Санкт-Петербургский филиал федерального государственного автономного образовательного учреждения высшего профессионального образования "Национальный исследовательский университет "Высшая школа экономики"

Факультет Санкт-Петербургская школа социальных и гуманитарных наук
Национального исследовательского университета
«Высшая школа экономики»

Рабочая программа дисциплины
«Интеллектуальный анализ данных и основы машинного обучения»

для направлений 38.03.04 «Государственное и муниципальное управление», 46.03.01 «История», 38.03.02 «Менеджмент» (ОП «Логистика и управление цепями поставок», ОП «Менеджмент»), 41.03.04 «Политология», 39.03.01 «Социология», 38.03.01«Экономика», 40.03.01 «Юриспруденция», 41.03.03 «Востоковедение и африканистика»
подготовки бакалавра (в рамках майнора «Обработка и анализ данных»)

Авторы программы:
Сироткин А.В., кандидат физико-математических наук, avsirotkin@hse.ru
Суворова А.В., кандидат физико-математических наук, asuvorova@hse.ru
Мусабиров И.Л., МА, магистр по направлению “Информатика и вычислительная техника”, ilya@musabirov.info

Утверждена академическим руководителем майнора,

А.В. Сироткин _____ «___» _____ 2016 г.

Санкт-Петербург, 2016

1 Область применения и нормативные ссылки

Настоящая рабочая программа дисциплины устанавливает минимальные требования к знаниям и умениям студента, а также определяет содержание и виды учебных занятий и отчетности.

Программа предназначена для преподавателей, ведущих данную дисциплину, учебных ассистентов и студентов направлений 38.03.04 «Государственное и муниципальное управление», 46.03.01 «История», 38.03.02 «Менеджмент», 41.03.04 «Политология», 39.03.01 «Социология», 38.03.01«Экономика», 40.03.01 «Юриспруденция», 41.03.03 «Востоковедение и африканистика» подготовки бакалавра, обучающихся по образовательным программам «Государственное и муниципальное управление», «История», «Менеджмент» «Логистика и управление цепями поставок», «Политология», «Социология», «Экономика», «Юриспруденция», «Востоковедение и африканистика» подготовки бакалавра, изучающих дисциплину «Интеллектуальный анализ данных и основы машинного обучения» в рамках майнора «Обработка и анализ данных».

Программа разработана в соответствии с:

- Образовательными стандартами НИУ ВШЭ по направлениям 38.03.04 «Государственное и муниципальное управление», 46.03.01 «История», 38.03.02 «Менеджмент», 41.03.04 «Политология», 39.03.01 «Социология», 38.03.01«Экономика», 40.03.01 «Юриспруденция», 41.03.03 «Востоковедение и африканистика» подготовки бакалавра (http://www.hse.ru/standards/standard);
- Образовательными программами «Государственное и муниципальное управление», «История», «Менеджмент» «Логистика и управление цепями поставок», «Политология», «Социология», «Экономика», «Юриспруденция» по направлениям подготовки 38.03.04 «Государственное и муниципальное управление», 46.03.01 «История», 38.03.02 «Менеджмент», 41.03.04 «Политология», 39.03.01 «Социология», 38.03.01«Экономика», 40.03.01 «Юриспруденция», 41.03.03 «Востоковедение и африканистика»;
- Рабочими учебными планами НИУ ВШЭ – Санкт-Петербург по направлениям подготовки 38.03.04 «Государственное и муниципальное управление», 46.03.01 «История», 38.03.02 «Менеджмент», 41.03.04 «Политология», 39.03.01 «Социология», 38.03.01«Экономика», 40.03.01 «Юриспруденция», 41.03.03 «Востоковедение и африканистика»

2 Цели освоения дисциплины

Целями освоения дисциплины «Интеллектуальный анализ данных и основы машинного обучения» являются изучение особенностей различных методов сбора и агрегации данных, формирование навыков планирования сбора и обработки данных, изучение пакетов ориентированных на обработку специфических данных, в частности, сетей.

