



МИНОБРНАУКИ РОССИИ

Федеральное государственное бюджетное образовательное учреждение
высшего образования

«МИРЭА - Российский технологический университет»

РТУ МИРЭА

Институт кибернетики

Кафедра высшей математики

ОТЧЁТ ПО Практике по получению первичных профессиональных умений
и навыков
(указать вид практики)

Тема практики: Разведовательный анализ наборов данных (kaggle.com)
приказ университета о направлении на практику
793 – С от 12.02.2019 г.

Отчет представлен к
рассмотрению:

Студент _____ группы
КМБО-01-18

Милешин А.Д.
(расшифровка подписи)
«6» нояб 2019г.

Отчет утвержден.
Допущен к защите:

Руководитель практики
от кафедры

Петрусеви́ч Д.А.
(расшифровка подписи)
«6» нояб 2019г.

Москва 2019



МИНОБРНАУКИ РОССИИ

Федеральное государственное бюджетное образовательное учреждение
высшего образования

«МИРЭА - Российский технологический университет»

РТУ МИРЭА

ЗАДАНИЕ НА Практику по получению первичных профессиональных умений и навыков

**Студенту 1 курса учебной группы КМБО-01-18 института кибернетики
Милешину Артему Дмитриевичу**

(фамилия, имя и отчество)

Место и время практики: Институт кибернетики, кафедра высшей математики

Время практики: с «5» ноября 2019 по «2» марта 2020

Должность на практике: практикант

1. ЦЕЛЕВАЯ УСТАНОВКА: изучение основ анализа данных и машинного обучения

2. СОДЕРЖАНИЕ ПРАКТИКИ:

2.1 Изучить: литературу и практические примеры по темам: 1) построение линейной регрессии, 2) использование метода главных компонент, 3) поиск и устранение линейной зависимости в данных, 4) основы нормализации данных, 5) методы классификации и кластеризации («решающее дерево», «случайный лес», «k ближайших соседей»), 6) обучение с учителем («градиентный спуск»)

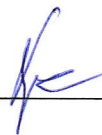
2.2 Практически выполнить: 1) снижение размерности исходных задач при помощи метода главных компонент при возможности; построение линейной регрессии для некоторого параметра, исключение регрессоров, не коррелирующих с объясняемой переменной; решение задачи классификации анализа данных на основе открытого набора данных с ресурса kaggle.com, посвященного пожарам

2.3 Ознакомиться: с применением метода главных компонент; методов классификации («решающего дерева», «случайного леса»); методов градиентного спуска («градиентным бустингом»); методов кластеризации («k ближайших соседей»).

3. ДОПОЛНИТЕЛЬНОЕ ЗАДАНИЕ:

4. ОГРАНИЗАЦИОННО-МЕТОДИЧЕСКИЕ УКАЗАНИЯ: построить классификацию на основе нескольких методов и произвести сравнение результатов классификации; сделать выводы о применимости использованных методов; сформировать выводы по результатам задачи из предметной области: на основе каких параметров нужно проводить классификацию типа лесного покрова?

Заведующий кафедрой
высшей математики



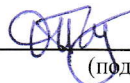
Ю.И.Худак

«6» нояб 2019 г.

СОГЛАСОВАНО

Руководитель практики от кафедры:

«6» ноября 2019 г.



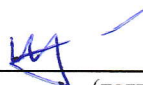
(подпись)

(Петрусевич Д.А.)

(фамилия и инициалы)

Задание получил:

«6»ноября 2019 г.



(подпись)

(Милешин А.Д.)

(фамилия и инициалы)

ИНСТРУКТАЖ ПРОВЕДЕН:

Вид мероприятия	ФИО ответственного, подпись, дата	ФИО студента, подпись, дата
Охрана труда	Петрусеви́ч Д.А.  «6» ноября 2019 г.	Милеши́н А.Д.  «6» ноября 2019 г.
Техника безопасности	Петрусеви́ч Д.А.  «6» ноября 2019 г.	Милеши́н А.Д.  «6» ноября 2019 г.
Пожарная безопасность	Петрусеви́ч Д.А.  «6» ноября 2019 г.	Милеши́н А.Д.  «6» ноября 2019 г.
Правила внутреннего распорядка	Петрусеви́ч Д.А.  «6» ноября 2019 г.	Милеши́н А.Д.  «6» ноября 2019 г.



