**Дипломная работа.**

**Сравнение различных библиотек для машинного обучения: scikit-learn, TensorFlow и PyTorch: Реализовать задачи классификации и регрессии с использованием scikit-learn, TensorFlow и PyTorch, сравнить их производительность и удобство использования.**

*Urban Univercity 2024.*

**Содержание дипломной работы.**

1. **Введение.**
2. **Где используется машинное обучение?**
3. **Что такое машинное обучение?**
4. **Что включают в себя библиотеки scikit-learn, TensorFlow и PyTorch?**
5. Примеры классификации и регрессии с использованием этих библиотек.
6. Сравнение Библиотек.
7. Рекомендации по использованию.

1.Введение .

Мы живем в эпоху научно технического прогресса, люди желая облегчить свой труд используют машины, различные компьютерные программы. Также изобретён Искусственный Интеллект (ИИ) который помогает во многих областях жизни.

Для того чтобы машины выполняли функции людей их нужно обучить. Для этого нужны специалисты по машинному обучению. В каких областях может найти себя такой специалист?

2.Где используется Машинное обучение?

1. **Финтех: кредитный скоринг, защита клиентов, поиск мошенников**

В финтехе и банках очень много решений сейчас принимается с помощью машинного обучения. Алгоритм помогает, например, решать задачи кредитного скоринга. Обрабатывая данные о клиенте из анкеты и открытых источников, система делает вывод о том, стоит выдавать кредит клиенту или нет. Это помогает менеджерам принимать решения и снижать риски для банка.

Также ML-модели могут отслеживать платежи и находить среди них мошеннические, когда кто-то украл карту клиента и пытается вывести с нее деньги. Машинное обучение помогает банкам искать клиентов в открытых источниках, например в реестрах юрлиц, отслеживать их отток, понимать, что клиент собирается уйти, и дает рекомендации, чтобы его удержать.

2. **Медицина: диагностика болезней на ранней стадии**

Диагностика — одно из важных направлений машинного обучения. Сейчас законодательно все устроено так, что поставить пациенту диагноз и назначить лечение может только человек. Все, что разрабатывается в сфере машинного обучения, — это системы помощи врачу.

Чаще всего для анализа при помощи моделей используются медицинские изображения: МРТ, КТ, флюорография. В текстовом анализе, как правило, написано число по определенному параметру (например сахар в крови) и границы допустимых значений. Такой результат может прочитать любой врач, они понимают, что означает выход за границу допустимых значений. С медицинскими изображениями все сложнее. Безусловно, технически мы можем сделать идеальное изображение всего, что у человека есть внутри. Но для этого нужно хорошо просветить тело радиацией, и человеку от этого будет очень плохо. Если же мы делаем изображения, которые не вредят здоровью, то качественную картинку получить невозможно. Она чаще всего зашумленная, и на ней сложно что-либо различить. Чтобы работать с такими изображениями, врачи специально учатся диагностике. К сожалению, специалистов, которые могут читать медицинские изображения, не хватает. Многие больницы сейчас покупают качественное оборудование, но врачей, которые могут интерпретировать результаты, просто нет.

Чем здесь могут помочь системы помощи медицинским специалистам? При помощи алгоритмов компьютерного зрения они анализируют изображение и выделяют области интереса: есть ли на картинке что-то нетипичное, на что стоит обратить внимание. Проанализировав область интереса, модель может определить, например, что на изображении — с высокой вероятностью туберкулез. После этого анализа изображение смотрит врач и подтверждает или опровергает диагноз, отправляет пациента на дообследование. Актуальное направление разработки сейчас — обследования, которые помогают видеть заболевания на ранней стадии развития, когда врач может еще не замечать их на медицинском изображении.

3. **Промышленность: контроль качества**

В промышленности очень много данных и очень много задач для алгоритмов: безопасность на производстве, автоматизация, контроль качества продукции. Трубы должны быть качественными, потому что, если их прорвет, это будет экологическая катастрофа. Ущерб от такого происшествия несопоставим со стоимостью трубы, поэтому для ТМК очень важно делать продукт высокого качества.