3 Компетенции обучающегося, формируемые в результате освоения дисциплины

В результате освоения дисциплины студент должен:

- Знать
  - ключевые понятия, цели и задачи использования машинного обучения;
  - методологические основы применения алгоритмов машинного обучения;
- Уметь
  - визуализировать результаты работы алгоритмов машинного обучения,
  - выбирать метод машинного обучения, соответствующий исследовательской задаче,

- интерпретировать полученные результаты;
- Иметь навыки (приобрести опыт):
- чтения и анализа академической литературы по применению методов машинного обучения
- построения и оценки качества моделей

Уровни формирования компетенций:

РБ - ресурсная база, в основном теоретические и предметные основы (знания, умения)
СД - способы деятельности, составляющие практическое ядро данной компетенции
МЦ - мотивационно-ценностная составляющая, отражает степень осознания ценности компетенции человеком и готовность ее использовать

В результате освоения дисциплины студент осваивает следующие компетенции:

Компетенция	Код по ОС ВШЭ	Уровень формирования компетенции	Дескрипторы – основные признаки освоения (показатели достижения результата)	Формы и методы обучения, способствующие формированию и развитию компетенции
Способен учиться, приобретать новые знания, умения, в том числе в области, отличной от профессиональной	УК-1	РБ	Способен на основе рекомендаций и самостоятельного поиска осваивать новые методы анализа данных	Практические и самостоятельные занятия по написанию программ для сбора и обработки информации
Способен вести исследовательскую деятельность, включая анализ проблем, постановку целей и задач, выделение объекта и предмета исследования, выбор способа и методов исследования, а также оценку его качества	УК-6	РБ	Способен выделять постановки задач для решения с использованием различных методов анализа данных, осознанно выбирать методы и инструментальные средства	Программный проект по анализу данных
Способен работать с информацией: находить, оценивать и использовать информацию из различных источников, необходимую для решения научных и профессиональных задач (в том числе на основе системного подхода)	УК-5	РБ	Владеет навыками написания программ для автоматизированного сбора и анализа информации из различных источников в глобальных компьютерных сетях, приемами и навыками построения моделей машинного обучения и их оценки	Практические занятия, Прохождение онлайн курса

4 Место дисциплины в структуре образовательной программы

Настоящая дисциплина относится к блоку дисциплин дополнительного профиля (майнора) «Обработка и анализ данных», обеспечивающих бакалаврскую подготовку.

Изучение дисциплины базируется на следующих дисциплинах:

- Программирование для анализа данных и воспроизводимые исследования
- Анализ данных и технологии работы с данными

Основные положения дисциплины должны быть использованы в дальнейшем при изучении следующих дисциплин майнора.

5 Тематический план учебной дисциплины

ОБЪЕМ ДИСЦИПЛИНЫ - 5 зачетных единиц

№	Название раздела	Всего часов	Аудиторные часы			Самостоятельная работа
			Лекции	Семинары	Практические занятия	
1	Введение в машинное обучение	48	4	5	6	33
2	Регрессия	24	1	3	8	12
3	Классификация	23	1	3	8	11
4	Применение методов машинного обучения	95	2	1	2	90
ИТОГО		**190**	**8**	**12**	**24**	**146**

6 Формы контроля знаний студентов

Тип контроля	Форма контроля	1 год				Параметры
		1	2	3	4	
Текущий	Контрольная работа		*			Письменная работа на 140 минут
	Домашнее задание	*				Программный проект по анализу данных
		*	*			Прохождение онлайн-тестов
			*			Прохождение онлайн-курса
Итоговый	Экзамен		*			Письменная работа на 140 минут

7 Критерии оценки знаний, навыков

Критерии оценки прохождения тестов

Одной из форм проверки усвоение теоретических навыков являются тесты в онлайн-платформе Stepik. Их прохождение играет формативную роль: количество попыток не ограничивается. Итоговая оценка за тесты: зачтено/ не зачтено. Оценка “зачтено” ставится при прохождении суммарно более 60% заданий в установленные и отображаемые на онлайн-платформе дедлайны, без ограничения числа попыток. При получении оценки “зачтено”, в формулу накопленной оценки выставляется 10, при получении оценки “не зачтено” -- 0.

Критерии заданий, включающих элементы анализа данных (в составе программного проекта, контрольных работ, экзамена)

- корректность применения методик анализа (в рамках знаний, полученных в курсе, смежных дисциплинах, домашнем чтении) и интерпретации результатов;
- уверенность использования языковых средств и структур данных, методов преобразования и агрегации данных в организации потока анализа данных, их ввода и вывода.

