Министерство образования и науки Российской Федерации

Калужский филиал
федерального государственного бюджетного образовательного
учреждения высшего образования
**«Московский государственный технический университет
имени Н.Э. Баумана
(национальный исследовательский университет)»**
(КФ МГТУ им. Н.Э. Баумана)

И.И. Кручинин
(к.т.н. доцент)

ЛАБОРАТОРНАЯ РАБОТА № 2
по курсу «Методы машинного обучения»

Метрические методы классификации многомерных объектов пересекающихся классов

Калуга

2018

Теоретические основы.

Метрические методы классификации

Во многих прикладных задачах измерять степень сходства объектов существенно проще, чем формировать признаковые описания. Если мера сходства объектов введена достаточно удачно, то, как правило, оказывается, что схожим объектам очень часто соответствуют схожие ответы. В задачах классификации это означает, что классы образуют компактно локализованные подмножества. Для формализации понятия «сходства» вводится функция расстояния в пространстве объектов X. Методы обучения, основанные на анализе сходства объектов, будем называть метрическими.

Метрический алгоритм классификации с обучающей выборкой X^ℓ относит объект u к тому классу $y \in Y$, для которого суммарный вес ближайших обучающих объектов $\Gamma_y(u,X^\ell)$ максимален:

$$a(u; X^\ell) = \arg\max_{y \in Y} \Gamma_y(u, X^\ell);$$

$$\Gamma_y(u, X^\ell) = \sum_{i=1}^{\ell} [y_u^{(i)} = y]\, w(i, u);$$

где весовая функция $w(i,u)$ оценивает степень важности i-го соседа для классификации объекта u. Функция $\Gamma_y(u,X^\ell)$ называется оценкой близости объекта u к классу y.

Обучающая выборка X^ℓ играет роль параметра алгоритма a. Настройка сводится к запоминанию выборки, и, возможно, оптимизации каких-то параметров весовой функции, однако сами объекты не подвергаются обработке и сохраняются «как есть». Алгоритм $a(u;X^\ell)$ строит локальную аппроксимацию выборки X^ℓ, причём вычисления откладываются до момента, пока не станет известен объект u. По этой

причине метрические алгоритмы относятся к методам ленивого обучения (lazy learning), в отличие от усердного обучения (eager learning), когда на этапе обучения строится функция, аппроксимирующая выборку.

Алгоритм *k* ближайших соседей (*k* nearest neighbors, *k*NN). Чтобы сгладить влияние выбросов (объектов, находящийся в окружении объектов чужого класса), будем относить объект u к тому классу, элементов которого окажется больше среди ближайших соседей.

$$w(i,u) = [i \leqslant k]; \qquad a(u; X^\ell, k) = \arg\max_{y \in Y} \sum_{i=1}^{k} [y_u^{(i)} = y].$$

При $k = 1$ этот алгоритм совпадает с предыдущим, следовательно, неустойчив к шуму. При $k = \ell$, наоборот, он чрезмерно устойчив и вырождается в константу.
Таким образом, крайние значения k нежелательны. На практике оптимальное значение параметра k определяют по критерию скользящего контроля с исключением объектов по одному (leave-one-out, LOO). Для каждого объекта $x_i \in X^\ell$ проверяется, правильно ли он классифицируется по своим k ближайшим соседям.

$$\text{LOO}(k, X^\ell) = \sum_{i=1}^{\ell} \Big[a\big(x_i; X^\ell \backslash \{x_i\}, k\big) \neq y_i \Big] \to \min_k .$$

Заметим, что если классифицируемый объект x_i не исключать из обучающей выборки, то ближайшим соседом x_i всегда будет сам x_i, и минимальное (нулевое) значение функционала LOO(k) будет достигаться при $k = 1$.

Рассмотрим реализацию метода K-ближайших соседей для решения задачи классификации в пакете R. Данные для исследования выбраны из медицинской отрасли. Пусть был составлен список характеристик пациентов прошедших медицинское обследование (до 500 строк).

	id	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean	concavity_mean	c
2	842302	M	17.99	10.38	122.8	1001	0.1184	0.2776	0.3001	0
3	842517	M	20.57	17.77	132.9	1326	0.08474	0.07864	0.0869	0
4	84300903	M	19.69	21.25	130	1203	0.1096	0.1599	0.1974	0
5	84348301	M	11.42	20.38	77.58	386.1	0.1425	0.2839	0.2414	0
6	84358402	M	20.29	14.34	135.1	1297	0.1003	0.1328	0.198	0

В таблице id – уникальный идентификатор пациента; diagnosis – диагноз (значение переменной отклика): M (англ.: malignant) – злокачественный, B (англ.: benign) – доброкачественный; остальные 30 столбцов содержат медицинские показатели. На языке R организуем ввод исходных данных:

```
wbcd <- read.csv("wbcd_data.txt", stringsAsFactors = FALSE)
```

Теперь мы можем обращаться к значениям столбцов по их именам. Например, wbcd$id – вектор из уникальных идентификаторов пациентов;
wbcd$diagnosis – вектор из диагнозов и т.д.
Подсчитаем количество обоих значений диагноза:
table(wbcd$diagnosis)
полученный результат:

```
  B   M
357 212
```

Реализованная в пакете R функция, выполняющая классификацию методом
kNN, предполагает, что переменная отклика (англ.: target feature) имеет тот же вид,
что и факторы. Поэтому перекодируем значения переменной отклика (diagnosis), выделив 2 уровня («M» и «B») и заменив эти буквы на «Malignant» и «Benign», соответственно:

```
wbcd$diagnosis <- factor(wbcd$diagnosis, levels = c("B", "M"),
labels = c("Benign", "Malignant"))
```

Подсчитаем проценты каждого из исходов, округлив результаты до десятых:

```
round(prop.table(table(wbcd$diagnosis))*100, digits = 1)
```

```
Benign Malignant
  62.7      37.3
```

Получим:

Поскольку идентификатор нам для исследования не нужен, удалим этот столбец и сохраним данные в той же переменной:
wbcd <- wbcd[-1]
Посмотрим на сводные характеристики некоторых показателей (напомним, что их всего 30), например, radius_mean, area_mean и smoothness_mean, т.е. «средний радиус», «средняя площадь» и «средняя гладкость»
summary(wbcd[c("radius_mean", "area_mean", "smoothness_mean")])
Для каждого из названных 3-х показателей мы тем самым получим минимальное и максимальное значения, выборочное среднее и выборочную оценку медианы, а также выборочные оценки 25%-го и 75%-го квантилей:

```
  radius_mean       area_mean      smoothness_mean
Min.   : 6.981   Min.   : 143.5   Min.   :0.05263
1st Qu.:11.700   1st Qu.: 420.3   1st Qu.:0.08637
Median :13.370   Median : 551.1   Median :0.09587
Mean   :14.127   Mean   : 654.9   Mean   :0.09636
3rd Qu.:15.780   3rd Qu.: 782.7   3rd Qu.:0.10530
Max.   :28.110   Max.   :2501.0   Max.   :0.16340
```

Нетрудно видеть, что диапазоны значений выбранных нами показателей сильно
различаются. Применим нормализацию данных. Для этого напишем функцию:
normalize <- function(x) {return ((x - min(x)) / (max(x) - min(x)))}
Теперь, чтобы применить эту функцию к списку, будем использовать функцию lapply (здесь буква «l» – от английского «list», т.е. список, «apply» - англ.:
«применить»). Кроме того, преобразуем результат в объект data.frame (это, как мы уже знаем, позволит обращаться непосредственно к строкам или столбцам мат-рицы) и сохраним результат в переменной: wbcd_n:
wbcd_n <- as.data.frame(lapply(wbcd[2:31], normalize))
Здесь мы указываем, что применяем нормализацию к столбцам с номерами 2 - 31.
(Напомним, что мы уже удалили первый столбец – id, теперь первый столбец содержит значения переменной отклика, а столбцы с номерами 2 — 31 содержат значения факторов (предикторов). После нормализации все пере-

менные будут иметь диапазон [0, 1]. В этом нетрудно убедиться, вновь вызвав функцию summary (но уже для новой переменной, т.е. wbcd_n).

К сожалению, у нас нет данных, «непомеченных» буквами «M» и «B» (а есть только данные, диагноз для которых уже поставлен). Поэтому для тестирования метода kNN, разобьём выборку на 2 части – одну из них будем использовать как обучаю-щую, а другую (несмотря на то, что мы знаем значения переменной отклика для каждой записи в ней) будем использовать для прогноза значения переменной от-клика. Потом мы сможем сравнить полученные результаты с истинным значением переменной отклика (диагнозом). Пусть в обучающую выборку (назовём её wbcd_train) войдут первые 469 записей

wbcd_train <- wbcd_n[1:469,]
В «подопытную» группу войдут остальные записи, т.е. те, что имеют номера от 470
до 569:
wbcd_test <- wbcd_n[470:569,]

Нам нужно убрать значения переменной отклика из наборов данных (т.е. оставить только значения факторов). Сохраним значения диагнозов в отдельных векторах:
wbcd_train_labels <- wbcd[1:469, 1]
Это будут диагнозы из обучающей выборки («1» означает, что мы берём их из первого столбца).
wbcd_test_labels <- wbcd[470:569, 1]
Это – диагнозы для тестовой выборки.

Выбор пакета в среде R осуществляется при помощи опции «Включить пакеты» меню «Пакеты» среды R. При выборе этой опции пользователю предлагается список установленных пакетов. Теперь нам нужна функция knn(). Она находится в пакете class, который необходимо подключить. После установки пакета нужно вызвать функцию library:
library("class")
Вызовем теперь функцию knn():
wbcd_test_pred <- knn(train = wbcd_train, test = wbcd_test,cl =
wbcd_train_labels, k=21)
Здесь параметр train – имя обучающей выборки, test – имя тестовой выборки, cl – имя файла со значениями переменной отклика для данных из обучающей выборки, k – число «опрошенных» соседей.

Распечатаем результат (мы сохранили его в переменной wbcd_test_pred)
Распечатаем истинные значения переменной отклика для тестовой выборки (мы ранее сохранили их в файле). На первый взгляд кажется, что эти списки совпадают, но при внимательном рассмотрении можно обнаружить отличия. Возникает вопрос: насколько точен прогноз, сделанный методом kNN?
Ответить на этот вопрос нам поможет функция CrossTable из пакета gmodels.

```
library("gmodels")
CrossTable(x = wbcd_test_labels, y = wbcd_test_pred, prop.chisq=FALSE)
```

```
Total Observations in Table:  100


                 | wbcd_test_pred
wbcd_test_labels |    Benign | Malignant | Row Total |
-----------------|-----------|-----------|-----------|
          Benign |        77 |         0 |        77 |
                 |     1.000 |     0.000 |     0.770 |
                 |     0.975 |     0.000 |           |
                 |     0.770 |     0.000 |           |
-----------------|-----------|-----------|-----------|
       Malignant |         2 |        21 |        23 |
                 |     0.087 |     0.913 |     0.230 |
                 |     0.025 |     1.000 |           |
                 |     0.020 |     0.210 |           |
-----------------|-----------|-----------|-----------|
    Column Total |        79 |        21 |       100 |
                 |     0.790 |     0.210 |           |
-----------------|-----------|-----------|-----------|
```

Здесь – левый верхний угол (Benign, Benign) содержит так называемые TN (англ.: True Negative) прогнозы, т.е. те случаи, когда рака у пациентки на самом деле не было, и метод kNN (правильно) решил, что его нет. Всего в тестовой выборке пациентов с доброкачественными образованиями было 77 (см. столбец «Row Total»), и метод kNN во всех 77 случаях правильно определил, что злокачественной опухоли нет. Иными словами, ни у одной из здоровых пациентов метод kNN не определил онкологического заболевания.
– правый верхний угол (Benign, Malignant) – FP (англ.: False Positive) – содержит количество ошибочно поставленных диагнозов «рак». Таких, как мы видим, не оказалось.

