

Белорусский государственный университет



Проректор по учебной работе и образовательным инновациям
О.Н.Здрок

«20» июня 2020__ г.

Регистрационный № УД-8895/уч.

Машинное обучение и анализ данных

**Учебная программа учреждения высшего образования
по учебной дисциплине для специальности:**

1-31 80 03 Математика и компьютерные науки

Профилизация: Математика

2020_ г.

Учебная программа составлена на основе ОСВО 1- 31 80 03-2019 и учебных планов рег. № G31-017/уч., G31з-018/уч. от 11.04.2019.

СОСТАВИТЕЛИ:

Радыно Евгений Мефодьевич, доцент кафедры функционального анализа и аналитической экономики Белорусского государственного университета, кандидат физико-математических наук.

РЕЦЕНЗЕНТЫ:

Пыжкова Ольга Николаевна, заведующий кафедрой высшей математики Учреждения образования «Белорусский государственный технологический университет», кандидат физико-математических наук

РЕКОМЕНДОВАНА К УТВЕРЖДЕНИЮ:

Кафедрой Функционального анализа и аналитической экономики
(протокол № 12 от 04.06.2020)

Научно-методическим Советом БГУ
(протокол № 5 от 17.06.2020)

Заведующий кафедрой



ПОЯСНИТЕЛЬНАЯ ЗАПИСКА

Цели и задачи учебной дисциплины

Цель учебной дисциплины «Машинное обучение и анализ данных» – ознакомление студентов магистратуры с основными современными способами автоматического обнаружения и описания закономерностей в данных из окружающего мира.

Обучение способам автоматизации деятельности на основе анализа данных и моделирования, без явного предварительного написания алгоритмов действий, подготовка специалистов, способных использовать фундаментальные математические знания в качестве основы при проведении прикладных исследований

Задачи учебной дисциплины:

1. Формирование у магистрантов способностей самостоятельно разрабатывать алгоритмы решения задач и их анализировать
2. Развивать и использовать инструментальные средства, информационные среды, автоматизированные системы
3. Использовать математические и компьютерные методы исследований при анализе современных естественнонаучных, экономических, социально-политических процессов;
4. Приобретение способностей самостоятельно расширять компьютерные математические знания с дальнейшим их использованием при анализе математических моделей широкого круга прикладных задач.

Место учебной дисциплины в системе подготовки специалиста с высшим образованием (магистра).

Учебная дисциплина «Машинное обучение и анализ данных» относится к модулю по выбору «Гармонический анализ и дифференциальные уравнения» компонент учреждения образования

Связи с другими учебными дисциплинами, включая учебные дисциплины компонента учреждения высшего образования, дисциплины специализации и др.

Учебная дисциплина связана со следующими учебными дисциплинами: «Теория вероятностей и математическая статистика».

Требования к компетенциям

Освоение учебной дисциплины «Машинное обучение и анализ данных» должно обеспечить формирование следующих **специализированных компетенций**:

СК-2. Быть способным использовать методы компьютерного моделирования на основе современных методик численного анализа прикладных дифференциальных задач,

СК-5. Быть способным применять современные методы гармонического анализа и дифференциальных уравнений в задачах естественных наук и экономики.

В результате освоения учебной дисциплины студент должен:

знать:

- классификацию видов данных имеющихся в окружающем мире;
- формулировку задач регрессии, классификации, ранжирования, оценки распределения, уменьшения размерности, кластеризации;
- основные подходы к решению задач вышеупомянутых задач регрессии, классификации, ранжирования, оценки распределения, уменьшения размерности, кластеризации;

уметь:

- определять вид имеющихся данных;
- проводить первичную обработку данных;
- формализовывать задачу, как задачу регрессии, классификации, ранжирования, оценки распределения, уменьшения размерности, кластеризации или их совокупность;
- решать упомянутые задачи с помощью современных программных средств;
- оценивать качество предложенного решения.

владеть:

способами автоматизации деятельности на основе анализа данных и моделирования, без явного предварительного написания алгоритмов действий.

Структура учебной дисциплины

Дисциплина изучается в 3 семестре дневной и заочной формы получения высшего образования. Всего на изучение учебной дисциплины «Машинное обучение и анализ данных» отведено:

– для очной формы получения высшего образования – 108 часов, в том числе 36 аудиторных часов, из них: лекции – 18 часов, лабораторные занятия – 18 часов.

– для заочной формы получения высшего образования – 8 аудиторных часов, из них: лекции – 4 часов, лабораторные занятия – 4 часа.

Трудоемкость учебной дисциплины составляет 3 зачетные единицы.

Форма текущей аттестации по учебной дисциплине – зачет.

СОДЕРЖАНИЕ УЧЕБНОГО МАТЕРИАЛА

Тема 1. Типы переменных

Непрерывные, категориальные, упорядоченные. Прецедентное обучение. Предикторы и целевые переменные.

Тема 2. Классификация моделей по типу обучения.

