Using Machine Learning to Organize the Community - 3

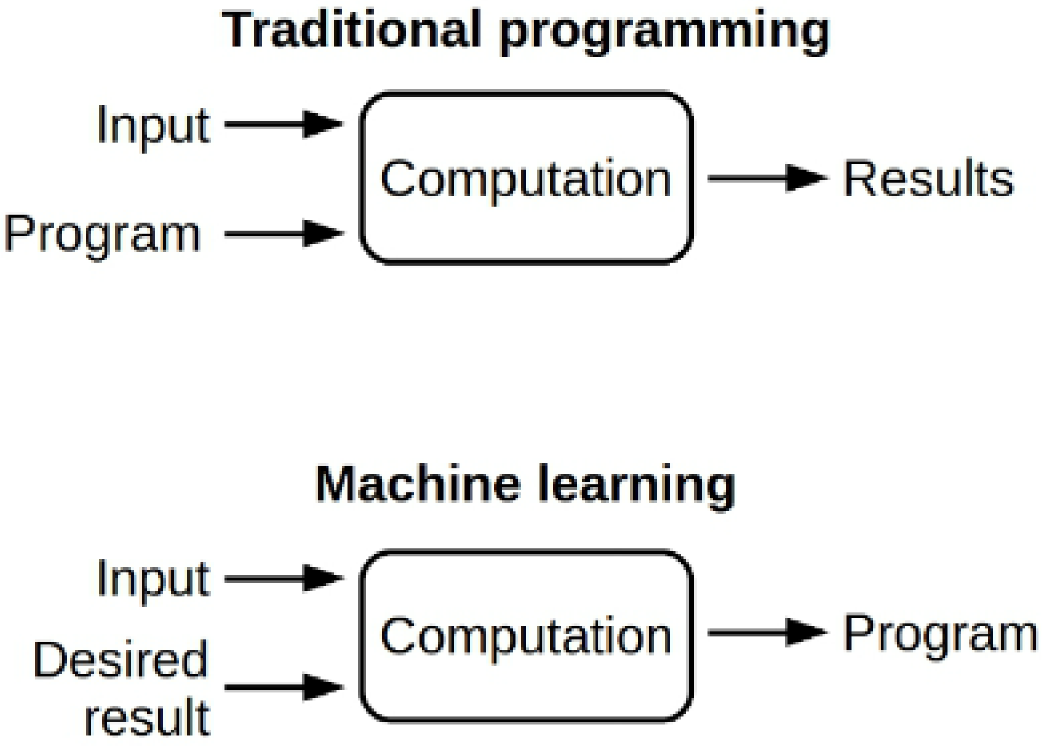
Article
Renato Banzai · Jul 19, 2020 3m read

Using Machine Learning to Organize the Community - 3

This is the third post of a series explaining how to create an end-to-end Machine Learning system.

Training a Machine Learning Model

When you work with machine learning is common to hear this work: training. Do you what training mean in a ML Pipeline?
Training could mean all the development process of a machine learning model OR the specific point in all development process
that uses training data and results in a machine learning model.



Source

So Machine Learning Models are not equal Common Applications?

In the very last point it looks like a normal application. But the concept is pretty different, because in a common application we don't use data in the same way. First of all, to create ML Models you will need real (production) data

at
all development process another difference is if you use different data you will have a different model..always...

Can you give some examples?

Sure! Imagine you are a system developer and create a system that controls services in a PetShop. The most simple way to
do this kind of system is create a DataBase do store data and create an application to interact with the user. When you
start develop it you don't need real data. You can create all system with fake data. This is impossible with machine learning models they need real data at the starting point of development and it will be used until the last point of development. When you create a machine learning model you provide data to an algorithm and the algorithm results in an
application fitted to those data, so if you use fake data, you will have a fake model.

What about Organize the Community with ML?

Getting back to the main subject. With all posts that I have I started exploring a sample and realize:

1. Most post were without any tag
2. Some posts were with wrong tags
3. Some posts (from sample) with tags were right tagged

How use this data?