МИНОБРНАУКИ РОССИИ

Федеральное государственное бюджетное образовательное учреждение
высшего образования


«МИРЭА - Российский технологический университет»

РТУ МИРЭА

РАБОЧИЙ ГРАФИК ПРОВЕДЕНИЯ Практики по получению
первичных профессиональных умений и навыков

студента Милешина А.Д. 1 курса группы КМБО-01-18 очной формы
обучения, обучающегося по направлению подготовки 01.03.02
«Прикладная математика и информатика»,
профиль «Математическое моделирование и вычислительная математика»


Неделя	Сроки выполнения	Этап	Отметка о выполнении
1	06.11.2019	Выбор темы практики/НИР. Пройти инструктаж по технике безопасности.	
1	06.11.2019	Вводная установочная лекция.	
3	20.11.2019	Построение и оценка линейной регрессии с помощью языка R	
5	04.12.2019	Использование метода главных компонент, выделение линейной зависимости в данных	
7	18.12.2019	Методы классификации и кластеризации; построение решающего дерева;	
9	10.01.2020	Концепция бэггинга, «случайный лес»; концепция бустинга; градиентные методы обучения и кластеризации	

17	02.03.2020	Представление отчётных материалов по практике/НИР и их защита. Передача обобщённых материалов на кафедру для архивного хранения.	
		Зачётная аттестация.	

Содержание практики и планируемые результаты согласованы с руководителем практики от профильной организации.

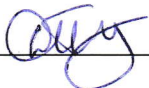
Согласовано:

Заведующий
кафедрой



/ ФИО / Худак Ю.И.

Руководитель
практики от кафедры



/ ФИО / Петрусевич Д.А.

Обучающийся



/ ФИО / Милешин А.Д.

Оглавление

Задание 3	2
Условие задания.....	2
Краткий Обзор	2
Решение задачи	2
<u>Информация о таблице и признаках</u>	<u>2</u>
<u>Анализ данных с использованием методов Pandas</u>	<u>3</u>
<u>Метод главных компонент</u>	<u>3</u>
<i>Описание метода</i>	<i>3</i>
<i>Работа с категориальными признаками</i>	<i>4</i>
<i>Нормализация переменных</i>	<i>4</i>
<i>Использование метода PCA</i>	<i>4</i>
Вывод	5

Задание 3

Условие задания

Исследовать данные из таблицы Forest Fires с помощью библиотеке Pandas языка Python.

Краткий Обзор

Данный раздел будет выполнен на языке python с использованием библиотек для обработки данных. А именно:

- numpy – для работы со значениями.
- pandas – для считывания данных с csv-файл.
- scikit-learn – для использования большинства алгоритмов машинного обучения.

Код программы в Приложении.

Решение задачи

Информация о таблице и признаках

Для начала прочтём информацию о таблице и данных. В данном случае нам представлена таблица с данными, где 517 элементов и 13 признаков. 13 признаков это:

- X, Y - пространственные координаты пожара на карте парка Монтезиньо, тип integer (X - от 1 до 9, Y – от 2 до 9)
- month – месяц, тип string
- day - день недели, тип string
- FFMC - FFMC-индекс из системы FWI, тип float (от 18.7 до 96.20)
- DMC - DMC-индекс из системы FWI, тип float (от 1.1 до 291.3 7)
- DC - DC-индекс от системы FWI, тип float (от 7.9 до 860.6)
- ISI - ISI-индекс из системы FWI, тип float (от 0 до 56.1)
- temp - температура (в градусах Цельсия), тип float (от 2.2 до 33.3)
- RH - относительная влажность воздуха (в %), тип integer (от 15 до 100)
- wind - скорость ветра (в км/ч), тип float (от 0.4 до 9.4)
- rain - скорость выпадения осадков (в мм/м2), тип float (от 0 до 6.4)
- area - выжженная площадь леса (в га), тип float (от 0 до 1090.84)

Очевидно, что в качестве категориальных признаков выступают day и month.

