Санкт-Петербургский государственный университет

Р А Б О Ч А Я П Р О Г Р А М М А
УЧЕБНОЙ ДИСЦИПЛИНЫ

Введение в науку о данных
An introduction to Data Science

Язык(и) обучения

русский

Трудоемкость в зачетных единицах: 1

Регистрационный номер рабочей программы: 058038

Санкт-Петербург

Раздел 1. Характеристики учебных занятий

1.1. Цели и задачи учебных занятий

Основной целью освоения дисциплины «Введение в науку о данных» является приобретение обучающимися знаний об основах предметной области через постановку и решение типичных задач, с которыми исследователь в области науки о данных может столкнуться в своей работе, а также практических навыков работы с инструментами анализа данных, применяемыми в разных сферах человеческой деятельности.

Поставленная цель достигается путем решения следующих задач курса:

1) Ознакомить студентов с основными задачами, решаемыми в области науки о данных, базовыми алгоритмами этой области, а также со сферами практического применения данных алгоритмов;
2) Развить практические навыки работы с реальными инструментами, применяемыми в области науки о данных;
3) Научить решать прикладные задачи по обработке и анализу данных на предмет выявления в них скрытых зависимостей, а также подбирать методы машинного обучения для этих задач.

1.2. Требования подготовленности обучающегося к освоению содержания учебных занятий (пререквизиты)

Дисциплина «Введение в науку о данных» является онлайн-курсом, разработанным Санкт-Петербургским государственным университетом.

Обучающийся по данной дисциплине, должен знать основы комбинаторики, теории вероятностей, линейной алгебры, дискретной математики, а также владеть базовыми навыками программирования на Python.

1.3. Перечень результатов обучения (learning outcomes)

Дисциплина способствует формированию компетенций, предусмотренных действующим Образовательным стандартом СПбГУ:

В результате освоения курса обучающиеся должны

знать:

- Методологию работы исследователя в области науки о данных (постановка целей исследования, сбор данных, обработка и преобразования данных, обследование данных, построение моделей и отбор методов, представление и визуализация результатов).
- Способы организации хранения данных.
- Методы и подходы к стандартизации и преобразованию данных.
- Методы машинного обучения (базовые методы классификации и кластеризации).

уметь:

- Решать прикладные задачи по обработке и анализу данных на предмет выявления в них скрытых зависимостей.
- Применять элементы теории вероятностей и математической статистики, лежащие в основе моделей и методов науки о данных
- Правильно подбирать методы машинного обучения для решения практических задач
- Организовывать рабочее окружение исследователя в области науки о данных (Jupyter).
- Использовать пакеты и библиотеки для машинного обучения (Matplotlib, SciPy/NumPy, Pandas, Scikitlearn)

владеть:

- Инструментарием для организации хранения данных.

- Навыками программной реализации на языке Pyton средств обработки и анализа данных.
- Навыками предобработки и визуализации данных

1.4. Перечень и объём активных и интерактивных форм учебных занятий

Активные и интерактивные формы занятий данного курса предоставляются студентам платформой онлайн-образования в виде интерактивного учебника, который содержит видеоматериалы, материалы для самостоятельной работы, тесты и проекты. К активным формам относятся:

- подготовка к видео урокам;
- изучение раздаточных материалов;
- выполнение индивидуальных заданий в виде тестов по реализации изученных методов науки о данных;

к интерактивным:

- экспертная оценка результатов решения индивидуальных заданий и коллективное обсуждение полученных результатов в рамках форума данного онлайн курса на платформе.
- обсуждение на форуме онлайн курса теоретических материалов и результатов тестов с авторами курса в режиме вопрос-ответ.

Общий объем активных и интерактивных форм учебных занятий составляет 1 зачетная единица.

