Министерство науки и высшего образования Российской Федерации
Федеральное государственное автономное образовательное учреждение
высшего образования «Новосибирский национальный исследовательский
государственный университет» (Новосибирский государственный университет, НГУ)

Экономический факультет

Согласовано
Декан ЭФ
_____ Богомолова Т.Ю.
подпись
«19» 10 2020 г.

РАБОЧАЯ ПРОГРАММА ДИСЦИПЛИНЫ

ВВЕДЕНИЕ В МАШИННОЕ ОБУЧЕНИЕ

Направление подготовки: 38.03.05 Бизнес-информатика
Направленность (профиль): Бизнес-информатика

Форма обучения: очная

Разработчики:

PhD Economics,
Рук. Отд. МО ЦФТ, Комаров И.В.

М.н.с., Механико-математический ф.,
Лаборатория аналитики потоковых данных и
машинного обучения, Бондаренко И.

Зав. кафедрой применения математических
методов в экономике и планировании
д.э.н., профессор Мкртчян Г.М.

Новосибирск
2020

Содержание

1. Перечень планируемых результатов обучения по дисциплине, соотнесенных с планируемыми результатами освоения образовательной программы

Цель дисциплины «Введение в машинное обучение»:

формирование представления о технологиях машинного обучения, их возможностях и ограничениях, изучение основных моделей и технологий анализа различных типов данных.

Задачи освоения дисциплины:

- изучить методы анализа данных, применяемые в машинном обучении;
- научить пользоваться библиотеками языка Python, приемами машинного обучения;
- привить практические навыки анализа данных и прогнозирования.

Планируемые результаты обучения по дисциплине, соотнесенные с планируемыми результатами освоения образовательной программы

Результаты освоения образовательной программы (компетенции)	В результате изучения дисциплины обучающиеся должны:		
	знать	уметь	владеть
ОПК-3 Способностью работать с компьютером как средством управления информацией, работать с информацией из различных источников, в том числе в глобальных компьютерных сетях	Особенности различных областей приложения методов машинного обучения	Использовать готовые библиотеки машинного обучения	Навыками разработки программ, реализующие алгоритмы машинного обучения
ПК-18 Способность использовать соответствующий математический аппарат и инструментальные средства для обработки, анализа и систематизации информации по теме исследования	Математические основы алгоритмов классификации, кластеризации и регрессии	Определить применимость алгоритмов машинного обучения к конкретной задаче	Основами применения базовых технологий машинного обучения в некоторых областях
ПК-19 Умение готовить научно-технические отчеты, презентации, научные публикации по результатам выполненных исследований	Методы визуализации результатов описания данных	Строить прогнозы на основе машинного обучения	Навыками описания и сравнения результатов применения различных подходов машинного обучения

2. Место дисциплины в структуре образовательной программы

Дисциплина «Введение в машинное обучение» является элективной и преподается в 6 семестре.

Дисциплины (практики), изучение которых необходимо для освоения дисциплины «Введение в машинное обучение»: «Линейная алгебра», «Математический анализ», «Теория вероятностей», «Информационные процессы, системы и сети», «Эконометрия», «Программирование».

3. Трудоемкость дисциплины в зачетных единицах с указанием количества академических часов, выделенных на контактную работу обучающегося с преподавателем (по видам учебных занятий) и на самостоятельную работу обучающегося

Для набора 2018-2020 года:

Трудоемкость дисциплины – 4 зачетных единиц, 144 часов.
Форма промежуточной аттестации: дифференцированный зачет.

Вид деятельности	Семестр
	6
Контактная работа, часов, в том числе:	66
лекции	28
практические занятия	28
групп. работа с преподавателем	6
контактная работа при аттестации	2
Самостоятельная работа, часов, в том числе:	78
самостоятельная работа во время занятий	72
самостоятельная работа во время контрольной аттестации	6
Всего, часов	144

Для набора 2017 года:

Трудоемкость дисциплины – 4 зачетных единиц, 144 часов.
Форма промежуточной аттестации: дифференцированный зачет.

