции Кульбака-Лейблера между двумя дискретными функциями плотности вероятности P и Q.
5. Существует несколько мер, с помощью которых можно определить, как оптимально разделить атрибуты в дереве решений. Перечислите три наиболее часто используемые меры и напишите их формулы.
6. Как определяется число кластеров в алгоритме К-средних?
7. Как проверить стационарность временной последовательности?
8. Объясните преимущества и недостатки методов бустинга
9. Какие метрики используются при обучении регрессионных моделей?
10. Как работает алгоритм оценки важности на основе перестановок?
11. Как вычисляются значения Шепли для задачи, связанной с табличными наборами данных?
12. Что такое непараметрический алгоритм?
13. В чём разница между ошибками первого и второго рода?
14. Что такое опорные вектора в SVM?

15. Методы поиска аномалий

16. Можно ли использовать PCA для уменьшения размерности нелинейного набора данных со многими переменными?

17. Есть ли какие-либо вероятностные результаты от SVM?

18. В каких многочисленных случаях модели машинного обучения вызывают недообучение?

19. В каких случаях модели машинного обучения могут оказаться переобученными?

20. Можно ли использовать логистическую регрессию для более чем двух классов?

21. Каково определение принципа Парето?

22. Что именно означает ядерный трюк?

23. Улучшит ли переклассификация категориальных переменных в непрерывные переменные прогностическую модель?

24. Какие стратегии выбора признаков доступны для выбора подходящих переменных для создания эффективных моделей прогнозирования?

25. Что делать, если набор данных содержит переменные с более чем 30% пропущенных значений?

26. Что отличает задачи анализа временных рядов от других задач регрессии?

27. Что такое перекрестная проверка и как она работает?

28. В чем разница между собственными векторами и собственными значениями?

29. Перечислите критерии переобучения и недообучения.

30. Какие стратегии используются для формирования выборок? В чем главное преимущество каждой стратегии?

31. Что означает наличие высоких и низких p-значений?

32. Какие характеристики модели показывает ROC-кривая?

33. Укажите разницу между перекрестной проверкой K-fold и стратифицированной перекрестной проверкой.

34. Если у нас есть ансамбль из K обученных моделей и ошибки некоррелированы, то путем их усреднения средняя ошибка модели уменьшается в K раз или нет?

35. Верно ли высказывание: Точность обобщения ансамбля увеличивается с количеством хорошо обученных моделей, из которых он состоит?

36. Верно ли высказывание: Обучение ансамбля единой монолитной архитектуры приводит к снижению разнообразия моделей и, возможно, снижению точности прогнозирования модели?

37. Что такое граница принятия решения?

38. Истинно ли высказывание: “Идеальный ансамбль состоит из максимально корректных классификаторов, различающихся между собой”?

39. При бэггинге мы повторно выбираем обучающий набор с повторами, и поэтому это может привести к тому, что некоторые экземпляры будут представлены много раз, а другие — не будут представлены вообще. Истинно ли это высказывание?

40. Как выбрать k в алгоритме KNN?

Информация о разработчиках

Аксёнов Сергей Владимирович, к.т.н., кафедра теоретических основ информатики (ТОИ) Института прикладной математики и компьютерных наук (ИПМКН) Национальный исследовательский Томский государственный университет (НИ ТГУ), доцент каф. ТОИ