**Программа учебной дисциплины «Теория информации»**

Утверждена

Академическим советом ООП

Протокол № от «\_\_»\_\_\_\_\_20\_\_ г.

|  |  |
| --- | --- |
| Автор  | **Н.К. Верещагин** nikolay.vereshchagin@gmail.com |
| Число кредитов  | 5 |
| Контактная работа (час.)  | 60 |
| Самостоятельная работа (час.)  | 130 |
| Курс  | 4 |
| Формат изучения дисциплины |  Очная, без использования онлайн курса |

1. **ЦЕЛЬ, РЕЗУЛЬТАТЫ ОСВОЕНИЯ ДИСЦИПЛИНЫ И ПРЕРЕКВИЗИТЫ**

Основная цель освоения дисциплины «Теория информации» - обучить студентов основным понятиям и методам теории информации, необходимым как в дальнейшем обучении, так и в работе по специальности.

В результате освоения дисциплины студент должен:

**знать:**

- основные понятия и методы теории информации, необходимые для дальнейшего изучения последующих дисциплин, предусмотренных базовым и рабочим учебными планами, а также для применения в профессиональной деятельности;

**уметь:**

- пользоваться основными методами теории информации для решения задач как в области теории информации, так и за ее пределами;

**владеть:**

- навыками формализации и решения практических задач методами теории информации.

Изучение дисциплины «Теория информации» базируется на следующих дисциплинах:

- математический анализ 1;

- дискретная математика 1.

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

* знать основные понятия дискретной математики, и математического анализа.

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

* Коды с исправлением ошибок;
* Машинное обучение и майнинг данных;
* Введение в теорию статистического обучения;
* Байесовские методы машинного обучения.

.

1. **СОДЕРЖАНИЕ УЧЕБНОЙ ДИСЦИПЛИНЫ**
* **Тема 1. Информация по Хартли.**

Информация по Хартли в сообщении неизвестного исхода (двоичный логарифм количества возможных исходов).

Информация в данном сообщении. Аддитивность информации при двух последовательных сообщениях.

Применение информации по Хартли для получения верхних и нижних оценок в задачах сортировки (нижняя оценка для n монет, верхняя оценка для 5 монет) и поиска фальшивой монетки на чашечных весах (нижняя и верхняя оценка для n монет, верхняя оценка для 12 монет).

* **Тема 2. Коммуникационная сложность.**

Коммуникационные протоколы. Разбиение матрицы функции на прямоугольники.

Метод трудных множеств , метод размера прямоугольника и метод ранга матрицы. Оценки этими методами коммуникационной сложности предикатов EQ, GT, DISJ, IP.

* **Тема 3. Классическая теория информации Шеннона.**

Определение энтропии Шеннона. Задача о префиксном кодировании.

Неравенство Крафта. Нижняя и верхняя оценка средней длины префиксного кода с помощь энтропии. Теорема Макмиллана. Сбалансированные коды.

Код Шеннона-Фано и арифметический код.

Совместно распределенные случайные величины. Условная энтропия. Теорема об энтропии пары (она не превосходит суммы энтропий). Независимость и энтропия.

Оценка среднего количества переданных бит в коммуникационных протоколах с помощью энтропии Шеннона.

Условная энтропия и ее свойства (она неотрицательна и не превосходит безусловной энтропии, она равна разности двух безусловных).

Понятие количества информации и его свойства.

Информационные неравенства: метод релятивизации, метод диаграмм.

Общая информация тройки слов и пример, когда она отрицательна.

* **Тема 4. Применение теории информации Шеннона.**

Базисные неравенства, неравенство треугольника. Цепное правило. Марковская цепь и ее свойство.

Теорема Шеннона об идеальном шифре. Неравенство Фано. Неравенство Фано для классификаторов. Неравенство Ромащенко-Каседа (для количества квадратов).

Количество слов с данными частотами. Сбалансированные слова и их количество. Кодирование, основанное на частотах диграмм. Стационарные источники.

