

Белорусский государственный университет

УТВЕРЖДАЮ

Проректор по учебной работе
и образовательным инновациям


О.Н. Здрок

«30» сентября 2020 г.

Регистрационный № УД- 8649 / уч.

МАШИННОЕ ОБУЧЕНИЕ НА БОЛЬШИХ МАССИВАХ ДАННЫХ

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

1-31 80 09 Прикладная математика и информатика

профилизация

Алгоритмы и системы обработки больших данных

2020 г.

Учебная программа составлена на основе ОСВО 1-31 80 09-2019 и учебного плана G31-072/уч. от 11.04.2019 г.

СОСТАВИТЕЛИ:

В.И. Отлига – старший преподаватель кафедры дискретной математики и алгоритмики факультета прикладной математики и информатики Белорусского государственного университета;

А.Б. Жиркевич – ассистент кафедры дискретной математики и алгоритмики факультета прикладной математики и информатики Белорусского государственного университета;

РЕЦЕНЗЕНТЫ:

Ю.Г. Василевский – кандидат физико-математических наук, начальник отдела разработки мобильных сервисов ООО «ЯндексБел».

РЕКОМЕНДОВАНА К УТВЕРЖДЕНИЮ:

Кафедрой дискретной математики и алгоритмики
(протокол № 14 от 19 марта 2020 года);

Научно-методическим Советом БГУ
(протокол № 4 от 25 марта 2020 года).

Заведующий кафедрой
дискретной математики и алгоритмики  В.М. Котов

ПОЯСНИТЕЛЬНАЯ ЗАПИСКА

Цели и задачи учебной дисциплины

Учебная дисциплина «Машинное обучение на больших массивах данных» знакомит студентов магистратуры с основными направлениями в разработке алгоритмов машинного обучения для прикладного анализа больших массивов данных и закладывает необходимую теоретическую базу для разработки новых алгоритмов машинного обучения, применимых для анализа больших данных на практике.

Цель учебной дисциплины – создание базы для применения современных методов машинного обучения на больших данных для решения практических задач и формирование у студентов магистратуры умения анализировать большие массивы данных с целью подбора наиболее оптимального алгоритма обучения. При изложении материала учебной дисциплины важно показать особенности работы с большими массивами данных и методики их учёта при построении алгоритмов машинного обучения.

Задачи учебной дисциплины:

1. Изучение особенностей больших массивов данных.
2. Изучение алгоритмов онлайн-обучения.
3. Изучение подходов к уменьшению размерности данных.

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

Учебная дисциплина относится к компоненту учреждения высшего образования и входит **в модуль «Избранные главы компьютерных наук»**.

Программа составлена с учетом **межпредметных связей** с учебными дисциплинами. Основой для изучения учебной дисциплины являются следующие учебные дисциплины II ступени высшего образования вузовского компонента «Методы машинного обучения» и «Нейронные сети» модуля «Модели и метод машинного обучения». Знания, полученные в учебной дисциплине, используются при изучении дисциплины «Информационный поиск», «Компьютерное зрение» и «Распознавание и синтез речи» модуля «Приложения машинного обучения».

Требования к компетенциям

Освоение учебной дисциплины «Машинное обучение на больших массивах данных» должно обеспечить формирование следующих специализированных и углубленных профессиональных компетенций:

специализированные компетенции:

СК-24. Использовать современные научные и технические достижения в области разработки эффективных алгоритмов для решения конкретных прикладных задач

СК-25. Развивать способность выбирать методологию и технологию проектирования компьютерных систем

углубленные профессиональные компетенции:

УПК-5. Владеть перспективными технологиями программирования.

В результате освоения учебной дисциплины студент магистратуры должен:

знать:

- особенности работы с большими массивами данных;
- основные подходы к разработке алгоритмов машинного обучения, учитывающих особенности больших данных;

уметь:

- подбирать подходящие модели машинного обучения с учётом особенностей больших данных;
- применять методы машинного обучения для решения прикладных задач на больших массивах данных;

владеть:

- программными средствами для разработки алгоритмов машинного обучения на больших массивах данных
- навыками разработки моделей машинного обучения на больших массивах данных.

Структура учебной дисциплины

Дисциплина изучается в 3-м семестре. Всего на изучение учебной дисциплины «Машинное обучение на больших массивах данных» отведено: – для очной формы получения высшего образования – 90 часов, в том числе 40 аудиторных часов, из них: лекции – 20 часов, практические занятия – 20 часов (в том числе – 10 часов дистанционного обучения).

Трудоемкость учебной дисциплины составляет 3 зачетные единицы.

Форма текущей аттестации по учебной дисциплине – зачет.

СОДЕРЖАНИЕ УЧЕБНОГО МАТЕРИАЛА

Раздел 1. Основы машинного обучения на больших массивах данных

Тема 1.1. Введение.

