**Тестовое задание по проекту “Антисон”**

Данное тестовое задание поможет нам оценить ваш уровень теоретических знаний в области статистики и теории вероятности, классического ML, NN (СV в частности), а также узнать про ваш практический опыт в этих сферах.

Тест состоит из следующих частей:

1. Общие вопросы по ML
2. Вопросы по NN и CV
3. Вопросы про опыт студента

Этот тест во многом важен нам для правильного выстраивания дальнейшего взаимодействия со студентами, так что мы ожидаем от вас самостоятельного выполнения и честных ответов. Удачи!

***Выделите правильный ответ любым понравившимся вам цветом, а на открытые вопросы попробуйте сформулировать ответ максимально кратко, но информативно.***

**Общие вопросы по ML:**

1. Какой из следующих методов наиболее часто используется для обучения в логистической регрессии
   1. LSE
   2. Метод максимального правдоподобия
   3. Джакартова близость
   4. Оба варианта a и b
2. Какая метрика неприменима для оценки качества логистической регрессии в сравнении с целевыми
   1. ROC\_AUC
   2. Точность
   3. Logloss
   4. MSE
3. Какой алгоритм используется для отбора признаков
   1. LASSO
   2. Ridge
   3. Оба
   4. Ни один из указанных алгоритмов
4. Важным преимуществом деревьев принятия решений является:
   1. Они строят легко интерпретируемые правила принятия решений
   2. Они довольно стабильны относительно изменений выборки
   3. Они не склонные к переобучению
   4. Они базируются на нормальности данных для устойчивого результата
5. Какое утверждение НЕ верно относительно иерархической кластеризации:
   1. Кластеризация часто используется для сегментации пользователей
   2. Дендрограмма может помочь с выбором оптимального количества кластеров
   3. Обязательно необходима нормировка (стандартизация) данных
   4. Метод устойчив к изменению метрик расстояния
6. Вы тестируете новую версию сайта с новой формой для регистрации. 60% абонентов показывают новый вариант, а 40% - старый. При этом из первых регистрируется 8% пользователей, а из вторых – 4%. Какова вероятность, если пользователь зарегистрировался, что ему показывали новый вариант сайта:
   1. 0,6
   2. 0,75
   3. 0,08
   4. 0,048
7. Правильную (выпадение каждого из 6 чисел- равновероятно) игральную кость бросают 6 раз. Какова вероятность выкинуть последовательность уникальных цифр? (можете написать ответ в виде числа с точностью до 4 знаков после запятой или формулой)

1 \* 5/6 \* 4/6 \* 3/6 \* 2/6 \* 1/6 = 5!/6^5 = 0.0154

1. Выберите правильные утверждения:
   1. При отсутствии зависимости между двумя переменными корреляция должна быть близка к нулю
   2. Нулевая корреляция означает отсутствие зависимости между двумя переменными
   3. Верны оба утверждения
   4. Оба утверждения неверны
2. Выберите верные утверждения про алгоритмы градиентного спуска (GD) и стохастического градиентного спуска (SGD):
   1. И в GD, и в SGD мы итеративно улучшаем параметры с целью минимизации функции ошибки
   2. В SGD мы обучаемся на всей выборке для оптимизации параметров на каждой итерации
   3. В GD мы можем использовать все данные или какое-то подмножество для корректировки параметров на каждой итерации
3. Добавление большого количества незначимых признаком может привести:
   1. Росту R-square
   2. Снижению R-square
4. Допустим, у вас есть переменные X, Y, Z и значения коэффициента корреляции Пирсона для пар (X, Y), (Y, Z) и (X, Z) - C1, C2, C3, соответственно. Теперь ко всем значениям переменной X мы прибавляем 2, а из всех значений Y вычитаем 2. Значения переменной Z остаются неизменными. Тогда если новые значения коэффициента корреляции D1, D2 и D3, как они соотносятся со старыми?
   1. D1= C1, D2 < C2, D3 > C3
   2. D1 = C1, D2 > C2, D3 > C3
   3. D1 = C1, D2 > C2, D3 < C3
   4. D1 = C1, D2 < C2, D3 < C3
   5. D1 = C1, D2 = C2, D3 = C3
   6. Нельзя сказать
5. Допустим, вы решаете задачу классификации. При этом к первому классу относится 99% выборки. Какие метрики справедливо использовать в данном случае (можно выбрать несколько вариантов ответа):
   1. Accuracy
   2. Precision and Recall
   3. F1-score
   4. AUC ROC
6. Выберите ВЕРНЫЕ утверждения (можно выбирать несколько вариантов ответа):
   1. И PCA, и SVD может использоваться для снижения размерности
   2. В случае, когда средние значения данных равны нулю, SVD и PCA будут выдавать одинаковые результаты
   3. PCA не может быть построено с помощью SVD
   4. Сингулярное разложение матрицы состоит из ортогональных и диагональной матрицы
7. Рассмотрим 2 независимые случайные величины X и Y , подчиненные нормальному распределению с параметрами N(a1, sigma1^2) и N(a2, sigma2^2). Выберите правильное утверждение:
   1. Распределение суммы X + Y НЕ является нормальным
   2. сумма X + Y имеет нормальное распределение с параметрами N(a1 + a2, sigma1^2 + sigma2^2)
   3. сумма X + Y имеет нормальное распределение, но нельзя сказать точно о параметрах распределения
   4. сумма X + Y имеет нормальное распределение с параметрами N((a1 + a2) / 2, ¼ \* (sigma1 + sigma2)^2)
8. Что из нижеперечисленного является популярным и эффективным алгоритмом, основанным на идее бэггинга?
   1. Решающее дерево
   2. Регрессия
   3. Классификация
   4. Случайный лес
9. Что из нижеперечисленного можно отнести к недостаткам решающих деревьев?
   1. Возможность проведения факторного анализа
   2. Неустойчивость к выбросам
   3. Склонность к переобучению
   4. Все из вышеперечисленного
10. Кратко перечислите основные приемы работы с пустыми значениями в выборке?

