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

Факультет Санкт-Петербургская школа экономики и менеджмента
Национального исследовательского университета
«Высшая школа экономики»

Департамент прикладной математики и бизнес-информатики

Рабочая программа дисциплины
Современные методы анализа данных

для образовательной программы «Анализ больших данных в бизнесе, экономике и обществе»
направления подготовки 01.04.02«Прикладная математика и информатика»
уровень магистратура

Разработчик(и) программы
Николенко С.И., доцент, snikolenko@hse.ru

Согласована менеджером ОП Анализ больших данных в бизнесе, экономике и обществе

«30»августа 2016г.

Е.С. Авдониной _____

Утверждена Академическим руководителем образовательной программы

А.В. Сироткин _____

«30»августа 2016г.

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 дисциплину «Современные методы анализа данных», учебных ассистентов и студентов направления подготовки 01.04.02 «Прикладная математика и информатика», обучающихся по образовательной программе «Анализ больших данных в бизнесе, экономике и обществе».

Рабочая программа дисциплины разработана в соответствии с:

- Образовательным стандартом НИУ ВШЭ http://www.hse.ru/data/2016/11/02/1111123560/01.04.02%20Прикладная%20математика%20и%20информатика.pdf;
- Образовательной программой «Анализ больших данных в бизнесе, экономике и обществе», направление подготовки 01.04.02 «Прикладная математика и информатика»;
- Объединенным учебным планом университета по образовательной программ «Анализ больших данных в бизнесе, экономике и обществе».

2 Цели освоения дисциплины

Целью освоения дисциплины «Современные методы анализа данных» является изучение основных аппаратов машинного обучения, эффективных алгоритмов обучения и применения обученных моделей, основ теории байесовского вывода. В результате изучения дисциплины у студента будет сформировано представление о современном состоянии дел в теории байесовского вывода. Студент получит также представление об основных методах машинного обучения, соответствующих алгоритмах вывода, вероятностных основах машинного обучения и соответствующих моделях. Изучение дисциплины будет способствовать как развитию вероятностной интуиции и разработке моделей и методов машинного обучения, так и практическому их применению.

3 Компетенции обучающегося, формируемые в результате освоения дисциплины

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	Форма контроля уровня сформированности компетенции
Системные компетенции					
Способен создавать новые теории, изобретать новые способы и инструменты профессиональной деятельности.	СК-2	РБ	Может модифицировать существующие и создавать новые алгоритмы анализа данных и модели представления данных.	Лекции, семинарские занятия	Текущий контроль, Экзамен
Способен совершенствовать и развивать свой интеллектуальный и культурный уровень, строить траекторию профес-	СК-4	РБ	Умеет исправлять свои ошибки, учитывать замечания и рекомендации и интегрировать их в свою работу.	Семинарские занятия	Экзамен

Компетенция	Код по ОС ВШЭ	Уровень формирования компетенции	Дескрипторы – основные признаки освоения (показатели достижения результата)	Формы и методы обучения, способствующие формированию и развитию компетенции	Форма контроля уровня сформированности компетенции
сионального развития и карьеры.					
Социально-личностные компетенции					
Способен использовать социальные и мультикультурные различия для решения проблем в профессиональной и социальной деятельности.	ПК-3	РБ	Может вести диалог и обосновывать принятые решения об использованных методах анализа.	Семинарские занятия	Текущий контроль
Способен определять, транслировать общие цели в профессиональной и социальной деятельности.	ПК-4	РБ	Способен ставить общую задачу анализа данных с учетом имеющихся ограничений.	Семинарские занятия	Текущий контроль
Инструментальные компетенции					
Способен публично представлять результаты профессиональной деятельности (в том числе с использованием информационных технологий).	ПК-12	РБ, СД	Обладает навыками презентации результатов своей работы. умеет сопровождать аналитические материалы соответствующим иллюстративным.	Семинарские занятия, самостоятельная работа студентов	Экзамен, текущий контроль.

4 Место дисциплины в структуре образовательной программы

Настоящая дисциплина относится к циклу дисциплин направления.

Изучение данной дисциплины базируется на следующих дисциплинах:

- «Линейная алгебра»,
- «Теория вероятностей и математическая статистика»,
- «Математический анализ».