Что такое большие данные в контексте машинного обучения. Примеры практических задач. Обзор программы курса. Estimation/Approximation/Optimization trade-off.

Тема 1.2. Онлайн-обучение линейных моделей.

Линейные модели. Функции потерь. Онлайн-обучение. Progressive validation. Полезные трюки для онлайн-обучения: учет весов объектов, адаптивный learning rate, учет размерности переменных. Онлайн бутстреп.

Тема 1.3. Оптимизация гиперпараметров.

Амёба, решетка, random search, байесовская оптимизация. Методы распараллеливания.

Раздел 2 Методы работы с большим количеством признаков

Тема 2.1. Хеширование в машинном обучении на больших массивах данных.

Хеширование признаков. Хеширование в обработке текстов. Хеш-ядро. Locality-sensitive hashing.

Тема 2.2. Методы работы с категориальными признаками.

One-hot encoding. Решающие деревья. Факторизационные машины. Field-aware Factorization Machines. DeepFM. Наивный байесовский классификатор. Счётчики.

Раздел 3. Методы распараллеливания алгоритмов

Тема 3.1. Метод BFGS.

BFGS и Limited-memory BFGS (L-BFGS). Сложность вычислений. Stochastic BFGS.

Тема 3.2. Блочнo-координатный спуск, ADMM

Блочнo-координатный спуск для generalized linear models (GLMs). Алгоритм GLMNET. Варианты распараллеливания (много ядер, распределенные системы). Метод ADMM. Consensus and sharing. Реализации ADMM для различных функций.

УЧЕБНО-МЕТОДИЧЕСКАЯ КАРТА УЧЕБНОЙ ДИСЦИПЛИНЫ

Дневная форма получения образования с применением дистанционных образовательных технологий

№ п/п	Название раздела, темы	Количество аудиторных часов				Количество часов УСР	Форма контроля знаний
		Лекции	Семинарские занятия	Практические занятия	Иное		
1	Основы машинного обучения на больших массивах данных	8		4 2(ДО)			
1.1	Введение	2					Устный опрос
1.2	Онлайн-обучение линейных моделей	4		2 2(ДО)			Отчеты по домашним практическим заданиям с их устной защитой.
1.3	Оптимизация гиперпараметров	2		2			Отчеты по домашним практическим заданиям с их устной защитой. Контрольная работа 1.
2	Методы работы с большим количеством признаков	8		2 4(ДО)			
2.1	Хеширование в машинном обучении на больших массивах данных	4					Коллоквиум
2.2	Методы работы с категориальными признаками	4		2 4(ДО)			Отчеты по домашним практическим заданиям с их устной защитой. Контрольная работа 2.

3.	Методы распараллеливания алгоритмов	4		4 4(ДО)			
3.1	Метод BFGS	2		2 2(ДО)			Устный опрос. Отчеты по домашним практическим заданиям с их устной защитой.
3.2	Блочно-координатный спуск, ADMM	2		2 2(ДО)			Устный опрос. Отчеты по домашним практическим заданиям с их устной защитой. Контрольная работа №3.

ИНФОРМАЦИОННО-МЕТОДИЧЕСКАЯ ЧАСТЬ

Перечень основной литературы

1. Suthaharan S. Machine Learning Models and Algorithms for Big Data Classification: Thinking with Examples for Effective Learning. – Springer, 2015. — 359 p.
2. Karim M. R., Kaysar M. M. Large Scale Machine Learning with Spark. – Packt Publishing Ltd, 2016. — 454 p.
3. Bekkerman R., Bilenko M., Langford J. Scaling up Machine Learning, Cambridge University Press, 2011. — 492 p. — ISBN-13: 978-0521192248.
4. Efficient Online Bootstrapping for Large Scale Learning Zhen Qin, Vaclav Petricek, Nikos Karampatziakis, Lihong Li, John Langford - 2013 - Big Learning Workshop at Neural Information Processing Systems 2013.
5. Leskovec J., Rajaraman A., Ullman J.D. Mining of Massive Datasets, Cambridge University Press, 2014. — 476 p. — ISBN-13: 978-1107077232.
6. Optimization Methods for Large-Scale Machine Learning Léon Bottou, Frank E. Curtis, Jorge Nocedal - 2018 - <https://arxiv.org/pdf/1606.04838.pdf>

Перечень дополнительной литературы

7. Bishop, C.M. Pattern Recognition and Machine Learning. — Springer, 2006. — 738 p.
8. Cristianini, N. & Shawe-Taylor, J. (2000), An Introduction to Support Vector Machines, Cambridge University Press.
9. Kearns, M. & Vazirani, U. (1994), An Introduction to Computational Learning Theory, MIT Press.
10. Hastie, T., Tibshirani, R., Friedman, J. The Elements of Statistical Learning, 2nd edition. — Springer, 2009. — 533 p.
11. Shalev-Shwartz Shai, Ben-David Shai. Understanding Machine Learning: From Theory to Algorithms, Cambridge University Press, 2014. — 409 p. — ISBN-13: 978-1107057135.

