

Здесь есть противоречия в кратком описании и формулировке задания. Стоит вернуть обработку текстов.

Задания на лабораторные работы по курсу “Нейро-нечеткие системы”

Общие положения

Для выполнения лабораторных работ вам потребуется выбрать некоторый набор данных. Можно использовать собственный набор данных, можно взять из одного из источников:

<https://www.kaggle.com/datasets>

<https://datasetsearch.research.google.com/>

<https://data.worldbank.org/>

<https://rosstat.gov.ru/opendata/>

и т.д.

Набор данных должен включать как минимум 5 параметров. Для лабораторной работы №2 вам потребуется размеченная целевая переменная. Для лабораторной работы № 3 вам потребуется новый набор данных, содержащих в себе тексты (могут быть загружены из открытых источников СМИ, фанфиков, отчетов, ..., либо взяты из одной из открытых коллекций текстов). Лабораторная работа № 4 может выполняться либо на новом наборе данных, либо на одном из предыдущих.

Лабораторная работа зачтена при разумном выполнении всех требований, описанных ниже. Реализовывать описанные методы не нужно, можно использовать любые библиотеки и любой язык программирования.

Лабораторная работа №1

Визуализация данных

Для выбранного набора данных необходимо провести его визуализацию. Следует сделать упор на выделение структуры данных, описании поведения параметров или совокупности параметров, возможности разделения данных на подгруппы. Необходимо провести осознанное преобразование признаков: изменение шкалы, преобразование к числовым значениям. Требуется отобразить данные с использованием методов PCA, MDS, t-SNE или UMAP или аналогичных. Требуется отобразить максимальный объем данных без использования методов снижения размерности пространства.

Лабораторная работа №2

Классификация, оценка точности классификации

Для выбранного набора данных необходимо провести сравнение результатов работы классификаторов из двух групп на выбор:

- регрессия (линейная, нелинейная, логистическая, метод опорных векторов), k-ближайших соседей, деревья принятия решений (деревья, случайный лес) (один из методов на выбор);

- методы с использованием стекинга, блендинга или бустинга (один из методов на выбор).

Для каждого метода необходимо построить матрицу ошибок и рассчитать одну из следующих метрик: f -мера, ROC AUC, ассигасу.

При обучении классификатора необходимо использовать кроссвалидацию. Необходимо визуализировать изменение точности работы метода на разных шагах кроссвалидации. Необходимо показать как меняется точность классификации при изменении гиперпараметров.

Лабораторная работа №3

Кластеризация, снижение размерности пространства признаков

Для выбранного набора данных необходимо провести кластеризацию при помощи двух методов, взятых из следующих разных групп: метода k -средних (или альтернативного из группы, основанной на расстояниях между точками), метода DBSCAN (или альтернативного из группы, основанной на плотности точек) или одним из методов иерархической кластеризации. Для метода из первой и третьей групп необходимо выбрать количество кластеров с использованием одного из методов (метод локтя, индекс Дана, индекс Дэвиса-Болдуина, индекс Калинского-Гарабача, метод силуэта, RAND - обязателен при наличии целевой переменной, кофенетическая корреляция). Для методов из второй группы необходимо выбрать ϵ и другие параметры метода.

Лабораторная работа №4

Плотные нейронные сети

Необходимо обучить персептрон на одном из использованных выше наборах данных и сравнить точность его работы с использованными ранее методами (см. л/р №2).

Лабораторная работа №5 (бонусная)

Глубинные нейронные сети

Необходимо обучить сверточную или рекуррентную нейронную сеть на одном из использованных выше наборах данных и сравнить точность его работы с использованными ранее методами (см. л/р №2).