Для освоения учебной дисциплины, студенты должны владеть следующими знаниями и компетенциями:

- Быть способными создавать новые теории, изобретать новые способы и инструменты профессиональной деятельности;
- Быть способными совершенствовать и развивать свой интеллектуальный и культурный уровень, строить траекторию профессионального развития и карьеры;
- Быть способными анализировать, верифицировать, оценивать полноту информации в ходе профессиональной деятельности, при необходимости восполнять и синтезировать недостающую информацию;
- Быть способным использовать социальные и мультикультурные различия для решения проблем в профессиональной и социальной деятельности;
- Быть способными определять, транслировать общие цели в профессиональной и социальной деятельности;

- Быть способными анализировать и воспроизводить смысл междисциплинарных текстов с использованием языка и аппарата прикладной математики;
- Быть способными публично представлять результаты профессиональной деятельности (в том числе с использованием информационных технологий).

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

- «Современные методы принятия решений»,
- «Распределенная обработка и анализ больших данных».

5 Тематический план учебной дисциплины

ОБЪЕМ ДИСЦИПЛИНЫ – 6 зачетных единиц.

№	Название раздела	Всего часов	Аудиторные часы		Самостоятельная работа
			Лекции	Семинары	
1	Введение. История искусственного интеллекта. Вспоминаем теорию вероятностей. Теорема Байеса и машинное обучение. Что умеет делать машинное обучение.	8	2	2	4
2	Правило Лапласа. Априорные распределения. Сопряжённые априорные распределения.	14	2	2	10
3	Наименьшие квадраты и ближайшие соседи. Линейная регрессия. Логистическая регрессия.	14	2	2	10
4	Статистическая теория принятия решений. Разложение bias-variance-noise. Оверфиттинг. Регуляризация: гребневая регрессия. Линейная регрессия по-байесовски.	14	2	2	10
5	Линейная регрессия: разные формы регуляризаторов. Лассо-регрессия. Эквивалентные ядра. Проклятие размерности.	14	2	2	10
6	Задачи классификации. Линейный дискриминант Фишера. Наивный байесовский классификатор: мультиномиальный и многомерный.	14	2	2	10
7	Логистическая регрессия: как обучать. Мультиклассовая логистическая регрессия. Аппроксимация по Лапласу. Пробит. Логистическая регрессия по-байесовски.	14	2	2	10
8	Метод опорных векторов (SVM). Трюк с ядрами.	14	2	2	10
9	Варианты SVM. SVM по-байесовски: relevance vector machines.	14	2	2	10
10	Кластеризация: иерархическая, методами теории графов. Алгоритм EM для кластеризации.	20	4	2	14
11	Скрытые марковские модели.	22	4	4	14
12	Комбинация моделей: усреднение, бутстрап, бэггинг. Бустинг: AdaBoost.	22	4	4	14

13	Обучение ранжированию: постановка задачи, RankBoost. LambdaRank.	22	4	4	14
14	Рекомендательные системы: метод ближайших соседей, сингулярное разложение матриц.	22	2	4	16
ИТОГО		**228**	**36**	**36**	**156**

6 Формы контроля знаний студентов

Тип контроля	Форма контроля	1 курс		Параметры **
		3 модуль	4 модуль	
Текущий	Контрольная работа		*	Письменная работа 80 минут
	Домашнее задание №1	*		Домашнее задание связанное с компьютерной реализацией указанных преподавателем алгоритмов.
	Домашнее задание №2		*	Домашнее задание связанное с компьютерной реализацией указанных преподавателем алгоритмов.
Итоговый	Экзамен		*	Письменный экзамен 120 мин.

7 Критерии оценки знаний, навыков

Критерии оценки **контрольной работы**:

Правильное решение предложенных задач по теме практического занятия, которое включает в себя корректное использование изученного математического аппарата к поставленным задачам.

Для предложенного набора задач указываются максимальные баллы получаемые за каждую задачу, общей суммой 10 баллов.

При оценки решения каждой задачи учитываются следующие факторы:

- правильность полученных ответов;
- корректность применяемых математических методов.

За решение каждой задачи выставляются накопительные баллы в соответствии с точностью и полнотой решения и описания процесса решения задачи.

Итоговое максимальное число баллов за одну контрольную работу – **10 баллов.**

Критерии оценки **за домашнее задание**:

Задание представляет собой программную реализацию заданных алгоритмов машинного обучение на языке программирования, согласованном с преподавателем. Результаты работы предоставляются в виде кода на выбранном языке.

