**МФК Статистика, сплайны, нейросети**

**(операции с «малыми» и «большими» данными)**

**Аннотация**

При обработке данных наблюдений физических, химических, биологических, экономических, социальных и других явлений или экспериментов оказывается, что нормальные распределения и линейные зависимости встречаются достаточно редко и известные математические методы малопригодны для большинства реальных случаев.

Используется новый подход к анализу переменных, как имеющих априорно «ненормальное» распределение, и к моделированию исследуемых зависимостей, как априорно нелинейных. Для моделирования парной зависимости используется сплайн-регрессия, а для множественной зависимости - многомерный ортогональный сплайн.

Новые принципы и алгоритмы характеризуются простотой вычислений и наглядностью результатов. Возможности новых методов демонстрируются на конкретных примерах, известных моделях и задачах, в том числе предложенных слушателями курса.

Рассматривается сплайн-модель искусственного нейрона (СМН), разработанная авторами на основе многомерных сплайнов, которая по разрешающей способности эквивалентна трехслойной сети известных нейронов и настраивается со скоростью в 1000 раз быстрее. При этом получаемые сплайн-составляющие СМН, как правило, наглядно отражают исследуемые причинно-следственные связи.

Использование СМН и их сетей значительно расширяет возможности моделирования нелинейных, парных и множественных зависимостей, распознавания образов, автоматической классификации (кластерного анализа), решения задач математической физики и других научно-практических задач.

Изложение курса адаптировано для слушателей, знакомых с математикой в объеме средней школы. Полученные знания могут быть успешно использованы при выполнении лабораторных, курсовых, дипломных и диссертационных работ, так же в дальнейшей научной деятельности.

**Программа**

1. Обзор содержания межфакультетского курса. Примеры сплайн-моделей и нейронных сетей (2 часа)

Требуемая подготовка слушателей. Принадлежности: широкие листы (в клетку), линейка, карандаши, резинка. Основные разделы и темы.

Статистика: бинарные модели, гистограммы и распределения. Особенности геостатистики. Оценка связи по разности распределений. Корреляция и регрессия. Нелинейная регрессия. Дисперсионный анализ, оценки Фишера и Пирсона. Оценка связи по разности двухмерных распределений. Множественные зависимости.

Сплайны, сущность и классификация. Сплайн-регрессия. Многомерный ортогональный сплайн и его настройка. Сплайн-модель нейрона и ее свойства. Примеры сплайн-моделей.

Нейросети: модель МакКалока-Питтса, перцептрон Розенблатта и их соединения. Интерпретация и ограничения линейных нейросетей. Сети из сплайн-моделей нейронов и их возможности.

Задачи к зачету для разного уровня математической подготовки. Зачеты по результатам решения задач, предложенных студентами.

2. Обработка данных. Переменные и распределения. Бинарные модели. Оценка связи по разности распределений. (2 часа)

Обработка данных – изучение, моделирование, управление. Наблюдение и эксперимент. Качественные и количественные переменные. Таблицы и матрицы данных. Анализ рядов наблюдений, сортировка и выявление ошибок. Гистограммы и распределения. Среднее квадратическое и среднее арифметическое отклонения. Метод средних и вариация его ошибки. Частотная связь - разность двух распределений.

3. Оценка зависимости между двумя переменными. Оценка связи по разности двухмерных распределений. (2 часа)

Распределение двух переменных и его визуализация равновероятными интервалами. Частотная связь двух переменных. Матричное представление распределений.

Модель линейного интервала и его настройка. Основные условия сходимости: симметричность вариации функции; псевдослучайный выбор, стохастическая аппроксимация.

Основные способы оценки корреляции: коэффициент линейной корреляции; частотная связь; коэффициент К.Пирсона; визуальная оценка ширины отклонений.

Примеры: Построение и анализ двумерных распределений. Вычисления коэффициентов частотной связи по разным сеткам. Оценка и вычисление коэффициента корреляции разными способами.

4. Сплайны, классификация, построение. Описание и настройка линейного интервала. Условия сходимости. (2 часа)

Формы зависимостей и их встречаемость. Ступенчатые и кусочно-линейные зависимости. Кусочно-параболические зависимости - «сплайны».

Классификация сплайнов: геометрические и функциональные; интерполяционные, сглаживающие и корреляционные. Описание сплайнов узловыми точками и алгоритмом их соединения.

Описание и настройка линейного интервала: неопределенная форма, частично определенная, определенная узловыми точками. Итерационный алгоритм настройки узловых точек.

Условия сходимости: симметричность вариации функции, псевдослучайный выбор, стохастическая аппроксимация.

5. Моделирование парной зависимости. Расчет параметров сплайн-регрессии. Условия сходимости. (2 часа)

 Построение сетки сплайна для дискретных и непрерывных переменных. Количество наблюдений и количество узлов. Расчеты ординат узловых точек: по ступенчатой функции; методом визуальной аппроксимации интервалов. Настройка ординат узловых точек. Расчеты и настройка за пределами сеток. Коэффициент корреляции и его ошибка.

Условия сходимости: симметричность вариации функции, псевдослучайный выбор, стохастическая аппроксимация.

