ВСЕРОССИЙСКАЯ ОЛИМПИАДА ПО ИСКУССТВЕННОМУ ИНТЕЛЛЕКТУ

МИНИСТЕРСТВО ПРОСВЕЩЕНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ

ПРОСВЕТ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ПРОСВЕЩЕНИЯ

Подготовка к решению заданий на заключительном этапе олимпиады по машинному обучению

Задания по машинному обучению

На заключительном этапе будет **два** задания, посвящённые различным областям применения машинного обучения. В частности, следует обратить внимание на следующие темы:

- Прогнозирование кликов
- Снижение размерности данных
- Задачи работы с разреженными данными
- Мультимодальное обучение
- Предобработка изображений
- Предобработка текста

Пример темы 1: прогнозирование кликов

CTR (Click-Through-Rate) - задача предсказания вероятности того, что пользователь нажмёт на рекламное объявление.

$$\text{CTR} = \frac{\text{Number of click-throughs}}{\text{Number of impressions}} \times 100(\%)$$

Особенности задачи:

- Высокая размерность
- Разреженность
- Взаимосвязь признаков

ВСЕРОССИЙСКАЯ ОЛИМПИАДА ПО ИСКУССТВЕННОМУ ИНТЕЛЛЕКТУ

МИНИСТЕРСТВО ПРОСВЕЩЕНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ

ПРОСВЕТ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ПРОСВЕЩЕНИЯ

CTR: проблема многомерности и разреженности данных

Высокая размерность и разреженность данных

Множество категориальных признаков -> преобразование в вектор однократного кодирования -> миллионы новых измерений, большая часть которых пустые.

Факторы риска:

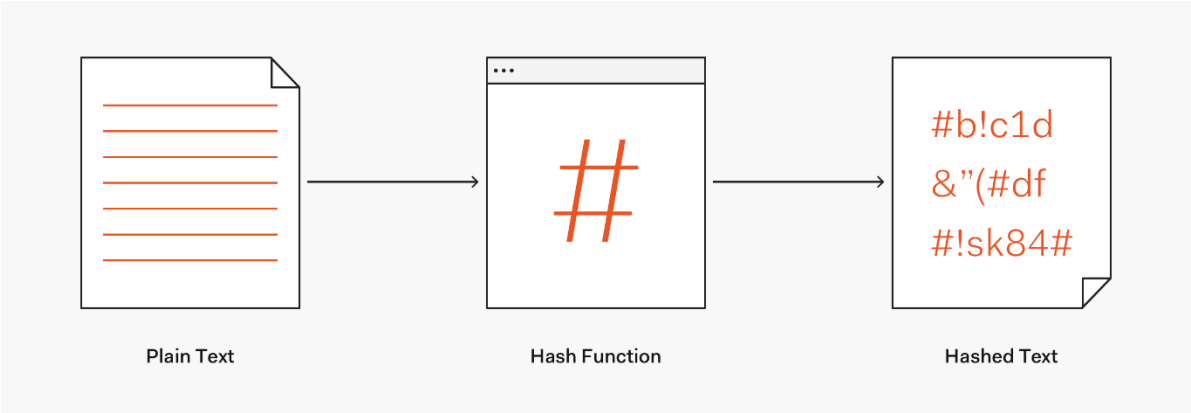
- One-hot Encoding
- Код подсчёта для частоты слов в документе
- Код TF-IDF

Основные подходы к решению

1. Hashing trick при кодировании категорий
2. Понижение размерности (t-sne и PCA)
3. Инкрементальное обучение

1. Подходы к решению: Hashing trick

Хэширование – алгоритм, получающие на вход данные (обычно строку) и возвращающий число.



Пример реализации - Scikit-learn HashingVectorizer

2. Подходы к решению: снижение размерности

T-SNE (t-distributed stochastic neighbor embedding) - стохастическое вложение соседей с t-распределением.