Критерии оценки прохождения онлайн-курса

Прохождение онлайн-курса является обязательной составляющей курса. Выбирая из предложенных преподавателями курсов или обсуждая с преподавателями найденные самостоятельно, студент сам выбирает уровень сложности курса. Оценка “Зачтено” выставляется по факту высылки сертификата о завершении курса или скриншота об успешном выполнении заданий курса, в соответствии с его требованиями. Если длительность курса превышает количество недель, имеющихся для его прохождения, по согласованию с преподавателем его отдельные модули могут быть пропущены студентом. Предъявление сертификата о завершении или скриншота об успешном выполнении курса ведет к оценке “зачтено”. При получении оценки “зачтено”, в формулу накопленной оценки выставляется 10, при получении оценки “не зачтено” -- 0. Дедлайн сдачи онлайн-курса -- за одну рабочую неделю до завершения 2-го модуля. В случае отсутствия сдачи онлайн-курса к этой дате, студенту выставляется оценка “не зачтено”
Дополнительные вопросы по онлайн-курсу могут быть добавлены в итоговый экзамен.

8 Содержание дисциплины

Раздел 1. Введение в машинное обучение

Задачи классификации и регрессии. Статистическая теория принятия решений. Разложение bias-variance-noise. Переобучение.

Литература по разделу

- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning. New York: Springer. (гл. 1, 2)
- Roger D. Peng and Elizabeth Matsui. The Art of Data Science. A Guide for Anyone Who Works with Data. (гл. 1, 2, 5-7)
- Roger D. Peng. Exploratory Data Analysis with R. (гл. 5)

Раздел 2. Регрессия.

Линейная регрессия: разные формы регуляризаторов. Лассо-регрессия. Гребневая регрессия
Регрессионные деревья.

Литература по разделу

- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning. New York: Springer. (гл. 3, 6-8)
- Roger D. Peng and Elizabeth Matsui. The Art of Data Science. A Guide for Anyone Who Works with Data. (гл. 7-9)

Раздел 3. Классификация.

Логистическая регрессия. Метод опорных векторов (SVM). Трюк с ядрами.
Метод ближайших соседей. Классификационные деревья.

Литература по разделу

- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning. New York: Springer. (гл. 4, 8, 9)
- Roger D. Peng and Elizabeth Matsui. The Art of Data Science. A Guide for Anyone Who Works with Data. (гл. 7-9)

Раздел 4. Применение методов машинного обучения
Рекомендательные системы. Сервисы машинного обучения. Обзор онлайн-платформ машинного обучения. Применение машинного обучения в бизнесе.
Литература по разделу

- Roger D. Peng and Elizabeth Matsui. The Art of Data Science. A Guide for Anyone Who Works with Data. (гл. 10, 1)
- Provost, Foster, and Tom Fawcett. 2013. Data Science for Business: What You Need to Know about Data Mining and Data-Analytic Thinking. 1 edition. Sebastopol, Calif.: O'Reilly Media.
- http://www.quora.com/Recommendation-Engines

7 Образовательные технологии

Преподавание языковых средств R и концепций эксплораторного анализа данных и статистического обучения осуществляется с использованием современных абстракций, с упором на понимание на концептуальном уровне и формальным введением по мере необходимости.

8. 1 Методические указания студентам по освоению дисциплины

Для обеспечения необходимого уровня уверенного владения инструментальными средствами (языком R и средой RStudio) предусмотрен сквозной компьютерный практикум по всем разделам курса, Кроме того, рабочая среда с веб-доступом позволяет прозрачно переносить работу между практикумом и самостоятельной работой студента. Поэтому для успешного освоения дисциплины студент должен пользоваться возможностью самостоятельной работы, дополнительными ресурсами, указанными в программе и на форуме курса. **Пользуйтесь возможностью самостоятельной работы!**
Свои знания по части материалов курса можно проверить используя онлайн-тесты. Хотя порог их прохождения на оценку достаточно низок, **мы советуем вам проходить все тесты и задавать вопросы на форуме.**
Учебные ассистенты, преподаватели, ваши однокурсники и старшекурсники часто делятся дополнительными ресурсами на форуме курса. **Пользуйтесь этой возможностью, задавайте и отвечайте на вопросы других.** Помимо возможного бонуса к оценке, это позволяет более глубоко усвоить материал.
Освоение одного из онлайн-курсов из списка одобренных преподавателями курса является обязательным! По факту завершения онлайн-курса от студента требуется выслать сертификат о завершении курса или скриншот об успешном выполнении заданий. По итогам контроля выполнения онлайн-курса выставляется часть накопительной оценки. Дополнительные вопросы по содержанию пройденного курса могут быть добавлены в итоговый экзамен.