Программы, не сданные в установленную дату, не принимаются. В таком случае выставляется оценка 0 баллов.

При оценке учитывается корректность работы программы, а так же понятность кода. Студенту могут задаваться вопросы с просьбой пояснить конкретные выбранные им способы решения задачи.

Максимальная оценка за домашнюю работу – **10 баллов.**

Критерии оценки **за экзамен:**

На экзамене содержится ряд задач, охватывающий все темы курса. За каждое правильно выполненное задание присваивается накопительные баллы, которые суммируются и переводятся в 10-балльную систему.

Экзамен **($О_{экз}$)** проводится в письменной форме. За каждое правильно выполненное задание присваивается накопительные баллы, которые суммируются.

Максимальная оценка за экзамен – **10 баллов.**

8 Содержание дисциплины

1. Введение. История искусственного интеллекта. Вспоминаем теорию вероятностей. Теорема Байеса и машинное обучение. Что умеет делать машинное обучение.

2. Правило Лапласа. Априорные распределения. Сопряжённые априорные распределения.

3. Наименьшие квадраты и ближайшие соседи. Линейная регрессия. Логистическая регрессия.

4. Статистическая теория принятия решений. Разложение bias-variance-noise. Оверфиттинг. Регуляризация: гребневая регрессия. Линейная регрессия по-байесовски.

5. Линейная регрессия: разные формы регуляризаторов. Лассо-регрессия. Эквивалентные ядра. Проклятие размерности.

6. Задачи классификации. Линейный дискриминант Фишера. Наивный байесовский классификатор: мультиномиальный и многомерный.

7. Логистическая регрессия: как обучать. Мультиклассовая логистическая регрессия. Аппроксимация по Лапласу. Пробит. Логистическая регрессия по-байесовски.

8. Метод опорных векторов (SVM). Трюк с ядрами.

9. Варианты SVM. SVM по-байесовски: relevance vector machines.

10. Кластеризация: иерархическая, методами теории графов. Алгоритм EM для кластеризации.

11. Скрытые марковские модели.

12. Комбинация моделей: усреднение, бутстрап, бэггинг. Бустинг: AdaBoost.

13. Обучение ранжированию: постановка задачи, RankBoost. LambdaRank.

14. Рекомендательные системы: метод ближайших соседей, сингулярное разложение матриц.

9 Образовательные технологии

9.1 Методические рекомендации преподавателю

Индивидуальные методические рекомендации преподавателю не требуются.

9.2 Методические указания студентам по освоению дисциплины

Самостоятельная работа студента включает:

- поиск и анализ информации по тематике курса;
- знакомство со специальной научной литературой по тематике курса;
- решение задач, включая программирование и визуализацию на компьютере, и анализ ситуаций, выданных преподавателем;
- подготовку к семинарским занятиям.

9.3 Перечень учебно-методического обеспечения для самостоятельной работы обучающихся по дисциплине

В качестве дополнительной самостоятельной подготовки студентам рекомендуется использовать сайт https://www.kaggle.com, для участия в соревнованиях по построению моделей, а так же для изучения разбора лучших решений, для завершенных соревнований.

10 Оценочные средства для текущего контроля и аттестации студента

10.1 Оценочные средства для оценки качества освоения дисциплины в ходе текущего контроля

Примеры заданий для контрольной работы:

1. Запишите апостериорное распределение линейной регрессии после получения N точек данных. Используя его как априорное, пересчитайте параметры апостериорного распределения после добавления ещё одной точки.

2. Для данной направленной графической модели вычислите вероятность p(x) (узел x и вероятности показаны на графе).

Пример домашнего задания № 1:

Сравнить поведение линейной регрессии с разными регуляризаторами на заданном датасете. Описать полученные результаты.

Пример домашнего задания № 2:

Реализовать EM-алгоритм кластеризации на заданном датасете, визуализировать результат при помощи t-SNE библиотеки.

10.2 Примеры заданий итогового контроля

1. Запишите градиент логарифма правдоподобия и матрицу Гессе для пробит-регрессии.

2. Что такое отложенная выборка, в чем её недостатки,, в оценке качества алгоритмов классификации.

3.Что такое регуляризация и для чего она используется. Приведите примеры разных типов регуляризации для линейной регрессии.

