Data Science

Yogesh Kulkarni

ML in Production - Deployment

(Ref: The most under-taught skill in machine learning - George Seif)

After Learning Machine Learning

- Once you create your machine learning algorithm, that research part is done.
- Then you really start the bulk of the work.
- How will the results of your model be delivered to the end user?
- In today's world you'll need some powerful hardware to be able to run it at a reasonable speed; that means running your machine learning API on the cloud.
- Thats called "putting it in Production" or "Deployment"

Whats needed

- In today's world you'll need some powerful hardware to be able to run it at a reasonable speed;
- That means running your machine learning API on the cloud.
- You run it on a cloud server and send the results back to the user!
- You automate your system pipeline and have it ready to automatically scale based on your user traffic!
- Cloud computing is the workhorse behind the real-world machine learning applications.

Cloud computing for machine learning

Cloud Players

- AWS: most popular, allows you to control and customize
- Azure: Offers you easy of use at the expense of a bit of control and customization.
- GCP is somewhere in the middle of the two with some abstraction but not too much.

AWS EC2

- Houses your machine learning servers.
- Set up your machine learning model on the server.
- When you want to run something on your model, you send the data you want processed to the server, your model processes it, and sends it back to the user!
- EC2 also offers auto-scaling so that you can automatically spawn more or less servers based on the demand

AWS Lambda

- With lambda, you can basically set up automated trigger functions which will only run when a certain condition is met.
- For example, have your lambda function send an email to your user only when a certain results comes up from your machine learning module, such as some critical situation

AWS S3

Very cheap, 99.9999999% up-time, with fast download and upload speeds!

AWS RDS

Managed Relational Database Service for MySQL, PostgreSQL, Oracle, SQL Server, and MariaDB. Organize all of your important data for your machine learning data, API, infrastructure, and model results here.

AWS CodeDeploy

Automatically have your code and new machine learning models deployed to your servers as soon as you commit them to GitHub

AWS Cloudwatch

Online logs to constantly monitor your machine learning system

Amazon Simple Queue Service (SQS)

A queue hosted in the cloud. Keep your machine learning jobs organised and in order using a cloud queue

AWS Mobile Hub

Build, test, and monitor your apps remotely using the cloud. Just log in to your AWS account without the hassle of pulling data from your app manually

Amazon API Gateway

Build, deploy, and manage your API at any scale in the cloud. Have all the information you need for this in one simple place

Amazon Sagemaker

Build, train, and test your machine learning models using a high-level easy to use interface

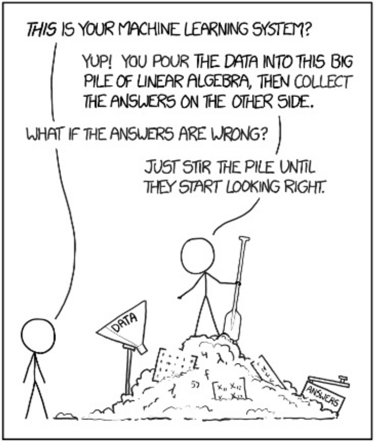
What Next?

- Coursera has a great one on GCP and Udemy has one on AWS!
- As always, it's a great idea to actually use the platform to learn it best.
- AWS offers a free tier for a year and their service aren't too expensive if you would like to play around with some of the non-free ones.
- GCP offers $300 of free credits for new accounts too!

Conclusion

What did you learn?

Machine Learning



From [xkcd]

ML Recap

(Ref: "Machine Learning Algorithm Cheat Sheet" - Laura Diane Hamilton)

Linear regression

Theme: Fitting Line

Pros

- Very fast (runs in constant time)
- Easy to understand the model
- Less prone to over-fitting

Cons

- Unable to model complex relationships
- Unable to capture nonlinear relationships without first transforming the inputs

Good at

- The first look at a dataset
- Numerical data with lots of features

Decision trees

Theme: Fitting a tree

Pros

- Fast
- Robust to noise and missing values
- Accurate

Cons

- Complex trees are hard to interpret
- Duplication within the same sub-tree is possible

Good at

- Medical diagnosis
- Credit risk analysis

Support Vector Machines

Theme: Partitioning Hyperplanes with wide margins

Pros

- Can model complex, nonlinear relationships
- Robust to noise (because they maximize margins)