Теорема Шеннона о бесшумном канале.

Теорема Вольфа-Слепяна.

Каналы с шумом и их пропускная способность. Теорема Шеннона о канале с шумом.

* **Тема 5. Игры и предсказания.**

Игры по предсказанию битов данной последовательности. Мартингалы. Теорема об определении мартингалов стратегиями.

Предсказания с экспертами. Случай логарифмического штрафа (ставки в казино) и случай единичного штрафа за ошибку).

Выпуклые функции штрафа. Условие Блеквела. Линейные, полиномиальный и экспоненциальный предсказатели.

* **Тема 6. Начала теории обучения.**

Pac learning и размерность Вапника – Червоненкиса. Лемма Зауэра – Шелаха.

Оценка качества эмпирически наилучшего предсказателя через количество концепций.

Выпуклые функции штрафа. Условие Блеквела. Линейные, полиномиальный и экспоненциальный предсказатели.

* **Тема 7. Алгоритмическая теория информации.**

Декомпрессоры. Колмогоровская сложность и теорема Колмогорова – Соломонова.

Оценка на количество слов колмогороской сложности не больше n. Неубывание колмогоровской сложности при алгоритмических преобразованиях.

Невычислимость колмогоровской сложности и теорема Геделя о неполноте в форме Чейтина.

Сложность других конструктивных объектов. Неравенство для сложности пары.

Условная сложность. Теорема Колмогорова – Левина о сложности пары. Количество информации. Его свойства. Релятивизация.

Сложность и энтропия Шеннона. Теорема Ромащенко о совпадении классов неравенств.

1. **ОЦЕНИВАНИЕ**

Оцениваются 6 домашних заданий и экзамен.

Оценка за каждое домашнее задание равна доле решенных задач, умноженной на 10. Общая оценка за домашние задания равна среднему арифметическому оценок за решение каждого из заданий. На решение каждого ДЗ дается 14 дней, решение ДЗ нужно сдавать семинаристу до начала семинара. Сдача домашних заданий после их срока невозможна.

Экзамен (письменный) оценивается по десятибалльной системе. На экзамене можно пользоваться любыми бумажными источниками и нельзя никакими электронными.

Оценка за домашние задания составляет накопленную оценку. Накопленная оценка и оценка за экзамен с коэффициентами 3/5 и 2/5 дают итоговую оценку. Таким образом, вес экзамена в итоговой оценке – 40%, а вес каждого из домашних заданий - 10%.

В вычислениях текущие оценки и промежуточные величины не округляются. Результат вычисляется точно и округляется только в момент выставления промежуточной и итоговой оценок. При выставлении итоговой и промежуточных оценок используется следующее правило округления: между 1 и 5 округление вниз, между 5 и 6 округление арифметическое, а в остальных случаях округление вверх. Т.е. 3,92 округляется до 3, 5,48 – до 5, 5,54 – до 6, 7,12 – до 8.

 Те, кто не смог прийти на экзамен по болезни, могут его сдать отдельно. Не набравшие в конце второго модуля нужное количество баллов (4) могут пересдать экзамен, а если и это не поможет, то сдавать экзамен комиссии. В последнем случае накопленная оценка аннулируется и оценка, полученная на экзамене, и является окончательной.

Для прохождения контроля студент должен продемонстрировать понимание основных определений, знание теорем и методов, умение применять изученные методы для решения задач.

За домашнее задание или экзамен допускается оценка 0 в тех случаях, когда студент решил ни одной задачи в работе или ничего не рассказал на коллоквиуме. Оценка 0 также ставится в случаях доказанного списывания на экзамене.

Комиссия проходит в форме устного экзамена, накопленная оценка не учитывается.