– левый нижний угол (Malignant, Benign) – FN (англ.: False Negative) – содержит количество «пропущенных» случаев заболевания. Иными словами, у 2-х пациентов (из 23) новообразование на самом деле было злокачественным, а метод kNN этого не заметил.
– правый нижний угол (Malignant, Malignant) – TP (англ.: True Positive) – содержит количество правильно диагностированных случаев онкологического заболевания – 21 (из 23).

Второй ряд содержит относительные величины – частные от деления элементов верхней строки на соответствующее число в строке «Row total».
Третий ряд содержит частные от деления элементов верхней строки на соответ-ствующее число в столбце «Column total».
Четвёртый ряд содержит частоты – частные от деления элементов верхней строки на объём тестовой выборки.
Таким образом, общее число ошибок метода kNN равно 2, что составляет 2% от общего числа прогнозов.

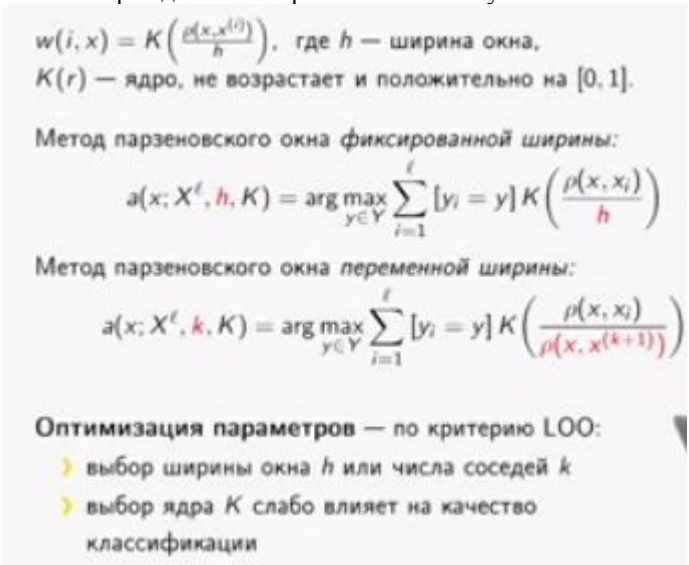
Метод парзеновского окна (Parzen Window)

Ещё один способ задать веса соседям — определить w_i как функцию от расстояния $\rho(u, x_u^{(i)})$, а не от ранга соседа i. Введём функцию ядра $K(z)$, невозрастающую на $[0,\infty)$. Положив $w(i,u) = K\left(\frac{1}{h}\rho(u, x_u^{(i)})\right)$ в общей формуле

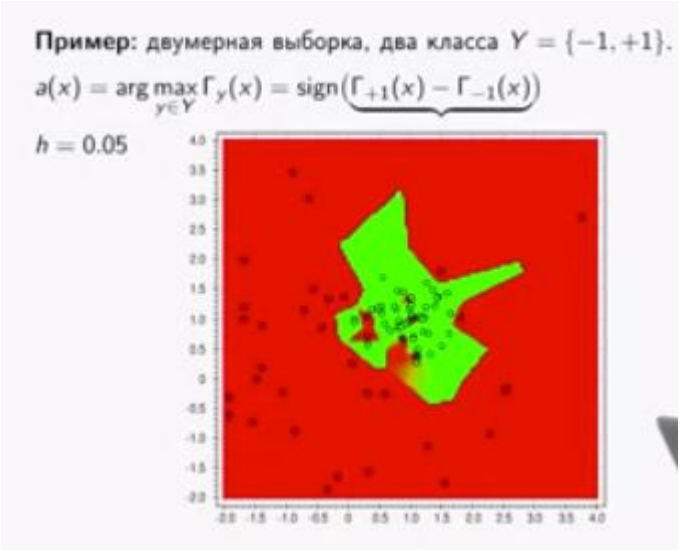
$$a(u; X^\ell, h) = \arg\max_{y \in Y} \sum_{i=1}^{\ell} [y_u^{(i)} = y]\, K\left(\frac{\rho(u, x_u^{(i)})}{h}\right).$$

Параметр h называется шириной окна и играет примерно ту же роль, что и число соседей k. «Окно» — это сферическая окрестность объекта u радиуса h, при попадании в которую обучающий объект x_i «голосует» за отнесение объекта u к классу y_i. Параметр h можно задавать априори или определять по скользящему контролю. Зависимость

LOO(h), как правило, имеет характерный минимум, поскольку слишком узкие окна приводят к неустойчивой классификации; а слишком широкие — к вырождению алгоритма в константу.

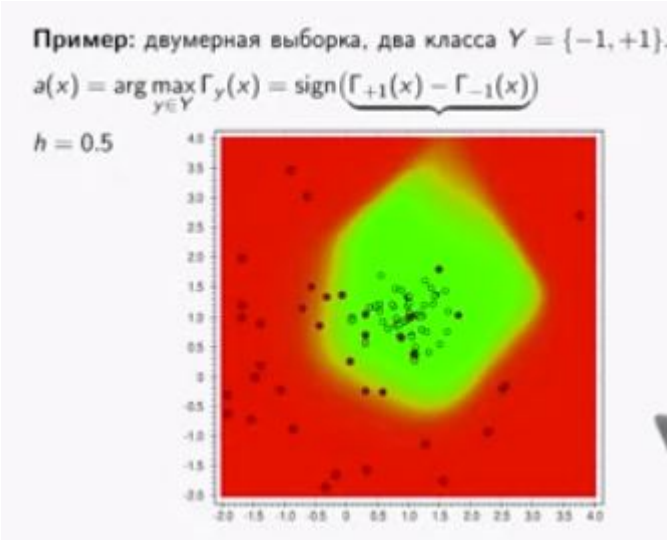


Фиксация ширины окна h не подходит для тех задач, в которых обучающие объекты существенно неравномерно распределены по пространству X. В окрестности одних объектов может оказываться очень много соседей, а в окрестности других — ни одного. В этих случаях применяется окно переменной ширины.