Вид деятельности	Семестр
	6
Контактная работа, часов, в том числе:	74
лекции	32
практические занятия	32
групп. работа с преподавателем	6
контактная работа при аттестации	2
Самостоятельная работа, часов, в том числе:	70
самостоятельная работа во время занятий	64
самостоятельная работа во время контрольной аттестации	6
Всего, часов	144

4. Содержание дисциплины, структурированное по темам (разделам) с указанием отведенного на них количества академических часов и видов учебных занятий

Содержание дисциплины «Введение в машинное обучение»:

	Содержание разделов
1	Зачем машинное обучение экономисту. *Определение машинного обучения. История машинного обучения. Отличие от эконометрики. Почему знания экономиста могут пригодиться. Примеры использования машинного обучения в экономических задачах.*
2	Линейная регрессия. *Постановка задачи. Примеры. Оптимизационная задача. Предпосылки линейной регрессии. Функция потерь. Регуляризация. Лассо и Гребень. Логистическая регрессия. Почему сигмоида?*
3	Логистическая регрессия. *Функция потерь. Градиентный спуск. Метод максимального правдоподобия. Почему cross entropy? Матрица ошибок. Точность и полнота, F1. Доля правильных ответов. Отсечение. ROC AUC. Gini.*
4	Выбор модели. *Отложенная выборка. Стратификация. Кросс-валидация. Валидация во времени. Смещение и разброс. Валидационная кривая. Кривая обучения.* Создание признаков. *Мешок слов. Время по кругу. Масштабирование.*
5	Проверка гипотез. *Классический подход. t-test. Бутстрап.* *Феномен уменьшения правды.* Закон Бенфорда.
6	Деревья решений. *Виды. Терминология. Алгоритм построения. Жадность и локальная оптимизация. Критерии разбиения. Bagging. Случайный лес. Интуиция создания градиентного бустинга на деревьях.*
7	Кластеризация. *k-means.* Снижение размерности. *PCA via SVD.* Визуализация. *t-SNE.*
8	Поиск аномалий. *Выбросы. Новизна. Примеры. Кластеризация как метод подмены задачи. Куртосис, Визуализация. Модель и отклонение. SVD. Модели описания данных. Пример: классификация ввода информации ФИО, не ФИО.*
9	Как стать дата саентистом. *Статистика. Программирование. Научный подход.*
10	Машинное обучение и бизнес. *Сравнение человека с алгоритмом. Автоматизация менеджмента. Функция потерь в рублях.*
11	Интерпретация моделей МО.
12	Нейронные сети. *Проблема представляемости. Проблема обучаемости.* *Обратное распространение ошибки.*
13	Сверточные нейронные сети.
14	Рекуррентные нейронные сети.
15	Компьютерное зрение.

Лекции (28 ч)

Наименование темы и их содержание	Объем, час

Модуль 1 Классическое машинное обучение.	
1. Машинное обучение и экономика.	1
2. Линейная регрессия.	2
3. Логистическая регрессия и метрики качества.	2
4. Выбор модели в машинном обучении.	2
5. Проверка гипотез и бутстрап.	2
6. От деревьев решений до градиентного бустинга на деревьях.	2
7. Обучение без учителя.	2
8. Поиск аномалий.	2
9. Подводим итоги: что еще нужно, чтобы стать дата саентистом?	1
10. Машинное обучение и бизнес.	2
11. Интерпретация моделей машинного обучения.	2
Модуль 2 Глубокое обучение	
12. Нейронные сети.	2
13. Сверточные нейросети.	2
14. Рекуррентные нейросети.	2
15. Компьютерное зрение.	2

Практические занятия (28 ч)

Содержание практического занятия	Объем, час
Настройка среды для анализа данных	2
Семинары по разбору домашних заданий	6
Решение задачи в соревновании	6
Создание признаков	2
Обучение нейронных сетей	4
Распознавание картинок	2

Самостоятельная работа студентов (78 ч)

Перечень занятий на СРС	Объем, час
Решение домашних заданий	20
Решение задачи в соревновании	40
Изучение теоретического материала, не освещаемого на лекциях	12
Подготовка к зачету	6

5. Перечень учебной литературы

5.1 Основная литература

1. Воронова, Л. И. Machine Learning: регрессионные методы интеллектуального анализа данных: учебное пособие / Л. И. Воронова, В. И. Воронов. — Москва: Московский технический университет связи и информатики, 2018. — 82 с. — ISBN 2227-8397. — Текст: электронный // Электронно-библиотечная система IPR BOOKS: [сайт]. — URL: http://www.iprbookshop.ru/81325.html (дата обращения: 07.11.2020). — Режим доступа: для авторизир. пользователей
2. Ракитский, А. А. Методы машинного обучения: учебно-методическое пособие / А. А. Ракитский. — Новосибирск: Сибирский государственный университет телекоммуникаций и информатики, 2018. — 32 с. — ISBN 2227-8397. — Текст:

электронный // Электронно-библиотечная система IPR BOOKS : [сайт]. — URL: http://www.iprbookshop.ru/90591.html (дата обращения: 07.11.2020). — Режим доступа: для авторизир. Пользователей
3. Неделько, В. М. Основы статистических методов машинного обучения: учебное пособие / В. М. Неделько. — Новосибирск: Новосибирский государственный технический университет, 2010. — 72 с. — ISBN 978-5-7782-1385-2. — Текст : электронный // Электронно-библиотечная система IPR BOOKS : [сайт]. — URL: http://www.iprbookshop.ru/45418.html (дата обращения: 07.11.2020). — Режим доступа: для авторизир. пользователей

5.2 Дополнительная литература

4. Сараев, П. В. Методы машинного обучения: методические указания и задания к лабораторным работам по курсу / П. В. Сараев. — Липецк: Липецкий государственный технический университет, ЭБС АСВ, 2017. — 48 с. — ISBN 2227-8397. — Текст: электронный // Электронно-библиотечная система IPR BOOKS: [сайт]. — URL: http://www.iprbookshop.ru/83183.html (дата обращения: 07.11.2020). — Режим доступа: для авторизир. пользователейЖурнал "Вычислительные технологии" // http://elibrary.ru/title_about.asp?id=8610
5. Журнал "Информатика и ее применения" // http://elibrary.ru/title_about.asp?id=26694
6. Журнал "Информатика и образование" // http://elibrary.ru/title_about.asp?id=8739

6. Перечень учебно-методических материалов по самостоятельной работе обучающихся

7. Kaggle InClass – данные и среда для самостоятельной работы с данными и вычислений по соревнованиям

8. YouTube – видео курса для изучения

7. Перечень ресурсов информационно-телекоммуникационной сети «Интернет», необходимых для освоения дисциплины

При освоении дисциплины используются следующие ресурсы:

– образовательные интернет-порталы:
https://habr.com/ru/company/ods/blog/322626/

– другие ресурсы информационно-телекоммуникационной сети «Интернет»:
https://vk.com/mlcourse2_ml - для оперативной связи, обмена лекциями, записями лекций, таблицами с рейтингом, заданиями

Google Calendar, Forms, Sheets, Docs – для обмена информацией по курсу

Взаимодействие обучающегося с преподавателем (синхронное и (или) асинхронное) осуществляется через личный кабинет студента в ЭИОС, электронную почту, социальные сети.

7.1 Современные профессиональные базы данных:

Современные профессиональные базы данных не используются.

7.2. Информационные справочные системы

- *https://www.coursera.org/learn/machine-learning*

8. Перечень информационных технологий, используемых при осуществлении образовательного процесса по дисциплине

Перечень программного обеспечения
Windows
Microsoft Office
Chrome
Возможно: Anaconda

9. Материально-техническая база, необходимая для осуществления образовательного процесса по дисциплине

Для реализации дисциплины «Введение в машинное обучение» используются специальные помещения:

1. Учебные аудитории для проведения занятий лекционного типа, занятий семинарского типа, групповых и индивидуальных консультаций, текущего контроля, промежуточной и итоговой аттестации;

2. Помещения для самостоятельной работы обучающихся;

Учебные аудитории укомплектованы специализированной мебелью и техническими средствами обучения, служащими для представления учебной информации большой аудитории.

Помещения для самостоятельной работы обучающихся оснащены компьютерной техникой с возможностью подключения к сети "Интернет" и обеспечением доступа в электронную информационно-образовательную среду НГУ.

Реализация дисциплины осуществляется с применением электронного обучения *Zoom* где обучение проводится на виртуальных аналогах, позволяющим достигать запланированных результатов по дисциплине.

Материально-техническое обеспечение образовательного процесса по дисциплине для обучающихся из числа лиц с ограниченными возможностями здоровья осуществляется согласно «Порядку организации и осуществления образовательной деятельности по образовательным программам для инвалидов и лиц с ограниченными возможностями здоровья в Новосибирском государственном университете».

