Алгоритмы машинного обучения 1
модуль 4, учебный год 2025–2026

Сергей Головань
Российская экономическая школа
sgolovan@nes.ru

ТА: Егор Горский (egorskii@nes.ru)

Описание курса

Курс «Алгоритмы машинного обучения 1» позволяет студентам совершенствовать навыки использования статистических и эконометрических методов, которые широко применяются в экономической науке, в частности, в финансах и макроэкономике. Основным предметом обсуждения курсе являются способы извлечения информации из накопленных данных, которые особенно полезны в областях, где данные собираются в больших объемах, в частности, в финансовых приложениях. Этот курс является курсом по выбору. Он состоит из 14 лекций и 7 семинаров.

Система оценивания и требования к выставлению итоговой оценки

Курс опирается на технику, освоенную при прослушивании стандартных курсов эконометрики, но кроме него не требует никаких предварительных знаний, кроме базовых курсов по математическому анализу, линейной алгебре и теории вероятностей.

В курсе будут предложены 4 домашних задания, которые составят 40% от окончательной оценки за курс. Остальные 60% приходятся на финальный экзамен.

Содержание курса

1. Введение в статистическое обучение
   (a) Что такое статистическое обучение?
   (b) Обучение с присмотром и без присмотра
   (c) Регрессия и классификация
2. Обучение под присмотром
   (a) Линейная регрессия (регуляризация: ридж-регрессия, лассо)
   (b) Линейная классификация (дискриминантный анализ, логистические модели)
   (c) Полиномиальные и непараметрические модели
   (d) Аддитивные модели
   (e) Модели, основанные на деревьях

   (f) Нейронные сети
   (g) Модели опорных векторов
   (h) Классификация по ближайшим соседям

3. Обучение без присмотра

   (a) Ассоциативные правила
   (b) Кластерный анализ
   (c) Факторный анализ

Структура и организация учебной дисциплины

Лекции будут следовать от мотивационных примеров и примеров экономических моделей к общим утверждениям и принципам. Кроме того, студентам будут выданы компьютерные упражнения, которые позволят освоить изучаемые методы на практике.

Примеры заданий и вопросов для самостоятельной работы и промежуточного контроля

1. Пусть дан набор данных, который мы разбили на два подмножества одинакового размера: обучающее и тестовое, и теперь применяем две разные классифицирующие процедуры. Сначала мы использовали логистическую регрессию и получили долю ошибок 20% на обучающем подмножестве и 30% на тестовом. Далее, мы использовали процедуру с 1 ближайшим соседом (т.е. $K = 1$) и получили среднюю долю ошибок (усредненную по обучающем и тестовом подмножествах), равную 18%. Основываясь на этих результатах, какой метод следует предпочесть для классификации нового наблюдения? Почему?

2. Пусть мы сгенерировали десять бутстраповских выборок из набора данных, содержащих красный и зеленый классы. Далее, мы применили метод классификации с помощью деревьев к каждой бутстраповской выборке и для определенного значения X получили 10 оценок вероятности P(Класс красный | X):

   $$0.1,\ 0.15,\ 0.2,\ 0.2,\ 0.55,\ 0.6,\ 0.6,\ 0.65,\ 0.7 \text{ и } 0.75.$$

   Существует два распространенных способа комбинирования этих результатов в одно предсказание. Первый способ — голосование. Второй способ — классификация, основанная на средней вероятности. Для приведенного случая каким получится результат классификации с помощью каждого из двух способов,

3. В данной задаче мы исследуем классификатор максимальной ширины полосы на игрушечном наборе данных.

   (a) Нам даны $n = 7$ наблюдений размерности $p = 2$. Каждому наблюдению приписана некоторая метка класса.

Набл.	X_1	X_2	Y
1	3	4	Красный
2	2	2	Красный
3	4	4	Красный
4	1	4	Красный
5	2	1	Синий
6	4	3	Синий
7	4	1	Синий

Нарисуйте эти наблюдения.

(b) Схематично изобразите оптимальную разделяющую гиперплоскость, найдите уравнение этой гиперплоскости.

(c) Опишите классифицирующее правило для классификатора максимальной ширины полосы. Описание может быть в таком духе: «Классифицировать в красный класс если $\beta_0 + \beta_1 X_1 + \beta_2 X_2 > 0$, и классифицировать в синий класс в противном случае». Найдите значения β_0, β_1 и β_2.

(d) Укажите на вашем рисунке полосу для гиперплоскости, соответствующей классификатору максимальной ширины полосы.

(e) Укажите опорные векторы для классификатора максимальной ширины полосы.

(f) Покажите, что небольшое перемещение седьмого наблюдения не повлияет на гиперплоскость, соответствующую классификатору максимальной ширины полосы.

(g) Нарисуйте еще одну гиперплоскость, не являющуюся оптимальной разделяющей гиперплоскостью. Найдите ее уравнение.

(h) Нарисуйте еще одно наблюдение таким образом, чтобы два класса стало невозможно разделить гиперплоскостью.

4. Рассмотрите нейронную сеть для зависимой переменной, принимающей K разных значений, для которой используется кросс-энтропийная функция потерь. Покажите, что если сеть не содержит скрытых слоев, то она эквивалентна логистической модели множественного выбора.

5. В данной задаче вам предлагается проделать процедуру кластеризации методом K-средних вручную при $K = 2$ на малой выборке размера $n = 6$ размерности $p = 2$. Наблюдения следующие:

Набл.	X_1	X_2
1	1	4
2	1	3
3	0	4
4	5	1
5	6	2
6	4	0

(a) Нарисуйте наблюдения.

(b) Случайно присвойте номер кластера каждому наблюдению. Выпишите номер кластера для всех наблюдений.

(c) Найдите центроид для каждого кластера.

(d) Отнесите каждое наблюдение к ближайшему к нему центроиду (в смысле евклидова расстояния). Выпишите номер кластера для каждого наблюдения.

(e) Повторяйте шаги (5c) и (5d) до тех пор, пока номера кластеров не перестанут меняться.

(f) На рисунке из пункта (5a) раскрасьте наблюдения согласно найденным кластерам.

Политика академической честности

Списывание, плагиат и другие нарушения академической этики в РЭШ недопустимы.