Вопрос в том, как это качество контролировать. На одном из этапов трубу просвечивают рентгеновским излучением. Снимок позволяет увидеть внутренние дефекты. Чтобы проконтролировать качество, нужен специалист, который будет сидеть и постоянно смотреть на эти изображения. Понятно, что человек не может проводить такой контроль стабильно качественно. Нужен прототип системы компьютерного зрения, которая отслеживает дефекты и сообщает о них сотруднику.

4. **Автомобили без водителей**

На сегодня в России законодательно разрешен проезд автомобилей без водителя в Москве, Московской области и Татарстане. По Иннополису уже ездят Яндекс. Такси без водителя. Проблема этого направления в том, что цена ошибки тут очень высока. Когда разработчики научатся решать проблемы, беспилотники появятся на улицах всех городов.

КАМАЗ занимается разработкой грузовиков без водителя, Россельмаш — беспилотных комбайнов, а Уралмаш начал думать о том, чтобы производить экскаваторы без водителя. Это очень перспективное направление.

5. **Фармакология: проектирование лекарств**

Один из последних прорывов машинного обучения в науках о жизни — при помощи нейронных сетей научились определять трехмерную структуру белка. Это упрощает процесс разработки новых лекарств. Например, когда мы расшифровываем геном какого-либо вируса, получаем одномерную последовательность, в то время как в реальности белки трехмерные. Чтобы определить трехмерную структуру генома, требуются длительные, на несколько месяцев эксперименты. А для того чтобы получить одно лекарство, таких экспериментов нужно несколько. Это долго и дорого.

Выяснилось, что при помощи нейросетей можно получить трехмерную структуру белка и определить, будет ли он взаимодействовать с вирусом нужным образом. Такую модель создала компания DeepMind. Для разработки алгоритма AlphaFold создатели использовали информацию о структурах 170 тысяч белков.

Эта разработка позволит быстрее проектировать лекарства и вакцины и с высокой точностью прогнозировать их эффективность.

6. **Информационная безопасность**

Как сейчас работают системы обнаружения вторжений? По так называемой сигнатуре атаки: когда вторжение произошло, производители антивирусов изучают, какие действия были произведены. Их записывают в сценарий антивируса: если происходят вот такие действия, то это атака вот такого типа и нужно ее остановить. Но это означает, что нет возможности остановить новые атаки — такие, которые раньше не происходили. Машинное обучение в информбезопасности используется как раз, чтобы обнаруживать атаки нового вида, которых раньше никто никогда не видел.

7.**Безопастность и охрана людей**

Система определения лиц и слежка за потенциальными преступниками. Очень хорошее подспорье для служб безопасности. Так же сюда можно добавить распознавание номеров автомобилей и моделей автомобилей.

Но для того чтобы машины выполняли эти действия их нужно обучить.

Для этого необходимо машинное обучение. Что это такое?

**3.Что такое машинное обучение?**

Из источника Urban University

Машинное обучение (ML) — это область искусственного интеллекта, которая занимается разработкой алгоритмов и статистических моделей, позволяющих компьютерам выполнять конкретные задачи без явного программирования на выполнение этих задач. Основная идея состоит в том, чтобы создать модель, способную "учиться" на данных и делать прогнозы или принимать решения на основе этого обучения.

Основные принципы машинного обучения

1. Обучение на данных: Модели машинного обучения обучаются на наборе данных. Этот процесс включает в себя использование алгоритмов, которые анализируют данные и выявляют в них закономерности.
2. Обобщение: После обучения на данных модель должна быть способна обобщать свои знания на новые, невиданные ранее данные. Это означает, что она должна давать правильные ответы не только на обучающих данных, но и на любых других данных из той же области.
3. Итеративное улучшение: Модели машинного обучения часто улучшаются итеративно. Это включает в себя оценку модели, выявление её недостатков и последующую настройку параметров или алгоритма для повышения точности.