Обучение с учителем (задачи регрессии, классификации и ранжирования, вероятностная постановка задачи классификации), обучение без учителя (кластеризация, уменьшение размерности, оценка распределения). Переобучение модели и способы борьбы с переобучением модели.

Тема 3. Энтропия и информация.

Плотность распределения. Метод максимального правдоподобия при оценке параметра.

Тема 4. Функция потерь при оценке качества работы модели.

Основные метрики для оценки качества в задачах классификации и регрессии.

Тема 5. Линейная регрессия и регуляризация (LASSO, Ridge).

Обобщенная линейная регрессия. Логистическая регрессия.

Тема 6. Классификация

Бинарная и множественная классификация. Сравнение бинарных классификаторов. ROC-кривая.

Тема 7. Деревья в задаче классификации и регрессии.

Метод ближайших соседей.

Тема 8. Метод опорных векторов.

Метод опорных векторов в задачах классификации и регрессии. Ядра.

Тема 9. Нейронные сети.

Перцептроны, рекуррентные сети, LSTM-сети.

Тема 10. Составные модели.

Случайный лес. Стэкинг.

Тема 11. Кластеризация

K-среднее. Ближайшие соседи.

Тема 12. Снижение размерности

Метод главных компонент.

УЧЕБНО-МЕТОДИЧЕСКАЯ КАРТА УЧЕБНОЙ ДИСЦИПЛИНЫ

Дневная форма получения образования

Номер раздела, темы	Название раздела, темы	Количество аудиторных часов						Форма контроля знаний
		Лекции	Практические занятия	Семинарские занятия	Лабораторные занятия	Иное	Количество часов УСР	
1	2	3	4	5	6	7	8	9
1.	Типы переменных	1			1			
2	Классификация моделей по типу обучения.	1			1			
3	Энтропия и информация.	1			1			Отчет по лабораторной работе с устной защитой
4	Функция потерь при оценке качества работы модели.	1			1			
5	Линейная регрессия и регуляризация (LASSO, Ridge).	1			1			Отчет по лабораторной работе с устной защитой
6	Классификация.	1			1			Отчет по лабораторной работе с устной защитой
7	Деревья в задаче классификации и регрессии. Составные модели (Bagging).	2			2			Отчет по лабораторной работе с устной защитой
8	Метод опорных векторов.	2			2			
9	Нейронные сети	2			2			Отчет по лабораторной работе с устной защитой
10	Составные модели	2			2			Отчет по лабораторной работе с устной защитой
11	Кластеризация	2			2			Отчет по лабораторной работе с устной защитой
12	Снижение размерности	2			2			Отчет по лабораторной работе с устной защитой
	Всего	18			18			

УЧЕБНО-МЕТОДИЧЕСКАЯ КАРТА УЧЕБНОЙ ДИСЦИПЛИНЫ

Заочная форма получения образования

Номер раздела, темы	Название раздела, темы	Количество аудиторных часов					Форма контроля знаний
		Лекции	Практические занятия	Семинарские занятия	Лабораторные занятия	Иное	
1	2	3	4	5	6	7	8
1.	Типы переменных.	1			1		Отчет по лабораторной работе с устной защитой, собеседование
2.	Классификация моделей по типу обучения.						
3.	Энтропия и информация.	1			1		Отчет по лабораторной работе с устной защитой, собеседование
4.	Функция потерь при оценке качества работы модели.						
5.	Линейная регрессия и регуляризация (LASSO, Ridge).						
6.	Классификация.						
7.	Деревья в задаче классификации и регрессии.	1			1		Отчет по лабораторной работе с устной защитой, собеседование
8.	Составные модели (Bagging). Метод опорных векторов.						
9.	Нейронные сети.	1			1		Отчет по лабораторной работе с устной защитой, собеседование
10.	Составные модели						
11.	Кластеризация.						
12.	Снижение размерности						
	Всего	4			4		

ИНФОРМАЦИОННО-МЕТОДИЧЕСКАЯ ЧАСТЬ

Перечень основной литературы

1. Вапник В.Н., Червоненкис А.Я. Теория распознавания образов. Статистические проблемы обучения. – Москва, 1974. – 416 с.
2. Trevor Hastie. Robert Tibshirani. Jerome Friedman. The Elements of Statistical Learning. Data Mining, Inference, and Prediction. – Springer, 2009. – 745 p
3. Рашка, С., Мирджалили, В. Python и машинное обучение: машинное и глубокое обучение с использованием Python, scikit-learn и TensorFlow, 2-е изд. / С. Рашка, В. Мирджалили. – СПб.: ООО «Диалектика», 2019 – 656 с.
4. Флах, П. Машинное обучение. Наука и искусство построения алгоритмов, которые извлекают знания из данных / П. Флах . – М.: ДМК Пресс, 2015 – 400 с.

