WaveAccess
We Make IT Easy

Машинное обучение на платформе .NET

Неволин Роман

roman.nevolin@waveaccess.ru

nevoroman

nevoroman/ml-dotnet

Вступление

Вступление : репозиторий доклада

Все, что есть в докладе (и немного сверху) - материалы, демки, бенчмарки и прочее добро - можно найти в Github-репозитории доклада

https://github.com/nevoroman/ml-dotnet

Вступление : из-за чего сыр-бор?

- Машинное обучение - это классно.
- .NET - это тоже замечательно
- По распространенному мнению, .NET не подходит для решения задач машинного обучения...

...но мы же не станем отступать из-за таких мелочей?

Машинное обучение

Машинное обучение : что это за зверь?

«Machine Learning is the field of study that gives computers the ability to learn without being explicitly programmed»

— Prof. Arthur Samuel

Машинное обучение : когда использовать?

- Когда трудно описать алгоритм решения задачи

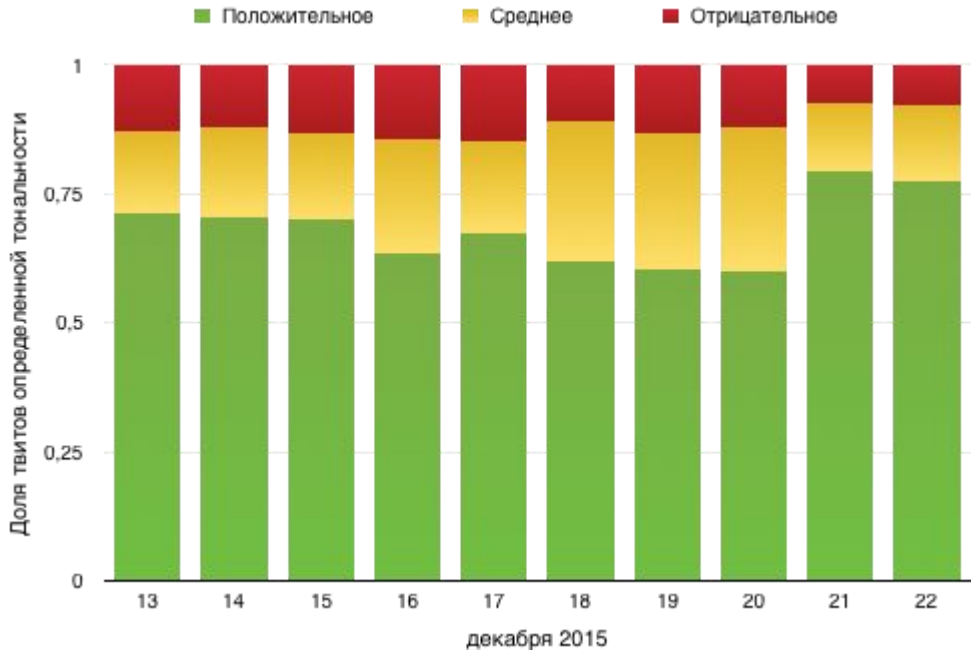
Машинное обучение : когда использовать?

- Когда трудно описать алгоритм решения задачи
- Когда нужно предугадать некоторые значения, имея большой набор данных

Машинное обучение : когда использовать?

- Когда трудно описать алгоритм решения задачи
- Когда нужно предугадать некоторые значения, имея большой набор данных
- Когда вы хотите улучшить работу имеющегося алгоритма за счет накопления опыта

Машинное обучение : зачем мне это?



Анализ отзывов фанатов о новых «Звездных войнах»

Решено в три дня и 10 строчек кода.

https://github.com/nevoroman/ml-dotnet/tree/master/MLIntro

Машинное обучение : зачем мне это?

- Позволяет легко решать трудные задачки

Машинное обучение : зачем мне это?

- Позволяет легко решать трудные задачки
- Базовые навыки легко осваиваются и полезны в других областях

Машинное обучение : зачем мне это?