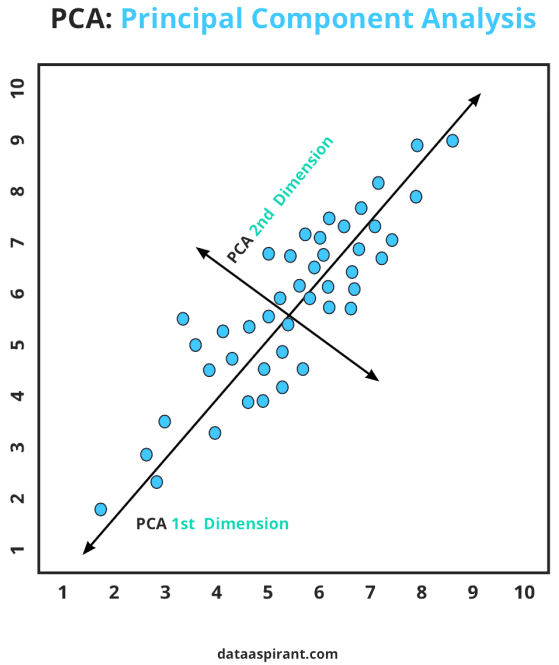
1. Считаем сходство точек в многомерном пространстве признаков на основе расстояний между ними.
2. Произвольным образом отображаем точки в более низкоразмерное пространство и считаем новую матрицу сходства.
3. Итеративно корректируем отображение, чтобы новая матрица был похожа похожа на исходную.
4. Гиперпараметр: перплексия - плотность соседей вокруг точки
5. Библиотека: sklearn.manifold, TSNE

2. Подходы к решению: снижение размерности

PCA (principal component analysis) - метод главных компонент

- Алгоритм отображает исходное пространство признаков в пространство меньшей размерности таким образом, чтобы потеря информации была минимальной

- Библиотека: sklearn.decomposition, PCA



3. Подходы к решению: инкрементальное обучение

Модель учится на небольшом количестве данных и постепенно получает новые, не переобучаясь с нуля.

Реализации в Scikit-learn

- Параметр Warm_start - позволяет сохранять состояние модели после ее обучения и использовать его в качестве начального при обучении на новых данных.
- Метод Partial_fit — позволяет обучать модель не на всех данных сразу, а по частям
- Применимость: SGDClassifier, SGDRegressor, PassiveAggressiveClassifier и PassiveAggressiveRegressor

CTR: Factorization Machines (FM)

Простейшие решения

- Регрессия на имеющихся признаках – не подходит: данные многомерны и разрежены, не учитывает взаимосвязь признаков (контекст)

- Константное решение на основе статистики ссылок

Factorization Machine (FM)

- Рекомендации на основе содержимого – учитываются признаки товаров, приобретённых пользователем (**content-based filtering**)

- Коллаборативная фильтрация – учитывается статистика приобретения товаров между похожими пользователями. Способ моделирования – матричная факторизация (**MF**)

Factorization Machine = content-based filtering + MF

Factorization Machine (FM)

Factorization Machine = content-based filtering + MF

$$\hat{y}(x) := \underbrace{w_0 + \sum_{i=0}^{n} w_i x_i}_{regression\ part} + \underbrace{\sum_{i=1}^{n} \sum_{j=i+1}^{n} \langle v_i, v_j \rangle x_i x_j}_{MF\ part} \quad w_0 \in \mathbb{R}; w \in \mathbb{R}^n; V \in \mathbb{R}^{n*k}; v_i, v_j \in \mathbb{R}^k$$