Возьмём финитное ядро — невозрастающую функцию $K(z)$, положительную на отрезке $[0,1]$, и равную нулю вне его. Определим h как наибольшее число, при котором ровно k ближайших соседей объекта u получают ненулевые веса: $h(u) = \rho(u, x_u^{(k+1)})$. Тогда алгоритм принимает вид

$$a(u; X^\ell, k) = \arg\max_{y \in Y} \sum_{i=1}^{k} [y_u^{(i)} = y] \, K\left(\frac{\rho(u, x_u^{(i)})}{\rho(u, x_u^{(k+1)})}\right).$$



(3.3)

Язык моделирования статистических данных R предлагает ряд встроенных функций для реализации метода окна Парзена.

Так функция parzen(x, bw = NULL, kernel = "gaussian", abc = FALSE, par = shorth(x), optim.method = "BFGS", ...) предлагает использовать несколько вариантов ядерных структур: "biweight", "cosine", "eddy", "epanechnikov", "gaussian", "optcosine", "rectangular", "triangular", "uniform"

Вызов функции для использования метода Парзена в среде разработки R осуществляется несколькими способами, например:

```
mlv(x, method = "kernel", ...)
mlv(x, method = "parzen", ...)
```

При этом можно использовать optim. method = c("Nelder-Mead", "BFGS", "CG", "L-BFGS-B", "SANN", "Brent")

```
M <- mlv(x, method = "kernel", kernel = "gaussian", bw = 0.3, par = shorth(x))
print(M)
plot(M)

x <- rbeta(1000,23,4)
mlv("beta", 23, 4)
```

```
mlv(x, method = "naive", bw = 1/3)
mlv(x, method = "grenander", p = 4)
mlv(x, method = "parzen", kernel = "gaussian")
mlv(x, method = "tsybakov", kernel = "gaussian")

M <- mlv(x, method = "kernel", boot = TRUE, R = 150)
print(M)
plot(M)
print(mean(M[["boot.M"]]))
```

Если выбрать тип ядра = "uniform", то запуск функций Парзена будет выглядеть следующим образом:

```
mean(naive(x, bw = 1/4))
M <- mlv(x, method = "naive", bw = 1/4)
print(M)
plot(M, xlim = c(0,2))
```

Задания к лабораторной работе

Вариант 1

1. Определим уровень финансовой устойчивости предприятия, как составной части общей устойчивости предприятия (при этом соблюдаются сбалансированность финансовых потоков, наличие средств, позволяющих организации поддерживать свою деятельность в течение определенного периода времени, в том числе обслуживая полученные кредиты и производя продукцию).
   При этом необходимо использовать показатели финансовой устойчивости: коэффициент автономии, Коэффициент финансового левериджа, Коэффициент обеспеченности собственными оборотными средствами, Коэффициент покрытия инвестиций, Коэффициент маневренности собственного капитала, Коэффициент мобильности имущества, Коэффициент мобильности оборотных средств, Коэффициент обеспеченности запасов, Коэффициент краткосрочной задолженности, коэффициент текущей ликвидности и коэффициент

быстрой ликвидности, коэффициент капитализации, коэффициент покрытия активов, коэффициент покрытия инвестиций.

Исходные данные следует организовать в виде таблицы:

RES	K1	K2	K3	K4	K5	…	KN
T	0.01						
F	0.022						
…	…	…	…	…	…	…	…
T	0.451						

В первом столбце заносится значение бинарной классификации – финансовой состояние предприятия - устойчиво или нет (True, False). В данном варианте признаками финансовой устойчивости будут: коэффициент автономии, Коэффициент финансового левериджа, коэффициент текущей ликвидности, коэффициент капитализации. Строк в таблице должно быть 110 (каждая строка - сведения по проверенному предприятию).

2. Используем метод К-ближайших соседей и метод Парзена. Сформировать обучающие и тестовые выборки. Полученные результаты визуализировать и сравнить. Представить значения параметров с минимальным уровнем ошибки. Для метода К соседей параметр К =18, для метода Парзена тип ядра выбрать "triangular", "uniform", а параметр optim. method ="Nelder-Mead", "BFGS". Проверить точность прогнозов.

Вариант 2

1. Определим уровень финансовой устойчивости предприятия, как составной части общей устойчивости предприятия (при этом соблюдаются сбалансированность финансовых потоков, наличие средств, позволяющих организации поддерживать свою деятельность в течение определенного периода времени, в том числе обслуживая полученные кредиты и производя продукцию).

Исходные данные следует организовать в виде таблицы:

RES	K1	K2	K3	K4	K5	…	KN
T	0.05						0.509
F	0.027			0.611			
…	…	…	…	…	…	…	…
T	0.458						

В первом столбце заносится значение бинарной классификации – финансовой состояние предприятия - устойчиво или нет (True, False). В данном варианте признаками финансовой устойчивости будут: Коэффициент обеспеченности собственными оборотными средствами, Коэффициент покрытия инвестиций, Коэффициент маневренности собственного капитала, Коэффициент мобильности имущества. Строк в таблице должно быть 140 (каждая строка - сведения по проверенному предприятию).

2. Используем метод К-ближайших соседей и метод Парзена. Сформировать обучающие и тестовые выборки. Полученные результаты визуализировать и сравнить. Представить значения параметров с минимальным уровнем ошибки. Для метода К соседей параметр К =20, для метода Парзена тип ядра выбрать "epanechnikov", "uniform", а параметр optim. method ="SANN", "BFGS". Проверить точность прогнозов.

Вариант 3

1. Определим уровень финансовой устойчивости предприятия, как составной части общей устойчивости предприятия (при этом соблюдаются сбалансированность финансовых потоков, наличие средств, позволяющих организации поддерживать свою деятельность в течение определенного периода времени, в том числе обслуживая полученные кредиты и производя продукцию).