8. 2 Учебно-методическая литература для самостоятельной работы студентов

- Grolemund, Garrett, and Hadley Wickham. 2016. *R for Data Science*. http://r4ds.had.co.nz/

- Roger D. Peng and Elizabeth Matsui. The Art of Data Science. A Guide for Anyone Who Works with Data. https://leanpub.com/artofdatascience
- Roger D. Peng. Exploratory Data Analysis with R. https://leanpub.com/exdata
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning. New York: Springer. http://www-bcf.usc.edu/~gareth/ISL/
- Provost, Foster, and Tom Fawcett. 2013. Data Science for Business: What You Need to Know about Data Mining and Data-Analytic Thinking. 1 edition. Sebastopol, Calif.: O'Reilly Media.
- Gandrud, C. (2013). Reproducible Research with R and R Studio. CRC Press.
- Davenport, Thomas H. 2014. Big Data at Work: Dispelling the Myths, Uncovering the Opportunities. Boston Massachusetts: Harvard Business Review Press.
- Siegel, Eric, and Thomas H. Davenport. 2013. Predictive Analytics: The Power to Predict Who Will Click, Buy, Lie, or Die. 1 edition. Hoboken, N.J: Wiley.

8. 3 Перечень учебно-методического обеспечения для самостоятельной работы обучающихся по дисциплине

Освоение одного из онлайн-курсов из списка одобренных преподавателями курса является обязательным! По факту завершения онлайн-курса от студента требуется выслать сертификат о завершении курса или скриншот об успешном выполнении заданий. По итогам контроля выполнения онлайн-курса выставляется часть накопительной оценки.

- Statistical Learning. https://lagunita.stanford.edu/courses/HumanitiesSciences/StatLearning/Winter2016/about
- Машинное обучение. https://www.coursera.org/learn/machine-learning
- Practical Machine Learning. https://www.coursera.org/learn/practical-machine-learning
- Machine Learning Foundations: A Case Study Approach; Machine Learning: Regression. https://www.coursera.org/learn/ml-foundations; https://www.coursera.org/learn/ml-regression
- Principles of Machine Learning; Applied Machine Learning. https://www.edx.org/course/principles-machine-learning-microsoft-dat203-2x-1#!; https://www.edx.org/course/applied-machine-learning-microsoft-dat203-3x#!
- Data Science Essentials. https://www.edx.org/course/data-science-essentials-microsoft-dat203-1x-1#!

9. Оценочные средства для текущего контроля и аттестации студента

9. 1 Вопросы для оценки качества освоения дисциплины

1. В чем принципиальное различие задач классификации и регрессии? Приведите содержательные примеры задач классификации и регрессии.
2. Как алгоритмом Conditional Inference Trees выбираются разбиения на каждом шаге? Какие существуют альтернативные способы выбора разбиений? В чем преимущества и недостатки этих вариантов?
3. В чем особенности задач обучения с учителем (supervised learning)?
4. Зачем нужно разделение на тестовую и обучающую выборки?
5. О чем говорит accuracy модели? Как она определяется?
6. В чем разница между Precision и Recall модели? Как они определяются? Для какого класса задач машинного обучения они применяются?

7. Объясните, что означают и как вычисляются: TRUE POSITIVE, TRUE NEGATIVE, FALSE POSITIVE, FALSE NEGATIVE. Для какого класса задач машинного обучения они применяются?

9.2 Пример заданий, входящих в тесты, контрольную работу и экзамен

1. Что из перечисленного является задачей обучения с учителем (supervised learning)?
(a) Выявить сочетания товаров, часто встречающихся вместе в покупках клиентов
(b) Выделить группы покупателей на основе сведений об истории их покупок
(c) Определить пол владельца аккаунта ВКонтакте, используя базу из 10000 аккаунтов с указанным полом
(d) Сформировать репрезентативную выборку для социологического опроса

2. Зачем нужно разделение на тестовую и обучающую выборки? Выберите все верные утверждения.
(a) Данных обычно слишком много, алгоритмы работают долго. Разделение нужно только для увеличения производительности
(b) На тестовой выборке проверяется окончательная модель, при построении модели она не используется
(c) Если обучать и тестировать модель на одних и тех же данных, то, скорее всего, мы получим модель, слишком хорошо подогнанную именно к этим данным и плохо обобщаемую
(d) Оценку качества построенной модели лучше прозводить на независимых данных, которые не использовались для обучения. Таким образом мы стараемся избежать переобучения

3. **Загрузите данные про ирисы и пакеты для работы с алгоритмами машинного обучения**

```
data(iris)
library(caret)
library(rpart)
Установите
set.seed(700)
```