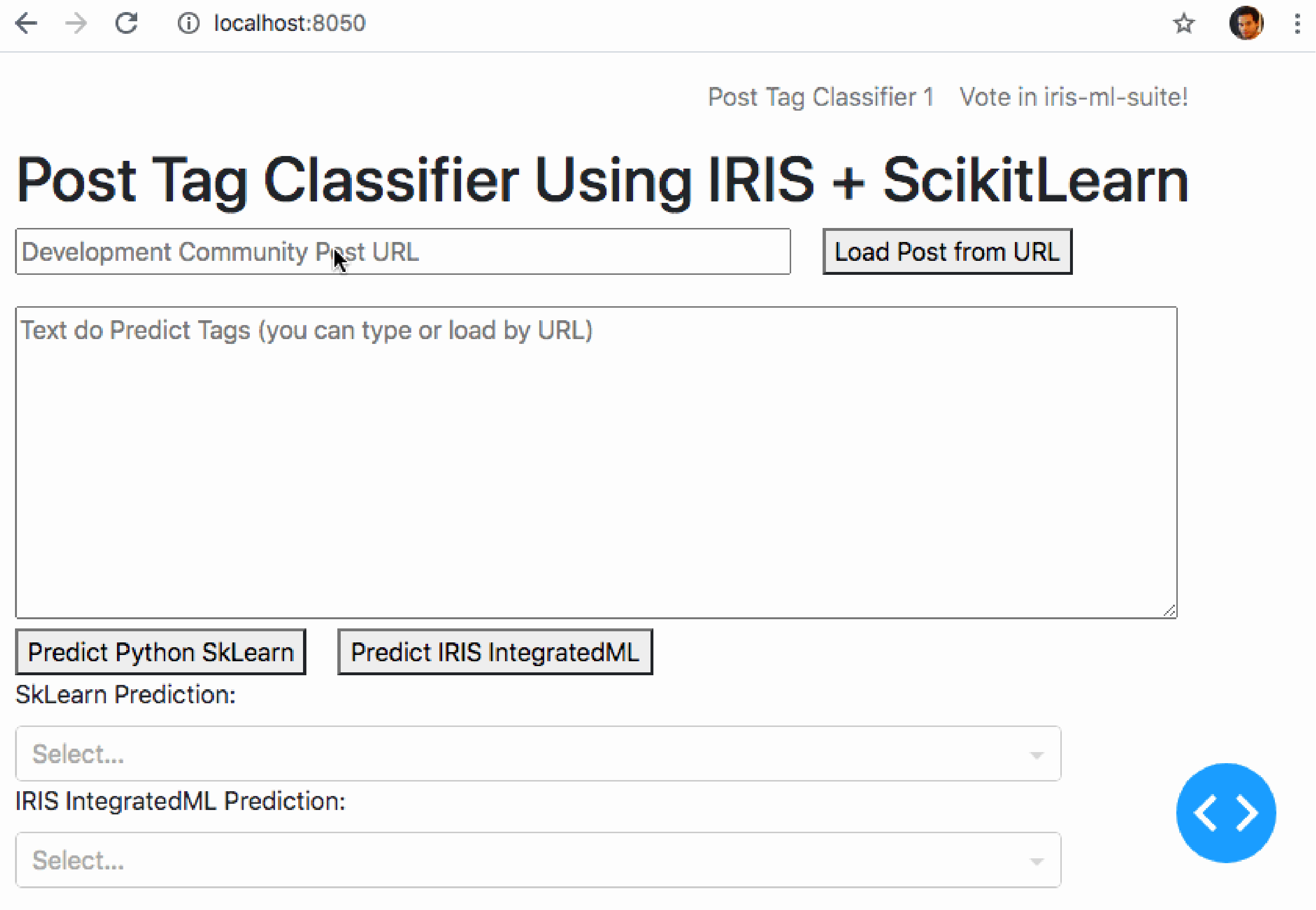
I take the option of create a BoW (Bag of Words) of the data and use the most popular classification algorithms with.
How do I featurize those data? My brief of attempts here:

1st Attempt - Training with only post data

Take all the data with tags, separate 80% for training and 20% to validate the model. The resulting model didn't perform
well. Most of time it was predicting all text were about Caché... (the BoW was with aroun 15000 features)

2nd Attempt - Training with only post data removing stop words

Thinking of most posts has the same words like: hi, for, to, why, in, on, the, do, don't... the model probably was taking
the wrong way to predict the tags. So I removed all stop words and vectorized the data, resulting in around 9000 features.
But the model continues to not perform well.


localhost:8050
Post Tag Classifier 1
Vote in iris-ml-suite!
Post Tag Classifier Using IRIS + ScikitLearn
Development Community Post URL
Load Post from URL
Text do Predict Tags (you can type or load by URL)
Predict Python SkLearn
Predict IRIS IntegratedML
SkLearn Prediction:
Select...
IRIS IntegratedML Prediction:
Select...

3rd Attempt - Training with Tags Description

Looking at tags table, I see a short description of each tag. So I give the chance to that small phrases be my guide to
the right way to tagging like in real life. It resulted in about 1500 features and using it the model starts performing well enough to show to other people =)

4th Attempt - Training with Tags Description

After all this quest I decide to test if IntegratedML could perform better than my model. So I converted the last Bag of
Words into a Table but I faced a problem: a bag of words with 1500 features means that I need a table with 1500 columns.
And reading the IRIS docs the limit of columns was 1000. So I changed my vectorizer to use only 900 features. It would
less but was the only way that I could think.

Classification Algorithms

I tried 3 algorithms with different params:

1. SVM
2. Random Forest
3. Logistic Regression

The best performer (at least in last days) was Logistic Regression in a strategy One vs Rest. One vs Rest consider that you have 1 class to predict versus a more than 1 other classes.

Next article: End-to-End Application

I hope you are enjoying. If you like the text and my application, please vote on https://openexchange.intersystems.com/contest/current in my application **iris-ml-suite**

#AI #Machine Learning #Python #InterSystems IRIS
40 1 1 10 133

Related posts

- Using Machine Learning to Organize the Community - 1
- Using Machine Learning to Organize the Community - 2
- Using Machine Learning to Organize the Community - 3

Source URL: https://community.intersystems.com/post/using-machine-learning-organize-community-3