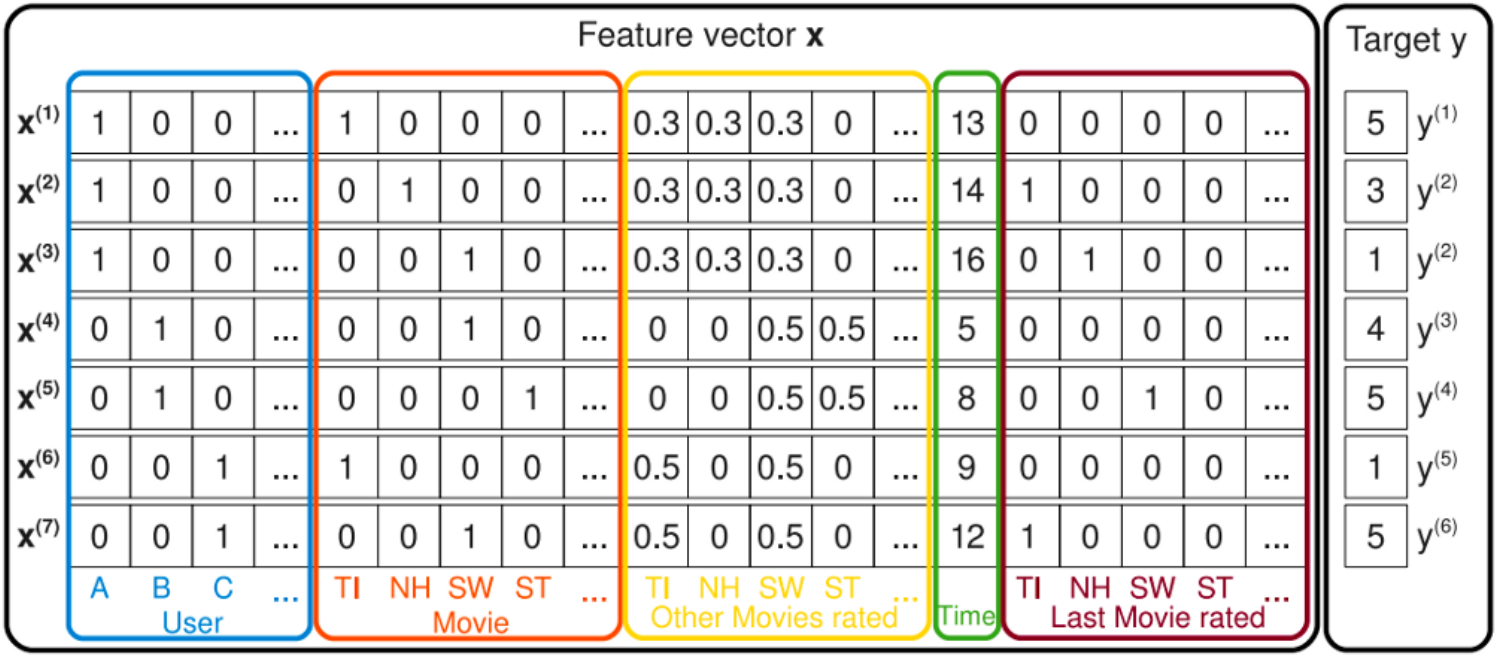
Factorization Machine (FM): пример

Предположим, у нас есть данные об обзорах фильмов, где пользователи ставят фильмам оценки в определенное время:

- Пользователь **u** из множества U = {Alice (A), Bob (B),...}
- Фильм **i** из множества I = {"Титаник" (TN), "Ноттинг Хилл" (NH), "Звездные Войны" (SW), "Стар Трек" (ST),...}
- Оценка **r** из {1, 2, 3, 4, 5}, поставленная во время **t**.

Factorization Machine (FM): пример

Получившийся вектор признаков:



Factorization Machine: можно ещё лучше

- FM
- FFM
- DeepFM

Ссылка на примеры ноутбуков с применением алгоритма:

Другие полезные ссылки

- Методы оптимизации алгоритмов
- Ещё примеры использования FM: здесь
- Пример использования алгоритмов PCA и T-Sne

ВСЕРОССИЙСКАЯ ОЛИМПИАДА ПО ИСКУССТВЕННОМУ ИНТЕЛЛЕКТУ

МИНИСТЕРСТВО ПРОСВЕЩЕНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ

ПРОСВЕТ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ПРОСВЕЩЕНИЯ

Задача: мультимодальное обучение

Мультимодальное обучение

Это обучение моделей сразу на нескольких источниках данных, имеющих разный формат, таких как текст, изображения, звук, видео и другие.