Разделите выборку на тестовую и обучающую в пропорции 30:70 (30% - на тестовую выборку). Обучите модель для предсказания сорта ирисов с помощью дерева решений (rpart()). В качестве ответа внесите точность предсказания на тестовой выборке (округлять не нужно)

4. Загрузите данные следующей командой.

```
library(readr)
money_data<-read_csv("~/shared/minor3_2016/data/money.csv")
```

Данные содержат информацию о общих доходах (income), общих расходах (spend), расходах по отдельным категориям (restaurant, cosmetic) и поле (gender), а также идентификатор каждого клиента (customer_id).
Для загруженных данных постройте линейную регрессию, предсказывающую доходы на основе остальных переменных. Не забудьте исключить поле с идентификатором (customer_id).

Какой коэффициент в полученной модели у переменной restaurant?

5. Загрузите данные

library(readr)
money_data<-read_csv("~/shared/minor3_2016/data/money.csv")
money_data$gender <- as.factor(money_data$gender)

Оставьте только те строки, где есть ненулевые расходы на рестораны и косметику (хотя бы на что-то одно из перечисленного)
Установите

set.seed(500)

Отделите случайно 20% строк в тестовую выборку

*test.ind = sample(seq_len(nrow(money_data)), size = nrow(money_data)*0.2)*

Для предсказания пола (gender) только по тратам на рестораны (restoraunt) и косметику (cosmetic) постройте две модели:
методом опорных векторов с линейным ядром,
методом 5-ти ближайших соседей.
Не забудьте загрузить нужные пакеты

library(e1071)
library(caret)
library(class)

В ответе укажите ту точность на тестовой выборке, что оказалась лучше (значение округлять не нужно)

11 Порядок формирования оценок по дисциплине

Накопленная оценка по дисциплине рассчитывается с помощью взвешенной суммы оценок за отдельные формы текущего контроля знаний следующим образом:

$$\mathbf{O_{накопленная} = 0.3*O_{текущий\ 1} + 0.25*O_{текущий\ 2} + 0.25*O_{текущий\ 3} + 0.2*O_{текущий\ 4}},$$

где:
$O_{текущий\ 1}$ – оценка за контрольную работу,
$O_{текущий\ 2}$ – оценка за домашнее задание - программный проект,
$O_{текущий\ 3}$ – оценка за домашнее задание - тесты.
$O_{текущий\ 4}$ – оценка за домашнее задание - онлайн-курс.

Результирующая оценка по дисциплине рассчитывается следующим образом:

$$O_{результ} = 0{,}70*O_{накопл} + 0{,}30*O_{экз}\text{, где}$$

$O_{накопл}$ – накопленная оценка по дисциплине
$O_{экз}$ – оценка за экзамен (включая вопросы по изученному студентом онлайн-курсу)

Способ округления экзаменационной и результирующей оценок: арифметический.
Активность на форуме курса приносит бонус к результирующей оценке до 20%.

12. Учебно-методическое и информационное обеспечение дисциплины

12. 1 Основная литература

- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning. New York: Springer. http://www-bcf.usc.edu/~gareth/ISL/
- Roger D. Peng and Elizabeth Matsui. The Art of Data Science. A Guide for Anyone Who Works with Data. https://leanpub.com/artofdatascience
- Roger D. Peng. Exploratory Data Analysis with R. https://leanpub.com/exdata

12. 2 Дополнительная литература

- Provost, Foster, and Tom Fawcett. 2013. Data Science for Business: What You Need to Know about Data Mining and Data-Analytic Thinking. 1 edition. Sebastopol, Calif.: O'Reilly Media.
- Gandrud, C. (2013). Reproducible Research with R and R Studio. CRC Press.
- Davenport, Thomas H. 2014. Big Data at Work: Dispelling the Myths, Uncovering the Opportunities. Boston Massachusetts: Harvard Business Review Press.
- Siegel, Eric, and Thomas H. Davenport. 2013. Predictive Analytics: The Power to Predict Who Will Click, Buy, Lie, or Die. 1 edition. Hoboken, N.J: Wiley.

12. 3 Программные средства

RStudio. Пакет MS Office/OpenOffice.org

13 Материально-техническое обеспечение дисциплины

Аудитория с проектором для лекций, компьютерные аудитории с современными версиями браузеров согласно требованиям RStudio Server к клиентам. Клиент ssh для консольного доступа к серверу. Сервер для работы студентов (спецификация в зависимости от количества записавшихся на майнор) с Ubuntu Linux, RStudio Server и другими пакетами, в зависимости от специфики проектов.