10. Оценочные средства для проведения текущего контроля и контрольной аттестации по дисциплине

Перечень результатов обучения по дисциплине «Введение в машинное обучение» и индикаторов их достижения представлен в виде знаний, умений и владений в разделе 1.

10.1 Порядок проведения текущего контроля и контрольной аттестации по дисциплине

Оценивание результатов обучения по дисциплине «Введение в машинное обучение» осуществляется по балльно-рейтинговой системе и включает следующие оценочные средства:

Текущий контроль успеваемости:

На курсе предусматривается проведение соревнований (конкурсов) по решению задач машинного обучения. В текущей версии курса предусматривается проведение 2-х конкурсов: «Элис» и «Аплифт».

Каждый конкурс максимально дает 30 баллов к финальной оценке. Конкурсы имеют длительность минимум в 2 недели.

Кроме конкурсов, знания оцениваются по 4 домашним заданиям. Задания подготовлены в виде форм с 10 вопросами и множественным выбором. Правильный ответ

на вопрос дает 1 балл. Максимально возможное количество баллов – 10 баллов за домашнее задание. Домашние задания как правило даются сроком на неделю.

При желании улучшить оценку по итогам текущего контроля, студент имеет право на контрольную аттестацию.

Контрольная аттестация:

При желании улучшить оценку возможно проведение контрольной аттестации студента. Контрольная аттестация возможна в двух видах:

- Устный зачет. 2 вопроса, теоретический и практический. 30 мин. на подготовку, можно пользоваться любыми источниками, в т.ч. Интернет.
- Контрольная работа. В течение недели необходимо выполнить и представить решение в виде Jupyter Notebook задания по выбору преподавателя. Например, распространение заболеваемости вирусом Cov-Sars-2 в Новосибирске на ближайший месяц.

В таблице представлены максимально возможные баллы по дисциплине с разбивкой по оценочным средствам.

Оценочные средства	**Баллы (максимум)**
Текущий контроль	
Домашнее задание №1 «Вопросы по конкурсу Элис»	10
Конкурс «Элис»	30
ДЗ № 2 «Вопросы по конкурсу Аплифт»	10
ДЗ № 3 «5 значимых признаков в конкурсе Аплифт»	10
Конкурс «Аплифт»	30
ДЗ № 4 «Воспроизвести код, добавив свои признаки для конкурса Аплифт»	10
или **Контрольная аттестация**	
Устный зачет	100
Итого	**100**

Перечень средств оценивания результатов обучения по дисциплине

«Введение в машинное обучение»

Таблица 10.1

Код компетенции	**Результат обучения по дисциплине**	**Оценочное средство**
ОПК-3	Знание особенностей различных областей приложения методов машинного обучения	Домашние задания Зачет
	Умение использовать готовые библиотеки машинного обучения	Конкурсы
	Владение навыками разработки программ, реализующие алгоритмы машинного обучения	Конкурсы
ПК-18	Знание математических основ алгоритмов классификации, кластеризации и регрессии	Конкурсы Зачет

	Умение определить применимость алгоритмов машинного обучения к конкретной задаче	Конкурсы
	Владение основами применения базовых технологий машинного обучения в некоторых областях	Конкурсы Зачет
ПК-19	Знание методов визуализации результатов описания данных	Домашние задания
	Умение строить прогнозы на основе машинного обучения	Конкурсы
	Владение навыками описания и сравнения результатов применения различных подходов машинного обучения	Конкурсы Зачет

