

# Программа для поступающих на продвинутый трек магистратуры кафедры Яндекса

## Раздел 1. Вероятность и статистика

- (1) Основы комбинаторики. Числа сочетания (с повторениями и без повторений), числа размещения (с повторениями и без повторений), перестановки. Бином Ньютона и биномиальные коэффициенты. Формула включений и исключений. Простейшие комбинаторные тождества. Знакопеременные тождества. Использование формулы включений и исключений для доказательства тождеств.
- (2) Основы теории вероятностей. Классическое определение вероятности. Условные вероятности и независимость событий (попарная и в совокупности). Формулы полной вероятности и Байеса.
- (3) Случайные величины в дискретном вероятностном пространстве, их распределения. Независимость случайных величин. Числовые характеристики распределений: математическое ожидание, дисперсия и ковариации. Основные свойства математического ожидания, дисперсии и ковариации.
- (4) Случайные величины в произвольном вероятностном пространстве: определение и основные действия над случайными величинами. Функция распределения случайной величины, ее основные свойства. Независимость случайных величин: эквивалентные определения.
- (5) Определение математического ожидания в общем случае: аппроксимация простыми случайными величинами. Свойства математического ожидания. Абсолютно непрерывные случайные величины и векторы, примеры. Формулы подсчета математических ожиданий для абсолютно непрерывного случая. Формула свертки.
- (6) Виды сходимостей случайных величин. Три предельных теоремы: закон больших чисел в форме Чебышёва, усиленный закон больших чисел в форме Колмогорова, центральная предельная теорема для независимых одинаково распределенных случайных величин.
- (7) Основы математической статистики. Основная задача математической статистики, понятие выборки. Эмпирическая функция распределения и её свойства. Теорема Гливленко-Кантелли. Теорема Колмогорова об эмпирической функции распределения. Критерий Колмогорова о проверке вида функции распределения.
- (8) Основные свойства точечных оценок: несмещённость, состоятельность, асимптотическая нормальность. Методы построения: метод моментов и метод максимального правдоподобия. Свойства оценок указанных методов.
- (9) Проверка статистических гипотез. Понятия статистического критерия, ошибок первого и второго рода, функции мощности и уровня значимости критерия. Состоятельность и несмещённость статистического критерия. Критерий Неймана-Пирсона для проверки простых гипотез. Асимптотический критерий хи-квадрат для полиномиальной схемы Бернулли, его свойства.
- (10) Проверка однородности нормальных выборок: критерий Фишера и t-критерий Стьюдента. Критерий Аспина-Уэлча и t-критерий Стьюдента для связанных выборок. Проверка гипотезы об отсутствии сдвига для связанных выборок: критерий знаков, критерий знаковых рангов Уилкоксона.

## Раздел 2. Оптимизация

- (1) Методы одномерной оптимизации без производной: методы золотого сечения, парабол и Брента.
- (2) Градиентные методы обучения. Свойство градиента о направлении наискорейшего убывания. Градиентный спуск. Методы оценивания градиента.
- (3) Метод Ньютона, его скорость сходимости.
- (4) Квазиньютоновские методы оптимизации. Метод L-BFGS.
- (5) Выпуклые множества и функции. Основные примеры. Операции, сохраняющие выпуклость. Простейшие свойства выпуклых функций.

- (6) Задачи условной оптимизации. Условия Каруша-Куна-Таккера. Двойственность.
- (7) Стандартные классы выпуклых задач: линейное программирование (LP), квадратичное программирование (QP, QCQP).
- (8) SGD и его основные модификации (momentum, RMSProp, Adam).

