

**Федеральное государственное автономное образовательное учреждение  
высшего образования  
"Национальный исследовательский университет  
"Высшая школа экономики"**

Факультет Компьютерных наук  
Департамент больших данных и информационного поиска  
Базовая кафедра Яндекс

**Рабочая программа дисциплины «Восстановление зависимостей  
из эмпирических данных»**

для направления 01.04.02 "Прикладная математика и информатика" подготовки  
магистра

для магистерской программы "Науки о данных"

Разработчик(и) программы  
Бабенко М.А., к.ф.-м.н. (maxim.babenko@gmail.com)

Одобрена на заседании базовой кафедры Яндекс  
«\_\_»\_\_\_\_\_ 2017 г.

Заведующий Кафедрой  
М.А. Бабенко \_\_\_\_\_

Утверждена Академическим советом образовательной программы  
«\_\_»\_\_\_\_\_ 2017 г., № протокола \_\_\_\_\_

Академический руководитель образовательной программы  
С.О. Кузнецов \_\_\_\_\_

Москва, 2018

*Настоящая программа не может быть использована другими подразделениями университета и другими вузами без разрешения подразделения разработчика программы.*



## **I. Область применения и нормативные ссылки**

Настоящая программа учебной дисциплины устанавливает требования к образовательным результатам и результатам обучения студента и определяет содержание и виды учебных занятий и отчетности.

Программа предназначена для преподавателей, ведущих дисциплину «Восстановление зависимостей с использованием эмпирических данных», учебных ассистентов и студентов направления подготовки 01.04.02 «Прикладная математика и информатика», обучающихся по образовательной программе «Науки о данных».

Программа учебной дисциплины разработана в соответствии с:

- Образовательным стандартом федерального государственного автономного образовательного учреждения высшего профессионального образования «Национального исследовательского университета «Высшая школа экономики»;
- Образовательной программой подготовки магистра по направлению 01.04.02 «Прикладная математика и информатика», специализации «Анализ Интернет - данных».
- Объединенным учебным планом университета по образовательной программе «Науки о данных», утвержденным в 2018 г.

## **2. Место дисциплины в структуре образовательной программы**

Курс посвящён проблеме восстановления зависимостей по эмпирическим данным. Зависимость понимается в широком смысле: как функция из заданного класса, как модель из класса моделей или в терминах имитации одного автомата другим.

Приводятся примеры конкретных теоретических постановок задачи восстановления зависимостей: регрессия, идентификация моделей, распознавание образов и их приложения. Рассматриваются понятия истинного риска и эмпирического риска как его возможный эквивалент. Обсуждаются основания, по которым замена истинного риска эмпирическим представляется разумной. Приводятся примеры, когда это не так.

Значительная часть курса посвящена методам минимизации эмпирического риска: методу наименьших квадратов, методу максимального правдоподобия для выбора модели, построению решающего правила, минимизирующего число ошибок на данных обучения в задачах распознавания образов.

Описываются методы построения линейных решающих правил, дискриминантной функции Фишера, перцептрон, методы построения потенциальных функций и нейронных сетей. Обсуждаются вычислительные возможности и существующие проблемы. В курсе будет также рассказано о стандартной процедуре машины опорных векторов (SVM).

В рамках курса будет дана критика подхода, основанного на минимизации среднего риска и предложено усовершенствование этого метода, основанного на свойстве равномерной сходимости эмпирического риска к истинному. Описываются критерии равномерной сходимости эмпирического риска к истинному, вводится понятие VC-размерности (размерность Вапника-Червоненкиса), описываются энтропийные критерии.

В конце курса рассматривается тема выбора оптимальной сложности модели.

Программа курса предусматривает лекции (26 часов) и практические занятия (38 часов).



### 3. Цели освоения дисциплины

Целью данного курса является ознакомление слушателей с основными методами восстановления зависимостей по эмпирическим данным. Приводятся примеры конкретных теоретических постановок задачи восстановления зависимостей, в частности, методы минимизации эмпирического риска. Рассматриваются понятия истинного риска и эмпирического риска как его возможный эквивалент. Все темы курса снабжены как теоретическими заданиями, позволяющими глубже понять суть рассматриваемых понятий и методов, так и практическими заданиями, призванными дать возможность сопоставить теорию с практикой.