Cons

- Need to select a good kernel function
- Model parameters are difficult to interpret

Good at

- Handwriting recognition
- Text classification

K-Nearest Neighbors

Theme: Partitioning Hyperplanes with wide margins

Pros

- Simple, Powerful
- No training involved ("lazy")
- Naturally handles multiclass classification and regression

Cons

- Expensive and slow to predict new instances
- Performs poorly on high-dimensionality datasets
- Must define a meaningful distance function

Good at

- Low-dimensional datasets
- Fault detection

Comparing ML Algorithms

- Power and Expressibility: ML methods differ in terms of complexity. Linear regression fits linear functions while NN define piecewise-linear separation boundaries. More complex models can provide more accurate models, but at the risk of over-fitting.
- Interpret-ability: some models are more transparent and understandable than others (white box vs. black box models)
- Ease of Use: some models feature few parameters/decisions (linear regression/NN), while others require more decision making to optimize (SVMs)
- Training Speed: models differ in how fast they fit the necessary parameters
- Prediction Speed: models differ in how fast they make predictions given a query

(pto . . .)

KNN Regression Example: 1d

Method	Power of Expression	Ease of Interpretation	Ease of Use	Training Speed	Prediction Speed
Linear Regression	5	9	9	9	9
Nearest Neighbor	5	9	8	10	2
Naive Bayes	4	8	7	9	8
Decision Trees	8	8	7	7	9
Support Vector Machines	8	6	6	7	7
Boosting	9	6	6	6	6
Graphical Models	9	8	3	4	4
Deep Learning	10	3	4	3	7

What Next?

Machine Learning Journey

To start with . . .

- If you want a more practical route then: Udacity Intro to Machine Learning
- If you want to learn Machine Learning in-depth then: Coursera Machine Learning Andrew Ng
- If you want other free courses, blogs, and books then: Phoenixts
- After being familiar with that, you should try learning:
  - How to Approach Almost Any ML Problem - Abhishek Thakur
  - Bias Variance Trade-Off http://scott.fortmann-roe.com/docs/BiasVariance.html
  - Measuring Errors http://scott.fortmann-roe.com/docs/MeasuringError.html
  - ROC Curve & AUC Explained https://www.youtube.com/watch?v=OAl6eAyP-yo

(Ref: "Simple 8 Step guide to learn Machine Learning with Python" - Randy Lao)

Machine Learning Learning Path

After you are done with Python ...

- Machine Learning in 20min: https://www.youtube.com/watch?v=MOdlp1d0PNA
- Skcikit-Learn Tutorial: https://www.youtube.com/watch?v=elojMnjn4kk
- Kaggle Machine Learning Tutorial: https://www.kaggle.com/learn/machine-learning
- Google Crash coursr Machine Learning
- Machine Learning at Berkeley https://ml.berkeley.edu/blog/tutorials/
- How to Learn Machine Learning in 6 Months https://www.youtube.com/watch?v=MOdlp1d0PNA&t=584s
- Learning Machine Learning & AI Guideline https://www.youtube.com/watch?v=PYKfXkd3t7c
- edX – Machine Learning (Columbia University, John Paisley) https://www.edx.org/course/machine-learning-columbiax-csmm-102x-0
- sentdex – Practical Machine Learning Tutorial (Youtube)

(Ref: "To start your DataScience Journey" - Randy Lao)

Transitioning into DataScience

Some amazing advice for those transitioning into DataScience: ...

- Kyle McKiou - DS Interview
- Sarah Nooravi - Personal Skills
- Beau Walker - How to Gain Experience
- Eric Weber - DS Companies
- Vin Vashishta - DS Interviews & Your Persona
- Kevin Tran - How to Land Your 1st DS Job
- David Langer - The 80/20 Rule of DS
- Favio Vázquez - Persistence
- Nic Ryan - Your Game Plan

(Ref: "Transitioning into DataScience" - Randy Lao)

Conculsion

Recipe Tour

Here, you will see 5 recipes of supervised classification algorithms applied to small standard datasets that are provided with the scikit-learn library. Each example is:

- Standalone: Each code example is a self-contained, complete and executable recipe.
- Just Code: The focus of each recipe is on the code with minimal exposition on machine learning theory.
- Simple: Recipes present the common use case, which is probably what you are looking to do.
- Consistent: All code example are presented consistently and follow the same code pattern and style conventions.