**Алгоритмы машинного обучения**

Машинное обучение делится на три основных типа, каждый из которых включает в себя множество алгоритмов:

1. Обучение с учителем (**Supervised Learning**): Модель обучается на размеченных данных, где каждой входной информации соответствует правильный выход. Примеры алгоритмов: линейная регрессия, логистическая регрессия, деревья решений, случайный лес (Random Forest), поддерживающие векторы (SVM), нейронные сети.
2. Обучение без учителя (**Unsupervised Learning**): Модель обучается на неразмеченных данных и пытается выявить скрытые структуры. Примеры алгоритмов: кластеризация (например, k-средних), ассоциативные правила, метод главных компонент (PCA).
3. Обучение с подкреплением (**Reinforcement Learning**): Модель обучается на основе взаимодействия с окружающей средой, получая вознаграждения или наказания за свои действия. Примеры алгоритмов: Q-learning, SARSA, глубокое обучение с подкреплением.

**Этапы работы с данными**

1. Сбор данных: Первый этап включает сбор данных из различных источников. Данные могут быть получены из баз данных, файлов, веб-скрапинга, API и других источников.
2. Очистка данных: На этом этапе данные очищаются от шума и ошибок. Это может включать удаление или исправление пропущенных значений, обработку выбросов, устранение дубликатов и исправление ошибок в данных.
3. Подготовка данных: Включает в себя приведение данных к такому виду, который может быть использован моделью. Это может включать нормализацию или стандартизацию данных, кодирование категориальных переменных, разбиение данных на обучающий и тестовый наборы.
4. Анализ данных: Предварительный анализ данных для понимания их структуры и свойств. Этот этап может включать визуализацию данных, вычисление статистических характеристик и выявление корреляций между переменными.

Для любой машины нужна программа для машинного обучения тоже.Для того чтобы облегчить процесс написания программы были созданы библиотеки: **scikit-learn, TensorFlow и PyTorch**

**4.Что включают в себя библиотеки scikit-learn, TensorFlow и PyTorch?**

**Scikit-learn (sklearn)** — это один из наиболее широко используемых пакетов Python для Data Science и Machine Learning. Он содержит функции и алгоритмы для машинного обучения: классификации, прогнозирования или разбивки данных на группы. Sklearn написана на языках Python, C, C++ и Cython.

**TensorFlow** — открытая программная библиотека для машинного обучения, разработанная компанией Google для решения задач построения и тренировки нейронной сети с целью автоматического нахождения и классификации образов, достигая качества человеческого восприятия. Применяется как для исследований, так и для разработки собственных продуктов Google. Основной API для работы с библиотекой реализован для Python, также существуют реализации для R, C#, C++, Haskell, Java, Go, JavaScript и Swift.

Является продолжением закрытого проекта DistBelief. Изначально TensorFlow была разработана командой Google Brain для внутреннего использования в Google, в 2015 году система была переведена в свободный доступ с открытой лицензией Apache 2.0.

**PyTorch** — это фреймворк для языка программирования Python, предназначенный для машинного обучения. Он включает в себя набор инструментов для работы с моделями, используется в обработке естественного языка, компьютерном зрении и других похожих направлениях.

Мы сравним эти библиотеки и расскажем плюсы и минусы.

Также мы проанализируем код который написан с использованием этих библиотек.

5.Реализация задач классификации и регрессии с использованием **scikit-learn, TensorFlow и PyTorch**, сравнить их производительность и удобство использования.

Для реализации задач классификации и регрессии с использованием библиотек scikit-learn, TensorFlow и PyTorch, мы можем создать примеры на основе простых наборов данных, таких как `Iris` для классификации и `Boston Housing` для регрессии. После этого мы можем сравнить производительность и удобство использования каждой библиотеки.