### 4. Тематический план дисциплины «Восстановление зависимостей с использованием эмпирических данных»

№	Название темы	Всего часов по дисциплине	Аудиторные часы		Самостоятельная работа
			Лекции	Сем. и практика	
1	Общая постановка задачи восстановления зависимостей	5	1	1	3
2	Метод максимального правдоподобия	6	1	2	3
3	Примеры конкретных задач восстановления зависимостей: регрессия, идентификация моделей, распознавание образов и их приложения	5	1	1	3
4	Построение непараметрических оценок распределений методом максимального правдоподобия	8	1	2	5
5	Метод наименьших квадратов для оценки регрессии. Метод максимального правдоподобия для выбора модели	7	1	1	5
6	Критерий отношения правдоподобия	7	1	2	4
7	Поиск решающего правила, минимизирующего число ошибок или среднее значение функции штрафа на данных обучения, в задачах распознавания образов	4	1	1	3
8	Многомерное линейное оценивание	6	1	2	3
9	Перцептрон. Потенциальные функции. Нейронные сети	6	1	1	4
10	Учёт априорной информации при линейном оценивании	6	1	2	3
11	Метод обобщённого портрета в задаче классификации	5	1	1	3
12	Байесовское оценивание	6	1	2	3



13	Машина опорных векторов (SVM)	8	2	2	4
14	Некоторые методы классификации	8	2	2	4
15	Критика метода минимизации эмпирического риска	5	1	1	3
16	Оптимальная гиперплоскость	6	1	2	3
17	Критерии равномерной сходимости частот к вероятностям. Функция роста. VC-размерность	5	1	1	3
18	Двойственная задача построения оптимальной гиперплоскости	6	1	2	3
19	Критерии равномерной сходимости частот к вероятностям. Связь с задачами обучения распознаванию образов	5	1	1	3
20	Построение непараметрической сплайн-регрессии	6	1	2	3
21	Критерии равномерной сходимости средних к математическим ожиданиям	5	1	1	3
22	Построение непараметрической ядерной регрессии	6	1	2	3
23	Проблема выбора оптимальной сложности модели	5	1	2	3
24	Различные виды регрессионных зависимостей	6	1	2	3
	Итого	144	26	38	80

### 5. Формы контроля и структура итоговой оценки

Текущий контроль - домашняя работа в первом модуле, контрольная работа в первом модуле.

Итоговый контроль – письменный экзамен (120 мин.)

Итоговая оценка вычисляется следующим образом:

$0,1 \cdot \text{оценка за домашнюю} + 0,2 \cdot \text{оценка за контрольную} + 0,7 \cdot \text{оценка за экзамен.}$

#### Таблица соответствия оценок по десятибалльной и системе зачет/незачет

Оценка по 10-балльной шкале	Оценка по 5-балльной шкале
1	Незачет
2	
3	
4	Зачет
5	
6	
7	
8	
9	

10	
----	--

**Таблица соответствия оценок по десятибалльной и пятибалльной системе**

<b>По десятибалльной шкале</b>	<b>По пятибалльной системе</b>
1 – неудовлетворительно 2 – очень плохо 3 – плохо	неудовлетворительно – 2
4 – удовлетворительно 5 – весьма удовлетворительно	удовлетворительно – 3
6 – хорошо 7 – очень хорошо	хорошо – 4
8 – почти отлично 9 – отлично 10 - блестяще	отлично – 5

## **6. Программа дисциплины**

### **Раздел I. Постановка задачи**

**Тема 1. Общая постановка задачи восстановления зависимостей** Интерпретация восстановления зависимости в терминах выбора функции из заданного класса. Интерпретация восстановления зависимости в терминах выбора модели из заданного класса моделей. Интерпретация восстановления зависимости в терминах имитации одного автомата другим.

### **Тема 2. Метод максимального правдоподобия.**

Обсуждается понятие функционала правдоподобия в случае зависимых и независимых наблюдений. Функционал правдоподобия выводится как эмпирическая оценка информационного расстояния Кульбака-Лейблера между истинным и исследуемым распределениями. Приводятся примеры построения оценок максимального правдоподобия для параметров равномерного распределения, экспоненциального распределения, распределения Пуассона, нормального распределения.

### **Тема 3. Примеры конкретных задач восстановления зависимостей: регрессия, идентификация моделей, распознавание образов и их приложения**

Вид риска при решении задач классификации, восстановления регрессии, оценки плотности. Истинный риск и эмпирический риск как его возможный эквивалент. Основания, по которым такая замена представляется разумной. Закон больших чисел. Основания, по которым такая замена является ошибочной. Проблема множественного сравнения.

