Hillsboro Python Machine Learning Meetup

Nov/2019

Ernest Bonat, Ph.D.

Senior Software Engineer

DAT Wi-Fi

Username: DAT Guest

Password: beaverton dat

- 6:00 – 6:40 pm: Pizza, **water only** and networking.

- 6:40 – 6:45 pm: Welcome message by Ernest Bonat, Ph.D.

- 6:45 – 8:00 pm: Presentation and open discussions.

- 8.00 pm – 9.00 pm: Coding and learning session. Bring your Python development laptop!

Why did I create this meetup?

1. Bad traffic to Portland downtown.

2. Very hard to find a parking lot.

3. Bad Python presentation code and old used Python tools.

Our Meetup Mission:

1. *"Come, Listen, Code and Learn"*

2. Finding and presenting best practices of Machine Learning using Python Data Ecosystem.

3. Create great networking place for Hillsboro-Beaverton Data Scientists.

Today Presentation

“Best Practices of Machine Learning Projects Workflow”

Ernest Bonat, Ph.D.

Senior Software Engineer

Senior Data Scientist

What is Machine Learning (ML)?

Very simple: ***Use business data, to answer business questions, to make business decisions.***

Three Types of Machine Learning Algorithms:

1. Supervised Learning – use labeled data to generate a function ((y_1, y_2,…y_n = F(x_1,x_2,…,x_n))

- **Regression**
- **Classification**
- Ranking

2. Unsupervised Learning – use un labeled data for group data clustering.

- **Clustering**
- Association Mining
- Segmentation
- Dimension Reduction

3. Reinforcement Learning – use to train the machine to make specific decisions.

- Decision Process
- Reward System
- **Recommendation System**

Read:

1. “What is Machine Learning? A Friendly Introduction for Aspiring Data Scientists and Managers”

https://www.analyticsvidhya.com/machine-learning/?utm_source=blog&utm_medium=commonly-used-machine-learning-algorithms

2. “10 Machine Learning Methods that Every Data Scientist Should Know”

https://towardsdatascience.com/10-machine-learning-methods-that-every-data-scientist-should-know-3cc96e0eeee9

1. **Regression**
2. **Classification**
3. **Clustering**
4. Dimensionality Reduction (DR)
5. **Ensemble Methods (Bagging, Boosting and Stacking)**
6. **Artificial Neural Network (ANN) and Deep Learning (DL)**
7. Transfer Learning
8. Reinforcement Learning
9. **Natural Language Processing (NLP)**
10. Word Embeddings

ML Project Data Types

- Numerical and Categorical Data – integers, floats, texts, etc.
- Image Array Data – integers between 0 and 255
- Time Series Data – integers related to date-time

Different ML algorithms are applied to different ML data types!

1. Clear Data Project Documentation

Name, Description, Data location and format, Specific data preprocessing requirements, Define features and targets (labels), etc.

Example: “Pima Indians Diabetes Project”

2. First Team Meeting "Project Requirements Review"

Teams: Data Engineers, Data Scientists, Subject Matter Experts, Project Managers, IT Developers and Systems Administrators, etc.) – online/onground meetings, no emails only!, dataset review (data format, how many rows, how many columns (features and target(s)), data physical meaning, etc.), purpose of the final ML model, deployment, retraining schedule, etc.

3. Data Load (IO tools (TXT, CSV, HDF5,..))

https://pandas.pydata.org/pandas-docs/stable/user_guide/io.html

Use pandas library to load the data:

- pd.read_csv() – use for comma-separated files

- read_hdf() – use for high definition files (images files)
- read_sql() – use to load data from database engines (MySQL, etc.)
- read_pickle() and to_pickle() combination – most use for big datasets
- Apply C++ extension to read multi-dimensional arrays – need to know C++ programing!

Use Case: Need to load big CSV files with millions of rows. What to do?

Step 1.

df = pd.read_csv("big_csv_file") – it takes less or more than 1 hour?

pd. to_pickle(df, pkl_file_path_name) – it takes less than 1 hour?