11 Порядок формирования оценок по дисциплине

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

$$О_{накопленная} = 0{,}3О_{дз1} + 0{,}3О_{дз2} + 0{,}4О_{кр}, \text{ где}$$

$О_{дз1}$ - оценка знаний студента за **домашнее задание № 1**;
$О_{дз2}$ - оценка знаний студента за **домашнее задание № 2**;
$О_{кр}$ – оценка знаний студента за **контрольную работу**;

Результирующая оценка по дисциплине (которая идет в диплом) рассчитывается следующим образом:

$$О_{результ} = 0{,}6О_{накопленная} + 0{,}4О_{экз}, \text{ где}$$

$О_{накопленная}$ - накопленная оценка по дисциплине;
$О_{экз}$ - оценка за экзамен.

В формулу для $\mathbf{O_{результ}}$ подставляются значения $\mathbf{O_{накопленная}}$ и $\mathbf{O_{экз}}$, округленные до ближайшего целого значения. $\mathbf{O_{результ}}$ округляется до ближайшего целого значения.

По усмотрению ведущего преподавателя, если это не противоречит действующим документам на момент экзамена, при получении накопленной оценки 8 баллов и более, студент может быть освобожден от экзамена. В таком случае, с согласия студента, ему выставляется результирующая оценка, равная накопленной.

Студент не получает возможность пересдать низкие результаты за домашнюю работу и/или контрольную работу, а также при пропуске соответствующих им учебных часов, при отсутствии уважительной причины пропуска.

При наличии уважительной причины пропуска (болезнь или другая документально подтвержденная причина), студент получает право пересдать пропущенные работы, при этом документы подтверждающие причину отсутствия, должны быть представлены в учебный офис, в течении недели со дня первого выхода на занятия после пропуска.

При получении неудовлетворительной оценки $\mathbf{O_{результ}}$ (значение после округления менее 4 баллов) выставляется оценка «НЕУДОВЛЕТВОРИТЕЛЬНО».

12 Учебно-методическое и информационное обеспечение дисциплины

12.1 Основная литература

1. Murphy K. P. Machine learning: a probabilistic perspective [Electronic Resource] / Kevin P. Murphy.- Cambridge University Press, 2012. - 1098 p. - Authorized access: http://site.ebrary.com/lib/hselibrary/detail.action?docID=10597102 (Online Digital Library "Ebrary").

12.2 Дополнительная литература

1. Hastie T. The Elements of Statistical Learning: Data Mining, Inference, and Prediction [Electronic Resource] / Trevor Hastie, Robert Tibshirani, Jerome Friedman. - Springer, 2009.- Authorized access: http://link.springer.com/book/10.1007/978-0-387-84858-7 (Online Digital Library "Springer eBooks").
2. Richert W. Building machine learning systems with python [Electronic Resource] / Willi Richert, Luis Pedro Coelho, Jonathan Chaffer.- Packt Publishing Ltd, 2013.- 290 p. Authorized access: http://site.ebrary.com/lib/hselibrary/detail.action?docID=10742638 (Online Digital Library "Ebrary").
3. Sammut C., Webb G. I. (ed.). Encyclopedia of machine learning [Electronic Resource] / Claude S., Geoffrey I. Webb/- – Springer Science & Business Media, 2011.- Authorized access: http://link.springer.com/referencework/10.1007/978-0-387-30164-8 (Online Digital Library "Springer eBooks").

12.3 Справочники, словари, энциклопедии

Sammut C., Webb G. I. (ed.). Encyclopedia of machine learning [Electronic Resource] / Claude S., Geoffrey I. Webb/- – Springer Science & Business Media, 2011.- Authorized access: http://link.springer.com/referencework/10.1007/978-0-387-30164-8 (Online Digital Library "Springer eBooks").

12.4 Ресурсы информационно-телекоммуникационной сети «Интернет»

Использование ресурсов информационно-телекоммуникационной сети «Интернет» не требуются.

12.5 Программные средства

Для успешного освоения дисциплины, студент использует следующие **программные средства:**

- Язык программирования Python

12.6 Информационные справочные системы

Информационные справочные системы для освоения дисциплины не требуются.

12.7 Дистанционная поддержка дисциплины

Дистанционная поддержка дисциплины также не требуется.

13 Материально-техническое обеспечение дисциплины

Семинары выполняются в компьютерном классе.
Необходим проектор для лекций.