Примеры: Анализ исходных данных. Построения сеток сплайна и визуальных аппроксимаций. Настройка ординат. Расчет коэффициента корреляции и его ошибки.

6. Моделирование множественной зависимости. Расчет параметров многомерного сплайна. Условия сходимости. (2 часа)

Моделирование множественной зависимости. Многомерный ортогональный сплайн. Нормирование числа аргументов. Веса аргументов. Нормирование значений функции. Моделирование дискретных функций. Примеры сплайн-моделей.

Задание и настройка параметров многомерного сплайна. Подготовка исходных данных. Выбор оптимального количества узлов сплайн-составляющих. Подготовка таблиц для вычислений. Пример расчетов 3 – 4 итераций.

Условия сходимости и устойчивости: симметричность вариации функции, псевдослучайный выбор, стохастическая аппроксимация, стабилизация постоянных составляющих.

Примеры: Расчеты функции по полной и неполной совокупности аргументов. Расчет параметров сплайн-модели по 15 - 20 наблюдениям, 2 - 3 аргументам, 2 - 3 узлам. Расчет коэффициента корреляции и его ошибки.

7. Основные свойства сплайн-моделей. Слияния, разбиения, корреляция аргументов (2 часа)

Сплайн-модель как взвешенная сумма парных зависимостей. Расчеты и настройка при неполных данных. Значения весов в зависимости от средних амплитуд сплайн-составляющих и коэффициентов корреляции. Слияние сплайн-моделей. Разбиение сплайн-моделей. Взаимная корреляция аргументов. Добавление узловых точек. Удаление узловых точек. Добавление рядов наблюдений. Автономные расчеты отдельных частей сплайн-моделей. Развертка многомерного пространства.

8. Интерпретация и возможности нейронных сетей. Сплайн-модель нейрона и ее сети (2 часа)

Модель нейрона МакКалока-Питтса, перцептрон Розенблатта. Нейронные сети и их настройка. Интерпретация и ограничения линейных нейронных сетей.

Сплайн-модель нейрона и ее настройка. Сравнение с линейными нейросетями. Сети из сплайн-моделей нейронов и их возможности.

9. Методика моделирования множественных зависимостей. Анализ временных и пространственных трендов. (2 часа)

Постановка задачи: выбор функции и возможных аргументов. оценка количества наблюдений, количества аргументов и количества узлов. Сбор и подготовка исходных данных.

Формирование и расчет выборочных сплайн-моделей: выбор информативных аргументов и оптимальной нелинейности. Введение новых переменных как относительных величин и свертки первичных данных.

Анализ трендов. Учет ретроспективы и перспективы – сдвиги по времени. Автокорреляционные и взаимно корреляционные, ретроспективные и перспективные, линейные и нелинейные функции.

Учет динамики функции и аргументов – введение производных.

Решение обратных задач. Выбор функции и аргументов. Генерация исходных данных.

10. Оценка различий по совокупности признаков. Распознавание образов и автоматическая классификация (2 часа)

Фазовое пространство и классы объектов. Различие и сходство объектов и классов: пространство Евклида, пространство Хемминга. Нормирование фазового пространства.

Распознавание образов. Модели распознавания: полная сеть точек; сферы; дискриминантные функции, комбинации. Самообучаемые системы. Оценка качества моделей.

Автоматическая классификация. Слияние точек. Матрица различий и ряд минимумов. Иерархическая система классификации; дерево слияний. Метод сфер и учет нелинейности классов. Оценка достоверности различий.

Разбиение классов. Самообучаемая модель с положительной обратной связью. Последовательность разбиений. Оценка достоверности различий. Линейные и нелинейные дискриминантные функции. Интерактивная система классификации на плоскости.

11. Моделирование физических полей пространственной сетью сплайн-моделей нейронов (2 часа)

Геометризация по данным равномерных сетей. Проблемы пересчета из неравномерной сети в равномерную. Метод триангуляции: малое число окружающих точек; неоднозначность построения сети.

Методы скользящего окна: осреднение. Учет удаленности точек. Учет перекрытий в угловых секторах. Метод «теней». Метод склейки локальных сплайн-поверхностей. Комбинированная сплайн-интерполяция. Примеры геометризации.

12. Обсуждение решения выбранных научно-практических задач, в том числе предложенных студентами. Рекомендации. Оценка точности моделирования. Интерпретация полученных результатов.

**Задания к зачету**

1. Построение 2-х сопряженных распределений и оценка их несовпадения по методу средних и частотной связи. Интерпретация результата.

2. Построение одномерной кусочно-линейной регрессии итерационным алгоритмом и расчет корреляционного отношения К.Пирсона. Интерпретация результата.

3. Построение одномерного корреляционного кусочно-линейного сплайна итерационным алгоритмом и оценка коэффициента нелинейной корреляции. Интерпретация результата.

4. Построение 2-3-мерного кусочно-линейного сплайна и оценка точности (моделирования, распознавания, классификации). Интерпретация результата.

5. Оценка точности (моделирования, распознавания, классификации) при работе с разными массивами данных. Интерпретация результата.