Тема 1. Методы машинного обучения. Дерево решений. Random Forest

Машинное обучение – раздел искусственного интеллекта, изучающий методы построения алгоритмов, способных обучаться. Машинное обучение может быть *с учителем* и *без учителя*. К задачам, решаемым при помощи обучения с учителем, относятся задачи *классификации и регрессии*. К задачам, решаемым при помощи обучения без учителя, относится задача *кластеризации*. Обучение производится на выборке объектов, каждый из которых характеризуется набором признаков (числовых или категориальных). Для обучения с учителем необходима *обучающая выборка*, содержащая объекты с известным распределением по классам (классификация) или с известными значениями зависимой переменной (регрессия). Алгоритм обучается на этой выборке распознавать класс объекта (значение зависимой переменной), исходя из значений признаков. Построенную таким образом модель можно использовать для прогноза класса (значения зависимой переменной) по значениям признаков для новых объектов, класс которых (значение зависимой переменной) не известен. Однако перед использованием построенной модели для решения практических задач необходимо оценить её качество при помощи *тестовой выборки*. Тестовая выборка содержит объекты, не входящие в состав обучающей, для которых также известны признаки и распределение по классам (значениям зависимой переменной). С помощью тестовой выборки можно оценить точность прогноза принадлежности к классу (значения зависимой переменной). Если точность прогноза высокая, то модель можно использовать.

При обучении *без учителя* обучающая выборка отсутствует, а есть лишь выборка объектов с известными значениями признаков (принадлежность к классам не известна). Задача кластеризации заключается в том, чтобы разбить объекты на группы (кластеры) со сходными значениями признаков.

Рассмотрим вначале методы классификации, которые можно условно разделить на логические, геометрические и вероятностные. В рамках данного занятия рассмотрим методы, основанные на использовании логики: дерево решений и Random Forest («Случайный лес»).

Для построения соответствующих моделей нам понадобятся следующие пакеты: **party** – для построения дерева решений, **randomForest** – для построения Random Forest, **caret** – для разбиения данных на обучающую и тестовую выборки, и оценки точности прогноза.

Дерево решений

Построим классификационную модель при помощи дерева решений для данных по опухолям молочной железы (соответствующие данные находятся в файле «Breast cancer.txt» в папке занятия). Импортируем данные в файл:

```
cancer <- read.delim("Breast cancer.txt")
```

Переменная cancer представляет собой таблицу из 10 столбцов, где содержатся 9 целочисленных признаков и метка класса: злокачественная или доброкачественная опухоль.

```
> colnames(cancer)
 [1] "Clump.Thickness"              "Uniformity.of.Cell.Size"
 [3] "Uniformity.of.Cell.Shape"     "Marginal.Adhesion"
 [5] "Single.Epithelial.Cell.Size" "Bare.Nuclei"
 [7] "Bland.Chromatin"              "Normal.Nucleoli"
 [9] "Mitoses"                      "Class"
```

Таблица содержит данные по 699 образцам опухоли (699 строк).

Признаки представляют собой целые числа от 1 до 10. Эти числа отражают интенсивность признака. Важно отметить, что дерево решений и Random Forest могут строить классификационные и регрессионные модели, использую признаки любого типа: номинальные, порядковые (как в нашем случае), числовые и смешанные.