- Позволяет легко решать трудные задачки
- Базовые навыки легко осваиваются и полезны в других областях
- Это чертовски весело!

ML и .NET

ML и .NET : мы этого точно хотим?

- .NET имеет кучу клевых инструментов для работы с данными

https://github.com/nevoroman/ml-dotnet/tree/master/ML&Dotnet

ML и .NET : мы этого точно хотим?

- .NET имеет кучу клевых инструментов для работы с данными
- Позволяет встраивать алгоритмы машинного обучения, не выходя из уютного дотнета

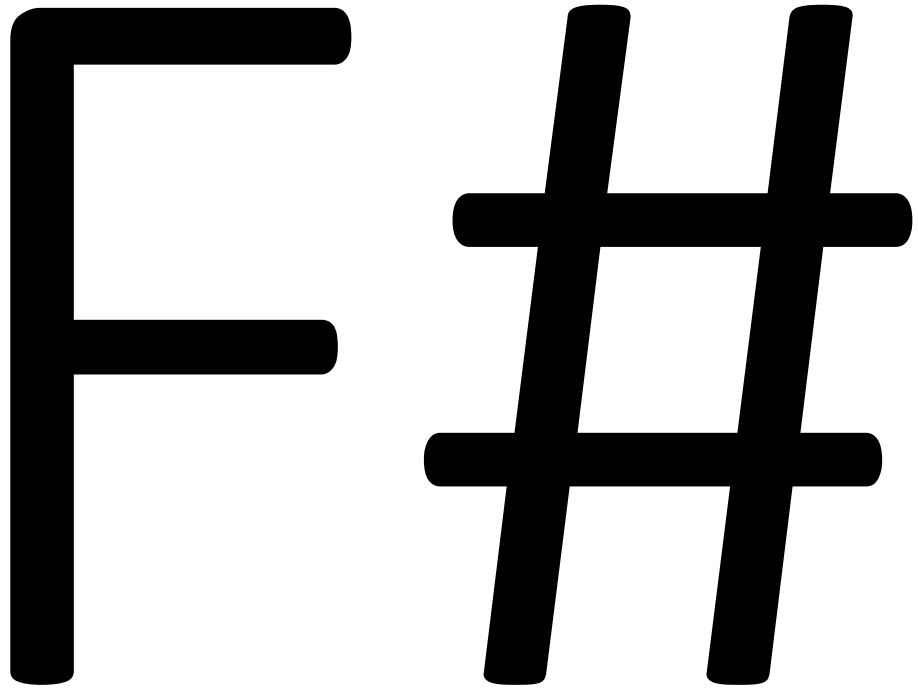
https://github.com/nevoroman/ml-dotnet/tree/master/ML&Dotnet

ML и .NET : мы этого точно хотим?

- .NET имеет кучу клевых инструментов для работы с данными
- Позволяет встраивать алгоритмы машинного обучения, не выходя из уютного дотнета
- F#

https://github.com/nevoroman/ml-dotnet/tree/master/ML&Dotnet

ML и .NET : мы этого точно хотим?



https://github.com/nevoroman/ml-dotnet/tree/master/ML&Dotnet

ML и .NET : препарируем F#

```
let rec quicksort list =
   match list with
   | [] ->                                // If the list is empty
        []                                // return an empty list
   | firstElem::otherElements ->          // If the list is not empty
        let smallerElements =             // extract the smaller ones
            otherElements
            |> List.filter (fun e -> e < firstElem)
            |> quicksort                  // and sort them
        let largerElements =              // extract the large ones
            otherElements
            |> List.filter (fun e -> e >= firstElem)
            |> quicksort                  // and sort them
        // Combine the 3 parts into a new list and return it
        List.concat [smallerElements; [firstElem]; largerElements]
```

ML и .NET : препарируем F#

ML и .NET : препарируем F#

```
let rec quicksort2 = function
    | [] -> []
    | first::rest ->
        let smaller,larger = List.partition ((>=) first) rest
        List.concat [quicksort2 smaller; [first]; quicksort2 larger]
```

Функциональное программирование - это просто!