### Задача 1: Классификация с использованием scikit-learn, TensorFlow и PyTorch

#### 1. Классификация с использованием scikit-learn

python

# Импорт необходимых библиотек

import numpy as np

from sklearn import datasets

from sklearn.model\_selection import train\_test\_split

from sklearn.ensemble import RandomForestClassifier

from sklearn.metrics import accuracy\_score

# Загрузка датасета Iris

iris = datasets.load\_iris()

X, y = iris.data, iris.target

# Разделение данных на обучающую и тестовую выборки

X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size=0.2, random\_state=42)

# Обучение модели

clf = RandomForestClassifier(n\_estimators=100)

clf.fit(X\_train, y\_train)

# Предсказания

y\_pred = clf.predict(X\_test)

# Оценка точности

accuracy = accuracy\_score(y\_test, y\_pred)

print(f'Accuracy (scikit-learn): {accuracy:.2f}')

#### 2. Классификация с использованием TensorFlow

python

import tensorflow as tf

from sklearn.preprocessing import OneHotEncoder

# Подготовка данных

X\_train\_tf, X\_test\_tf = tf.convert\_to\_tensor(X\_train, dtype=tf.float32), tf.convert\_to\_tensor(X\_test, dtype=tf.float32)

y\_train\_tf = y\_train.reshape(-1, 1)

y\_test\_tf = y\_test.reshape(-1, 1)

# One-hot encoding

encoder = OneHotEncoder(sparse=False)

y\_train\_tf = encoder.fit\_transform(y\_train\_tf)

y\_test\_tf = encoder.transform(y\_test\_tf)

# Создание модели

model = tf.keras.Sequential([

tf.keras.layers.Dense(10, activation='relu', input\_shape=(X\_train\_tf.shape[1],)),

tf.keras.layers.Dense(3, activation='softmax')

])

# Компиляция модели

model.compile(optimizer='adam', loss='categorical\_crossentropy', metrics=['accuracy'])

# Обучение модели

model.fit(X\_train\_tf, y\_train\_tf, epochs=100, verbose=0)

# Оценка модели

loss, accuracy = model.evaluate(X\_test\_tf, y\_test\_tf)

print(f'Accuracy (TensorFlow): {accuracy:.2f}')

#### 3. Классификация с использованием PyTorch

python

import torch

import torch.nn as nn

import torch.optim as optim

from torch.utils.data import DataLoader, TensorDataset

# Преобразование данных в тензоры

X\_train\_pt = torch.FloatTensor(X\_train)

X\_test\_pt = torch.FloatTensor(X\_test)

y\_train\_pt = torch.LongTensor(y\_train)

# Создание датасета и загрузчика

train\_dataset = TensorDataset(X\_train\_pt, y\_train\_pt)

train\_loader = DataLoader(train\_dataset, batch\_size=16, shuffle=True)

# Определение модели

class SimpleNN(nn.Module):

def \_\_init\_\_(self):

super(SimpleNN, self).\_\_init\_\_()

self.hidden = nn.Linear(X\_train.shape[1], 10)

self.output = nn.Linear(10, 3)

def forward(self, x):

x = torch.relu(self.hidden(x))

x = self.output(x)

return x

# Инициализация модели, оптимизатора и функции потерь

model\_pt = SimpleNN()

optimizer = optim.Adam(model\_pt.parameters(), lr=0.01)

criterion = nn.CrossEntropyLoss()

# Обучение модели

for epoch in range(100):

for batch\_X, batch\_y in train\_loader:

optimizer.zero\_grad()

outputs = model\_pt(batch\_X)

loss = criterion(outputs, batch\_y)

loss.backward()

optimizer.step()

# Оценка модели

with torch.no\_grad():

y\_pred\_pt = model\_pt(X\_test\_pt)

\_, predicted = torch.max(y\_pred\_pt, 1)

accuracy = (predicted.numpy() == y\_test).mean()

print(f'Accuracy (PyTorch): {accuracy:.2f}')

### Задача 2: Регрессия с использованием scikit-learn, TensorFlow и PyTorch

#### 1. Регрессия с использованием scikit-learn

python

from sklearn.datasets import load\_boston

from sklearn.linear\_model import LinearRegression

from sklearn.metrics import mean\_squared\_error

# Загрузка датасета

boston = load\_boston()

