**Санкт-Петербургский государственный университет**

**Р А Б О Ч А Я П Р О Г Р А М М А**

**УЧЕБНОЙ ДИСЦИПЛИНЫ**

Вычислительная статистика и статистические задачи машинного обучения

Computational Statistics and Statistical Methods for Machine Learning

**Язык(и) обучения**

русский

Трудоемкость в зачетных единицах: 2

Регистрационный номер рабочей программы: 055197

**Раздел 1. Характеристики учебных занятий**

**1.1. Цели и задачи учебных занятий**

Целью изучения дисциплины «Вычислительная статистика и статистические задачи машинного обучения» является освоение обучающимися основных методов построения стохастических моделей и обработки реальных статистических данных.   
В результате изучения дисциплины обучающиеся должны знать содержание дисциплины и иметь общее представление о принципах построения стохастических моделей и статистической обработке данных.

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

Дисциплина изучается в шестом семестре обучения. Обучающиеся должны обладать математической культурой, обычно приобретаемой в течение первых пяти семестров обучения.

**1.3. Перечень результатов обучения (learning outcomes)**

В результате освоения дисциплины обучающиеся должны владеть основными понятиями и методами построения стохастических моделей и обработки реальных статистических данных, указанными в разделе 2.2.

Кроме того, в результате обучения обучающиеся должны:   
знать и уметь использовать основные современные понятия, и подходы к моделированию стохастических динамических процессов;

получить навыки построения статистических моделей и анализа данных;

знать и уметь учитывать основные тенденции развития современного естествознания;

обладать способностью активно применять общенаучные базовые знания естественных наук, математики и информатики в области прикладной математики и информатики;

обладать способностью работы с информацией из различных источников, включая сетевые ресурсы сети Интернет, для решения профессиональных задач.

**1.4. Перечень и объём активных и интерактивных форм учебных занятий**

Ответы на вопросы вопросы учащихся во время занятий.

Обсуждение полученных на занятии результатов.

Совместное решение задач, связанных с изученным материалом.

Выполнение индивидуальных заданий

Личные и групповые консультации.

Совместный поиск ошибок в заведомо неправильных рассуждениях.

**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 | | | | | | | | | | | | | | | | | | |
| Форма обучения: очная | | | | | | | | | | | | | | | | | | |
| Семестр 7 | 30 |  | 2 |  |  |  |  |  | 2 |  |  |  | 10 |  | 28 |  | 2 | 2 |
|  | 2-100 |  | 2-100 |  |  |  |  |  | 2-100 |  |  |  | 1-1 |  | 1-1 |  |  |  |
| ИТОГО | 30 |  | 2 |  |  |  |  |  | 2 |  |  |  | 10 |  | 28 |  |  | 2 |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Виды, формы и сроки текущего контроля успеваемости и промежуточной аттестации | | | | | | |
| Код модуля в составе дисциплины, практики и т.п. | Формы текущего контроля успеваемости | | Виды промежуточной аттестации | | Виды итоговой аттестации  (только для программ итоговой аттестации и дополнительных образовательных программ) | |
| Формы | Сроки | Виды | Сроки | Виды | Сроки |
| ОСНОВНАЯ ТРАЕКТОРИЯ | | | | | | |
| Форма обучения: очная | | | | | | |
| Семестр 7 |  |  | экзамен, устно, традиционная форма | по графику промежуточной аттестации |  |  |

**2.2. Структура и содержание учебных занятий**

Период обучения (модуль): **Семестр 7**

|  |  |  |  |
| --- | --- | --- | --- |
| **№ п/п** | **Наименование темы (раздела, части)** | **Вид учебных занятий** | **Кол-во часов** |
| 1 | Стохастическое моделирование | лекции | 8 |
| самостоятельная работа по методическим материалам | 2 |
| 2 | Введение в статистическую обработку данных. Первичная обработка данных. | лекции | 6 |
| самостоятельная работа по методическим материалам | 2 |
| 3 | Вычислительные задачи статистики | лекции | 8 |
| самостоятельная работа по методическим материалам | 2 |
| 4 | Машинное обучение | лекции | 8 |
| самостоятельная работа по методическим материалам | 4 |
| 5 | Промежуточная аттестация | консультации | 2 |
| самостоятельная работа | 28 |
| экзамен | 2 |
| Итого | |  | 72 |

**Тема 1. Математическое Моделирование.**

Основные понятия. Компьютерное моделирование. Статистическое моделирование (Метод Монте-Карло). Общая схема метода Монте-Карло. Моделирование случайных величин, имеющих равномерное распределение. Сведение моделирования любого распределения к равномерному. Моделирование экспоненциального и нормального распределения. Вычисление интеграла по вероятностной мере. Методы понижения дисперсии.

Системы и сети массового обслуживания как пример математической модели. Структура систем массового обслуживания.