Logistic Regression

Logistic regression fits a logistic model to data and makes predictions about the probability of an event (between 0 and 1).

```
from sklearn import datasets
from sklearn import metrics
from sklearn.linear_model import LogisticRegression

dataset = datasets.load_iris()

model = LogisticRegression()
model.fit(dataset.data, dataset.target)
print(model)

expected = dataset.target
predicted = model.predict(dataset.data)

print(metrics.classification_report(expected, predicted))
print(metrics.confusion_matrix(expected, predicted))
```

Naive Bayes

Naive Bayes uses Bayes Theorem to model the conditional relationship of each attribute to the class variable. This recipe shows the fitting of an Naive Bayes model to the iris dataset.

```
from sklearn import datasets
from sklearn import metrics
from sklearn.naive_bayes import GaussianNB

dataset = datasets.load_iris()

model = GaussianNB()
model.fit(dataset.data, dataset.target)
print(model)

expected = dataset.target
predicted = model.predict(dataset.data)

print(metrics.classification_report(expected, predicted))
print(metrics.confusion_matrix(expected, predicted))
```

k-Nearest Neighbor

The k-Nearest Neighbor (kNN) method makes predictions by locating similar cases to a given data instance (using a similarity function) and returning the average or majority of the most similar data instances.

```
from sklearn import datasets
from sklearn import metrics
from sklearn.neighbors import KNeighborsClassifier

dataset = datasets.load_iris()

model = KNeighborsClassifier()
model.fit(dataset.data, dataset.target)
print(model)

expected = dataset.target
predicted = model.predict(dataset.data)

print(metrics.classification_report(expected, predicted))
print(metrics.confusion_matrix(expected, predicted))
```

Classification and Regression Trees

Classification and Regression Trees (CART) are constructed from a dataset by making splits that best separate the data for the classes or predictions being made.

```python
from sklearn import datasets
from sklearn import metrics
from sklearn.tree import DecisionTreeClassifier

dataset = datasets.load_iris()

model = DecisionTreeClassifier()
model.fit(dataset.data, dataset.target)
print(model)

expected = dataset.target
predicted = model.predict(dataset.data)

print(metrics.classification_report(expected, predicted))
print(metrics.confusion_matrix(expected, predicted))
```

Support Vector Machines

Support Vector Machines (SVM) are a method that uses points in a transformed problem space that best separate classes into two groups.

```
from sklearn import datasets
from sklearn import metrics
from sklearn.svm import SVC

dataset = datasets.load_iris()

model = SVC()
model.fit(dataset.data, dataset.target)
print(model)

expected = dataset.target
predicted = model.predict(dataset.data)

print(metrics.classification_report(expected, predicted))
print(metrics.confusion_matrix(expected, predicted))
```

Benefits and drawbacks of scikit-learn

Benefits

- Consistent interface to machine learning models
- Provides many tuning parameters but with sensible defaults
- Exceptional documentation
- Rich set of functionality for companion tasks
- Active community for development and support

Potential drawbacks

- Harder (than R) to get started with machine learning
- Less emphasis (than R) on model interpret-ability

Essential Machine Learning Algorithms

After you are done with Python ... Here is a list of 10 Essential Algorithms that you should know to understand the basics of MachineLearning:

- Logistic Regression https://lnkd.in/gJ2BwhD
- Linear Regression https://lnkd.in/gdZDbT5
- Decision Trees https://lnkd.in/gwadA-p
- Random Forests https://lnkd.in/gRYHcvt
- Neural Networks https://lnkd.in/gZQhWyv
- Bayesian Techniques https://lnkd.in/gY3qVYP
- Support Vector Machines https://lnkd.in/gWJKRyn
- Gradient Boosting Machine https://lnkd.in/gv85yDV
- K-Nearest Neighbors https://lnkd.in/gsiyqcM
- Regularized Linear Models https://lnkd.in/g3fn3cr

(Ref: "Essential Machine Learning Algorithms" - Randy Lao)