X\_reg, y\_reg = boston.data, boston.target

# Разделение данных

X\_train\_reg, X\_test\_reg, y\_train\_reg, y\_test\_reg = train\_test\_split(X\_reg, y\_reg, test\_size=0.2, random\_state=42)

# Обучение модели

reg\_model = LinearRegression()

reg\_model.fit(X\_train\_reg, y\_train\_reg)

# Предсказания и оценка

y\_pred\_reg = reg\_model.predict(X\_test\_reg)

mse = mean\_squared\_error(y\_test\_reg, y\_pred\_reg)

print(f'Mean Squared Error (scikit-learn): {mse:.2f}')

#### 2. Регрессия с использованием TensorFlow

python

# Подготовка данных

X\_train\_tf\_reg = tf.convert\_to\_tensor(X\_train\_reg, dtype=tf.float32)

y\_train\_tf\_reg = tf.convert\_to\_tensor(y\_train\_reg, dtype=tf.float32)

X\_test\_tf\_reg = tf.convert\_to\_tensor(X\_test\_reg, dtype=tf.float32)

y\_test\_tf\_reg = tf.convert\_to\_tensor(y\_test\_reg, dtype=tf.float32)

# Создание модели

model\_reg = tf.keras.Sequential([

tf.keras.layers.Dense(64, activation='relu', input\_shape=(X\_train\_reg.shape[1],)),

tf.keras.layers.Dense(1)

])

model\_reg.compile(optimizer='adam', loss='mean\_squared\_error')

# Обучение модели

model\_reg.fit(X\_train\_tf\_reg, y\_train\_tf\_reg, epochs=100, verbose=0)

# Оценка модели

mse\_reg = model\_reg.evaluate(X\_test\_tf\_reg, y\_test\_tf\_reg)

print(f'Mean Squared Error (TensorFlow): {mse\_reg:.2f}')

#### 3. Регрессия с использованием PyTorch

python

# Преобразование данных в тензоры

X\_train\_pt\_reg = torch.FloatTensor(X\_train\_reg)

y\_train\_pt\_reg = torch.FloatTensor(y\_train\_reg).view(-1, 1)

X\_test\_pt\_reg = torch.FloatTensor(X\_test\_reg)

y\_test\_pt\_reg = torch.FloatTensor(y\_test\_reg).view(-1, 1)

# Создание датасета и загрузчика

train\_dataset\_reg = TensorDataset(X\_train\_pt\_reg, y\_train\_pt\_reg)

train\_loader\_reg = DataLoader(train\_dataset\_reg, batch\_size=16, shuffle=True)

# Определение модели

class LinearRegressionModel(nn.Module):

def \_\_init\_\_(self):

super(LinearRegressionModel, self).\_\_init\_\_()

self.linear = nn.Linear(X\_train\_reg.shape[1], 1)

def forward(self, x):

return self.linear(x)

# Инициализация модели, оптимизатора и функции потерь

model\_reg\_pt = LinearRegressionModel()

optimizer\_reg = optim.Adam(model\_reg\_pt.parameters(), lr=0.01)

criterion\_reg = nn.MSELoss()

# Обучение модели

for epoch in range(100):

for batch\_X, batch\_y in train\_loader\_reg:

optimizer\_reg.zero\_grad()

outputs = model\_reg\_pt(batch\_X)

loss = criterion\_reg(outputs, batch\_y)

loss.backward()

optimizer\_reg.step()

# Оценка модели

with torch.no\_grad():

y\_pred\_pt\_reg = model\_reg\_pt(X\_test\_pt\_reg)

mse\_reg\_pt = mean\_squared\_error(y\_test\_reg, y\_pred\_pt\_reg.numpy())

print(f'Mean Squared Error (PyTorch): {mse\_reg\_pt:.2f}')

6. Сравнение их производительности и удобства использования.