Алгоритм обслуживания. Дисциплина обслуживания. Абсолютный и относительный приоритеты.

Символика Кендалла. Распределение Эрланга.

Процесс Пуассона, его свойства и методы его моделирования.

Сети массового обслуживания, ее узлы матрица маршрутизации, расширенная матрица маршрутизации. Примеры. Методы моделирования сетей массового обслуживания. Основные этапы моделирования сетей массового обслуживания. Метод особых событий и метод дельта t.

МСМС-метод и его применения.

**Тема 2. Введение в статистическую обработку данных. Первичная обработка данных**

Исторические корни математической статистики. Ее связь с другими науками. Основные числовые характеристики выборки и общий принцип построения их оценок. Выборочные среднее и дисперсия, начальный и центральный моменты.Оценки параметров положения. Медиана, мода, выборочное среднее. Оценки параметров рассеяния. Выборочная дисперсия, стандартное отклонение, абсолютное отклонение, квантили, квартили, персентили, интерквартильный размах, размах. Оценки параметров формы плотности. Асимметрия и эксцесс. Роль нормального распределения в статистике.Теорема об асимптотической нормальности функций от асимптотически нормальных случайных величин. Асимптотическая нормальность выборочных дисперсии, асимметрии и корреляции. Многомерное нормальное распределение для произвольной ковариационной матрицы. Его плотность и график. Болезнь размерности. Арифметика нормального распределения. Независимость выборочных средних и дисперсии. Распределения связанные с нормальным: хи-квадрат, Стьюдента, Фишера. Где каждое из них применяется. Семипараметрический подход к оцениванию и использование его для анализа погрешности.

**Тема 3. Вычислительные задачи математической статистики**

Теория статистического оценивания. Общая постановка задачи. Определения статистики, статистической оценки, квадратичной функции риска и ее связь с дисперсией оценки. Примеры статистических оценок. Доверительное оценивание (зачем оно нужно?). Доверительный интервал, уровень доверия. Доверительный интервал для математического ожидания нормального распределения (дисперсия известна и неизвестна). Доверительный интервал для дисперсии нормального распределения. Доверительный интервал для биномиального распределения. Построение доверительного интервала на основе асимптотически нормальных статистических оценок. Доверительный эллипсоид. Требования к статистическим оценкам. Метод моментов. Обобщенный метод моментов. Метод наименьших квадратов и его геометрическая интерпретация. Метод хи-квадрат. Байесовский подход к теории оценивания. Априорная и апостериорная плотности распределения. Апостериорный байесовский риск. Интерпретации байесовского подхода. Роль байесовского подхода в современных методах обработки данных. Примеры байесовских оценок. Оценки максимума апостериорной вероятности. Функция правдоподобия. Оценки максимума правдоподобия (ОМП).

Корреляционный и регрессионный анализ. Выборочная ковариация и корреляция. Уметь считать. Функция регрессии. Многомерная линейная регрессия. Метод наименьших квадратов. Матричная запись метода наименьших квадратов. Геометрическая интерпретация метода наименьших квадратов. Многомерные гауссовские наблюдения и модель линейной регрессии. Простая линейная регрессия. Свойства оценок МНК простой линейной регрессии. Коэффициент детерминации, его содержательный смысли связь с выборочным коэффициентом корреляции. Основное тождество "SST= SSE+SSR", его содержательный смысл. Нелинейная регрессия. МНК. Его геометрическая интерпретация. Асимптотическая нормальность оценок МНК (зачем она нужна). Оценки МНК для регрессии как ОМП. Логистическая регрессия. Оценка максимума правдоподобия. Jackknife. Оценка дисперсии. Бутстрап . Построение несмещенных оценок. Бутстрап параметрический и непараметрический. ЕМ-алгоритм суть дела. Сходимость ЕМ-алгоритма. ЕМ-алгоритм как метод покоординатного спуска. FDR.

**Тема 4. Машинное обучение**

Сравнение рисков байесовского классификатора и произвольного

Plugin метод

Связь задач классификации и регрессии

Задача минимизации эмпирического риска Избыточный риск, обобщающая способность Анализ погрешности ERM на основе равномерных теорем Гливенко-Кантелли

SVM, AdaBoost и нейронные сети как задачи ERM

**Раздел 3. Обеспечение учебных занятий**

**3.1. Методическое обеспечение**

**3.1.1 Методические указания по освоению дисциплины**

Успешное освоение дисциплины возможно благодаря посещению лекций, участию в обсуждении вопросов, подготовленных к занятию, самостоятельной работе, включающей в себя чтение специальной литературы по разделам темы.

**3.1.2 Методическое обеспечение самостоятельной работы**

Самостоятельная работа учащихся в рамках данной дисциплины является важным компонентом обучения, предусмотренным компетентностно-ориентированным учебным планом и рабочей программой учебной дисциплины.