Исходные данные следует организовать в виде таблицы:

RES	K1	K2	K3	K4	K5	…	KN
T	0.05						0.539
F	0.027	0.240		0.671			
…	…	…	…	…	…	…	…
T	0.458			0.683			

В первом столбце заносится значение бинарной классификации – финансовой состояние предприятия - устойчиво или нет (True, False). В данном варианте признаками финансовой устойчивости будут: Коэффициент краткосрочной задолженности, коэффициент текущей ликвидности и коэффициент быстрой ликвидности, коэффициент капитализации, коэффициент покрытия активов, коэффи-

циент покрытия инвестиций. Строк в таблице должно быть 130 (каждая строка - сведения по проверенному предприятию).

2. Используем метод К-ближайших соседей и метод Парзена. Сформировать обучающие и тестовые выборки. Полученные результаты визуализировать и сравнить. Представить значения параметров с минимальным уровнем ошибки. Для метода К соседей параметр К = 16, для метода Парзена тип ядра выбрать "gaussian", "optcosine" а параметр optim. method ="CG", "L-BFGS-B". Проверить точность прогнозов.

Вариант 4

1. Определим уровень финансовой устойчивости предприятия, как составной части общей устойчивости предприятия (при этом соблюдаются сбалансированность финансовых потоков, наличие средств, позволяющих организации поддерживать свою деятельность в течение определенного периода времени, в том числе обслуживая полученные кредиты и производя продукцию).

Исходные данные следует организовать в виде таблицы:

RES	K1	K2	K3	K4	K5	…	KN
T	0.252	0.301					0.539
F	0.327	0.240		0.671			
…	…	…	…	…	…	…	…
T	0.458			0.683			

В первом столбце заносится значение бинарной классификации – финансовой состояние предприятия - устойчиво или нет (True, False). В данном варианте признаками финансовой устойчивости будут: Коэффициент мобильности имущества, Коэффициент мобильности оборотных средств, Коэффициент обеспеченности запасов, Коэффициент краткосрочной задолженности. Строк в таблице должно быть 120 (каждая строка - сведения по проверенному предприятию).

2. Используем метод К-ближайших соседей и метод Парзена. Сформировать обучающие и тестовые выборки. Полученные результаты визуализировать и сравнить. Представить значения параметров с минимальным уровнем ошибки. Для метода К соседей параметр К

= 22, для метода Парзена тип ядра выбрать "rectangular", "triangular" а параметр optim. method = "Brent", "Nelder-Mead". Проверить точность прогнозов.

Вариант 5

1. Определим уровень финансовой устойчивости предприятия, как составной части общей устойчивости предприятия (при этом соблюдаются сбалансированность финансовых потоков, наличие средств, позволяющих организации поддерживать свою деятельность в течение определенного периода времени, в том числе обслуживая полученные кредиты и производя продукцию).

   Исходные данные следует организовать в виде таблицы:

RES	K1	K2	K3	K4	K5	…	KN
T	0.252	0.301					0.539
F	0.327	0.240		0.695	0.671		
…	…	…	…	…	…	…	…
T	0.458			0.683			

   В первом столбце заносится значение бинарной классификации – финансовой состояние предприятия - устойчиво или нет (True, False). В данном варианте признаками финансовой устойчивости будут: коэффициент автономии, коэффициент покрытия инвестиций, коэффициент быстрой ликвидности, Коэффициент мобильности имущества. Строк в таблице должно быть 127 (каждая строка - сведения по проверенному предприятию).

2. Используем метод К-ближайших соседей и метод Парзена. Сформировать обучающие и тестовые выборки. Полученные результаты визуализировать и сравнить. Представить значения параметров с минимальным уровнем ошибки. Для метода К соседей параметр К = 24, для метода Парзена тип ядра выбрать "biweight", "cosine", "eddy" а параметр optim. method = "CG". Проверить точность прогнозов.

Вариант 6

Характеристики грибов	К1	К2	К3	К4
Калории	34	9	29	7
Белки	3.7	0.24	3.1	0.12
Углеводы	1.1	0.32	0.97	0.04
Диаметр шляпки	25	21	14	13
Диаметр ножки	10	11	8	9
Длина ножки	20	18	13	12
Цвет шляпки	коричневый	Светло-коричневый	Рыжий	Грязно-рыжий
Цвет ножки	Белый	Грязно-желтый	Рыжий	Рыжий
Вид ножки	Трубчатая сплошная	Трубчатая	Трубчатая	Трубчатый
Содержание железа	5.2	0.88	4.4	0.53
Содержание кобальта	6	0.7	5.1	0.72
Содержание калия	4.6	0.5	3.8	0.84
Тип ножки	Гладкая	Неровная	Гладкая	Шероховатая
Слой под шляпкой	трубчатый	трубчатый	Трубчатый	Трубчатый
Жиры	1.7	0.7	1.2	0.2
Тип среза шляпки	Светлый	Грязно- розовый	Светло – желтый	Грязно-желтый
Тип среза ножки	Светлый	Грязно-розовый	Светло – желтый	Грязно-желтый
Название	Белый боровик	Желчный гриб (ложный белый)	Лисичка	Ложная лисичка

Используем классификацию грибов по вкусовым качествам:
1 категория: белый гриб, польский гриб, рыжик, груздь

2 категория: подосиновик, подберезовик, масленок, шампиньон, волнушка, белянка, лисичка
3 категория: моховик, козляк, сморчок
4 категория: горькушка, скрипица, рядовка

По строению шляпки можно выделить грибы: трубчатые, пластинчатые, сумчатые. К трубчатым можно отнести грибы из 1 и 2 вкусовой категории. К пластинчатым можно отнести грибы из 3 и 4 вкусовой категории. К сумчатым грибам относят сморчки, строчки и трюфельные грибы (т.е. из 3 вкусовой категории)

1. Необходимо разработать метрический классификатор, который функционирует на предметной области исследования лесной растительности. Различные виды и категории грибов часто употребляются в пищу в столовых, ресторанах, в домашних условиях. При этом необходимо использовать характеристические показатели качества грибов (см. сводную таблицу). В данном варианте надо использовать признаки грибов: содержание калорий, жиров, углеводов, белков, тип слоя под шляпкой, длина ножки и шляпки, цвет шляпки.
Исходные данные следует организовать в виде таблицы:

RES	K1	K2	K3	K4	K5	…	KN
T	0.01			11			
F	0.028		5.1		4.3		
…	…	…	…	…	…	…	…
T	0.451			12			

В первом столбце заносится значение бинарной классификации – съедобный или не съедобный гриб (True, False). Строк в таблице должно быть 80 (каждая строка - сведения по выбранному грибу).