1. **ПРИМЕРЫ ОЦЕНОЧНЫХ СРЕДСТВ**

# Оценочные средства для текущего контроля студента

# Шесть домашних заданий. Примеры задач:

1. Дано шесть монет разного веса и известно, что первая монета легче второй, третья легче четвертой и пятой. Имеются весы, позволяющие сравнить по весу любые две монеты. Доказать, что для того, чтобы упорядочить по весу все монеты, необходимо 7 взвешиваний.
2. Доказать, что более того, необходимо 8 взвешиваний.
3. Имеются 60 монет, среди которых ровно одна фальшивая (неизвестно какая). Все настоящие монеты одного веса, а фальшивая легче или тяжелее. На чашечных весах можно сравнивать по весу любые две группы монет. Нужно найти фальшивую монету и выяснить, легче она или тяжелее. Докажите, что необходимо сделать 5 взвешиваний.
4. Доказать, что в предыдущей задаче 5 взвешиваний достаточно.
5. Алиса задумала целое число от 1 до $2^n$. Сколько денег нужно иметь, чтобы отгадать задуманное число, задавая вопросы с ответами ДА,НЕТ, если за ответ ДА приходится платить 2 рубля, а за ответ НЕТ --- 5 рублей? (Ответ нужен с точностью $o(n)$.)
6. У Алисы имеется подмножество $x$ множества $\{1,\dots,n\}$, а у Боба подмножество $y$ множества $\{1,\dots,n\}$. Сколько бит необходимо передать (от Алисы к Бобу и обратно), чтобы найти мощность объединения $x$ и $y$? (Ответ нужен с точностью $o(n)$.)
7. Построить детерминированный коммуникационный протокол, который вычисляет функцию $GT\_n$, передавая в среднем константу битов. Функция $GT\_n(x,y)$ определена на парах $x,y$ целых чисел в интервале $\{1,\dots,2^n\}$ и принимает значение 1, если $x>y$, и значение 0, иначе. Говоря о среднем, мы имеем в виду, что $x,y$ выбираются случайно и независимо среди всех чисел указанного интервала с равномерным распределением.
8. У Алисы имеется целое число $x$ в интервале $\{0,\dots,2^n-1\}$, а у Боба целое число $y$ в том же интервале. Докажите, что для того, чтобы найти разность этих чисел, необходимо всего передать (от Алисы к Бобу и обратно) $2n$ бит.
9. Пусть в слове длины 1000 количества букв a,b,c,d,e,f,g равны, соответственно, 50, 100, 150, 150, 150, 200, 200. Найдите энтропию Шеннона для этих частот. Постройте код Хаффмана для алфавита a,b,c,d,e,f,g, и этих частот. Найдите длину кода этого слова, полученного применением кодирования Хаффмана к каждой букве. Найдите длину арифметического кода этого слова с точностью до плюс минус одного бита (арифметический код вычисляется от всего слова целиком, а не побуквенно).
10. Пусть случайная величина $\alpha$ имеет распределение $p(0)=1/9$, $p(1)=1/6$, $p(2)=1/6$, $p(3)=5/9$. Положим $\beta=\alpha\pmod 2$. Найти $H(\alpha)$, $H(\beta)$, $H(\alpha,\beta)$.
11. Веса 2 монеток выбираются случайно и независимо среди чисел $1,\dots,4$. Какова энтропия Шеннона случайной величины, равной результату сравнения на чашечных весах весов первой и второй монетки? Какова энтропия случайной величины, равной разности весов первой и второй монетки?
12. Пусть случайная величина $\alpha$ имеет распределение $1/3,2/3$, а случайная величина $\beta$ имеет распределение $1/2,1/2$. В каких пределах может изменяться $H(\alpha,\beta)$?
13. Построить распределение вероятностей на парах чисел $x,y$, обладающее следующим свойством. Любой коммуникационный протокол, вычисляющий функцию GT$\_n$, в среднем передает не менее $n$ битов.
14. 14. Рассматриваются трехзначные числа с неубывающими цифрами. Спрашивается: (2 балла) Какая из цифр (первая,вторая,третья) несет больше информации о таком числе? (2 балла) Сколько информации в первой цифре такого числа о его второй цифре? о третьей цифре? (2 балла) Зависит ли первая цифра числа от двух последних?
15. (4 балла) На входе классификатора объекты трех классов, количество объектов в классах относятся как 2:2:1. Точность классификатора равна 0.8, покрытие 0.7. Какова минимальная информативность классификатора? Задачу можно решать численно.
