2.2 A general framework for data science

Thursday, October 7, 2021 2:13 PM

Key points:

1. The epistemology: how to produce new knowledge
2. Three phases of machine learning: conceptual, mathematical, applied
3. Learning theory
    a. Hypotheses and optimization
    b. Bias vs variance

1. The epistemology: how to know the unknown (known known, known unknown, unknown unknown)
   The common goal: to know
    a. Declarative:
        i. We are told what is true: y is y (or x is y) >> facts
        ii. We assume that something is true to serve a basis for reasoning
        iii. We memorize that something is true.
        iv. Encyclopedia

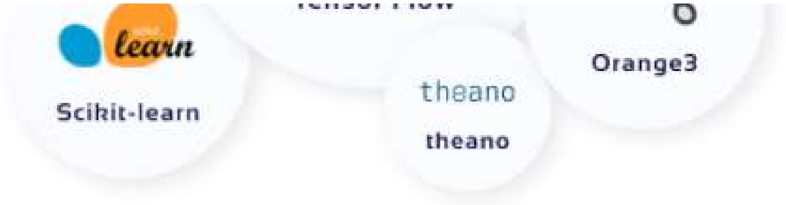
    a. Learning:
        i. we explore, check, and hopefully approach to the truth: from x working our way to know y
           X: observations or data
           y: what we are interested to know: a class or a value
        ii. Learning from experiences (what we have seen or known)
            1) Limited by experiences
            2) Consistent with experiences
            3) Inconsistent with experiences
        iii. Learning from trials and errors
            1) Heuristic (instead of exact) solutions
            2) Optimal (instead of the best) solutions

2. Three phases of machine learning:
    a. Conceptual:
        i. The purpose:
            1) Classification
            2) Regression
        ii. What is the logical basis to "map" input X to output y?

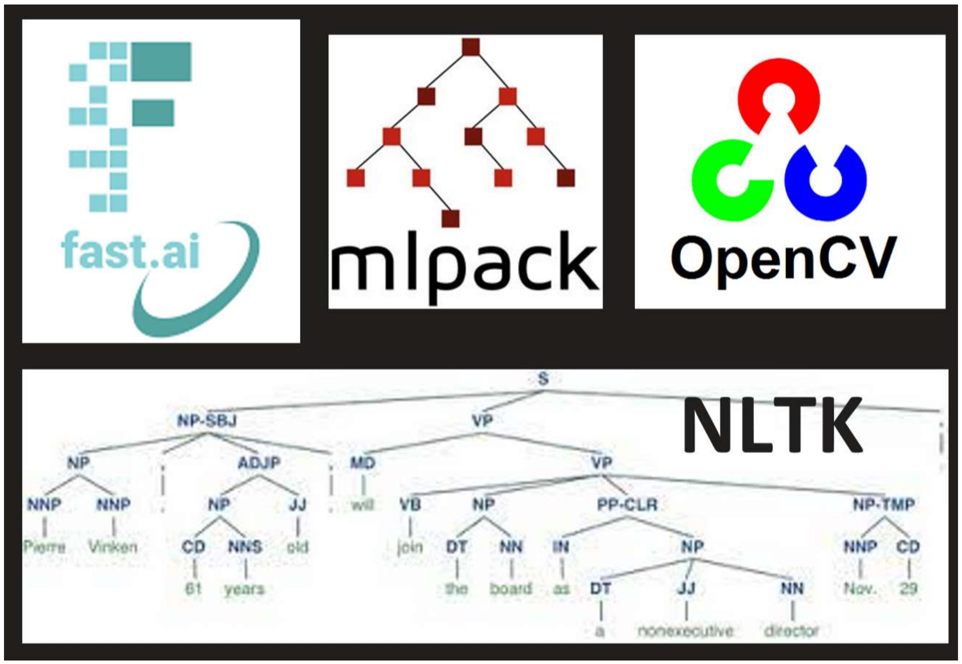
iii. What are the strategic options to determine an optimal mapping?

b. Mathematical:
   i. What is the mathematical formula for the mapping function?
   ii. How to formulate the optimalization function? (a maximum likelihood function, loss function, or reward function)
   iii. How many hyper-parameters to set initial conditions?
   iv. How many parameters we need to estimate?
   v. Relay heavily on linear algebra and calculus (derivatives)

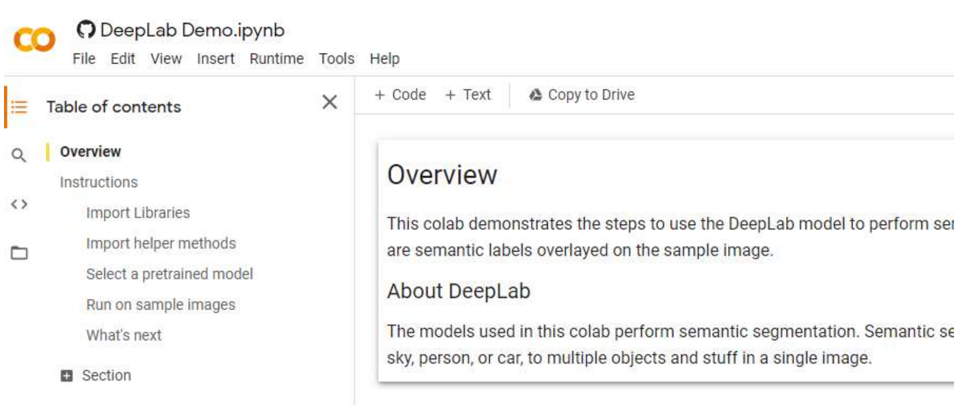
c. Applied:
i. What programming languages?
   1) Python
   2) R
   3) Matlab
   4) C++
   5) Java
   6) JavaScript
   7) Julia
   i. What libraries?