ВСЕРОССИЙСКАЯ ОЛИМПИАДА ПО ИСКУССТВЕННОМУ ИНТЕЛЛЕКТУ

МИНИСТЕРСТВО ПРОСВЕЩЕНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ

ПРОСВЕТ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ПРОСВЕЩЕНИЯ

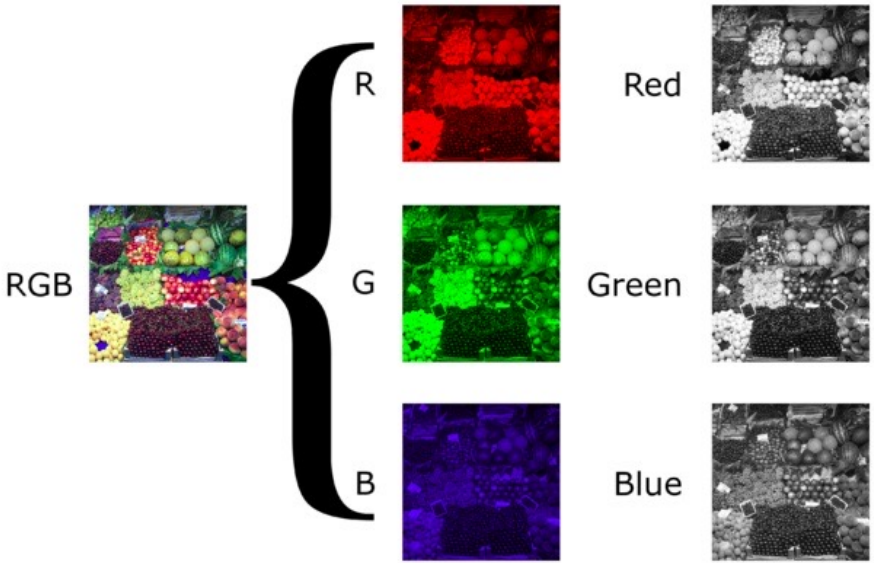
Задача: работа с изображениями

Обработка изображений: цели

1. Очистка данных
2. Дополнение данных (аугментация)
3. Изменение формата данных на подходящий для модели