Разберёмся с индексами из системы FWI – Системы оценки погодного индекса. Узнаем, что означают данные нам индексы:

- FFMC (Fire Fuel Moisture Code) – индекс содержания влаги в горючих материалах. Данный индекс показывает лёгкость возгорания лесных горючих материалов.
- DMC (Duff Moisture Code) – индекс содержания влаги в лесной подстилке.
- DC (Drought Code) – индекс влажности глубоких слоёв почвы.
- ISI (Initial Spread Index) – индекс ожидаемых темпов распространения огня. Первичный анализ данных

Теперь рассмотрим данные из таблицы. Таблица не содержит пропусков, а аномальными значениями можно считать $area > 110$ (2% данных), $FFMC < 80.7$ (2% данных), $rain > 0.64$ (1% данных), $ISI > 22.44$ (менее 1% данных).

При просмотре значений можно заметить, что почти в половине случаев $area = 0$. Это связано с тем, что эта переменная приближена к 0.

Анализ данных с использованием методов Pandas

Поскольку мы просмотрели данные, для дальнейшего удобства далее сгруппируем всё по клеткам. Для начала создадим новую таблицу, где занесём количество возгораний в данной клетке, средние значения по всем числовым индексам и укажем месяц и дату, в которые чаще всего случались возгорания.

Карта парка Монтезиньо

	2	3	4	5	6	7	8	9
1	19	10	15	4				
2	25	1	27	20				
3		1	43	7	4			
4		22	36	25	8			
5			23	3	4	25		
6		25	9	49	3			
7		2	45	11	2			
8		3	1	4	52		1	
9			4	2	1			6

По итогу получим, что чаще всего пожары происходили в клетке (6, 8) (52 раза), чаще всего происходили они в августе, а самый частый день пожаров – воскресенье.

Визуализируем количество пожаров в каждой клетке используя Excel. Заметим существование “очагов” пожаров, таких как (3, 4) или (6, 5). Так же заметим существование “аномальных” пожаров, таких как в (9, 9) или (8, 8).

Рисунок 1. Карта частоты возгораний парка Монтезиньо

Метод главных компонент

Описание метода

Теперь используем метод PCA для очагов возгорания. Это будет 4 очага:

1. Клетка (8, 6)
2. Клетки (3, 4) и (4, 4)
3. Клетка (6, 5)
4. Клетка (7, 4)

Остальные клетки нам будут не интересны поскольку количество случаев возгорания у них < 30 , а поскольку метод PCA основан на машинном обучении, это слишком маленькая выборка для обучения. Клетку (4, 4) мы будем рассматривать вместе с клеткой (3, 4). Таким образом мы рассматриваем первые 5 клеток по количеству возгораний.

Работа с категориальными признаками

Поскольку метод PCA работает только с числами, то преобразуем признаки month и day двумя разными методами: Label Encoding и One-hot Encoding.

Label Encoding – метод, который преобразует исходный столбец с категориальным признаком, присвоив каждому значению признака некое число. Таким образом Label Encoding преобразует столбец из string в int64. Тем самым метод PCA сможет работать только с числами и при этом категориальный признак будет участвовать.

One-hot Encoding – метод, который добавляет таблицу, где признаками выступают все возможные значения категориального признака. Значения в самой таблице – 0 или 1, которые означают, был ли данный признак у значения или нет.

Метод One-hot Encoding будем использовать для month что бы увидеть влияние отдельного месяца (если это влияние существует), Label Encoding будем использовать для day поскольку влияние отдельного дня на изменение погоды и возникновение пожара мало, поэтому мы будем проверять просто влияние изменения дня на погоду.

Нормализация переменных

Теперь, для правильной работы PCA нормализуем значения в таблице. Для этого скоррелируем данные.