### Раздел 3. Алгоритмы и структуры данных

- (1) Время и память как основные ресурсы. Модели вычислений: RAM, разрешающие деревья. Сложность на заданном входе, сложность в худшем случае, сложность в среднем случае, рандомизированная сложность. Нижняя оценка на число сравнений при сортировке в модели разрешающих деревьев.
- (2) Динамическое программирование: общие принципы, свойства задач, эффективно решаемых при помощи динамического программирования. Рекурсивная реализация с мемоизацией и итеративная реализация. Примеры: вычисление редакционного расстояния, решение задачи о рюкзаке.
- (3) Жадные алгоритмы: общая идея, критерии применимости. Примеры: построение остовов минимального веса в графах, жадные алгоритмы в задачах планирования. Жадный алгоритм на матрице.
- (4) Разновидности алгоритмов сортировки: inplace, stable. Быстрая сортировка (Quick-Sort). Способы выбора разделяющего элемента. Элиминация хвостовой рекурсии.
- (5) Порядковые статистики. Рандомизированный алгоритм Quick-Select. Детерминированный алгоритм поиска (метод "медианы медиан").
- (6) Сортировка слиянием (Merge-Sort).
- (7) Кучи: основные определения и свойства. Операции Sift-Down и Sift-Up. Бинарные и к-ичные кучи. Построение кучи за линейное время. Алгоритм сортировки Heap-Sort.
- (8) Хеш-функции. Коллизии. Разрешение коллизий методом цепочек. Гипотеза простого равномерного хеширования, оценка средней длины цепочки.
- (9) Универсальные семейства хеш-функций, оценка средней длины цепочки. Построение универсального семейства для целочисленных ключей. Совершенные хеш-функции. Построение совершенной хеш-функции методом двухуровневого хеширования.
- (10) Фильтр Блума (Bloom filter). Оценка вероятности ложноположительного срабатывания.
- (11) Графы: основные определения и способы представления в алгоритмах. Обход в ширину. Обход в глубину.
- (12) Проверка на ацикличность и топологическая сортировка.
- (13) Деревья поиска. Вставка и удаление элементов. Inorder-обход дерева.
- (14) Сильно связанные компоненты и топологическая сортировка конденсации.
- (15) Дерево отрезков.
- (16) Кратчайшие пути в графах. Оценки расстояний и их релаксация. Алгоритмы Беллмана-Форда, Флойда и Дijkstra.
- (17) Остовы минимального веса. Лемма о минимальном ребре в разрезе. Алгоритмы Краскала и Прима.
- (18) Системы непересекающихся множеств. Реализация с использованием леса. Ранги вершин, эвристика ранга. Логарифмическая оценка ранга через количество элементов. Эвристика сжатия путей. Оценка учетной стоимости операций (без доказательства).

### Раздел 4. Машинное обучение

- (1) Линейная модель регрессии. Аналитическое решение для среднеквадратичной ошибки (с выводом). Градиентное обучение линейной регрессии.
- (2) Регуляризация линейных моделей. Разреженные модели и L1-регуляризация.
- (3) Линейная модель классификации. Отступ. Обучение линейных классификаторов через верхнюю оценку на долю ошибок. Примеры верхних оценок.
- (4) Функционалы ошибки для классификации: матрица ошибок, accuracy, precision, recall, F-мера. ROC-кривая и AUC-ROC. Precision-recall-кривая и AUC-PR.

- (5) Логистическая регрессия. Оценивание вероятностей. Вывод логистической функции потерь из метода максимального правдоподобия.
- (6) Метод опорных векторов. Вывод постановки задачи для разделимого и неразделимого случаев. Ядра и ядерной переход.
- (7) Решающие деревья: определение и жадный алгоритм обучения. Функционал качества при выборе предиката. Общий вид критерия информативности (через функцию потерь) и конкретные примеры для регрессии (дисперсия) и классификации (критерий Джини и энтропийный критерий).
- (8) Разложение ошибки на смещение и разброс.
- (9) Бэггинг и случайные леса.
- (10) Градиентный бустинг. Обучение базовых алгоритмов для произвольной дифференцируемой функции потерь. Сокращение шага. Выбор прогнозов в листьях деревьев в градиентном бустинге.
- (11) Стекинг и блендинг.
- (12) Задача кластеризации. Метрики качества. Методы: K-Means, графовые методы, иерархическая кластеризация.
- (13) Задача визуализации, t-SNE.
- (14) Обучение представлений слов и word2vec.
- (15) Понижение размерности. Метод главных компонент: постановка задачи и решение.
- (16) User-based и item-based рекомендательные системы.
- (17) Модели со скрытыми переменными (latent factor model, LFM) для построения рекомендаций. Обучение LFM: стохастический градиентный спуск, ALS, HALS.
- (18) Неявная информация в рекомендательных системах, implicit ALS.
- (19) Факторизационные машины. FFM. Способы обучения моделей.
- (20) Тематическое моделирование. Постановка задачи. Методы LSA, PLSA и LDA.
- (21) Нейронные сети. Метод обратного распространения ошибки.
- (22) Свёрточные и рекуррентные нейронные сети. Примеры архитектур, особенности обучения.
- (23) Backpropagation through time, пример вычисления градиентов, проблема затухающих градиентов.
- (24) Параллельное обучение линейных моделей (на одной машине и на кластере).
- (25) Концепция MapReduce.
- (26) Байесовский подход к теории вероятностей. Оценка параметров в байесовском и частотном подходе. Примеры байесовских рассуждений.
- (27) Вероятностная модель линейной регрессии. Метод релевантных векторов для задачи регрессии.
- (28) EM-алгоритм в общем виде. Примеры применения. Разделение смеси с помощью EM-алгоритма.