Раздел 2. Организация, структура и содержание учебных занятий

2.1. Организация учебных занятий

2.1.1 Основной курс

Трудоёмкость, объёмы учебной работы и наполняемость групп обучающихся																		
Код модуля в составе дисциплины, практики и т.п.	Контактная работа обучающихся с преподавателем												Самостоятельная работа				Объём активных и интерактивных форм учебных занятий	Трудоёмкость
	лекции	семинары	консультации	практические занятия	лабораторные работы	контрольные работы	коллоквиумы	текущий контроль	промежуточная аттестация	итоговая аттестация	под руководством преподавателя	в присутствии преподавателя	сам. раб. с использованием методических материалов	текущий контроль (сам.раб.)	промежуточная аттестация (сам.раб.)	итоговая аттестация (сам.раб.)		
ОСНОВНАЯ ТРАЕКТОРИЯ																		
очная форма обучения																		
5 недель									2				24		10		36	1
ИТОГО									2				24		10		36	1

Формы текущего контроля успеваемости, виды промежуточной и итоговой аттестации			
Период обучения (модуль)	Формы текущего контроля успеваемости	Виды промежуточной аттестации	Виды итоговой аттестации
ОСНОВНАЯ ТРАЕКТОРИЯ			
очная форма обучения			
5 недель		зачет	

2.2. Структура и содержание учебных занятий

Период обучения**: 5 недель**

№ п/п	Наименование темы (раздела, части)	Вид учебных занятий
1	Введение	видео уроки
		тесты
		по методическим материалам
2	Математический инструментарий науки о данных	видео уроки
		тесты
		по методическим материалам
3	Программный инструментарий науки о данных	видео уроки
		тесты
		по методическим материалам
4	Машинное обучение: обучение с учителем	видео уроки
		тесты
		по методическим материалам
5	Машинное обучение: обучение без учителя	видео уроки
		тесты
		по методическим материалам

Содержание учебных занятий

1. Введение. В этом модуле рассматривается предметная область, приводятся примеры данных и сфер для которых возникают задачи анализа данных, способы хранения данных и их виды, форматы и модели.
2. Математический инструментарий науки о данных. Даётся краткий обзор наиболее важных и часто встречаемых в задачах анализа данных терминов из теории вероятностей и математической статистики. Понятие собственного вектора и собственного числа матрицы, экстремальное свойство собственных чисел симметричной матрицы. Сингулярное разложение матрицы и его применение для аппроксимации матрицы посредством матриц небольшого ранга. Статистическая интерпретация сингулярного разложения.
3. Программный инструментарий науки о данных. Модуль посвящен основам программирования на языке Python для анализа данных, рассматривается набор базовых библиотек - NumPy, matplotlib, Pandas и sklearn. В завершении приводятся демонстрация работы некоторых алгоритмов на примере классификации текстов.
4. Машинное обучение: обучение с учителем. Рассматривается межотраслевой стандартный процесс для исследования данных, введение в машинное обучение и постановка задачи обучения с учителем, задачи классификации и регрессии. Приводятся различные оценки качества классификации и рассматриваются методы опорных векторов и алгоритмические композиции: boosting, stacking, bagging.
5. Машинное обучение: обучение без учителя. В данном модуле мы рассмотрим задачу обучения без учителя - кластеризацию. Разберем базовые методы решения данной задачи и отметим способы определения качества группировки объектов в группы-кластеры. В последнем уроке уделим внимание методу обработки естественного языка Латентно-семантическому анализу (LSA), который решают задачу дистрибутивной семантики

Раздел 3. Обеспечение учебных занятий

3.1. Методическое обеспечение

3.1.1. Методические указания по освоению дисциплины

Обучающийся автоматически зачисляется на образовательную платформу и на онлайн-дисциплину. Освоение онлайн-дисциплины возможно только с корпоративной почты @student.spbu.ru.

Обучающемуся необходимо войти на курс, используя логин выданной корпоративной электронной почты (stXXXXXX@student.spbu.ru) по следующей инструкции:

1. Войти на платформу по той ссылке, указанной в расписании.
2. Нажать «забыли пароль» и указать адрес своей корпоративной почты, на адрес которой придет ссылка-инструкция по восстановлению пароля.
3. В личном кабинете открыть вкладку «Мои курсы», в которой представлен перечень тех онлайн-курсов, которые указаны в расписании, с указанием группы.
4. Нажать «Перейти к материалам курса».

Обучающийся должен:

- ознакомиться со всеми инструкциями, данными в онлайн-курсе;
- регулярно посещать личный кабинет на платформе, где размещен онлайн-курс;
- просматривать видеоматериалы курса, изучать дополнительные материалы и выполнять контрольные задания, данные после каждого модуля.