Kaggle Datasets and Projects

- Binary Classification
  - Indian Liver Patient Data https://bit.ly/2OvwYtm
  - Financial Data - Fraud Detection https://bit.ly/2Iyg2x0
  - Predict Product Backorders? https://bit.ly/2OqMwie
  - Adult Census Income https://bit.ly/2zLAKXB
- Multi-Classification
  - Iris https://bit.ly/2xS1xQn
  - Fall Detection https://bit.ly/2IxrxEP
  - Biomechanical Features of Ortho Patients https://bit.ly/2Hqv9ep
- Regression
  - Video Game Sales https://bit.ly/2qsu2OR
  - NYC Property Sales https://bit.ly/2AKijRz
  - Gas Sensors https://bit.ly/2OvYIhm
- NLP
  - The Enron Email Dataset https://bit.ly/2xS3gVR
  - Ubuntu Dialogue Corpus https://bit.ly/2ygxBx1
  - Old Newspapers https://bit.ly/2NXCDcg
  - Blog Authorship Corpus https://bit.ly/2y4t4xr
- Time Series
  - Crypto Historical Data https://bit.ly/2y5tzYb
  - Exoplanet Hunting in Deep Space https://bit.ly/2RoOD4K
- Image Processing
  - YouTube Faces https://bit.ly/2OvLYoF

Machine Learning Projects

Last one is a MUST DO! . . .

- Pokemon - Weedle's Cave
- Titanic ML
- Housing Prices Prediction
- Instacart Market Basket Analysis
- Quora Question Pairs
- Human Resource Analytics
- Analyzing Soccer Player Faces
- Recruit Restaurant Visitor Forecasting
- TensorFlow Speech Recognition
- Yourself: The BEST project you'll ever work on is you.

(Ref: "10 GREAT DataScience Projects to work on" - Randy Lao)

Machine Learning Dataset sites

Got Data? ...

- UCI Machine Learning Repository https://archive.ics.uci.edu/ml/datasets.html
- Kaggle: https://www.kaggle.com/datasets
- Quandl: The premier source for financial, economic, and alternative datasets https://www.quandl.com/
- KDD Cup Archives: Archives of the Data Mining and Knowledge Discovery competition http://www.kdd.org/kdd-cup
- Data Driven: Datasets where data science can be used to create a social impact https://lnkd.in/gGtpN9q
- Data Gov The home of the U.S. Government's open data https://www.data.gov/

(Ref: "Top 6 websites to get datasets for Machine Learning" - Randy Lao)

Machine Learning Websites

Favorite websites to learn from . . .

- FastML - Machine Learning Made Easy
- Analytics Vidhya - Learning Everything About Analytics
- Machine Learning Mastery - Title Explains It
- KDNuggets - One of the most popular Data Science blogs
- Data Science Central - Online Resource for Big Data Practictioners
- Data at Quora - Where Data Scientists Share What They've Learned
- Towards Data Science - Sharing Data Science Concepts, Ideas, and Codes

(Ref: "7 Favorite websites to learn from" - Randy Lao)

Machine Learning Youtube channels

Favorite video lists to learn from ...

- 3Blue1Brown Essence of Linear Algebra
- StatQuest (Joshua Starmer) Statistics Made EASY
- Siraj Raval Fun Machine Learning & AI
- Analytics University Anything Analytics & Machine Learning
- AlphaOpt Optimization & Gradient Descent (Short and easy explanation)
- 3Blue1Brown Simple Visual Explaination on Neural Networks
- Two Minute Papers Awesome AI Research For Everyone

(Ref: "My Favorite resources on Youtube" - Randy Lao)

References

Many publicly available resources have been refereed for making this presentation. Some of the notable ones are:

- Introduction to Machine Learning with scikit.learn - West of Ireland Data Science
- STAT 365/665: Data Mining and Machine Learning - Taylor Arnold
- CSC 600: Data Mining - Richard Burns
- Data Science Simplified - Pradeep Menon
- Learn Data Science - Nitin Borwankar
- IAML: Decision Trees - Victor Lavrenko and Charles Sutton
- Data Science Notebooks
- Analytics Vidhya Blogs
- Machine Learning - Brett Wujek , SAS Institute
- Introduction to Entropy for Data Science - Mike Schulte

Thanks . . .

- Search **"Yogesh Haribhau Kulkarni"** on Google and follow me on LinkedIn and Medium
- Office Hours: Saturdays, 2 to 5pm (IST); Free-Open to all; email for appointment.
- Email: yogeshkulkarni at yahoo dot com