ML и .NET : препарируем F#

```
[<Literal>]
let connectionString = "Data Source=DESKTOP-QM2SSL0\SQLEXPRESS; Initial
Catalog=MyDatabase; Integrated Security=True;"

type Sql = SqlDataConnection<connectionString>
let db = Sql.GetDataContext()

// find the number of customers with a gmail domain
query {
    for c in db.Customers do
    where (c.Email.EndsWith("gmail.com"))
    select c
    count
    }
```

https://github.com/nevoroman/ml-dotnet/tree/master/ML&Dotnet

ML и .NET : немного уличной магии

```
// Required package to save charts
open RProvider.grDevices

// Create path to an image testimage.png on the Desktop
let desktop = Environment.GetFolderPath(Environment.SpecialFolder.Desktop)
let path = desktop + @"\testimage.png"

// Open the device and create the file as a png.
// R.bmp, R.jpeg, R.pdf, ... will generate other formats.
R.png(filename=path, height=200, width=300, bg="white")
// Create the chart into the file
R.barplot(widgets)
// Close the device once the chart is complete
R.dev_off ()
```

ML и .NET : витаем в облаках

```
let downloadAsync (url : string) = async {
    use webClient = new System.Net.WebClient()
    let! html = webClient.AsyncDownloadString(System.Uri url)
    return html.Split('\n')
}


cloud {
    let! t1 = downloadAsync "http://www.nessos.gr/" |> Cloud.OfAsync
    let! t2 = downloadAsync "http://www.mbrace.io/" |> Cloud.OfAsync
    return t1.Length + t2.Length
}
```

ML и .NET : ближе к делу

Постановка задачи :

Необходимо без использования библиотек и бубна реализовать спам-фильтр, написав для этих целей как можно меньше и кода, не жертвуя, по возможности, читаемостью.

Решение:

F#, Байесовский классификатор и немного функциональщины

ML и .NET : что за классификатор?

Наивный байесовский классификатор - простой вероятностный классификатор, основанный на применении теоремы Байеса.

$$P(A \mid B) = \frac{P(A \cdot B)}{P(B)}$$

Инструментарий

Инструментарий : чем думать будем?

- Accord Framework
- numl
- Encog
- Azure ML

https://github.com/nevoroman/ml-dotnet/tree/master/MLTools

Инструментарий : Accord Framework

- Великолепно документирован
- Огромный арсенал всевозможных алгоритмов
- Гибкий
- Требует некоторого времени на освоение

https://github.com/nevoroman/ml-dotnet/tree/master/MLTools

Инструментарий : numl

- Просто осваивается
- Не требует долгого развертывания; для базовой реализации алгоритма достаточно и минуты
- Базовый набор алгоритмов
- Отличный «тестовый стенд» - для смены алгоритма необходимо поменять буквально одну строчку кода

https://github.com/nevoroman/ml-dotnet/tree/master/MLTools

Инструментарий : Encog

- Довольно гибок
- Неплохая встроенная работа с данными
- Имеет Java реализацию
- Работе с ним посвящена неплохая книга
- Некоторые алгоритмы он реализует просто великолепно.

https://github.com/nevoroman/ml-dotnet/tree/master/MLTools

Инструментарий : Azure ML

- Прост в использовании : ваш ребенок может случайно стать Data Scienist'ом
- Это клевые модные облачные вычисления
- Функционал легко расширяется самописными модулями
- Легкая и приятная работа с данными