	X	Y	FFMC	DMC	DC	ISI	temp	RH	wind	rain	area
X	1.000000	0.539548	-0.021039	-0.048384	-0.085916	0.006210	-0.051258	0.085223	0.018798	0.065387	0.063385
Y	0.539548	1.000000	-0.046308	0.007782	-0.101178	-0.024488	-0.024103	0.062221	-0.020341	0.033234	0.044873
FFMC	-0.021039	-0.046308	1.000000	0.382619	0.330512	0.531805	0.431532	-0.300995	-0.028485	0.056702	0.040122
DMC	-0.048384	0.007782	0.382619	1.000000	0.682192	0.305128	0.469594	0.073795	-0.105342	0.074790	0.072994
DC	-0.085916	-0.101178	0.330512	0.682192	1.000000	0.229154	0.496208	-0.039192	-0.203466	0.035861	0.049383
ISI	0.006210	-0.024488	0.531805	0.305128	0.229154	1.000000	0.394287	-0.132517	0.106826	0.067668	0.008258
temp	-0.051258	-0.024103	0.431532	0.469594	0.496208	0.394287	1.000000	-0.527390	-0.227116	0.069491	0.097844
RH	0.085223	0.062221	-0.300995	0.073795	-0.039192	-0.132517	-0.527390	1.000000	0.069410	0.099751	-0.075519
wind	0.018798	-0.020341	-0.028485	-0.105342	-0.203466	0.106826	-0.227116	0.069410	1.000000	0.061119	0.012317
rain	0.065387	0.033234	0.056702	0.074790	0.035861	0.067668	0.069491	0.099751	0.061119	1.000000	-0.007366
area	0.063385	0.044873	0.040122	0.072994	0.049383	0.008258	0.097844	-0.075519	0.012317	-0.007366	1.000000

Рисунок 2. Корреляция данных

Исходя из данных, нормализовать будем по параметрам ‘rain’ и ‘area’ двумя способами: нормализация стандартным отклонением и MinMax.

Способ нормализации стандартным отклонением вычисляется по формуле:

$$z = \frac{x - \mu}{\sigma}$$

где μ - значение, σ – стандартное отклонение.

Способ нормализации MinMax вычисляется по формуле:

$$z = \frac{x - \min(x)}{\max(x) - \min(x)}$$

После нормализации все значения лежат в интервале (-1:1), что означает что нормализация прошла успешно.

Использование метода PCA

Перед запуском метода PCA данные разделяют на тестовую сборку и тренировочную. Поскольку мы сделаем 4 независимых вычисления с помощью PCA, перед каждым разом мы будем создавать свою тренировочную и тестовую сборки. Теперь приступим к запуску самого метода.

По итогам запуска получаем:

- Для очага (8, 6) 90% дисперсии объясняет 1 компоненты, и в эту компоненту больший вклад вносят значения из столбца DC – индекс влажности глубоких слоёв почвы, или индекс засухи. Из этого можно предположить, что причиной возникновения пожаров является засуха. Так же на этот вывод подталкивает то, что немалый вклад в первую компоненту вносят значения из столбца DMC – индекс засухи верхних слоёв почвы. Из этого следует то, что пожары в данном очаге возникали из-за засухи.

For X = 8, Y = 6:
1 component: 92.51% of initial variance
0.000 x day + -0.002 x FFMC + -0.229 x DMC + -0.972 x DC + 0.002 x ISI + -0.019 x temp + 0.043 x RH + 0.002 x wind + 0.000 x rain + -0.000 x area + 0.000 x apr + -0.001 x aug + 0.000 x dec + 0.000 x feb + 0.000 x jan + -0.000 x jul + 0.000 x jun + 0.001 x mar + 0.000 x may + 0.000 x nov + -0.000 x oct + -0.001 x sep

на этот вывод подталкивает то, что немалый вклад в первую компоненту вносят значения из столбца DMC – индекс засухи верхних слоёв почвы. Из этого следует то, что пожары в данном очаге возникали из-за засухи.

- Для очага (6, 5) 90% дисперсии объясняет 1 компонента. В первую компоненту наибольший вклад так же вносят значения из столбца DC. Но, в отличии от первого случая, у данного очага сильно меньший вклад вносят переменные DMC и FFMC, и наибольший вклад из месяцев даёт март и сентябрь. Значит, что засухи в этом регионе немного отличались от предыдущего очага, но всё равно пожары в данном очаге происходили из-за засух.

