ИНСТРУМЕНТЫ МАТЕМАТИЧЕСКОГО МОДЕЛИРОВАНИЯ В МАШИННОМ ОБУЧЕНИИ

СВЕТЛАНА ГОВОР

ФЕДЕРАЛЬНЫЙ ТЬЮТОР ПО ПРОГРАММАМ СОВРЕМЕННЫХ МАТЕМАТИЧЕСКИХ МЕТОДОЛОГИЙ ДЕПАРТАМЕНТА ОБРАЗОВАТЕЛЬНЫХ ПРОГРАММ ДЛЯ ДЕТЕЙ И МОЛОДЕЖИ, ТЕХНОПАРКИ «КВАНТОРИУМ», РУКОВОДИТЕЛЬ СЕКЦИИ ПО МАТЕМАТИЧЕСКОМУ ОБРАЗОВАНИЮ В НАПРАВЛЕНИИ «ИНФОРМАЦИОННАЯ АНАЛИТИКА» МГТУ ИМ. Н.Э. БАУМАНА

ОБЛАСТИ ПРИМЕНЕНИЯ МАШИННОГО ОБУЧЕНИЯ

1. Искусственный интеллект
2. Бизнес (фондовые рынки)
3. Распознавание сигналов, текстов
4. Автоматизация процессов
5. Система безопасности
6. Прогнозирование

МАТЕМАТИЧЕСКАЯ БАЗА

Теория графов

- Блок-схемы

Дифференцирование

- Основные понятия
- Численное дифференцирование
- Дифференцирование временного ряда

Интегрирование

- Основные понятия
- Численное интегрирование
- Интегрирование временного ряда
- Кумулятивная сумма. Скользящая средняя

Основы линейной алгебры

- Векторы, матрица

- Матричная арифметика

Теория вероятностей

- Основные понятия
- Нормальное распределение
- Мат. ожидание, стандартное отклонение, ковариация, корреляция
- Гипотезы. Условная вероятность. Теорема Байеса

Модели регрессии

- Линейная регрессия
- Логистическая регрессия
- Множественная регрессия

ОБЪЕКТ ИССЛЕДОВАНИЯ

Мы имеем некоторый объект исследования или «черный ящик». На него воздействуют какие-то факторы, и при помощи каких-то функциональных зависимостей мы и имеем функцию отклика.

Основные характеристики статистики:

1. Генеральная совокупность
2. Выборка (связанная выборка — просто разные объекты, не связанные между собой; несвязанная — каждому объекту из одной выборки соответствует ровно один объект из другой)
3. Наблюдение

ГИПОТЕЗА В СТАТИСТИКЕ

Гипотеза в статистике — есть некое научное предположение, которое необходимо проверить и далее принять или отвергнуть:

1. Нулевая гипотеза (H_0) — это гипотеза о том, что две совокупности, которые сравниваются по одному или нескольким признакам не отличаются
2. Альтернативная гипотеза (H_1) — это единственное утверждение, являющееся логическим отрицанием нулевой гипотезы

Мода — наиболее часто встречающееся значение.

Медиана — середина упорядоченного ряда значений.

МЕРЫ ЦЕНТРАЛЬНОЙ ТЕНДЕНЦИИ

Математическое ожидание — сумма значений, деленная на их количество.

Доверительный интервал — то, какие границы мы допускаем относительно своего среднего значения.

Дисперсия — разброс значений относительно среднего значения.

Корреляция — статистическая взаимосвязь двух или более случайных величин.

Нормальное распределение — характеризуется мат. ожиданием, дисперсией, основной альтернативной гипотезой, функцией и плотностью распределения.

РЕГРЕССИОННЫЙ АНАЛИЗ «ВОЛШЕБНЫЙ ЭЛИКСИР»

Он позволяет измерить величину зависимости между какой-то переменной и исходом:

$$y = a + bx$$

Задачи, решаемые с помощью регрессионного анализа:

1. Выявить зависимость появления сердечно-сосудистых заболеваний от занимаемой должности в фирме?
2. Какая зависимость между весом и ростом человека?
3. Определить зависимость обратившихся к врачу пациентов от сделанных прививок?

Мы наблюдаем зависимость относительно функциональной зависимости огромного количества значений.

Регрессионную модель можно усложнять, добавляя туда еще некоторые факторы.

Процесс исследования регрессионного анализа

1. Знак
2. Величина
3. Значимость

Если мы исследуем вес человека, то результат, зависит, например, от того, чей именно вес мы исследуем — мужчины или женщины. Но мы не учли, зависит ли вес от образования или от того, насколько часто человек занимается спортом, а также влияет ли на вес расовая принадлежность.

ЛОГИСТИЧЕСКАЯ РЕГРЕССИЯ

Это статистическая модель, у которой значение функции является вероятность того, что данное исходное значение принадлежит к определенному классу. Результат содержится в интервале от [0, 1].

Между выборками есть такая поверхность, которая называется линейным дискриминантом.

Дисперсионный анализ

Тоже используется в машинном обучении. Дисперсионный анализ можно исследовать в Excel.

Например, у нас есть задача — исследовать зависимость пола от дБ. И первый вопрос, на который мы ответим: существует ли зависимость между полом и силой звука? Дисперсионный анализ отвечает, что эта зависимость есть.

КЛАСТЕРНЫЙ АНАЛИЗ

Сбор данных, содержащих информацию о выборке объектов, и затем упорядочивающая объекты в сравнительно однородные группы. Между этими группами есть расстояние. Его можно измерить при помощи Евклидова расстояния, максимума/минимума или нахождения разности. Этих способов очень много.