2. Используем метод К-ближайших соседей и метод Парзена. Сформировать обучающие и тестовые выборки. Полученные результаты визуализировать и сравнить. Представить значения параметров с минимальным уровнем ошибки. Для метода К соседей параметр К = 38, для метода Парзена тип ядра выбрать "eddy", "epanechnikov", а параметр optim. method ="Nelder-Mead", "BFGS". Проверить точность прогнозов.

Вариант 7

1. Необходимо разработать метрический классификатор, который функционирует на предметной области исследования лесной растительности. Различные виды и категории грибов часто употребляются в пищу в столовых, ресторанах, в домашних условиях. При этом необходимо использовать характеристические показатели качества грибов (см. сводную таблицу). В данном варианте надо использовать признаки грибов: содержание кобальта, железа, калия, диаметр ножки и шляпки и типы срезов ножки и шляпки.

Исходные данные следует организовать в виде таблицы:

RES	K1	K2	K3	K4	K5	…	KN
T	0.01						
F	0.028		5.1		4.3		
…	…	…	…	…	…	…	…
T	0.451						

В первом столбце заносится значение бинарной классификации – съедобный или не съедобный гриб (True, False). Строк в таблице должно быть 70 (каждая строка - сведения по выбранному грибу).

2. Используем метод К-ближайших соседей и метод Парзена. Сформировать обучающие и тестовые выборки. Полученные результаты визуализировать и сравнить. Представить значения параметров с минимальным уровнем ошибки. Для метода К соседей параметр К = 36, для метода Парзена тип ядра выбрать "biweight", "cosine", а параметр optim. method ="CG", "L-BFGS-B". Проверить точность прогнозов.

Вариант 8

1. Необходимо разработать метрический классификатор, который функционирует на предметной области исследования лесной растительности. Различные виды и категории грибов часто употребляются в пищу в столовых, рестора-

нах, в домашних условиях. При этом необходимо использовать характеристические показатели качества грибов (см. сводную таблицу). В данном варианте надо использовать признаки грибов: диаметр ножки и шляпки, длина ножки и шляпки, тип слоя под шляпкой, содержание кобальта.

Исходные данные следует организовать в виде таблицы:

RES	K1	K2	K3	K4	K5	…	KN
T	0.01						
F	0.028		5.1		4.3		
…	…	…	…	…	…	…	…
T	0.451						

В первом столбце заносится значение бинарной классификации – гриб первой или второй категории (True, False). Строк в таблице должно быть 75 (каждая строка - сведения по выбранному грибу).

2. Используем метод К-ближайших соседей и метод Парзена. Сформировать обучающие и тестовые выборки. Полученные результаты визуализировать и сравнить. Представить значения параметров с минимальным уровнем ошибки. Для метода К соседей параметр К = 34, для метода Парзена тип ядра выбрать "gaussian", "optcosine", а параметр optim. method ="SANN", "Brent". Проверить точность прогнозов.

Вариант 9

1. Необходимо разработать метрический классификатор, который функционирует на предметной области исследования лесной растительности. Различные виды и категории грибов часто употребляются в пищу в столовых, ресторанах, в домашних условиях. При этом необходимо использовать характеристические показатели качества грибов (см. сводную таблицу). В данном варианте надо использовать признаки грибов: диаметр ножки и шляпки, цвет ножки и шляпки, тип слоя под шляпкой, содержание железа, жиров и углеводов.

Исходные данные следует организовать в виде таблицы:

RES	K1	K2	K3	K4	K5	…	KN
T	0.01						
F	0.028		5.1		4.3		
…	…	…	…	…	…	…	…
T	0.451						

В первом столбце заносится значение бинарной классификации – гриб сумчатый или пластинчатый (True, False). Строк в таблице должно быть 85 (каждая строка - сведения по выбранному грибу).

2. Используем метод К-ближайших соседей и метод Парзена. Сформировать обучающие и тестовые выборки. Полученные результаты визуализировать и сравнить. Представить значения параметров с минимальным уровнем ошибки. Для метода К соседей параметр К = 29, для метода Парзена тип ядра выбрать "rectangular", "triangular", "uniform", а параметр optim. method ="BFGS". Проверить точность прогнозов.

Вариант 10

1. Необходимо разработать метрический классификатор, который функционирует на предметной области исследования лесной растительности. Различные виды и категории грибов часто употребляются в пищу в столовых, ресторанах, в домашних условиях. При этом необходимо использовать характеристические показатели качества грибов (см. сводную таблицу). В данном варианте надо использовать признаки грибов: содержание белков, жира, кобальта, диаметр шляпки, длина ножки, тип среза шляпки, цвет ножки.

Исходные данные следует организовать в виде таблицы:

RES	K1	K2	K3	K4	K5	…	KN
T	0.01						
F	0.028		5.1		4.3	8.1	
…	…	…	…	…	…	…	…
T	0.451						

В первом столбце заносится значение бинарной классификации – гриб трубчатый или пластинчатый (True, False). Строк в таблице должно быть 72 (каждая строка - сведения по выбранному грибу).

2. Используем метод К-ближайших соседей и метод Парзена. Сформировать обучающие и тестовые выборки. Полученные результаты визуализировать и сравнить. Представить значения параметров с минимальным уровнем ошибки. Для метода К соседей параметр К = 33, для метода Парзена тип ядра выбрать , "gaussian", "triangular", "uniform", а параметр optim. method ="BFGS", "SANN". Проверить точность прогнозов.