Если говорить коротко, то у нас есть два пути: либо игнорировать такие значения, либо заполнять пропуски. Перечислю некоторые методы конкретнее:

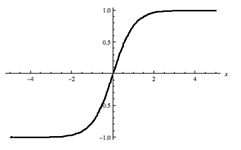
1) Удалить переменную, если в ней много пропущенных значений, и если она не является каким-либо значимым фактором в выборке. 2) Удаление строк/столбцов, содержащих сравнительно большое кол-во пропущенных значений, но так, чтобы модель не потеряла достоверность после удаления, и не появилась погрешность. 3) Заменить медианой, средним или модой, если переменная не сильно влияет на результат, или вариация данных невелика. 4) Метод k-ближайших соседей: пропущенному значению присваивается среднее значение k ближайших к нему точек, значения которых уже известны, на основании функции расстояния (чаще всего – евклидова метрика), и рассчитывается их взвешенное среднее.

1. Какая из нижеперечисленных последовательностей работы с данными наиболее корректна при решении задачи классификации или регрессии?
   1. Нормализация данных -> PCA -> обучение
   2. PCA -> Нормализация результатов PCA -> обучение
   3. Нормализация данных -> PCA -> Нормализация результатов PCA -> обучение
   4. Ничего из вышеперечисленного
2. Что из нижеперечисленного является верным утверждением про Naive Bayes?
   1. Предполагается, что все признаки в датасете одинаково сильно влияют на целевую переменную
   2. Предполагается, что признаки независимые
   3. Справедливы оба варианта a и b
   4. Ни одно из утверждений не относится к Naive Bayes
3. В каких случаях из перечисленных алгоритм K-means склонен приводить к плохим результатам: 1) в данных есть выбросы; 2) кластеры обладают разной плотностью точек 3) кластеры не являются выпуклыми
   1. 1 и 2
   2. 2 и 3
   3. 1, 2, и 3
   4. 1 и 3
4. Какие методы целесообразно использовать для выбора оптимального количества компонент k в PCA:
   1. Выбрать наименьшее из возможных k, которое отражает как минимум 99% от вариативности исходного набора
   2. Выбирать k близким к m (например, k = 0.99\*m с округлением до ближайшего целого числа), где m - исходное количество признаков
   3. Выбрать максимальное значение k, при котором сохраняется 99% вариативности датасета
   4. Методом elbow

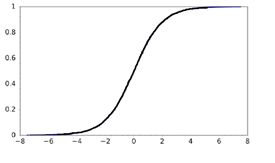
**Вопросы по NN и CV:**

1. Ниже три графика с различными функциями активации. Выберите из представленных опций правильное утверждение:

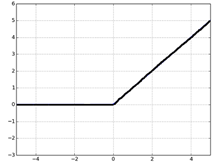
**1.**

****

**2.**

****

**3.**

****

* 1. 1 – функция гиперболического тангенса, 2 – ReLU функция, 3 – сигмоида
  2. 1 – сигмоида, 2 – ReLU функция, 3 - функция гиперболического тангенса
  3. 1 – ReLu функция, 2 - функция гиперболического тангенса, 3 - сигмоида
  4. 1 – функция гиперболического тангенса, 2 – сигмоида, 3 – ReLU функция

1. Ответьте кратко, с какими проблемами на практике сталкиваются при работе с функциями активации sigm и tahn?