В случае возникновения вопросов по содержанию онлайн-курса, обучающийся может обращаться на форум онлайн-курса в раздел «Обсуждения».

Обучающийся проверяет свою успеваемость в разделе «Прогресс».

Самостоятельный просмотр видео уроков дисциплины и проработка изученных на них материалов; изучение методических и раздаточных материалов курса; выполнение тестов по изучаемым разделам; пользование форумом курса на платформе онлайн-образования для обсуждения учебных материалов с авторами и другими участниками; использование рекомендованной авторами курса литературы и ресурсов в сети интернет.

3.1.2. Методическое обеспечение самостоятельной работы

Самостоятельная работа обучающихся в рамках дисциплины «Введение в науку о данных» является необходимым компонентом обучения. Методическое обеспечение самостоятельной работы осуществляется ресурсами платформы онлайн-образования.

Освоение курса осуществляется в процессе аудивизуального знакомства с содержанием онлайн-лекций и систематической самостоятельной работы, подразумевающей тщательное изучение содержания.

Методическое обеспечение самостоятельной работы включает в себя дополнительные материалы, размещенные к каждому модулю

3.1.3. Методика проведения текущего контроля успеваемости и промежуточной аттестации и критерии оценивания

Освоение дисциплины рассчитано на 5 недель, каждая из которых посвящена подробному изучению одной темы из п. 2.2.

Промежуточная аттестация проходит в форме индивидуального тестирования по изученным в течение недели материалам (видео уроки, раздаточные материалы).

Выполнение контрольных заданий по каждому модулю является обязательным. Обучающийся проверяет свою успеваемость в разделе «Отметки» («Прогресс»). Текущая успеваемость по итогам освоения модулей влияет на допуск к промежуточной аттестации по дисциплине.

Учет успеваемости обучающихся производится централизованно и передается в Учебное управление.

Промежуточная аттестация по дисциплине является обязательной.

Зачет проводится в очном (оффлайн) формате.

Допуск к промежуточной аттестации: не менее 40 % за выполнение оцениваемых контрольных заданий (КЗ) по курсу (подсчет автоматический).

Оценка «зачтено» выставляется при условии выполнения обучающимся итогового теста не менее чем на 60 % или 180 баллов.

Для выполнения тестового задания отводится до 45 минут (1 академический час).

3.1.4. Методические материалы для проведения текущего контроля успеваемости и промежуточной аттестации (контрольно-измерительные материалы, оценочные средства)

Полный перечень теоретических вопросов, практических заданий и тестов доступен на платформе онлайн-образования

Образцы тестовых заданий для проведения текущего контроля

1. На пересечении каких областей знаний лежит Наука о данных:
   a. Компьютерные науки и анализ данных
   b. Компьютерные науки, математическое моделирование, специальные знания предметной области.
   c. Анализ данных и специальные знания предметной области.
2. В какой сфере экономики Наука о данных применяется сегодня максимально интенсивно:
   a. Здравоохранение и медицина
   b. Финансовая и банковская сферы
   c. Бизнес-аналитика
3. Влияет ли специфика задачи и данных на выбор способов хранения и обработки информации?
   a. Да
   b. Нет
4. В каких случаях обычно прибегают к вычислениям на видеокартах (GPU)?
   a. Слишком много данных, их негде хранить
   b. Нужно ускорить некоторые вычислительные операции
   c. Исключительно, когда задача связана с графическими данными (обработка изображений и т.д.)
5. *Задача:* Проводится эксперимент по выбору двух случайных чисел от 1 до 100. Имеются следующие события:

   1) Оба числа четные;
   2) Числа взаимопросты;
   3) Хотя бы одно из чисел нечетно;
   4) Первое число не превосходит второе;
   5) Сумма чисел больше ста.

   Пусть эксперимент закончился с исходом (13, 28). Для каких из событий этот исход оказался благоприятным:
   [выбор одного из вариантов]
   a) 2, 3, 4
   b) 1, 2, 3, 4
   c) 3, 4
   d) 2, 3, 5
   e) 1, 4, 5
6. *Задача:* Какие из пар событий в предыдущей задаче являются несовместными?