Step 2.

df = pd.read_pickle(pkl_file_path_name) – it takes minutes only? Always # use this line to load the data into the dataframe.

4. Data Preprocessing (Data Cleansing) – it represents 60% - 80% work of the whole ML project. It provided by the Data Engineers team.

"Without clean data your talk has no meaning"

- Need a clear Data Cleansing Requirements Document

5. Data Profiling (Data Statistical Analysis and Data Visualization) – standard data stats analysis, correlation analysis, etc., data charts visualization, etc.

Use Pandas Profiling – very important!

https://github.com/pandas-profiling/pandas-profiling

6. Second Team Meeting "Project Requirements Review"

Many questions need to be answered: Is the data stats analysis makes sense with the project requirements? is the right physical input parameters correlation? do we need more data? do need new features? do we create new features (Feature Engineering – an act?), continue or stop the project? many more.

7. Data Split in three sets (training% | validation% | test%) – all the data sets need to be available. Need to try three most common combinations:

- **80% | 10% | 10%**
- 70% | 15% | 15%
- 60% | 20% | 20%

8. Data Scaling – it definitely is a good practice. Scaling data is the process of increasing or decreasing the magnitude according to a fixed ratio, in simpler words you change the size but not the shape of the data.

Read: “Data science: Scaling of Data in python”

https://medium.com/@stallonejacob/data-science-scaling-of-data-in-python-ec7ad220b339

Most common used data scaling:

1. **Centring** – $X_{centring} = X - X_{mean}$ (general used for numerical and categorical data)

2. **Standardization** – $X_{standardization} = (X - X_{mean}) / X_{std}$ (general used for numerical and categorical data)

3. **Normalization** – $X_{normalization} = (X - X_{min}) / (X_{max} - X_{min})$ (general used for image processing data including array of 0…255 integers numbers)

9. Select and apply classic and modern models – most common mistake done by Data Scientists today! They want to use modern Deep Learning models to do everything! Why? No used of classic models at all!

Two types of main models use today:

1. Ensembles models (Random Forest, AdaBoost, GBM, XGBoost, LightGBM, CatBoost, etc.) – based on tree-based ensembles algorithms…

2. Artificial Neural Network models (MLP, CNN, FNN, RNN, etc.)

Read: “5 Easy Questions on Ensemble Modeling Everyone Should Know”

https://www.analyticsvidhya.com/blog/2015/09/questions-ensemble-modeling/

Sequence of models I recommend to use:

1. Random Forest (RF) with scikit-learn framework
2. eXtreme Gradient Boosting (XGBoost)
3. Multi-layer Perceptron (MLP) with scikit-learn framework
4. Convolutional Neural Network (CNN) with Keras-TensorFlow

If Random Forest model for regression or classification result metrics are bad, you need to stop the project right away. Don’t waste your time!

10. Define Hyperparameters Model

- GridSearchCV()
- RandomizedSearchCV()
- **Hyperopt: Distributed Hyperparameter Optimization – very fast!**

No hyperparameters model guessing PLEASE!

11. Model Fitting

Most of the time slow process. It runs at night in general.

12. Model Prediction

Apply predict() method to get label y_predicted.

13. Calculate Model Metrics

For Classification:

- Classification Accuracy Score
- Classification Report
- Classification Confusion Matrix
- Area Under ROC Curve

For Regression:

- Mean Absolute Error
- Mean Squared Error
- R Squared (R^2)

Read: “Metrics to Evaluate Machine Learning Algorithms in Python”

https://machinelearningmastery.com/metrics-evaluate-machine-learning-algorithms-python/

14. Third Team Meeting “Project Requirements Review”

Are the final calculated project metrics good? How to deploy the model? How and what application(s) will be using the model? When to schedule model retraining and how? Develop API web services?

15. Model Production Deployment, Security and Retraining Schedule

- Network (serialize file)
- Intranet (API web services)
- Internet (API web services)