Про sigm: при приближении к границам диапазона значения начинают слабо реагировать на изменение аргумента, и градиент на этих участках становится близок к 0. Также, если исходные веса имеют слишком большие значения, большинство нейронов перейдет в состояние насыщения, тогда нейросеть перестает обучаться или делает это очень медленно.

Про tahn: функции гиперболического тангенса также свойственна проблема исчезновения градиента, его затухания или увеличения (у нее производная круче, чем у сигмоиды, а значит и градиент больше).

1. Какие варианты интерполяции в сверточных сетях вы знаете?

Интерполируем с помощью приближения полиномами (можем, по теореме Вейерштрасса), тригонометрической интерполяцией (пользуясь дискретным преобразованием Фурье), интерполируя по Лагранжу или Ньютону.

1. Напишите 3-4 примера наиболее известных архитектур сверточных сетей, которые вы знаете

Residual Network (ResNet), AlexNet, GoogLeNet

1. Опишите кратко идею автокодировщиков, а также их возможные модификации:

Автокодировщик – это метод машинного обучения без учителя. Он копирует входные данные и сжимает их, и с помощью декодера восстанавливает исходные данные, при этом, входной и выходной слои автокодировщика имеют одинаковое кол-во нейронов. Когда он копирует данные, он пытается сохранить в своем скрытом представлении только наиболее важные признаки объекта, также, мы можем таким способом находить закономерности в данных. В визуализации данных его применяют для сглаживания шума или снижения размерности.

Модификации: регуляризованный автокодировщик(Sparse Autoencoders); Stacked Autoencoders(несколько слоёв Sparse Autoencoders), Deep Autoencoders(состоит из двух идентичных глубоких сетей).

1. Опишите своими словами как вы понимаете transfer learning

Это метод обучения, которые позволяет использовать опыт, накопленный при решении одной задачи, для решения другой, например, можно умение распознавать цвета на фотографии использовать в качестве помощи при построении прогнозов для другой модели, которая решает другую похожую задачу. Transfer learning становится популярен, так как для некоторых задач стартового набора данных крайне не хватает или нужные датасеты вовсе отсутствуют, кроме того, обучение сложных и больших моделей с нуля бывает требует очень большого времени и ресурсов, тут то мы и можем воспользоваться уже готовым решением от другой модели.

**Расскажите нам про ваш опыт:**

1. Работали ли вы ранее с реальными проектами (с внедрением) в области ML? Если да, то очень кратко опишите, в чем была суть задачи:

У меня был опыт разработки рекомендательной системы для Московской электронной школы на реальных данных. Задача была предложена Департаментом информационных технологий Москвы в рамках программы «Умный город» в ОЦ «Сириус».

Суть задачи заключалась в следующем: у нас имелся портал для учителей, мы имеет 7 датасетов с данными о них и их деятельности, интересах и тп, и нам нужно на основе этого построить алгоритм, который будет выдавать подборку материалов, актуальную для каждого пользователя. Сначала нужно было очистить данные, нормализовать их, и, основываясь на очищенных датасетах, смерджить их, чтобы получить 3 датасета, на которых можно уже работать. Далее, мы начали анализ, строили графики и разрабатывали гипотезы: как нам рекомендовать пользователям в зависимости от того, насколько много мы о них знаем. Так, мы составили алгоритм, в подумали о пользователях с «большой историей», тех, о ком знаем мало, но есть ключики, по которым можем их распознать(тут использовали метод ближайших соседей, и пытались рекомендовать то же, что и «самому близкому» соседу), и решили проблему «холодного старта» для новых пользователей.

1. С каким стэком вы привыкли работать над ML-задачами:
   1. Python
   2. R
   3. Оба
2. Был ли у вас ранее опыт решения задач с применением техник нейронных сетей? Если да, то кратко опишите его:

Еще не было опыта в решении такого рода задач.

1. Сталкивались ли вы с задачами из области компьютерного зрения? Аналогично, если да, то кратко опишите задачу

К сожалению нет, но очень хотелось бы погрузиться в эту сферу.

1. С какими инструментами/библиотеками в области CV вы знакомы?

Работал с библиотеками NumPy, Pandas, matplotlib для работы с данными. Не имея опыта в CV, мне доводилось только слышать о библиотеках OpenCV, scikit-learn, TensorFlow.