learn
Scikit-learn
theano
theano
Orange3



fast.ai
mlpack
OpenCV
NLTK



DeepLab Demo.ipynb
File Edit View Insert Runtime Tools Help
Table of contents
Overview
Instructions
Import Libraries
Import helper methods
Select a pretrained model
Run on sample images
What's next
Section
+ Code + Text
Copy to Drive
Overview
This colab demonstrates the steps to use the DeepLab model to perform se
are semantic labels overlayed on the sample image.
About DeepLab
The models used in this colab perform semantic segmentation. Semantic se
sky, person, or car, to multiple objects and stuff in a single image.

https://colab.research.google.com/github/tensorflow/models/blob/master/research/deeplab/deeplab_demo.ipynb

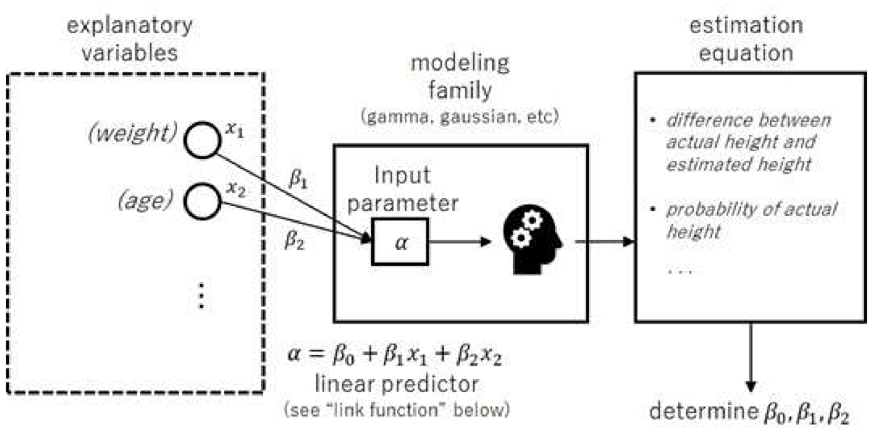
1. Learning Theory
   a. Assumptions to map X to y:
      i. There is a data generation process that generates a data distribution D
         (X, y) are sampled from the data distribution: (X, y) ~ D
         Our training data and testing data are from the same data distribution
         The data generation process persists and allows machine learning

      ii. Data X are independent, identically distributed (iid) random samples from the data distribution
         X includes one or more variables. Each variable is an iid random variable.
         X is a vector of features (or a feature vector)

   b. Our hypothesis (h*) is set to be the data generation process.

      The hypothesis is ….

      Data samples --> learning algorithm --> estimated hypothesis ($\hat{h}$)



https://tsmatz.wordpress.com/2017/08/30/glm-regression-logistic-poisson-gaussian-gamma-tutorial-with-r/

   C. Bias and variance are peroperties of an algorithm given sample size m
      i. From data's perspective

Q 1: Ultimately, what is machine learning set to learn? C

A. The sample distribution
B. The population distribution
C. The data generation process
D. All of the above

Q 2: Which of the following is a constant (i.e. have a fixed value): A

A. h*
B. Data samples
C. $\hat{h}$
D. None of the above

Q 3: Why should we worry about bias and variance in machine learning? A

A. If the result has a high bias, our model is underfitting the data
B. If the result has a high variance, our model is underfitting the data
C. If the result has a high bias, our model can either underfit or overfit the data.
D. If we have to choose, we prefer a model that has a low bias but a high variance.

ii. From parameters' perspective

Strategies to reduce high variance:

i. Increase sample size (increase m; get more data in a sample and more samples)
ii. Regularization (add penalty: L1, L2)

Strategies to reduce high bias

i. Enlarge the hypothesis space
ii. Change hypothesis family (algorithm class)

Generalized Linear Model (GLM):
Gaussian family: for continuous data (float)
Poisson family: for counts (positive integers)
Binominal (Bernouli, logistic) family: for binary data
Gamma family: left-limited, time or duration of event occurrences

D. Hypothesis space, errors, bias, and variance

B

Q 4: Which of the following statement is correct:

A. G is the best possible hypothesis , so the estimate error = 0
B. h* is the best possible hypothesis in the hypothesis family under consideration, so h* must be closest to G in the entire hypothesis family
C. $\hat{h}$ is what the algorithm learnt, so it is the best hypothesis in the hypothesis family
D. All of the above

Q 5: What is irreducible error? A

A. Baye's error
B. Approximation error
C. Estimation error
D. Empirical error

Continue to Lab package 2