[множественный выбор]

a) 1 и 2,
b) 1 и 3,
c) 1 и 4,
d) 1 и 5,
e) 2 и 3,
f) 2 и 4,
g) 2 и 5,
h) 3 и 4,
i) 3 и 5,
j) 4 и 5.

7. Сколько этапов стандарте CRISP-DM (CRoss Industry Standard Process for Data Mining)?
   a. Трех
   b. Пяти
   c. Шести
   d. Двух
8. Какая из ниже приведенных техник не относится к выбору модели в анализе данных?
   a. Cross correlation
   b. K-fold cross-validation
   c. Leave one out
   d. 5-fold cross-validation

3.1.5. Методические материалы для оценки обучающимися содержания и качества учебного процесса

Для оценки обучающимися содержания и качества учебного процесса используется система рейтингов и отзывов, которая является частью платформы онлайн-образования.

3.2. Кадровое обеспечение

3.2.1. Образование и (или) квалификация преподавателей и иных лиц, допущенных к проведению учебных занятий

Модерацию/содержательное сопровождение дисциплины осуществляют научно-педагогические работники, имеющие ученую степень, главные и ведущие специалисты в данной предметной области, а также обучающиеся в аспирантуре (под руководством научного руководителя) для прохождения педагогической практики.

3.2.2. Обеспечение учебно-вспомогательным и (или) иным персоналом

Не требуется.

3.3. Материально-техническое обеспечение

3.3.1. Характеристики аудиторий (помещений, мест) для проведения занятий

При проведении зачета в очной форме и для самостоятельной работы требуется стандартно оборудованный компьютерный класс.

3.3.2. Характеристики аудиторного оборудования, в том числе неспециализированного компьютерного оборудования и программного обеспечения общего пользования

Стандартно оборудованные учебные аудитории и стандартно оборудованный компьютерный класс для самостоятельной работы.

3.3.3. Характеристики специализированного оборудования

Специализированного программного обеспечения не требуется.

3.3.4. Характеристики специализированного программного обеспечения

Нет специальных требований.

3.3.5. Перечень и объёмы требуемых расходных материалов

Не требуется.

3.4. Информационное обеспечение

3.4.1. Список обязательной литературы

1. Laura Igual, Santi Seguí Introduction to Data Science A Python Approach to Concepts, Techniques and Applications, Springer – 2017
2. Nelli, Fabio Python Data Analytics Data Analysis and Science Using Pandas, matplotlib, and the Python Programming Language - Berkeley, CA : Apress : Imprint: Apress, 2015.
3. Хорн Р., Джонсон Ч. Матричный анализ. М.Мир.1989

3.4.2. Список дополнительной литературы

1. Hastie T., Tibshirani R., Friedman J. The Elements of Statistical Learning. Springer, 2014.
2. Bishop C. M. Pattern Recognition and Machine Learning. — Springer, 2006.
3. Коэльо Л.П., Ричарт В. Построение систем машинного обучения на языке Python. 2016.
4. Gan G., Ma G., Wu J. Data Clustering: Theory, Algorithms, and Application. 2007
5. Hastie T., Tibshirani R.The Elements of Statistical learning: Data Mining, Inference, and Prediction. Second Edition - Springer Series in Statistics – 2016
6. Кристофер Д. Маннинг, Прабхакар Рагхаван, Хайнрих Шютце Введение в информационный поиск, 2014
7. Марк Лутц, «Изучаем Python», Символ-Плюс , 2011
8. Sebastian Raschka, «Python Machine Learning», Packt Publishing, 2015

3.4.3. Перечень иных информационных источников

https://www.coursera.org/learn/vvedeniye-v-nauku-o-dannykh

Раздел 4. Разработчики программы

Фамилия, имя, отчество	Учёная степень	Учёное звание	Должность	Контактная информация (служебный адрес электронной почты, служебный телефон)
Блеканов Иван Станиславович	к.т.н		доцент	i.blekanov@spbu.ru
Севрюков Сергей Юрьевич	-	-	старший преподаватель	s.sevryukov@spbu.ru