For X = 6, Y = 5:
1 component: 97.32% of initial variance
0.001 x day + -0.005 x FFMC + -0.178 x DMC + -0.984 x DC + -0.003 x ISI + -0.011 x temp + -0.004 x RH + 0.002 x wind + -0.000 x rain + -0.000 x area + 0.000 x apr + -0.000 x aug + 0.000 x dec + 0.001 x feb + -0.000 x jan + -0.000 x jul + 0.000 x jun + 0.001 x mar + -0.000 x may + -0.000 x nov + -0.000 x oct + -0.001 x sep

Значит, что засухи в этом регионе немного отличались от предыдущего очага, но всё равно пожары в данном очаге происходили из-за засух.

- Для очага (7, 4) 90% дисперсии объясняет так же 1 компонента. В первую компоненту наибольший вклад так же вносят значения из столбца DC. И вообще по общим значениям данный очаг очень похож на предыдущий.

For X = 7, Y = 4:
1 component: 93.83% of initial variance
-0.000 x day + -0.004 x FFMC + -0.194 x DMC + -0.981 x DC + 0.005 x ISI + -0.007 x temp + 0.002 x RH + 0.001 x wind + -0.000 x rain + -0.000 x area + 0.000 x apr + -0.000 x aug + -0.000 x dec + 0.001 x feb + -0.000 x jan + 0.000 x jul + 0.000 x jun + 0.001 x mar + -0.000 x may + -0.000 x nov + -0.000 x oct + -0.001 x sep

Отличия в месяцах: в данном очаге вклад вносит ещё февраль, а также отличия в вкладе столбцов давления и ветра – они ниже, чем у предыдущего очага. Но в данном очаге вклад столбца DMC чуть выше, что говорит о том, что засухи в этом регионе были более продолжительными.

- Для очага (3, 4) и (4, 4) 90% дисперсии объясняет 1 компонента. Большой вклад в первую компоненту вносят значения из столбца DC, и вообще по значениям данный очаг сильно похож на первый очаг (8, 6). Сильно отличаются вклады столбцов DMC и давления. Кроме того, вклад так же вносит месяц февраль. В остальном данный результат очень похож на первый очаг.

For X = 3, Y = 4 and X = 4, Y = 4:
1 component: 98.05% of initial variance
-0.001 x day + -0.010 x FFMC + -0.172 x DMC + -0.985 x DC + -0.010 x ISI + -0.012 x temp + -0.020 x RH + 0.002 x wind + -0.000 x rain + -0.000 x area + 0.000 x apr + -0.001 x aug + 0.000 x dec + 0.000 x feb + -0.000 x jan + 0.000 x jul + -0.000 x jun + 0.001 x mar + -0.000 x may + -0.000 x nov + -0.000 x oct + -0.001 x sep

Сильно отличаются вклады столбцов DMC и давления. Кроме того, вклад так же вносит месяц февраль. В остальном данный результат очень похож на первый очаг.

Вывод

После использования метода РСА увеличилось представление о самих очагах и причинах возникновения пожара. А именно слишком сухая почва на глубине, что обозначает частые периодические засухи в данных регионах. Но при этом всё засушливость во всех рассмотренных регионах разная, что видно по итоговым цифрам. Так пожары в регионе (8, 6) возникают так же и из-за температуры, поскольку она вносит больший вклад 1 компоненту, чем во всех остальных случаях.

Из этого мы можем сделать вывод, что пожары в данной местности возникают зачастую из-за засух и, как следствие, слишком сухой глубокой почвы. Этот же вывод доказывает результат метода РСА, запущенный для всех случаев пожара. Как видно, 90% дисперсии объясняет 1 компонента и наибольший вклад в эту компоненту вносят значения из столбца DC.

1 component: 96.22% of initial variance
0.001 x X + 0.000 x Y + -0.001 x day + -0.000 x FFMC + -0.186 x DMC + -0.982 x DC + -0.005 x ISI + -0.010 x temp + 0.001 x RH + 0.001 x wind + -0.000 x rain + -0.000 x area + 0.000 x apr + -0.001 x aug + 0.000 x dec + 0.000 x feb + 0.000 x jan + 0.000 x jul + 0.000 x jun + 0.001 x mar + 0.000 x may + 0.000 x nov + -0.000 x oct + -0.001 x sep

доказывает результат метода РСА, запущенный для всех случаев пожара. Как видно, 90% дисперсии объясняет 1 компонента и наибольший вклад в эту компоненту вносят значения из столбца DC.