16. Найти пропускную способность следующего двоичного канала с шумом. Если входная буква равна 0, то на выходе канала появляются 0 и 1 с вероятностями 1/3 и 2/3, соответственно.
17. Если входная буква равна 1, то на выходе канала появляются 0 и 1 с вероятностями 1/6 и 5/6, соответственно.
18. Найти наименьшую вероятность ошибки при передаче по этому каналу одного бита при кодировании его двумя буквами. (Вероятность ошибки при данной схеме кодирования-декодирования $E,D$ определяется, как максимум по входам $x\in\{0,1\}$ вероятности события $D(W(E(x)))\ne x$.)
19. На множестве слов длины 4 в двухбуквенном алфавите 0,1 задана бернуллиева мера с вероятностями 0 и 1 равными 1/3 и 2/3, соответственно. Какова минимально возможная вероятность ошибки при кодировании элементов этого множества трехбитовыми последовательностями.
20. 19. Найдите номер слова abcbcccbab в алфавитном порядке среди всех слов длины 10 в трёхбуквенном алфавите с тем же буквенным составом (нумерация слов начинается с 1).
21. Известно, что в слове длины 1004 в трехбуквенном алфавите $\{a,b,c\}$, начинающемся с буквы $a$, количества вхождений двухбуквенных сочетаний $aa,bb,cc,ab,bc,ca,ac$ равны, соответственно, 100, 50, 50, 201, 201, 301, 100. Оцените количество таких слов с точностью до множителя $100$.
22. Игрок в казино может поставить на любой из двух исходов (0,1) бросания монетки любую ставку в пределах его текущего капитала. Сделанная ставка удваивается в случае выигрыша и теряется иначе. Пусть игроку стало известно, что в первых шести бросаниях будет ровно 3 единицы. Придумайте стратегию игрока, гарантирующую возрастание
23. начального капитала в 3.2 раза. Какую ставку эта стратегия делает после того, как в первом бросании выпал 0? Какую ставку эта стратегия делает после того, как в первых двух бросаниях выпали нули? Докажите, что число 3.2 в этой задаче нельзя увеличить.
24. 22. Пусть в предыдущей задаче нужно только предсказывать следующий бит последовательности, а денежная ставка при этом не делается. Игрок получает по одному рублю за каждое верное предсказание (а за неверное предсказание не штрафуется). Какую максимальную сумму можно гарантированно выиграть в такой игре?
25. 23. Нам надо предсказывать результаты бросания кубика с шестью гранями,
26. всего $n$ бросаний. Каждое предсказание является \emph{действительным} числом от 1 до 6. На каждом бросании мы проигрываем $(x-y)^4$ рублей,
27. где $x$ --- выпавшая цифра, а $y$ --- сделанное предсказание. У нас имеется 6 советчиков, каждый из которых
28. перед каждым бросанием сообщает свое предсказание. Построить алгоритм, который проигрывает не больше, чем на $O(\sqrt n)$, чем проиграл бы каждый советчик, если бы играл сам (как бы ни выпадал кубик).
29. 24. Найти размерность Вапника-Червоненкиса семейства линейных булевых функций от $n\ge2$ переменных. (Пояснение, булева функция --- это отображение из $\{0,1\}^n$ в $\{0,1\}$. Такая функция называется линейной, если она имеет вид
30. $a\_1x\_1\oplus\dots\oplus a\_nx\_n\oplus b$, где $a\_1,\dots a\_n, b\in\{0,1\}$.)
31. 25. В модели PAC learning придумать полиномиальный алгоритм изучения неизвестной булевой функции $f$ от $n$ переменных, являющейся конъюнкцией некоторых из этих $n$ переменных. (Алгоритм получает на вход пары вида $(x,f(x))$, в которых $x$ выбирается
32. случайно с некоторым неизвестным распределением на $\{0,1\}^n$. После получения некоторого полиномиального от $n$ количества таких пар, алгоритм должен с вероятностью 99\% (по крайней мере)
33. выдать некоторую конъюнкцию $g$, значение которой совпадает со значением $f$ на 99\% (по крайней мере) входов $x$ по этому распределению.)