Перечень рекомендуемых средств диагностики и методика формирования итоговой оценки

Для диагностики компетенции в рамках учебной дисциплины рекомендуется использовать следующие формы:

1. Устная форма: устный опрос, коллоквиум.
2. Письменная форма: контрольные работы.
3. Устно-письменная форма: отчеты по домашним практическим заданиям с их устной защитой, оценивание на основе проектного метода.

Формой текущей аттестации по дисциплине «Машинное обучение на больших массивах данных» – зачет.

При формировании итоговой оценки используется рейтинговая оценка знаний студента, дающая возможность проследить и оценить динамику процесса достижения целей обучения. Рейтинговая оценка предусматривает использование весовых коэффициентов для текущего контроля знаний студентов по дисциплине.

Примерные весовые коэффициенты, определяющие вклад текущего контроля знаний и текущей аттестации в рейтинговую оценку (формирование оценки за текущую успеваемость):

- отчёты по практическим домашним заданиям с их устной защитой – 60 %;
- контрольные работы – 20 %;
- коллоквиум – 10 %;
- устный опрос – 10%;

Рейтинговая оценка по дисциплине рассчитывается на основе оценки текущей успеваемости и зачетной оценки с учетом их весовых коэффициентов. Вес оценки по текущей успеваемости составляет 30 %, зачетная оценка – 70%.

Примерная тематика практических занятий

Занятие 1-2. Библиотека Vowpal Wabbit.

Занятие 3. Библиотека HyperOpt

Занятие 4-5. Локально-чувствительное хеширование

Занятие 6. Библиотека LightGBM

Занятие 7-8. Подход “Parameter server”

Занятие 9-10. Массовое прогнозирование временных рядов

Описание инновационных подходов и методов к преподаванию учебной дисциплины

При организации образовательного процесса большинства практических занятий используется *практико-ориентированный подход*, который предполагает освоение содержания учебного материала через решение практических задач, а также приобретение навыков эффективного выполнения разных видов профессиональной деятельности.

Кроме этого, при организации образовательного процесса используется комбинация *методов группового обучения, проектного обучения и учебной дискуссии*. Комбинация методов предполагает: ориентацию на генерирование идей, приобретение навыков для решения исследовательских, творческих и коммуникационных задач, появление нового уровня понимания изучаемой темы, применение знаний (теорий, концепций) при решении проблем, определение способов их решения.

Методические рекомендации по организации самостоятельной работы обучающихся, подготовка к экзамену

Для организации самостоятельной работы студентов магистратуры по учебной дисциплине следует использовать информационно коммуникационные технологии:

Образовательный портал БГУ (<https://edufpmi.bsu.by>);

Образовательный портал InsightRunner (<https://acm.bsu.by>);

система AnyTask (<https://anytask.org/school/bsu>);

разместить в сетевом доступе комплекс учебных и учебно-методических материалов (учебно-программные материалы, учебное издание для теоретического изучения дисциплины, презентации лекций, методические указания к практическим занятиям, электронные версии домашних заданий, материалы текущего контроля и текущей аттестации, позволяющие определить соответствие учебной деятельности обучающихся требованиям образовательных стандартов высшего образования и учебно-программной документации, в том числе вопросы для подготовки к экзамену, задания, вопросы для самоконтроля, список рекомендуемой литературы, информационных ресурсов и др.).

Рекомендуемая тематика контрольных работ

- 1) Контрольная работа №1. *Оптимизация гиперпараметров.*
- 2) Контрольная работа №2. *Методы работы с категориальными признаками.*
- 3) Контрольная работа №3. *Блочно-координатный спуск, ADMM.*

ПРОТОКОЛ СОГЛАСОВАНИЯ УЧЕБНОЙ ПРОГРАММЫ

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)
Технологии проектирования и разработки высоконагруженных веб-систем	Дискретной математики и алгоритмики	Нет	Оставить содержание учебной дисциплины без изменения (протокол № 14 от 19 марта 2020 г.).

ДОПОЛНЕНИЯ И ИЗМЕНЕНИЯ К УЧЕБНОЙ ПРОГРАММЕ

на ____ / ____ учебный год

№№ ПП	Дополнения и изменения	Основание

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

Заведующий кафедрой

(ученая степень, звание)

(подпись)

(И.О. Фамилия)

УТВЕРЖДАЮ
Декан факультета

(ученая степень, звание)

(подпись)

(И.О.Фамилия)