https://github.com/nevoroman/ml-dotnet/tree/master/MLTools

Инструментарий : немного побенчмаркаем

https://github.com/nevoroman/ml-dotnet/tree/master/MLTools

Инструментарий : немного побенчмаркаем

```
public double DecisionTreeTest()
{
    var attributes = DecisionVariable.FromCodebook(_trainingData.CodeBook,
_trainingData.InputColumnNames.ToArray());

    var classificationDecisionTree = new DecisionTree(attributes,
_trainingData.OutputPossibleValues);
    new C45Learning(classificationDecisionTree).Run(_trainingData.InputData,
_trainingData.OutputData);

    var testingDataCount = _testingData.InputData.Length;
    double error = 0;
    for (var i = 0; i < testingDataCount; i++)
    {
        var input = _testingData.InputData[i];
        var result = classificationDecisionTree.Compute(input);
        if (result != _testingData.OutputData[i]) error++;
    }
    return error / testingDataCount;
}
```

https://github.com/nevoroman/ml-dotnet/tree/master/MLTools

Инструментарий : немного побенчмаркаем

Iris dataset

Method	Tool	Median (us)	Correct
K Nearest Neighbors	numl	5006.5	100%
K Nearest Neighbors	Accord Framework	123.2	100%
K Nearest Neighbors	scikit-learn (Python)	965.5	99%
Decision Tree	numl	2338.1	86.66%
Decision Tree	Accord Framework	687.7	96.66%
Decision Tree	scikit-learn (Python)	501.9	93.33%
Naive Bayes	numl	1657.8	80%
Naive Bayes	Accord Framework	93.2	100%
Naive Bayes	scikit-learn (Python)	1056.3	96.66%

https://github.com/nevoroman/ml-dotnet/tree/master/MLTools

Инструментарий : немного побенчмаркаем

Skin dataset

Method	Tool	Median (ms)	Correct
K Nearest Neighbors	numl	12783892.2	99.94%
K Nearest Neighbors	Accord Framework	1445.6	99.94%
K Nearest Neighbors	scikit-learn (Python)	1249.1	99.94%
Decision Tree	numl	2663.4	78.92%
Decision Tree	Accord Framework	3740.9	92.76%
Decision Tree	scikit-learn (Python)	317.8	99.93%
Naive Bayes	numl	1664.4	92.55%
Naive Bayes	Accord Framework	46	92.35%
Naive Bayes	scikit-learn (Python)	183.3	92.35%

https://github.com/nevoroman/ml-dotnet/tree/master/MLTools

Инструментарий : переваривая результаты

- Accord отлично показал себя на всех выбранных алгоритмах
- Некоторые алгоритмы могут работать ОЧЕНЬ медленно, но в большинстве случаев - это проблема реализации.
- На больших объемах данных Accord и numl показали сравнимые результаты, однако Accord все еще точнее
- Скорость и точность, которую показал Accord, сравнима с Scikit.

Инструментарий : ну и зачем нам numl?

```
public double DecisionTreeTest()
{
    var generator = new DecisionTreeGenerator();
    var model = generator.Generate(_description, _trainingData);
    return Estimate(model);
}

public double KNNTest()
{
    var generator = new KNNGenerator(2);
    var model = generator.Generate(_description, _trainingData);
    return Estimate(model);
}

public double NaiveBayesTest()
{
    var generator = new NaiveBayesGenerator(2);
    var model = generator.Generate(_description, _trainingData);
    return Estimate(model);
}
```

Инструментарий : FsLab

Объединяет в себе все лучшие инструменты манипулирования данными, созданные для F#

```
let usdata =
  frame [ "gdp" => series us.``GDP (current US$)``
          "uni" => series us.``School enrollment, tertiary (% gross)``
          "co2" => series us.``CO2 emissions (kt)`` ]
  |> Frame.dropSparseRows

let result = R.lm(formula = "gdp~uni+co2", data = usdata)
R.summary(result)
R.plot(result)
```

https://github.com/nevoroman/ml-dotnet/tree/master/MLTools

С чего начать?

- Mathias Brandewinder - Machine Learning Projects for .NET Developers
- Matthew Kirk - Thoughtful Machine Learning
- Tomas Petricek - Analyzing and Visualizing Data with F#

У вас найдется минутка поговорить о ML?

 roman.nevolin@waveaccess.ru

 nevoroman

 nevoroman/ml-dotnet

Спасибо за внимание!