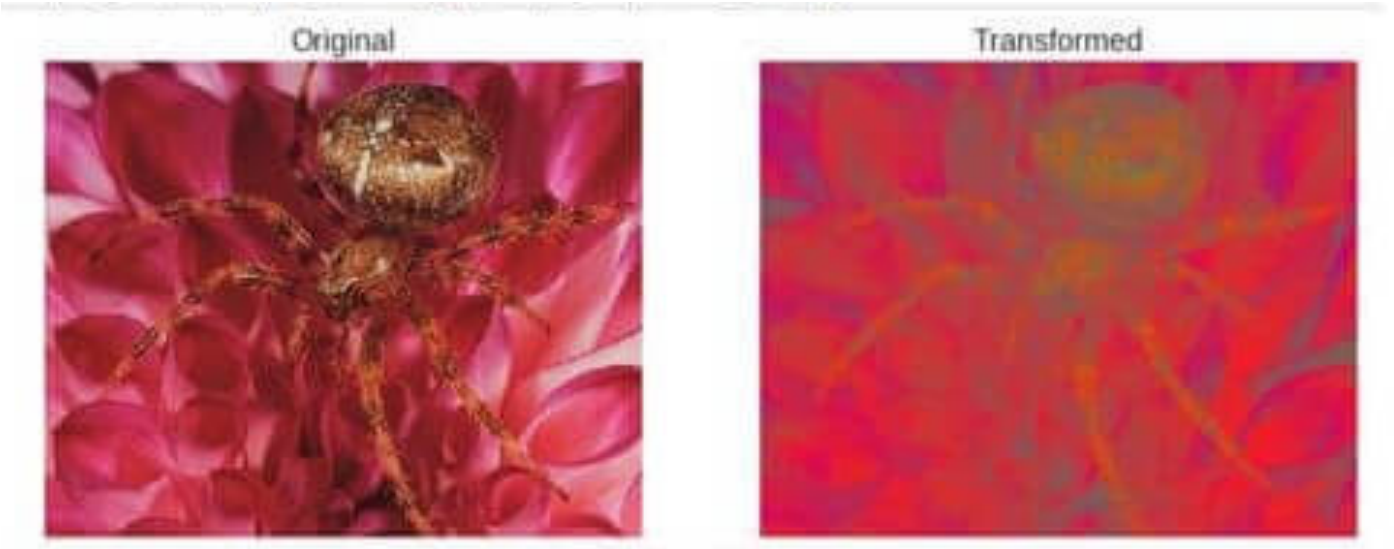
Перемасштабирование

Изменение диапазона для параметров пикселей.



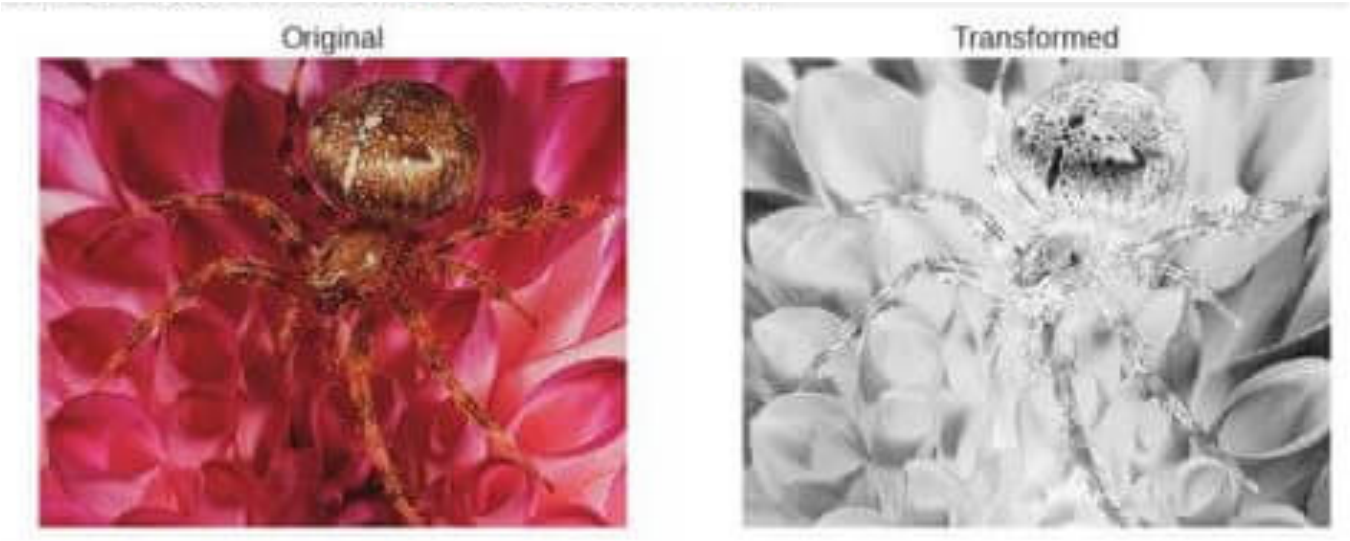
Центрирование

Изменение диапазона так, чтобы среднее равнялось 0.



Перевод в оттенки серого

Переход от цветного изображения к чёрно-белому.



Центрирование признаков

Приведение значений пикселей изображения к нормальному распределению.

Преобразования для аугментации

Аугментация – увеличение выборки данных через их модификацию.