Таблица 10.2

Критерии оценивания результатов обучения	**Шкала оценивания**
Топ-25% в классе на лидерборде конкурса: Положение оценивается по данным, которые недоступны для построения модели. При невозможности обеспечить недоступность данных – оценивается решение в Jupyter Notebook. Решения допускаются при отсутствии ликов – когда тестовые данные используются для предсказания. **Решение домашнего задания:** Ответ на 8 или больше вопросов правильно. **Зачет:** – отсутствие ошибок при ответе на тестовые вопросы, – полнота ответа на теоретический вопрос, – умение высказать свое мнение, – наличие исчерпывающих ответов на дополнительные вопросы. При изложении ответа на теоретический вопрос обучающийся мог допустить непринципиальные неточности. **Контрольная работа**: Цель задания достигнута и студент может объяснить как получено решение.	*Отлично* (100 – 75 % максимальной оценки)
Топ-50% в классе на лидерборде конкурса (см. выше) **Решение домашнего задания:** Ответ на 6 или больше вопросов правильно. **Зачет:** – не менее 80% ответов на тестовые вопросы должны быть правильными, – полнота ответа на теоретический вопрос, – наличие полных ответов на дополнительные вопросы с возможным наличием ошибок. **Контрольная работа**: Цель задания может быть достигнута использованием незначительной доработки и студент может объяснить как получено решение.	*Хорошо* (75 – 50 % максимальной оценки)
Топ-70% в классе на лидерборде конкурса (см. выше) **Решение домашнего задания:** Ответ на 4 или больше вопросов правильно. **Зачет:**	*Удовлетворительно*

– не менее 60% ответов на тестовые вопросы должны быть правильными, – наличие неполного ответа на теоретический вопрос, – наличие неполных и / или содержащих существенные ошибки ответов на дополнительные вопросы. **Контрольная работа**: Цель задания может быть достигнута использованием значительной доработки и студент может с трудом объяснить, как получено решение.	(50 – 30 % максимальной оценки)
Выше более 70% класса на лидерборде конкурса (см. выше) **Решение домашнего задания:** Ответ на 3 или меньше вопросов правильно. **Зачет:** – присутствие многочисленных ошибок (более 60% ответов содержат ошибки) на тестовые вопросы, – фрагментарный ответ на теоретический вопрос, – отсутствие ответов на дополнительные вопросы. **Контрольная работа**: Цель задания не может быть достигнута и студент не может объяснить, как получено или может быть получено решение.– отсутствие ответов на дополнительные вопросы.	*Неудовлетворительно* (менее 30 % максимальной оценки)

Баллы, набранные за выполнение заданий текущего контроля и/или контрольной аттестации, конвертируются в оценку по дисциплине следующим образом:

При отличном выполнении домашних заданий и выигрыша в соревнованиях, студент максимально может набрать 100 баллов, минимально 75.

Для хорошей оценки, необходимо получить более 30 баллов за конкурсы и более 20 за домашние задания, т.е. 50 – 75 баллов.

От 30 до 50 баллов – оценка удовлетворительно. Половина баллов должна быть от участия в конкурсах.

Менее 30 баллов – оценка неудовлетворительно.

Типовые контрольные задания и иные материалы, необходимые для оценки результатов обучения

Теоретические вопросы

1. Смещение (bias) и разброс (variance).
2. Принципиальное отличие байесовского и классического подхода к статистике
3. Методология CRISP-DM
4. Интерпретация моделей машинного обучения
5. Деревья решений и CART алгоритм
6. “Случайный лес” (Random Forest)
7. Градиентный бустинг
8. Гребневая регрессия
9. Логистическая регрессия
10. LASSO
11. Регуляризация
12. Выявление аномалий
13. Алгоритм классификации kNN
14. Метрики качества регрессии MAE RMSE MAPE

15. Метрики качества классификации перекрестная кросс-энтропия, accuracy, точность, полнота, F-мера.
16. Алгоритмы снижения размерности
17. Алгоритмы кластеризации
18. Ошибки первого и второго рода, уровень значимости и мощность
19. p-value - что означает и как интерпретировать

Практические вопросы

1. Процедура кросс-валидации
2. Кросс-валидация для временных рядов
3. Решение проблемы дисбаланса классов
4. Решение проблемы пропущенных значений
5. Кривые валидации и обучения.
6. Стекинг
7. Кодирование категориальных признаков
8. Биннинг признаков
9. Отбор признаков
10. Извлечение признаков из текста
11. Стандартизация признаков
12. Проблемы, вызванные скоррелированными признаками

Оценочные материалы по текущему контролю и промежуточной аттестации, предназначенные для проверки соответствия уровня подготовки по дисциплине «Введение в машинное обучение», планируемым результатам освоения образовательной программы (в соответствии с образовательными стандартами), хранятся на кафедре-разработчике РПД в печатном и электронном виде.

Лист актуализации рабочей программы дисциплины
«Введение в машинное обучение»

№	Характеристика внесенных изменений (с указанием пунктов документа)	Дата и № протокола Ученого совета ЭФ	Подпись ответственного