Вариант 11

Больные увеличением щитовидной железы общим числом 50 человек были разделены на две группы.

Группа 1. Лечение оказалось успешным; пациент здоров.

Группа 2. Лечение безуспешно, состояние больного осталось без изменения.

По результатам обследования 50 пациентов имеются следующие измерения:

y6 – йод, регистрируемый через 2 часа после принятия испытательной дозы;

y9 – йод, регистрируемый через 16 часов после принятия испытательной дозы;

y10 – содержание в крови белковосвязанного йода ($PB^{131}J$) через 32 часа;

kl – номер группы.

Таблица

№	Kl	y6	y9	y10
1	1	14.4	25.1	0.20
2	1	20.1	40.1	0.11
3	1	24.1	32.1	0.17
4	1	11.1	16.9	0.12
5	1	16.3	32.1	0.36

6	1	40.5	64.4	0.21
7	1	52.7	50.0	0.53
8	1	20.8	22.3	0.13
9	1	14.0	3.1	0.18
10	1	27.0	41.7	0.19
11	1	44.3	63.8	0.22
12	1	47.5	50.1	0.29
13	1	54.0	57.0	0.19
14	1	16.1	20.6	0.22
15	1	57.5	74.5	0.49
16	1	37.8	63.0	0.32
17	2	55.8	48.0	2.74
18	2	75.0	60.0	1.37
19	2	72.0	65.0	0.70
20	2	70.6	45.0	1.40
21	2	24.1	45.0	0.22
22	2	33.2	55.0	0.01
23	2	30.4	44.6	0.09
	…	…		
50	2		61.1	

Необходимо разработать метрический классификатор, который функционирует на предметной области исследования медицинских анализов. Три характеристических показателя надо взять из таблицы. Используем метод К-ближайших соседей и метод Парзена. Сформировать обучающие и тестовые выборки. Полученные результаты визуализировать и сравнить. Представить значения параметров с минимальным уровнем ошибки. Для метода К соседей параметр К = 11, для метода Парзена тип ядра выбрать , "gaussian", , "uniform", а параметр optim. method = "SANN". Проверить точность прогнозов.

Вариант 12

Больные увеличением щитовидной железы общим числом 55 человек были разделены на две группы.

Группа 1. Лечение оказалось успешным; пациент здоров.

Группа 2. Лечение безуспешно, состояние больного осталось без изменения.

По результатам обследования 55 пациентов имеются следующие измерения:

Y3 – йод, регистрируемый через 1 час после принятия испытательной дозы;

Y5 – йод, регистрируемый через 5 часов после принятия испытательной дозы;

Y7 – содержание в крови белковосвязанного йода ($PB^{131}J$) через 17 часа;

Y8 – йод, регистрируемый через 8 часов после принятия испытательной дозы;

kl – номер группы.

Таблица

№	Kl	Y3	Y5	Y7	Y8
1	1	14.4	25.1	0.25	32
2	1	20.1	40.1	0.17	54
3	1	24.1	32.1	0.19	27
4	1	13.1	16.9	0.14	11
5	1	17.3	32.1	0.39	14
6	1	40.5	64.4	0.22	51
7	1	52.7	50.0	0.56	7
8	1	20.8	22.3	0.17	9
9	1	14.0	3.1	0.19	10
10	1	27.0	41.7	0.13	21
11	1	44.3	63.8	0.26	43
12	1	47.5	50.1	0.21	56
13	1	54.0	57.0	0.17	12
14	1	16.1	20.6	0.29	34
15	1	87.5	74.5	0.48	12
16	1	37.8	63.4	0.31	17
17	2	35.8	48.0	2.74	14
18	2	65.0	60.0	1.37	17
19	2	62.0	65.0	0.70	23
20	2	71.6	45.0	1.40	10
21	2	24.1	45.0	0.22	9
22	2	37.2	55.0	0.01	31
23	2	35.4	44.6	0.09	65

	…	…			28
55	2		61.1		42

Необходимо разработать метрический классификатор, который функционирует на предметной области исследования медицинских анализов. Четыре характеристических показателя надо взять из таблицы. Используем метод К-ближайших соседей и метод Парзена. Сформировать обучающие и тестовые выборки. Полученные результаты визуализировать и сравнить. Представить значения параметров с минимальным уровнем ошибки. Для метода К соседей параметр К = 16, для метода Парзена тип ядра выбрать , "cosine", , "uniform", а параметр optim. method = "CG", "L-BFGS-B". Проверить точность прогнозов.

Вариант 13

Больные увеличением щитовидной железы общим числом 48 человек были разделены на две группы.

Группа 1. Лечение оказалось успешным; пациент здоров.

Группа 2. Лечение успешно на 85 %, возможен рецидив.

По результатам обследования 48 пациентов имеются следующие измерения:

Y3 – йод, регистрируемый через 2 часа после принятия испытательной дозы;

Y5 – йод, регистрируемый через 10 часов после принятия испытательной дозы;

Y7 – содержание в крови белковосвязанного йода ($PB^{131}J$) через 19 часов;

Y8 – йод, регистрируемый через 21 час после принятия испытательной дозы;

Y9 – содержание в крови белковосвязанного йода ($PB^{131}J$) через 26 часов;

kl – номер группы.

