**Контрольная работа по методам классификации**

Контрольная работа состоит из нескольких заданий. От варианта студента зависят входные данные, для которых будут решаться задания, а также некоторые особенности выполнения заданий.

Всем студентам предоставляются наборы данных, в зависимости от варианта, для этих данных необходимо:

1. Построить классификатор на основе метода ближайших k соседей и определить класс тестового значения (описание метода можно найти по ссылке <http://www.machinelearning.ru/wiki/index.php?title=%D0%9C%D0%B5%D1%82%D0%BE%D0%B4_%D0%B1%D0%BB%D0%B8%D0%B6%D0%B0%D0%B9%D1%88%D0%B5%D0%B3%D0%BE_%D1%81%D0%BE%D1%81%D0%B5%D0%B4%D0%B0> ). От варианта зависят весовая функция и значение k.
2. Построить классификатор на основе алгоритма CART построения дерева принятия решений (методические указания и пример будут приведены ниже). От варианта зависит выбор обучающей выборки. Дополнительная теория по решающим деревьям: <https://habrahabr.ru/company/ods/blog/322534/>
<https://habrahabr.ru/post/116385/>

**Методические указания по построению дерева классификации**

Пусть дана обучающая выборка из n объектов, каждый из которых описывается m признаками $\{\left(X\_{11},X\_{11},…,X\_{1m},Y\_{1}\right),…\left(X\_{n1},X\_{n2},…,X\_{nm},Y\_{n}\right)\}$. При этом $X\_{i}\in X$ - вектор признаков i-го объекта, а $Y\_{i}$ – класс i-го объекта. Будем считать для простоты, что все признаки являются числовыми характеристиками.

Опишем алгоритм построения дерева:

Входные данные: указатель на вершину дерева, множество объектов $(X,Y)'$

*Если все объекты* $(X,Y)'$ *относятся к одному классу, то вершина становится терминальной и дальше не обрабатывается, иначе:*

1. *Для каждого признака* $i\in (1..m)$ *найдём такое значение* $S\_{i}$*, что при разделении всех объектов* $(X,Y)'$ *на два подмножества* $(X,Y)\_{1}^{'},(X,Y)\_{2}^{'}$ *по правилу* $X\_{1i}^{'}\leq S\_{i}, X\_{2i}^{'}>S\_{i}$ *(то есть, объекты разделяются на 2 подмножества по i-му признаку), в одном из подмножеств будет максимально выделен один класс. Выберем признак, который даёт наилучшее разбиение.*
2. *Подмножество* $\left(X,Y\right)\_{1}^{'}$ *поместим в левое поддерево, а* $\left(X,Y\right)\_{2}^{'}$ *- в правое. Применим рекурсивно этот же алгоритм для каждого из поддеревьев.*

Рассмотрим этот алгоритм на примере:

Дано множество объектов {(1,1,1),(1,2,1),(2,1,1),(2,2,1),(3,3,2),(3,4,2),(2,3,2),(2,4,2),(3,2,2)}



В каждом объекте первые два числа – признаки, координаты точки, третье число – класс, к которому точка относится.

Построим для данного примера дерево:

1. В корневую вершину передаётся всё множество объектов. Рассмотрим признаки и найдём для них значения чисел $S\_{i}$. $S\_{1}=2.5, S\_{2}=2.5$. Изобразим эти разделяющие прямые на рисунке



Обе прямые разделяют по классам одинаково, но выберем $S\_{2}$, так как эта прямая делит на более близкие по количеству подмножества.

{(1,1,1),(1,2,1),(2,1,1),(2,2,1),(3,3,2),(3,4,2),(2,3,2),(2,4,2),(3,2,2)}

$$y\leq 2.5$$

{(1,1,1),(1,2,1),(2,1,1),(2,2,1), (3,2,2)}

В правом поддереве все объекты относятся к одному классу, поэтому вершина становится терминальной и будет относить любую попавшую в неё точку ко 2му классу (синие).

К левому поддереву необходимо применить аналогичный алгоритм. Найдём значения $S\_{i}$ для каждого признака. $S\_{1}=2.5, S\_{2}=1.5$



Лучшее разделение по классам нам даёт первый признак ($S\_{1}=2.5)$. Следовательно, построим дерево следующим образом: В левом поддереве у нас окажутся только объекты из 1го класса, в правом поддереве только из второго. В итоге получим готовое дерево принятия решений для рассмотренной обучающей выборки:

{(1,1,1),(1,2,1),(2,1,1),(2,2,1),(3,3,2),(3,4,2),(2,3,2),(2,4,2),(3,2,2)}

$$y\leq 2.5$$

{(1,1,1),(1,2,1),(2,1,1),(2,2,1), (3,2,2)}

$$x\leq 2.5$$

**Варианты для контрольной работы:**

Для всех вариантов в методе ближайших соседей значение k (количество соседей) необходимо брать на 1 больше, чем количество классов в обучающей последовательности. Варианты обучающих последовательностей и тестового объекта:

1. (X,Y)={(1,8,1), (1,3,1), (3,5,1), (1,1,1), (2,7,1), (3,8,1), (2,4,1), (8,7,2), (11,12,2), (12,14,2), (8,13,2)}: тестовый объект x’=(5,8)
2. (X,Y)={ (2,7,1), (6,6,1), (8,6,1), (7,5,1), (5,9,1), (9,9,2), (11,2,2), (6,4,2), (10,9,2), (8,6,3), (1,10,3), (9,7,3), (5,13,3), (2,14,3)}: тестовый объект x’=(10,8)
3. (X,Y)={ (1,7,1), (3,2,1), (6,8,1), (4,7,1), (9,8,1), (4,5,1), (1,2,1), (14,10,2), (8,12,2), (14,12,2), (11,10,2), (13,8,2), (13,6,2)}: тестовый объект x’=(6,7)
4. (X,Y)={ (7,9,1), (2,5,1), (5,6,1), (8,6,1), (7,6,1), (7,9,2), (14,7,2), (14,2,2), (6,7,2), (10,3,2), (11,9,2), (9,1,2)}: тестовый объект x’=(12,12)
5. (X,Y)={ (2,9,1), (9,3,1), (3,9,1), (5,5,1), (8,9,1), (5,1,1), (6,4,1), (12,8,2), (12,14,2), (12,11,2), (11,6,2), (13,7,2)}: тестовый объект x’=(4,10)
6. (X,Y)={ (7,8,1), (6,7,1), (2,1,1), (2,4,1), (9,9,1), (8,4,1), (4,7,1), (11,13,2), (6,11,2), (14,8,2), (11,7,2)}: тестовый объект x’=(6,1)
7. (X,Y)={ (4,7,1), (4,3,1), (4,8,1), (8,6,2), (14,5,2), (9,4,2), (3,13,3), (8,10,3), (2,7,3)}: тестовый объект x’=(4,9)
8. (X,Y)={ (5,9,1), (2,9,1), (3,7,1), (8,8,2), (14,4,2), (10,1,2), (12,4,2), (7,7,2), (12,7,2), (9,13,3), (2,14,3), (1,7,3), (5,14,3), (6,6,3), (9,6,3)}: тестовый объект x’=(5,6)
9. (X,Y)={ (8,4,1), (9,4,1), (4,2,1), (4,1,1), (8,1,1), (11,7,2), (9,4,2), (12,5,2), (14,1,2), (6,12,3), (7,14,3), (9,9,3), (1,12,3), (4,13,3), (8,9,3)}: тестовый объект x’=(6,10)
10. (X,Y)={ (6,5,1), (1,9,1), (1,9,1), (1,6,1), (8,4,1), (14,11,2), (13,13,2), (6,7,2), (12,11,2), (13,9,2), (9,7,2)}: тестовый объект x’=(7,3)
11. (X,Y)={ (7,2,1), (8,1,1), (8,7,1), (8,2,1), (9,9,1), (6,8,1), (13,8,2), (6,1,2), (11,8,2), (4,12,3), (7,14,3), (1,8,3), (9,6,3)}: тестовый объект x’=(13,10)

Варианты весовых функций:

1. ![w(i,u) = [i\leq k]]() — метод  ближайших соседей;
2. ![w(i,u) = [i\leq k] q^i]() — метод  экспоненциально взвешенных ближайших соседей, где предполагается ; $q=0.75$
3.  — [метод парзеновского окна](http://www.machinelearning.ru/wiki/index.php?title=%D0%9C%D0%B5%D1%82%D0%BE%D0%B4_%D0%BF%D0%B0%D1%80%D0%B7%D0%B5%D0%BD%D0%BE%D0%B2%D1%81%D0%BA%D0%BE%D0%B3%D0%BE_%D0%BE%D0%BA%D0%BD%D0%B0) фиксированной ширины ; $h=0.1$
4.  — [метод парзеновского окна](http://www.machinelearning.ru/wiki/index.php?title=%D0%9C%D0%B5%D1%82%D0%BE%D0%B4_%D0%BF%D0%B0%D1%80%D0%B7%D0%B5%D0%BD%D0%BE%D0%B2%D1%81%D0%BA%D0%BE%D0%B3%D0%BE_%D0%BE%D0%BA%D0%BD%D0%B0) переменной ширины;

Здесь  — заданная неотрицательная монотонно невозрастающая функция на , $K\left(r\right)=e^{-r}$

Как определить свой вариант? У каждого студента есть некоторый номер N, который его уникально идентифицирует.

Вариант выборки для методы ближайших соседей определяется по формуле:

$$N\_{в}=(\left(N+13\right) mod 11)+1 $$

Вариант весовой функции определяется по формуле:

$$N\_{вф}=(\left(N+7\right) mod 4)+1$$

Вариант выборки для метода построения решающего дерева определяется по формуле:

$$N\_{вд}=\left(\left(N\*N+2\right) mod 11\right)+1$$

В отчёте по контрольной работе необходимо указать свой номер и результаты расчётов по указанным выше формулам. После чего уже описывать подробный ход решения. Для метода построения решающих деревьев необходимо изобразить графически получившееся в итоге дерево (можно как в примере).