Контрольные и тестовые вопросы по ПР5

«Анализ данных методом решающих деревьев» по вариантам с ответами

Вариант 1

1. Какой критерий чаще всего используется при построении регрессионных деревьев решений?
   A) Gini Index
   B) Entropy
   C) Least Square
   D) Information Gain

Ответ: C

2. Как параметр Maximal Depth влияет на поведение модели решающего дерева?
   A) Определяет количество признаков
   B) Задает максимальное количество узлов
   C) Ограничивает количество уровней дерева, контролируя сложность модели
   D) Указывает минимальный размер выборки

Ответ: C

3. Что делает параметр Minimal Gain в дереве решений?
   A) Определяет число соседей
   B) Отсекает слабые признаки при обучении
   C) Устанавливает минимальное улучшение качества, необходимое для создания нового разбиения
   D) Устанавливает минимальное значение ошибки

Ответ: C

4. В чем заключается преимущество использования предварительной обрезки (prepruning) дерева?
   A) Сокращает количество эпох
   B) Повышает качество визуализации
   C) Предотвращает переобучение за счет ограничения избыточных

ветвлений
D) Ускоряет кросс-валидацию

Ответ: C

5. Почему RMSE считается хорошей метрикой в задачах регрессии?
   A) Показывает отношение к максимальному значению
   B) Линейно оценивает качество
   C) Наказывает большие ошибки сильнее
   D) Применим только к классификации

Ответ: C

6. Какой параметр решающего дерева указывает минимальное количество объектов в листе?
   A) Minimal Gain
   B) Maximal Depth
   C) Minimal Leaf Size
   D) Prepruning

Ответ: C

7. Как влияет использование Set Role в RapidMiner при анализе дерева решений?
   A) Позволяет задать целевую переменную для обучения
   B) Удаляет выбросы
   C) Разделяет данные на обучающую и тестовую выборки
   D) Отбирает важные признаки

Ответ: A

8. Какие методы можно использовать для предотвращения переобучения модели решающего дерева?
   Ответ: Обрезка дерева (prepruning и postpruning), ограничение глубины (max_depth), увеличение минимального размера листа (min_leaf_size), использование ансамблевых методов (например, Random Forest).

9. Почему корреляция между предсказанными и реальными значениями важна при анализе модели?
   Ответ: Корреляция отражает силу и направление линейной связи между предсказанными и фактическими значениями, что позволяет оценить согласованность модели с данными.

10. В чем смысл использования статистического анализа перед обучением решающего дерева?
    Ответ: Позволяет выявить распределение значений признаков, наличие выбросов, масштаб признаков и сделать предварительные выводы о возможных влияниях на целевую переменную.

Вариант 2

1. Какая особенность решающих деревьев делает их особенно удобными для интерпретации?
   A) Глубокая архитектура
   B) Способность обучаться без нормализации
   C) Визуализируемая структура с логическими условиями
   D) Поддержка многомерных выходов

Ответ: C

2. Какой тип переменной лучше всего подходит для предсказания с помощью регрессионного дерева?
   A) Категориальная
   B) Логическая
   C) Числовая непрерывная
   D) Символьная

Ответ: C

3. Как влияет увеличение значения Minimal Leaf Size на модель?
   A) Повышает чувствительность
   B) Увеличивает количество узлов
   C) Упрощает дерево, снижая переобучение
   D) Делает модель более сложной

Ответ: C

4. Что происходит, если в дереве решений не ограничить глубину и не применить обрезку?
   A) Увеличивается точность на тестовой выборке
   B) Модель становится слишком простой
   C) Повышается риск переобучения
   D) Модель не сможет обучиться

Ответ: C

5. Какой способ разбиения данных в RapidMiner обеспечивает репрезентативность обучающей и тестовой выборок?
   A) Holdout
   B) Sequential Sampling
   C) Shuffled Sampling
   D) Stratified Sampling

Ответ: C

6. Какой оператор в RapidMiner оценивает производительность модели в задачах регрессии?
   A) Performance (Classification)
   B) Apply Model
   C) Performance (Regression)
   D) Correlation Matrix

Ответ: C

7. Что означает показатель "Squared Correlation" в оценке модели?
   A) Абсолютное отклонение
   B) Процент объясненной дисперсии
   C) Ошибка классификации
   D) Коэффициент переобучения

Ответ: B

8. Как можно оценить важность признаков при использовании дерева решений?
   Ответ: С помощью таблицы значимости атрибутов (Attribute Weights), где каждому признаку присваивается количественная мера его вклада в модель.

9. Почему использование RMSE предпочтительнее MSE для оценки ошибки?
   Ответ: RMSE выражает ошибку в тех же единицах, что и предсказываемая переменная, что делает её более интерпретируемой, особенно при анализе отклонений.

10. Как влияет большое количество пропущенных занятий на итоговую оценку согласно анализу дерева?
    Ответ: Увеличение пропусков обычно снижает итоговую оценку, так как отсутствие на занятиях негативно влияет на успеваемость.

