Сколковский институт науки и технологий (Сколтех)

Рабочая программа дисциплины	**Введение в анализ данных**

Преподаватель	Панов Максим Евгеньевич, кандидат физико-математических наук

Аннотация

Описание курса
Курс дает введение в основные темы современного анализа данных, такие как классификация, регрессия, кластеризация и снижение размерности. Каждая тема сопровождается обзором ключевых алгоритмов машинного обучения, решающих данную задачу, и проиллюстрирована набором примеров. Основная цель курса - дать широкий обзор основных методов машинного обучения. Особое внимание уделяется современным библиотекам анализа данных, которые позволяют эффективно решать указанные выше задачи.

1. Основная информация

Академический уровень курса	Магистратура Аспирантура
Количество кредитов	3

Предварительные требования к курсу / рекомендации

Линейная алгебра, математический анализ, алгоритмы.
Необходимы навыки программирования как минимум на среднем уровне! Во время курса вы будете писать простые программы на Python, подобные этой. http://scikit-learn.org/stable/auto_examples/plot_cv_predict.html

Тип оценки - дифференцированная

Отображение оценок в процентах

A:	85
B:	75
C:	55
D:	40
E:	25
F:	0

2. Содержание курса

Тема	Краткое содержание	Лекции (час)	Семинары (час)	Лабораторные занятия (час)
Общее введение	Определение науки о данных, реальные примеры применения науки о данных, обзор основных тем машинного обучения.	4		
Решение задач машинного обучения на Python	Почему Python, обзор Библиотеки Python: пакет scikit-учиться, панды, Сиборн, визуальное исследование. Практический пример: изучения Набор данных "Титаника"	1	1	1
Элементы многомерной статистики	Многомерное, нормальное, условно нормальное, Wishart распределения; Дисперсионный анализ; Многомерный дисперсионный анализ; Коррекция при многократном тестировании; Гистограммы; Плотность ядра Оценка	1	1	1
Регрессия, перекрестная проверка	Контролируемое обучение, k ближайших соседей, линейная регрессия, регуляризация L1 и L2, концепции переоснащения и недооснащения (смещение-дисперсия Компромисс). Практический пример: набор данных о спросе на совместное использование велосипедов	1	1	1
Классификация, показатели качества	Задачи классификации, логистическая регрессия, SVM, функции потерь, точность и отзыв, ROC-кривая. Практический пример: набор данных Titanic (продолжение)	1	1	1
Деревья принятия решений	Обзор, обработка пропущенных значений, расчет важности объектов, сложности алгоритмов, визуализация. Практический пример: набор данных Iris	1	1	1
Создание ансамбля	Bagging, Boosting, Random Forest, Gradient Boosting, Библиотека XGBoost. Практический пример: Тип лесного покрова Предсказание	1	1	1
Особенности проектирования и выбора	Подходы к выбору объектов: оболочки, фильтры, встроенные методы; категориальные объекты, текстовые объекты, временные ряды.	1	1	1

	Практический пример: Амазонка Доступ сотрудников			
Размерность Сокращение	Главный Компонент Анализ, обзор Нлинейных методов (Isomap), LTSA, tSNE. Практические примеры: DR для аэродинамических профилей и создание новых аэродинамических профилей, генетическая характеристика Еврейского происхождения	1	1	1
Кластеризация	K-средние, Gaussian Mixture Модель, Иерархическая кластеризация Спектральная кластеризация. Практический пример: кластеризация текстовых документов	1	1	1
Основы нейронной сети	Стохастический Градиентный спуск, Многослойный персептрон, функции активации (ReLU, tanh), отсев, обучение и валидация. Рано останавливающиеся сверточные сети; библиотека Keras. Практический пример: проблемы с игрушками, распознавание ключевых точек лица	1	1	1
Масштабируемые алгоритмы	Парадигма MapReduce; Коллаборативная фильтрация. Практический пример: Netflix	1	1	1

3. Результаты обучения

Результаты обучения в Сколтехе указаны в соответствии со структурой **результатов обучения в Сколтехе**

Знания
Формулировки всех основных проблем машинного обучения. **Математические подробности о наиболее важных методах и алгоритмах анализа данных.**

Навыки
Выберите подходящий метод для решения конкретных задач анализа данных. **Выполните базовую обработку данных и визуальный анализ, создайте функции для последующего машинного обучения.** **Примените библиотеки машинного обучения, выберите гиперпараметры алгоритма.** **Критически оцените полученные результаты и перепроектируйте конвейеры обработки данных.**

Опыт
Решайте реальные задачи в области обработки данных, используя современные методы машинного обучения.

4. Задания и выставление оценок

Тип назначения	Краткое содержание задания	% от итоговой оценки за курс
Домашние задания	Два домашних задания. Домашнее задание 1 охватывает лекции 1-4, домашнее задание 2 охватывает лекции 5-8.	70
Командный проект	В течение семестра студенты предлагают командный проект и представляют его в заключительной презентации.	30

5. Критерии оценки

Задание 1 Типа	Домашние задания

Пример задания 1

Домашнее задание:
1. Реализуйте метод k ближайших соседей на Python.
2. Оцените смещение и дисперсию как функцию размера окрестности.
3. Оцените качество прогнозирования kNN в двух сценариях: а) данные используются как есть, б) данные нормализуются заранее.

Критерии оценки для задания 1

1) Общая грамотность и стиль изложения — 20%;
2) Правильность применения метода - 30%;
3) Правильность процедуры оценки эффективности и экспериментов — 45%;
4) Выводы — 15%

Задание 2 Типа	Финальный проект

Пример задания 2

Глубокий анализ реальной задачи, связанной с обработкой данных: классификация походов за покупками на основе анализа потребительской корзины.
Выполните следующий анализ:
1. Проанализируйте данные из файла.
2. Выполните визуальный анализ данных.
3. Создайте процедуру перекрестной проверки.
4. Предложите и оцените несколько методов генерации признаков, основанных на особых характеристиках набора данных.
5. Сравните алгоритмы классификации (включая различные наборы гиперпараметров).
6. Оцените эффективность наилучшей модели с точки зрения бизнеса.

Основные критерии оценки:
1) Общая грамотность и стиль доклада — 10%;

2) Методы и подходы в области науки о данных — 20%;
3) Глубина понимания предмета — 45%;
4) Презентация и ответы на вопросы — 25%.

6. Учебники и интернет-ресурсы

Необходимые учебники	ISBN-13 (or ISBN-10)
The Elements of Statistical Learning, 2nd edition by Hastie, Tibshirani and Friedman, Springer-Verlag, 2008	9780387848570
Pattern Recognition and Machine Learning by Bishop, Springer, 2006	9780387310732
Рекомендуемые учебники	
Machine Learning: A Probabilistic Perspective by Kevin P. Murphy, MIT Press, 2012.	9780262018029
Bayesian Reasoning and Machine Learning by David Barber, Cambridge University Press, 2012.	9780521518147
Deep Learning by Yoshua Bengio, Ian Goodfellow, and Aaron Courville.	9780262035613

Веб-ресурсы (ссылки)	Описание
http://scipy-lectures.org	Tutorials on the scientific Python ecosystem.

7. Оборудование

Программное обеспечение
Python 3.10+

Оборудование
Ноутбук с предустановленным python