1. Производительность:

- Все модели могут показать конкурентоспособные результаты на простых задачах, но более сложные модели и большие наборы данных могут требовать больше времени для обучения в TensorFlow и PyTorch по сравнению с scikit-learn.

- TensorFlow и PyTorch предоставляют более мощные инструменты для настройки архитектуры модели, что позволяет достигать лучшей производительности на сложных задачах.

2. Удобство использования:

- scikit-learn: Отличается простым и интуитивно понятным интерфейсом. Идеально подходит для быстрого прототипирования и работы с классическими алгоритмами машинного обучения.

- TensorFlow: Обеспечивает высокий уровень гибкости для создания сложных моделей глубокого обучения, но требует больше кода и времени для настройки.

- PyTorch: Более удобен в использовании, особенно для исследования, благодаря динамической вычислительной графике и гибкости. Он также упрощает отладку.

Scikit-learn против TensorFlow

Scikit-learn — широко используемая библиотека машинного обучения с открытым исходным кодом для Python. Она построена на основе и интегрируется с такими часто используемыми библиотеками, как NumPy, SciPy, Matplotlib и pandas, что делает ее доступной и универсальной.

TensorFlow , также библиотека машинного обучения с открытым исходным кодом, специализируется на глубоком обучении и нейронных сетях. TensorFlow поддерживает несколько языков программирования, таких как Python, C/C++, Java и Javascript, и другие.

Рассмотрите scikit-learn, если вы новичок в машинном обучении или разрабатываете что-то с использованием алгоритмов, не основанных на нейронных сетях.

Рассмотрите TensorFlow, если вы хотите использовать подход глубокого обучения в сочетании с аппаратным ускорением с помощью графических процессоров и TPU или на кластере компьютеров (что scikit-learn изначально не поддерживает).

PyTorch против scikit-learn

PyTorch — это библиотека программного обеспечения глубокого обучения для Python, C++ и Julia. PyTorch в основном используется для сквозного построения и обучения глубоких нейронных сетей с возможностью создания пользовательских моделей и алгоритмов обучения.

Scikit-learn — это библиотека для традиционных алгоритмов машинного обучения, используемых для кластеризации, классификации, регрессии и т. д. Природа Scikit-learn как «черного ящика» делает ее более доступной для тех, кто относительно недавно начал заниматься машинным обучением.

Рассмотрите PyTorch, если вы разрабатываете приложения, в которых есть задачи, требующие больших вычислительных затрат, такие как обработка естественного языка, компьютерное зрение и т. д. С Pytorch вы также можете использовать ускорение графического процессора в своих интересах.

Рассмотрите scikit-learn, если вы разрабатываете небольшой исследовательский проект, не требующий значительного объема данных. С scikit-learn вы будете фокусироваться не на настройке, а на скорости и удобстве использования алгоритмов машинного обучения.

PyTorch против TensorFlow

PyTorch — это фреймворк глубокого обучения с питоническим и объектно-ориентированным подходом. PyTorch имеет больше возможностей отладки и тестирования, чем TensorFlow.

TensorFlow — это низкоуровневая библиотека глубокого обучения, которая обеспечивает рабочие процессы для высокоуровневых API, таких как Keras, хотя и с меньшей вычислительной мощностью. TensorFlow в настоящее время используется более широко, чем PyTorch.

Если Python играет центральную роль в вашей разработке, рассмотрите PyTorch с его многочисленными возможностями отладки.

Рассмотрите TensorFlow, если вам нужна библиотека, совместимая с различными языками программирования, такими как C/C++, Java, JavaScript, Go и т. д. TensorFlow также имеет обширную поддержку нескольких платформ.

1. Заключение

Сценарий выбора библиотеки зависит от задачи. Если вы ищете простоту и скорость, scikit-learn может быть лучшим выбором для традиционных методов. Если ваши задачи связаны с глубоким обучением, TensorFlow и PyTorch предпочтительнее, причём PyTorch может быть удобнее на этапе prototyping и исследований.