Перечень дополнительной литературы

1. Вандер Плас, Дж. Python для сложных задач: наука о данных и машинное обучение / Дж. Вандер-Плас. – СПб.: Питер, 2018 – 576 с.
2. Маккинни, У. Python и анализ данных / У. Маккинни. – М.: ДМК Пресс, 2015 – 482 с.
3. Ш. Бастиан. Крупномасштабное машинное обучение вместе с Python. – М.: ДМК Пресс, 2017

Перечень рекомендуемых средств диагностики и методика формирования итоговой оценки

Для диагностики компетенций используется устно-письменная форма в виде отчета по лабораторной работе с устной защитой, устных докладов и отчетов по самостоятельно разрабатываемым темам, устных и письменных отчетов по самостоятельно разрабатываемому проекту, зачет в устно-письменной форме.

Формой текущей аттестации по дисциплине «Машинное обучение и анализ данных» учебным планом предусмотрен зачет.

Зачет по дисциплине проходит в устной или письменной форме.

Описание инновационных подходов и методов к преподаванию учебной дисциплины (проективный, практико-ориентированный)

При организации образовательного процесса используются:
практико-ориентированный подход, который предполагает ориентацию на генерирование идей, реализацию групповых студенческих проектов, развитие предпринимательской культуры;
метод группового обучения в форме организации малых групп (команд), работающих как над общими заданиями, так и над специфическими учебными заданиями.

Методические рекомендации по организации самостоятельной работы обучающихся, кроме подготовки к экзамену, подготовка к зачету

При изучении учебной дисциплины «Машинное обучение и анализ данных» рекомендуется использовать следующие формы самостоятельной работы:

- выполнение домашнего задания;
- изучение материала, вынесенного на самостоятельную проработку;
- разработка самостоятельно выбранной темы курса, в том числе поиск (подбор) и обзор литературы и электронных источников, изучение материала, подготовка демонстрации;
- демонстрация по самостоятельно выбранной теме курса.

Примерный перечень вопросов к зачету

А. Вопросы, требующие устного ответа.

- A1. Типы переменных (непрерывные, категорные, упорядоченные и т.д.).
- A2. Прецедентное обучение. Предикторы и целевые переменные.
- A3. Задачи регрессии и классификации. Вероятностная постановка задачи классификации.
- A4. Переобучение модели. Способы борьбы с переобучением модели.
- A5. Энтропия и информация.
- A6. Плотность распределения. Метод максимального правдоподобия при оценке параметра.
- A7. Сравнение бинарных классификаторов. ROC-кривая.
- A8. Функция потерь при оценке качества работы модели.
- A9. Шаги анализа данных CRISP-DM.

Б. Вопросы, требующие демонстрации

- Б1. Линейная регрессия

- Оценка качества работы линейной регрессии. Метрика R^2
- L1 регуляризация (LASSO-regression), L2 регуляризация (Ridge Regression)
- http://scikit-learn.org/stable/modules/linear_model.html#ordinary-least-squares

Б2. Обобщенная линейная регрессия

- логистическая регрессия, предсказание целочисленных величин
- http://scikit-learn.org/stable/modules/linear_model.html#logistic-regression

Б3. Деревья в задаче классификации и регрессии (CART-trees)

- <http://scikit-learn.org/stable/modules/tree.html>

Б4. Ближайшие соседи

- <http://scikit-learn.org/stable/modules/neighbors.html>

Б5. Метод опорных векторов

- Обычный метод опорных векторов и с использованием ядер
- <http://scikit-learn.org/stable/modules/svm.html>

Б6. Метод главных компонент / Principal Component Analysis

- <http://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html>

Б7. Составные модели (Bagging)

- Случайный лес (Random forest)
- <http://scikit-learn.org/stable/modules/ensemble.html>

Б8. Нейронные сети

- Перцептрон
- http://scikit-learn.org/stable/modules/neural_networks_supervised.html

Б9. Кластеризация

- http://scikit-learn.org/stable/tutorial/statistical_inference/unsupervised_learning.html

ПРОТОКОЛ СОГЛАСОВАНИЯ УЧЕБНОЙ ПРОГРАММЫ УВО

Название учебной дисциплины, с которой требуется согласование	Название кафедры	Предложения об изменениях в содержании учебной программы учреждения высшего образования по учебной дисциплине	Решение, принятое кафедрой, разработавшей учебную программу (с указанием даты и номера протокола)
Математические модели в информационных технологиях	Кафедра веб-технологий и компьютерного моделирования	нет	Изменений не требуется, протокол № 12 от 04.06.2020

**ДОПОЛНЕНИЯ И ИЗМЕНЕНИЯ К УЧЕБНОЙ ПРОГРАММЕ ПО
ИЗУЧАЕМОЙ УЧЕБНОЙ ДИСЦИПЛИНЕ**

на ____ / ____ учебный год

№ п/п	Дополнения и изменения	Основание

Учебная программа пересмотрена и одобрена на заседании кафедры
_____ (протокол № ____ от _____ 201_ г.)

Заведующий кафедрой

УТВЕРЖДАЮ
Декан факультета