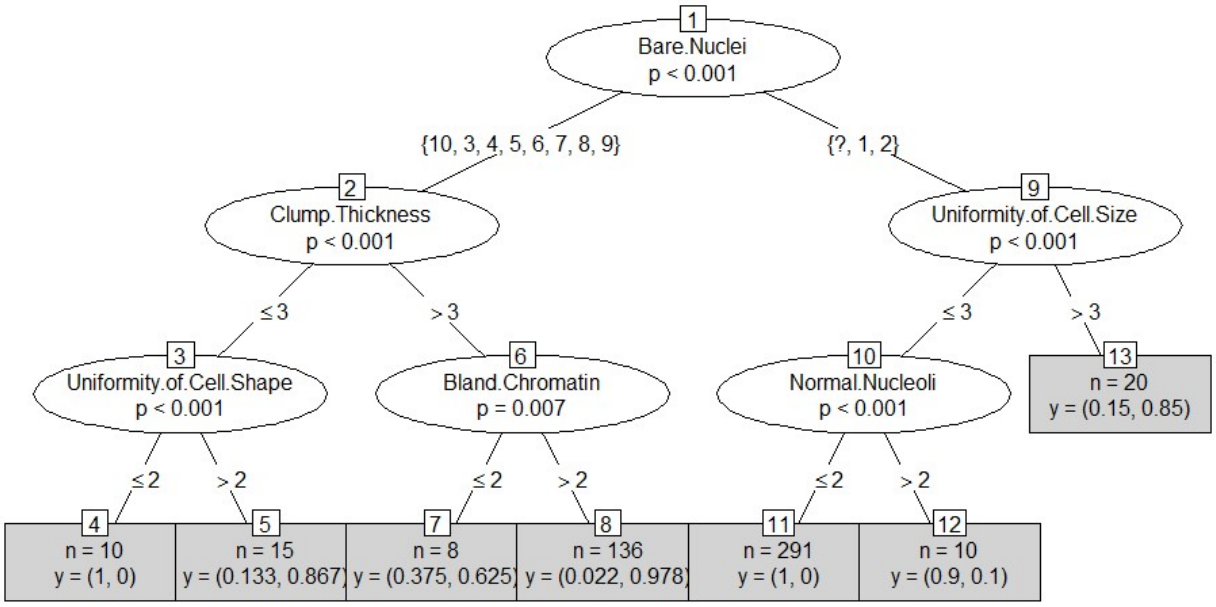
Разделим данные на обучающую (tr) и тестовую (test) выборки в соотношении 7:3 при помощи функции **createDataPartition** из пакета **caret**:

```
inTr <- createDataPartition(cancer$Class, p=0.7, list=FALSE)
tr <- cancer[inTr,]
test <- cancer[-inTr,]
```

Результатом выполнения данной функции являются индексы строк таблицы, которые будут отнесены к обучающей выборке (p=0.7 означает, что 70% строк пойдут в обучающую выборку). Подставляя эти индексы с положительным и отрицательным знаком в переменную cancer, можно получить таблицы с обучающей и тестовой выборками.

Построим дерево решений при помощи функции **ctree** из пакета **party** и визуализируем его при помощи функции **plot**:

```
cancer_ctree <- ctree(Class ~ ., data=tr)
plot(cancer_ctree, type="simple")
```



Наиболее важный признак находится в корне дерева. Как видно из рисунка, таким признаком для классификации опухолей на злокачественные и доброкачественные является Bare Nuclei (наличие «голых» ядер разрушенных клеток).

В листах дерева указаны количества образцов с соответствующей комбинацией значений признаков (n) и соотношения образцов разных классов (y) в листе. Во внутренних узлах также указаны значения p, которые рассчитываются на основе перестановочного теста для оценки связи двух категориальных переменных: известный класс объектов и принадлежность объекта к левому или правому дочернему узлу дерева. Величины 1-p используются для отбора признака, который лучше всего позволяет разделить объекты по классам в данном узле дерева. Если значение p больше 0.05 для всех признаков, то рост дерева в данном узле останавливается.