### **Тема 4. Построение непараметрических оценок распределений методом максимального правдоподобия.**

Обсуждается вид функционала правдоподобия в случае отсутствия информации о параметрическом виде распределения, выводится формула множительной оценки Каплана-Мейера. Рассматривается построение оценок максимального правдоподобия при наличии цензурирования. Приводится пример смеси распределений, в котором метод максимального правдоподобия не позволяет получить оценку параметров.

## **Раздел II. Методы минимизации эмпирического риска**

### **Тема 5. Метод наименьших квадратов для оценки регрессии. Метод максимального правдоподобия для выбора модели**

Правдоподобие независимой выборки. Правдоподобие для экспоненциального семейства распределений. Оценки максимального правдоподобия. Связи методов максимального правдоподобия и наименьших квадратов. Стандартные процедуры построения регрессии и максимизации правдоподобия.

### **Тема 6. Критерий отношения правдоподобия.**

Теорема Андерсона об отношении правдоподобия, лемма Пирсона. Построение правила классификации по отношению правдоподобия.

Иллюстрация метода классификации на примере данных «Ирисы Фишера». ROC кривые.

### **Тема 7. Поиск решающего правила, минимизирующего число ошибок или среднее значение функции штрафа на данных обучения, в задачах распознавания образов**

Линейные решающие правила. Сравнение с дискриминантной функцией Фишера. Процедуры нахождения дискриминантной функции и линейного программирования.

### **Тема 8. Многомерное линейное оценивание**

Построение оценок максимального правдоподобия для параметров многомерной линейной регрессии при различных гипотезах о ковариационной матрице нормально распределённой помехи. Вычисление ковариационной матрицы оценок максимального правдоподобия.

### **Тема 9. Персептрон. Потенциальные функции. Нейронные сети**

Обсуждаются классические методы классификации, их вычислительные возможности и проблемы реализации.

### **Тема 10. Учёт априорной информации при линейном оценивании**

Вид правдоподобия при наличии априорной информации. Трансформация оценок максимального правдоподобия при учёте априорной информации. Дисперсии оценок максимального правдоподобия с априорной информацией.

### **Тема 11. Метод обобщённого портрета в задаче классификации** Линейный классификатор. Оптимальная разделяющая гиперплоскость.

Обобщенный портрет (ОП). Двойственная задача квадратичной оптимизации при линейных

ограничениях.

### **Тема 12. Байесовское оценивание**

Понятие апостериорного распределения. Апостериорное среднее как решение оптимизационной задачи минимизации апостериорного среднего риска с квадратичной функцией потерь.

### **Тема 13. Машина опорных векторов (SVM)**

Линейный классификатор в спрямляющем пространстве. Машина опорных векторов (SVM) – ядра вместо скалярных произведений (совмещение потенциалов с ОП). Виды ядер, параметры.

### **Тема 14. Некоторые методы классификации**

Метод наивного Байеса, метод потенциальных функций. Примеры решения задач.

### **Тема 15. Критика метода минимизации эмпирического риска**

Примеры, когда минимизация эмпирического риска не работает. Проблема равномерной сходимости эмпирического риска к истинному (или частот вероятностям, или средних к математическим ожиданиям).

### **Тема 16. Оптимальная гиперплоскость**

Определение оптимальной гиперплоскости. Графическая интерпретация оптимальной гиперплоскости. Построение оптимальной гиперплоскости из геометрических соображений.

### **Тема 17. Критерии равномерной сходимости частот к вероятностям.**

#### **Функция роста. VC-размерность**

Алгоритмы с полной памятью. Функция роста. Комбинаторная размерность.

### **Тема 18. Двойственная задача построения оптимальной гиперплоскости**

Понятие двойственной задачи квадратичного программирования с линейными ограничениями. Понятие опорных векторов. Примеры составления двойственной задачи для построения оптимальной гиперплоскости.

### **Тема 19. Критерии равномерной сходимости частот к вероятностям.**

#### **Связь с задачами обучения распознаванию образов**

Условия равномерной сходимости частот к вероятностям по классу событий. Основная лемма. Достаточное условие равномерной сходимости частот к вероятностям. Энтропийный критерий.

