

Оценочные материалы для текущего контроля и промежуточной аттестации по дисциплине

Анализ данных

Код, направление подготовки	01.03.02 ПРИКЛАДНАЯ МАТЕМАТИКА И ИНФОРМАТИКА
Направленность (профиль)	Технологии программирования и анализ данных
Форма обучения	очная
Кафедра-разработчик	Кафедра прикладной математики
Выпускающая кафедра	Кафедра прикладной математики

Типовые задания для контрольной работы:

Вариант 1

1. Время решения тестовой задачи составило (сек):
59, 50, 54, 52, 54, 56, 58, 52, 51, 48, 59, 62, 53, 60, 43, 48, 45, 59.
Построить гистограмму частот и график выборочной функции распределения, разбив данные на 6 интервалов. Вычислить точечные оценки математического ожидания и дисперсии.
2. Для каждого из трех видов рекламных щитов отобрано случайным образом по 6 магазинов, расположенных в соответствующем районе. Объемы продаж за месяц (млн руб.) приведены в таблице. Влияет ли расположение рекламных щитов на объем продаж товара?

Магазин	Расположение щитов		
	1	2	3
1	8,1	3,9	4,8
2	7,5	4,4	6,0
3	5,8	2,8	4,4
4	6,6	3,4	5,8
5	5,9	4,8	6,2
6	4,8	3,6	4,9

Вариант 2

1. Вычислить выборочное среднее, выборочную дисперсию, моду и медиану выборки:
7, 3, 3, 6, 4, 5, 1, 2, 1, 3.
2. Компания, производящая спортивные товары, сравнивает расстояние, которое пролетают мячи, изготовленные по четырем разным технологиям. По каждой технологии произведено по 8 мячей. Мячи испытаны в течение короткого промежутка времени при одинаковых погодных условиях. Результаты испытаний в м:

Мяч	Технология			
	1	2	3	4
1	206	203	217	213
2	226	223	230	231
3	208	206	221	221
4	224	223	227	222
5	206	205	218	229
6	229	234	231	235
7	204	204	224	213
8	228	219	225	228

Зависит ли дальность полета мяча от технологии производства?

Типовые вопросы к экзамену:

1. Коэффициент детерминации (R^2), информационные критерии (AIC, BIC), метрики (MSE, MAE, MAPE).
2. Гауссовская линейная модель - доверительные интервалы для коэффициентов модели и для отклика, гипотезы о незначимости коэффициента и группы коэффициентов, общая линейная гипотеза, сравнение моделей.
3. Дисперсия остатков линейной модели в условиях гетероскедастичности, визуальный анализ.
4. Критерии проверки на гомоскедастичность: Бройша-Пагана и Голдфелда-Квандта. Преобразование Бокса-Кокса.
5. Устойчивые оценки дисперсии Уайта, асимптотическая нормальность
6. Байесовский классификатор.
7. Квадратичный и линейный дискриминантный анализ. Оценки параметров, вид разделяющей поверхности. Наивный байесовский классификатор.
8. Робастная регрессия. Методы на основе ближайшего соседа - kNN, взвешенный kNN, их свойства.
9. Непараметрическая регрессия, локальное усреднение, оценка Надарая-Ватсона.
10. Условия сходимости оценки Надарая-Ватсона, выбор ширины ядра, доверительная лента.
11. Локальная линейная регрессионная модель.
12. Метод главных компонент (PCA) как выбор направлений с максимальной дисперсией, формулы перехода в сжатое пространство и обратно.
13. Дисперсии образа, выбор размерности сжатого пространства на основе доли необъясненной дисперсии
14. Методы SNE и t-SNE: первоначальный вариант SNE, симметричный SNE, проблема скученности, метод t-SNE как решение проблемы.
15. Метод UMAP. Постановка задачи: графы, функционал качества (KL).
16. Таблицы сопряженности 2x2, точный тест Фишера, меры взаимосвязи, определение количества наблюдений.
17. Общий случай таблиц сопряженности, типы вероятностных моделей, критерий хи-квадрат.
18. Влияние признаков на целевую переменную: корреляции, подход с помощью решающих деревьев – важность признаков на основе Mean Decrease Impurity, Permutation feature importance, Drop Column feature importance.

19. Критерии проверки однородности для бернуллиевских выборок, доверительные интервалы для разности (простые и Уилсона).
20. Проверка на равенство средних нормальных выборок (t-test, 3 сл.), проверка равенства дисперсий, проверка однородности нормальных выборок.
21. АВ-тестирование. Принципы разбиения, особенности. АА-тесты. Разбиение на тестовые группы, сроки теста, проверка нескольких гипотез.
22. Критерии Смирнова и Розенблатта.
23. Критерий Уилкоксона-Манна-Уитни, его свойства, связанная с ним оценка параметра сдвига.
24. F-критерий и критерий Бартлетта, их применимость.
25. Критерий Краскела-Уоллиса и Джонкхиера.
26. Post-hoc анализ: LSD Фишера, HSD Тьюки, критерии Неменья и Данна, оценка контраста.
27. Однофакторный дисперсионный анализ для случая связанных выборок. F-критерий, критерии Фридмана и Пейджа. Post-hoc анализ.
28. Двухфакторный дисперсионный анализ, случай дополнительной контрольной группы.
29. Парадокс Симпсона, примеры и выводы. Контрафактивная модель, причинно-следственный эффект, статистическая связь, утверждение о том, что связь не есть причинность.
30. Равенство величины причинно-следственного эффекта и статистической связи при случайном назначении воздействия.
31. Марковское распределение на графе. Условная независимость и ее свойства. Оценка распределений в графе методом максимального правдоподобия. Интервенция, средний условный эффект как способ оценки причинно-следственного эффекта по графу.
32. Марковское свойство. Свойства d-разделимости и d-связности, теорема об условной независимости на множестве вершин.
33. Построение причинно-следственных графов по данным: метод индуктивной причинности. Оценка условной независимости: частная корреляция, причинность по Грейнджеру.