Выполним прогноз для тестовой выборки при помощи функции **predict**:

```
pred<-predict(cancer_ctree,newdata=test)
```

Переменная pred содержит метки класса, предсказанные при помощи построенного на обучающей выборке дерева решений. Рассчитаем показатели точности прогноза при помощи функции **confusionMatrix**:

```
> confusionMatrix(data=pred,reference=test$Class,positive="malignant")
Confusion Matrix and Statistics

           Reference
Prediction  benign malignant
  benign       127         1
  malignant     10        71

               Accuracy : 0.9474
                 95% CI : (0.9078, 0.9734)
```

```
    No Information Rate : 0.6555
    P-Value [Acc > NIR] : < 2e-16

                  Kappa : 0.8868
 Mcnemar's Test P-Value : 0.01586

            Sensitivity : 0.9861
            Specificity : 0.9270
         Pos Pred Value : 0.8765
         Neg Pred Value : 0.9922
             Prevalence : 0.3445
         Detection Rate : 0.3397
   Detection Prevalence : 0.3876
      Balanced Accuracy : 0.9566

       'Positive' Class : malignant
```

В начале выведенной в консоль информации представлена таблица сопряженности 2 x 2, описывающая, сколько образцов опухолей в тестовой выборке было классифицировано правильно и неправильно. Если один класс представить как положительный (в данном случае «malignant»), а второй – как отрицательный, то значения в ячейках таблицы можно обозначить как

	Reference	
Prediction	benign	malignant
benign	TN	FN
malignant	FP	TP

TP – истинно-положительные примеры

TN – истинно-отрицательные примеры

FP – ложноположительные примеры

FN- ложноотрицательные примеры

Соответствующие показатели точности прогноза можно рассчитать, исходя из этих значений:

Accuracy (точность) = (TP+TN)/(TP+TN+FP+FN)

Sensitivity (чувствительность) = TP/(TP+FN)

Specificity (специфичность) = TN/(TN+FP)

Prevalence (предсказательная ценность) = (TP+FN)/(TP+TN+FP+FN)

PPV (положительная предсказательная ценность) = TP/(TP+FP)

NPV (отрицательная предсказательная ценность) = TN/(TN+FN)

Detection Rate (частота выявления) = TP/(TP+TN+FP+FN)

Detection Prevalence (частота распространения) = (TP+FP)/(TP+TN+FP+FN)

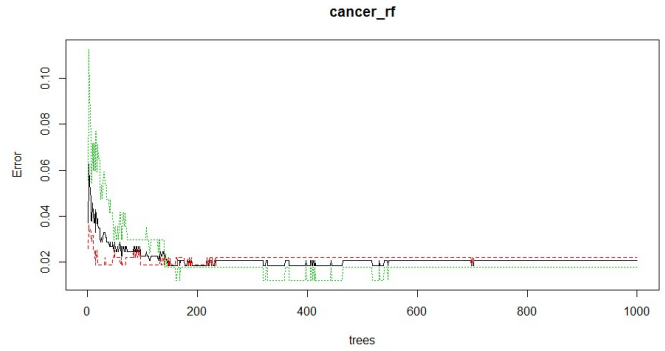
Balanced Accuracy (сбалансированная точность) = (sensitivity+specificity)/2

Высокие значения точности (0.947), чувствительности (0.986) и специфичности (0.927) свидетельствуют о применимости построенной модели для классификации образцов опухолей на злокачественные и доброкачественные.

Random Forest

Random Forest («случайный лес») представляет собой ансамбль деревьев, каждое из которых построено на выборке объектов, сгенерированной при помощи бутстреп-анализа из исходных данных, и части признаков, случайно отобранных из всех имеющихся. Построим классификационную модель при помощи функции **randomForest** пакета **randomForest** и изобразим на графике зависимость точности прогноза от числа построенных деревьев при помощи функции **plot**:

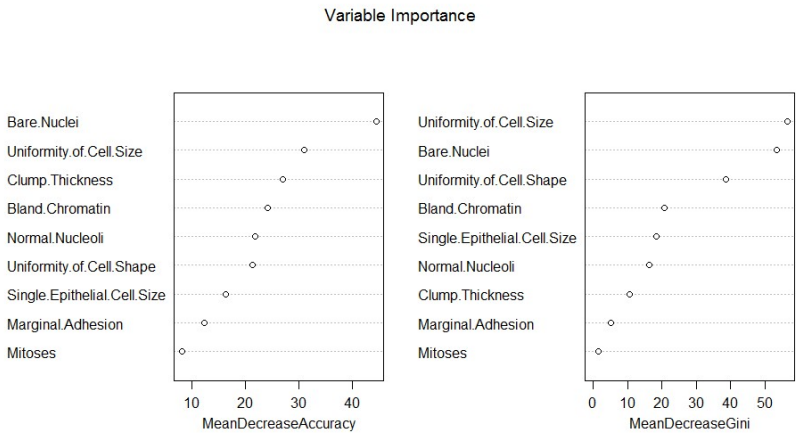
```
cancer_rf <- randomForest(Class ~ ., data=tr, ntree=1000, importance=T)
plot(cancer_rf)
```



На оси абсцисс графика показано количество деревьев, а на оси ординат – ошибка прогноза. Эта ошибка рассчитывается на основе прогноза, полученного с использованием только части деревьев, для тех объектов, которые не входили в соответствующие бутстреп-выборки. Такие образом, отбор параметров алгоритма, включая число деревьев, с использование полученных величин ошибки прогноза не приведет к переобучению. Как видно из графика точность перестает меняться при увеличении количества построенных деревьев больше 500. Это означает, что для построения модели достаточно 500 деревьев.

Параметр importance=TRUE показывает, что будет вычисляться вклад каждого признака в точность классификации. Вклад признака оценивается при помощи «перемешивания» его значений между объектами и расчета падения точности классификации. Точность классификации может быть оценена по соответствующей формуле, приведенной выше, или по индексу Джини. Выполнить соответствующую оценку можно при помощи функции **varImpPlot**:

```
varImpPlot(cancer_rf, main="Variable Importance")
```



Из графика видно, что наибольшее падение точности происходит при рандомизации переменных Bare Nuclei и Uniformity of cell size, что совпадает с результатами, полученными с помощью дерева решений. Соответствующие результаты, можно также сохранить в таблицу:

```
> var.imp <- data.frame(importance(cancer_rf, type=1))
> head(var.imp)
                             MeanDecreaseAccuracy
Clump.Thickness                          26.95132
Uniformity.of.Cell.Size                  31.01060
Uniformity.of.Cell.Shape                 21.36863
Marginal.Adhesion                        12.38887
Single.Epithelial.Cell.Size              16.42712
Bare.Nuclei                              44.38043
```

Посчитаем точность прогноза на тестовой выборке:

```
> pred<-predict(cancer_rf,newdata=test)
> confusionMatrix(data=pred,reference=test$Class,positive="malignant")
Confusion Matrix and Statistics

           Reference
Prediction  benign malignant
  benign       129         1
  malignant      8        71

               Accuracy : 0.9569
                 95% CI : (0.9198, 0.9801)
    No Information Rate : 0.6555
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.9068
 Mcnemar's Test P-Value : 0.0455

            Sensitivity : 0.9861
            Specificity : 0.9416
         Pos Pred Value : 0.8987
         Neg Pred Value : 0.9923
             Prevalence : 0.3445
         Detection Rate : 0.3397
   Detection Prevalence : 0.3780
      Balanced Accuracy : 0.9639

       'Positive' Class : malignant
```

Как видно из результатов, точность прогноза чуть выше, чем при использовании дерева решений.

Практическое задание

В таблице из файла Diabetes.txt представлены данные по пациентам, страдающим диабетом (переменная Class = Y), и пациентам без диабета (переменная Class = N). Каждый пациент охарактеризован 8 признаками, имеющими отношение к диабету. Постройте классификационные модели для диабета, используя эти признаки, с помощью двух методов: дерево решений и Random Forest. Какие из признаков наиболее важны для классификации? Посчитайте точность прогноза на тестовой выборке для двух методов. Какой метод более точен? Проверьте, меняется ли точность прогноза при использовании Random Forest, если строить разное число деревьев.

Вопросы

1. Какие существуют типы машинного обучения?
2. Что представляет собой дерево решений?
3. Что представляет собой метод Random Forest?