Настоящей программой предусмотрены формы самостоятельной работы, связанные решением задач, выполнением индивидуального задания, подготовкой к контрольным работам, использованием методических материалов и чтением дополнительной литературы, указанной в программе. В лекциях указывается дополнительный материал для освоения каждой темы и приводится соответствующий список литературы.

Методическое обеспечение самостоятельной работы обеспечивается наличием в библиотеке необходимых учебников и монографий, наличием электронных копий другой методической литературы, а также электронных копий комментариев к наиболее сложным частям материала, написанным преподавателем.

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

Аппарат контроля за усвоением материала включает в себя решение задач в течение семестра, индивидуальное задание и одну контрольные работы. Текущий контроль успеваемости складывается из результатов проверки качества решения задач студентами по изучаемым темам и определения уровня понимания материала путем проведения дискуссий и обсуждений.

**Критерии оценивания по ECTS:**

Оценка A:

Студент выполнил 10 задач, выдаваемых в течение семестра, успешно написал проверочную работу\* в конце семестра и правильно ответил на 5 теоретических вопросов на финальном зачете.

Оценка B:

Студент выполнил 9 задач, выдаваемых в течение семестра, успешно написал проверочную работу\* в конце семестра и правильно ответил на 4 теоретических вопроса из 5.

Оценка C:

Студент выполнил 8 задач, выдаваемых в течение семестра, успешно написал проверочную работу\* в конце семестра и правильно ответил на 4 теоретических вопроса из 5.

Оценка D:

Студент выполнил 8 задач, выдаваемых в течение семестра, успешно написал проверочную работу\* в конце семестра и правильно ответил на 3 теоретических вопроса из 5.

Оценка E:

Студент выполнил не менее 7 задач, выдаваемых в течение семестра, удовлетворительно написал проверочную работу\* в конце семестра и правильно ответил на 3 теоретических вопроса из 5.

Оценка F:

Студент выполнил менее 7 задач, выдаваемых в течение семестра, и(или) плохо (не) написал проверочную работу\* в конце семестра и (или) неправильно ответил на 3 теоретических вопроса из 5.

\*Для зачета необходимо получить за проверочную работу оценку выше F. Оценка в целом, не может быть выше, чем оценка за проверочную работу.

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

На практических занятиях предполагается добиться максимального усвоения учащимися методов решения статистических задач на уровне западных бизнес школ. Контрольные работы и тестовая задача на экзамене проверяет этот уровень усвоения.

Примерный список вопросов к экзамену состоит из вопросов, перечисленных в разделе 2.2.

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

Оценка обучающимися содержания и качества учебного процесса по дисциплине осуществляется в установленном в СПбГУ порядке.

**3.2. Кадровое обеспечение**

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

К преподаванию дисциплины могут быть допущены преподаватели, имеющие диплом о высшем образовании по соответствующему направлению.

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

Не предусмотрено.

**3.3. Материально-техническое обеспечение**

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

Стандартно оборудованные лекционные аудитории для проведения интерактивных лекций: видеопроектор, экран, др. оборудование.

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

На некоторых практических занятиях в аудитории может потребоваться проекционное оборудование.

**3.3.3 Характеристики специализированного оборудования**

Для подготовки к некоторым занятиям может потребоваться принтер, чтобы распечатать раздаточные материалы.

**3.3.4 Характеристики специализированного программного обеспечения**

Специальных требований нет.

**3.3.5 Перечень и объёмы требуемых расходных материалов**

В соответствии с разделом 3.3.3 может потребоваться белая бумага формата А4 для печати на принтере.

**3.4. Информационное обеспечение**

**3.4.1 Список обязательной литературы**

1. Ермаков С.М. Метод Монте-Карло в вычислительной математике. Невский Диалект.2009.
2. Ивченко Г.И., Медведев Ю.И. Введение в математическую статистику. ЛКИ 2010.

**3.4.2 Список дополнительной литературы**

1. Ермаков С.М. Метод Монте-Карло и смежные вопросы, М., Наука, 1975.
2. Fishman G.S. Monte-Carlo Concepts, Algorithms, and Applications, Springer-Verlag, New York-Berlin-Heidelberg. 1999.
3. Wasserman L. All of Statistics Springer 2004.
4. Bickel P., Doksum K. Mathematical Statistics. Prentice Hall, 2001.
5. Дж. Дуб. Вероятностные процессы. М., ИЛ, 1956.

**3.4.3 Перечень иных информационных источников**

http://statmod.ru/wiki/, ресурсы сети Интернет.

**Раздел 4. Разработчики программы**

Ермаков Михаил Сергеевич, доктор физико-математических наук, профессор, Кафедра статистического моделирования, m.ermakov@spbu.ru