### **Тема 20. Построение непараметрической сплайн-регрессии**

Определение сплайн-функции как решения оптимизационной задачи с ограничениями на разрывность производных. Связь степени полиномиального сплайна с порядком разрывной производной. Физический смысл сплайн-функции. Различные виды краевых условий. Сглаживающие сплайны.

### **Тема 21. Критерии равномерной сходимости средних к математическим ожиданиям**

Оценка вероятности равномерного по классу абсолютного отклонения для равномерно ограниченных функций. Примеры применения.

### **Тема 22. Построение непараметрической ядерной регрессии**

Вывод ядерной регрессии через оценку плотности Парзена-Розенблатта. Определение и виды ядер. Асимптотические свойства ядерной регрессии. Проблема выбора эффективной ширины ядра. Правило подстановки, cross-validation (скользящий контроль).

### **Тема 23. Проблема выбора оптимальной сложности модели**

Связь сложности модели с точностью её идентификации по эмпирическим данным. Статистическая мера сложности. Оценка среднего риска через эмпирический риск с учётом сложности модели.

### **Тема 24. Различные виды регрессионных зависимостей**

Обсуждается понятие регрессии как способ описания влияния различных факторов с целевой переменной. Рассматриваются линейная многомерная регрессия, логистическая регрессия, Пуассоновская регрессия, Кокс-регрессия и методы их построения.

## **7. Учебно-методическое и информационное обеспечение дисциплины**

### **Список литературы**

#### **Основная литература**

1. Гмурман В.Е., Теория вероятностей и математическая статистика: Учеб. пособие для вузов, М.:Высш. шк. (2003):480 с.
2. А.И. Кибзун, Е.Р. Горяинова, А.В. Наумов, А.Н. Сиротин, Теория Вероятностей и математическая статистика/ Базовый курс с примерами и задачами, Москва:ФИЗМАТЛИТ (2002):224 с.
3. Д.А. Коршунов, Н.И. Чернова, Сборник задач и упражнений по математической статистике, Российская академия наук сибирское отделение институт математики им. С.Л. Соболева Новосибирск:Издательство института математики (2004):128 с.
4. М.Б. Лагутин, Наглядная математическая статистика, М:БИНОМ. Лаборатория знаний (2009):474 с.
5. Кокс Д.Р., Оукс Д., Анализ данных типа времени жизни, Пер. с англ. М:Финансы и статистика (1988):191с.
6. E. L. Kaplan, Paul Meier, Nonparametric Estimation from Incomplete Observations, JASA 1958, 53: 457-481
7. Феллер В., Введение в теорию вероятностей и её приложения (Том 2), Москва,«Мир» (1967):765 с.
8. С.А. Айвазян, В.М. Бухштабер, И.С. Енюков, Л.Д. Мешалкин, ПРИКЛАДНАЯ СТАТИСТИКА: Классификация и снижение размерности, М.:Финансы и статистика (1989):607 с.



9. С.А. Айвазян, И.С. Енюков, Л.Д. Мешалкин, Исследование зависимостей М:Финансы и статистика (1985):489 с.
10. С.А. Айвазян, И.С. Енюков, Л.Д. Мешалкин, Основы моделирования и первичная обработка данных, М:Финансы и статистика (1983):472 с.
11. Вапник В.Н., Червоненкис А.Я., Теория распознавания образов, М:Наука (1974):416с.
12. Д. Письменный, Конспект лекций по теории вероятностей и математической статистике, М:Айрис-пресс (2004):256 с.
13. Поляк Б.Т., Введение в оптимизацию, М.:Наука (1983):383 с.
14. Айзерман М.А., Браверман Э.М., Розоноэр Л.И., Метод потенциальных функций в теории обучения машин, М:Наука (1970):384 с.
15. В.В. Вьюгин, Элементы математической теории машинного обучения, М:МФТИ (2010):341 с.
16. К.Л. Самаров, Аналитическая геометрия, ОООРезольвента (2009):33 с.
17. Мерков А.Б., Введение в методы статистического обучения (2009):140 с.
18. В.Н. Вапник, Восстановление зависимостей по эмпирическим данным, М:Наука (1979):449
19. Тихонов А.Н., Арсенин В.Я., Методы решения некорректных задач, М:Наука. Главная редакция физико-математической литературы (1979):142 с.
20. Червоненкис А.Я., Компьютерный анализ данных. Яндекс, 2009.