# Тематика экзамена (второй модуль):

1. Метод трудных множеств оценки коммуникационной сложности.
2. Метод размера прямоугольников оценки коммуникационной сложности.
3. Метод ранга матрицы оценки коммуникационной сложности.
4. Неравенство Крафта. Связь энтропии Шеннона и задачи о префиксном кодировании.
5. Теорема Макмиллана.
6. Теорема об энтропии пары (она не превосходит суммы энтропий).
7. Условная энтропия и её свойства (для энтропии Шеннона).
8. Понятие количества информации и его свойства (для энтропии Шеннона).
9. Неравенство Фано. Неравенство Фано для классификаторов.
10. Количество слов с данными частотами. Сбалансированные слова и их количество.
11. Теорема Шеннона о бесшумном канале.
12. Теорема Вольфа-Слепяна.
13. Теорема Шеннона о канале с шумом.
14. Теорема о связи мартингалов и стратегий игры в казино.
15. Предсказания с экспертами: линейный предсказатель.
16. Предсказания с экспертами: экспоненциальный предсказатель.
17. Лемма Зауэра - Шелаха.
18. Колмогоровская сложность: определение и теорема Колмогорова-Соломонова. Оценка на число слов колмогоровской сложности не больше n.
19. Условная сложность. Теорема Колмогорова - Левина о сложности пары.
20. **РЕСУРСЫ**
	1. **Основная литература**
	2. [Н.К. Верещагин, Е.В. Щепин. Информация, кодирование и предсказание.](https://wiki.school.yandex.ru/shad/groups/2017/semester1/informationtheory/.files/Sch_Ver.pdf) Москва, МНЦМО 2012.

 **V.2 Дополнительная литература.**

1. [А.M. Яглом, И.М. Яглом. Вероятность и информация.](http://www.mccme.ru/free-books/shen/shen-probability.pdf)
2. Кернер, Чисар. Теория информации.
3. А. Шень. Игры и стратегии с точки зрения математики (c1) 2-е изд., М.: МЦНМО, 2008, 40 с., ISBN 978-5- 94057-432-3 <http://www.mccme.ru/free-books/shen/shen-games.pdf>
4. В.А. Успенский, Н.К. Верещагин, А. Шень.
5. Колмогоровская сложность.
6. <http://www.mccme.ru/free-books/shen/kolmbook.pdf>
7. Nicolo Cesa-Bianchi, Gabor Lugosi. Prediction, learning, and games. Cambridge University Press, 2006
	1. **Программное обеспечение**

|  |  |  |
| --- | --- | --- |
| **№ п/п** |  |  |
|   1. |  |  |
|  |  |  |

* 1. **Профессиональные базы данных, информационные справочные системы,
	интернет-ресурсы (электронные образовательные ресурсы)**

|  |  |  |
| --- | --- | --- |
|  |  |  |
|  |  |
|  |  |  |
|  |  |  |
|  |  |
|  |  |  |

* 1. **Материально-техническое обеспечение дисциплины**

Учебные аудитории для лекционных занятий по дисциплине обеспечивают использование и демонстрацию тематических иллюстраций, соответствующих программе дисциплины в составе:

ПЭВМ с доступом в Интернет (операционная система, офисные программы, антивирусные программы);

мультимедийный проектор с дистанционным управлением.