Вариант 3

1. Какова роль критерия Information Gain в построении дерева решений?
   A) Он оценивает расстояние между кластерами
   B) Он определяет, насколько хорошо разбиение уменьшает энтропию
   C) Он определяет глубину дерева
   D) Он используется только в регрессии

Ответ: B

2. Какое преимущество имеет использование Gini Index по сравнению с Entropy в деревьях классификации?
   A) Более высокая точность
   B) Более быстрые вычисления
   C) Улучшенная визуализация
   D) Возможность использования в регрессии

Ответ: B

3. Что произойдет, если параметр Minimal Gain установлен слишком высоким?
   A) Дерево будет переобучено
   B) Будет построено избыточное количество узлов
   C) Дерево будет слишком простым и недостаточно точным
   D) Модель не сможет классифицировать категориальные данные

Ответ: C

4. Как влияет выбор метода оценки качества модели на интерпретацию её эффективности?
   A) Только влияет на визуализацию
   B) Позволяет понять, какие ошибки делает модель
   C) Не оказывает значительного влияния
   D) Нужен только для выбора алгоритма

Ответ: B

5. В чем заключается ключевое отличие между Classification Tree и Regression Tree?
   A) Classification Tree применяется только к числовым данным
   B) Regression Tree предсказывает категориальные значения
   C) Regression Tree работает с непрерывными значениями, а Classification Tree — с категориальными
   D) Classification Tree использует RMSE как метрику

Ответ: C

6. Что делает оператор "Weight by Information Gain" в RapidMiner?
   A) Строит дерево
   B) Преобразует категории в числа
   C) Присваивает каждому признаку вес на основе снижения энтропии
   D) Стандартизирует числовые значения

Ответ: C

7. Какой показатель регрессионной модели показывает среднюю величину ошибки?
   A) R^2
   B) Squared Correlation
   C) RMSE
   D) Gain Ratio

Ответ: C

8. Чем может быть вызвано переобучение дерева при работе с учебной выборкой?
   Ответ: Слишком большая глубина дерева, слишком малый размер листа, отсутствие предварительной обрезки и использование слишком большого количества признаков.

9. Как выбор предиктора с максимальным Information Gain влияет на дерево?
   Ответ: Он приводит к наиболее эффективному разбиению данных на каждом уровне дерева, что способствует построению более точной модели с минимальной энтропией.

10. Почему важно оценивать структуру дерева после обучения?
    Ответ: Чтобы убедиться в интерпретируемости модели, обнаружить потенциальное переобучение, избыточные ветви и оценить логичность разбиений.

Вариант 4

1. Что произойдет, если в дереве решений не ограничить параметр Minimal Leaf Size?
   A) Листья будут игнорироваться
   B) В дереве не будет ветвлений

C) Возможно создание листьев, содержащих слишком мало данных, что увеличивает переобучение
D) Ошибка компиляции

Ответ: C

2. Какова основная причина использования деревьев решений в задачах анализа данных?
   A) Высокая точность на шумных данных
   B) Простота интерпретации и визуализации логики модели
   C) Возможность работы только с бинарными признаками
   D) Требование нормализации данных

Ответ: B

3. Как влияет установка значения Minimal Gain на чувствительность дерева к данным?
   A) Никак не влияет
   B) Чем выше значение, тем больше ветвлений
   C) Чем выше значение, тем менее чувствительным становится дерево к незначительным улучшениям
   D) Повышает вероятность переобучения

Ответ: C

4. Что показывает метрика "Relative Error" при регрессионном анализе?
   A) Отношение количества узлов к глубине
   B) Разницу между RMSE и MAE
   C) Отношение ошибки модели к ошибке наивного прогноза
   D) Отношение обучающей и тестовой ошибок

Ответ: C

5. Какое преимущество дает обрезка дерева после его построения (postpruning)?
   A) Увеличивает скорость работы модели
   B) Уменьшает количество признаков
   C) Убирает лишние ветви и повышает обобщающую способность
   D) Повышает глубину дерева

Ответ: C

6. Что измеряет параметр Squared Correlation в регрессии?
   A) Ошибку прогноза

B) Уровень шума
C) Долю дисперсии, объясненной моделью
D) Количество правильно классифицированных наблюдений

Ответ: C

7. Какое влияние на структуру дерева оказывает значительная корреляция между признаками?
   A) Уменьшает глубину
   B) Делает дерево бинарным
   C) Может привести к выбору избыточных предикторов, дублирующих информацию
   D) Повышает точность без последствий

Ответ: C

8. Как можно определить, что дерево переобучено?
   Ответ: Если модель показывает высокую точность на обучающих данных, но низкую на тестовых — это явный признак переобучения.

9. Почему важно сравнивать RMSE с дисперсией целевой переменной?
   Ответ: Это позволяет понять, насколько значима ошибка модели по сравнению с естественной вариативностью данных — высокая RMSE относительно дисперсии указывает на слабую модель.

10. В чем преимущество дерева решений перед логистической регрессией при работе с категориальными переменными?
    Ответ: Деревья автоматически обрабатывают категориальные признаки без необходимости кодирования, в отличие от логистической регрессии, требующей one-hot encoding.