Таблица

№	Kl	Y3	Y5	Y7	Y8	Y9
1	1	14.4	25.1	0.25	32	1.77
2	1	20.1	40.1	0.17	54	2.86
3	1	24.1	32.1	0.19	27	3.11

4	1	13.1	16.9	0.14	11	1.48
5	1	17.3	32.1	0.39	14	12
6	1	40.5	64.4	0.22	51	40.3
7	1	52.7	50.0	0.56	7	1.97
8	1	20.8	22.3	0.17	9	13.86
9	1	14.0	3.1	0.19	10	15.23
10	1	27.0	41.7	0.13	21	24.08
11	1	44.3	63.8	0.26	43	…
12	1	47.5	50.1	0.21	56	
13	1	54.0	57.0	0.17	12	…
14	1	16.1	20.6	0.29	34	
15	1	87.5	74.5	0.48	12	
16	1	37.8	63.4	0.31	17	
17	2	35.8	48.0	2.74	14	
18	2	65.0	60.0	1.37	17	
19	2	62.0	65.0	0.70	23	
20	2	71.6	45.0	1.40	10	
21	2	24.1	45.0	0.22	9	
22	2	37.2	55.0	0.01	31	
23	2	35.4	44.6	0.09	65	
	…	…			28	
48	2		61.1		42	0.98

Необходимо разработать метрический классификатор, который функционирует на предметной области исследования медицинских анализов. Пять характеристических показателей надо взять из таблицы. Используем метод К-ближайших соседей и метод Парзена. Сформировать обучающие и тестовые выборки. Полученные результаты визуализировать и сравнить. Представить значения параметров с минимальным уровнем ошибки. Для метода К соседей параметр К = 44, для метода Парзена тип ядра выбрать , "epanechnikov", , "uniform", а параметр optim. method = "Brent". Проверить точность прогнозов.

Вариант 14

Больные увеличением щитовидной железы общим числом 70 человек были разделены на две группы.

Группа 1. Лечение оказалось успешным; пациент здоров.

Группа 2. Лечение успешно на 75 %, возможен рецидив.

По результатам обследования 70 пациентов имеются следующие измерения:

Y3 – йод, регистрируемый через 3 часа после принятия испытательной дозы;

Y5 – йод, регистрируемый через 12 часов после принятия испытательной дозы;

Y7 – содержание в крови белковосвязанного йода ($PB^{131}J$) через 20 часов;

Y9 – содержание в крови белковосвязанного йода ($PB^{131}J$) через 33 часа;

kl – номер группы.

Таблица

№	Kl	Y3	Y5	Y7	Y9
1	1	14.4	25.1	0.25	1.77
2	1	20.1	40.1	0.17	2.86
3	1	24.1	32.1	0.19	3.11
4	1	13.1	16.9	0.14	1.48
5	1	17.3	32.1	0.39	12
6	1	40.5	64.4	0.22	40.3
7	1	52.7	50.0	0.56	1.97
8	1	20.8	22.3	0.17	13.86
9	1	14.0	3.1	0.19	15.23
10	1	27.0	41.7	0.13	24.08
11	1	44.3	63.8	0.26	…
12	1	47.5	50.1	0.21	
13	1	54.0	57.0	0.17	…
14	1	16.1	20.6	0.29	
15	1	87.5	74.5	0.48	
16	1	37.8	63.4	0.31	
17	2	35.8	48.0	2.74	
18	2	65.0	60.0	1.37	
19	2	62.0	65.0	0.70	
20	2	71.6	45.0	1.40	

21	2	24.1	45.0	0.22	
22	2	37.2	55.0	0.01	
23	2	35.4	44.6	0.09	
	…	…			
70	2		61.1		0.98

Необходимо разработать метрический классификатор, который функционирует на предметной области исследования медицинских анализов. Четыре характеристических показателя надо взять из таблицы. Используем метод К-ближайших соседей и метод Парзена. Сформировать обучающие и тестовые выборки. Полученные результаты визуализировать и сравнить. Представить значения параметров с минимальным уровнем ошибки. Для метода К соседей параметр К = 37, для метода Парзена тип ядра выбрать , "epanechnikov", "eddy" , а параметр optim. method = "Nelder-Mead". Проверить точность прогнозов.

Вариант 15

Больные увеличением щитовидной железы общим числом 63 человека были разделены на две группы.

Группа 1. Лечение оказалось успешным; пациент здоров.

Группа 2. Лечение успешно на 90 %, возможен рецидив.

По результатам обследования 63 пациентов имеются следующие измерения:

Y3 – йод, регистрируемый через 4 часа после принятия испытательной дозы;

Y5 – йод, регистрируемый через 16 часов после принятия испытательной дозы;

Y7 – содержание в крови белковосвязанного йода ($PB^{131}J$) через 19 часов;

Y8 – йод, регистрируемый через 31 час после принятия испытательной дозы;

Y9 – содержание в крови белковосвязанного йода ($PB^{131}J$) через 46 часов;

kl – номер группы.

Таблица

№	Kl	Y3	Y5	Y7	Y8	Y9
1	1	14.4	25.1	0.25	32	1.77
2	1	20.1	40.1	0.17	54	2.86
3	1	24.1	32.1	0.19	27	3.11
4	1	13.1	16.9	0.14	11	1.48
5	1	17.3	32.1	0.39	14	12
6	1	40.5	64.4	0.22	51	40.3
7	1	52.7	50.0	0.56	7	1.97
8	1	20.8	22.3	0.17	9	13.86
9	1	14.0	3.1	0.19	10	15.23
10	1	27.0	41.7	0.13	21	24.08
11	1	44.3	63.8	0.26	43	…
12	1	47.5	50.1	0.21	56	
13	1	54.0	57.0	0.17	12	…
14	1	16.1	20.6	0.29	34	
15	1	87.5	74.5	0.48	12	32.08
16	1	37.8	63.4	0.31	17	
17	2	35.8	48.0	2.74	14	
18	2	65.0	60.0	1.37	17	
19	2	62.0	65.0	0.70	23	
20	2	71.6	45.0	1.40	10	21.75
21	2	24.1	45.0	0.22	9	
22	2	37.2	55.0	0.01	31	22.56
23	2	35.4	44.6	0.09	65	
	…	…			28	
63	2		61.1		42	0.98

Необходимо разработать метрический классификатор, который функционирует на предметной области исследования медицинских анализов. Пять характеристических показателей надо взять из таблицы. Используем метод К-ближайших соседей и метод Парзена.

Сформировать обучающие и тестовые выборки. Полученные результаты визуализировать и сравнить. Представить значения параметров с минимальным уровнем ошибки. Для метода К соседей параметр К = 8, для метода Парзена тип ядра выбрать"gaussian", "optcosine", а параметр optim. method = "Brent", "CG". Проверить точность прогнозов.