- Вращение
- Горизонтальный и вертикальный сдвиг
- Обрезка
- Приближение и удаление
- Отображение по горизонтали или вертикали

ВСЕРОССИЙСКАЯ ОЛИМПИАДА ПО ИСКУССТВЕННОМУ ИНТЕЛЛЕКТУ

МИНИСТЕРСТВО ПРОСВЕЩЕНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ

ПРОСВЕТ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ПРОСВЕЩЕНИЯ

Задача: работа с текстом

Обработка текста

Существуют различные подходы к выделению данных из текстового формата:

- Bag of words
- TF-IDF
- Word Embeddings (word2vec, Global Vectors)

Word Embeddings

Данный подход позволяет предсказывать вероятность слова по его окружению – вероятность, присваиваемая моделью, близка к вероятности встретить его в таком контексте в реальном тексте.

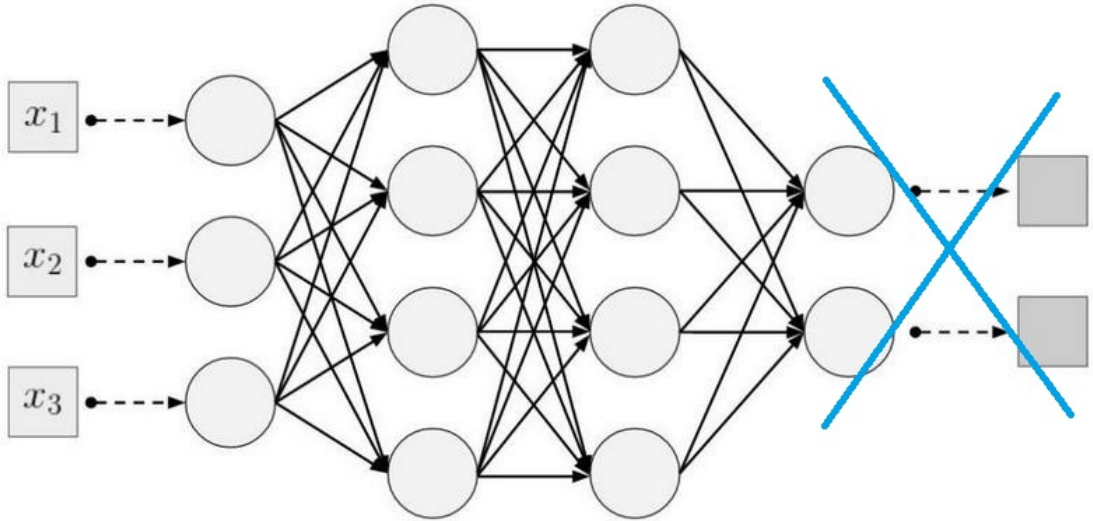
word2vec — способ построения множества векторов слов с помощью нейронной сети с использованием косинусного сходства.

Мультимодальное обучение: пример

1. Преобразование текста в вектор (word embeddings)
2. Преобразование изображения в вектор

Мультимодальное обучение: пример

1. Преобразование текста в вектор (word embeddings)
2. Преобразование изображения в вектор



Мультимодальное обучение: пример

1. Преобразование текста в вектор (word embeddings)
2. Преобразование изображения в вектор (нейросеть-классификатор)
3. Сравнение двух векторов на основе косинусного сходства

Мультимодальное обучение

Другие идеи:

- AutoGluon
- TorchMultimodal

Полезные ссылки

- Ноутбуки с использованием word2vec
- Пример использования AutoGluon
- Пример обучения нейросети с изображениями и текстом
- Обзор нейросетей для классификации

ВСЕРОССИЙСКАЯ ОЛИМПИАДА ПО ИСКУССТВЕННОМУ ИНТЕЛЛЕКТУ

МИНИСТЕРСТВО ПРОСВЕЩЕНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ

ПРОСВЕТ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ПРОСВЕЩЕНИЯ

Ответы на вопросы