Data Science

Lesson 4: Introduction to Machine Learning, Classification with K-Nearest Neighbors

Unit: Machine Learning

OBJECTIVE -

By the end of this class the students will be able to:

- Explain what machine learning is
- Determine if a question is a supervised or unsupervised learning problem
- Determine the appropriate approach for the type of problem (regression, classification, clustering, dimension reduction)
- Create their own KNN algorithm using Python

QUICK RUNDOWN

(Note: Timing/Topics breakdown are up to the instructor - these are general guidelines for a three hour lesson)

Timing	Topic	Activity
10 min	Opening framing	• Review previous lecture's content (cleaning and exploring data). What was the most interesting thing you learned last time? What was the hardest? • Introduce today's learning objectives • Connect to a greater learning goal/describe how this fits into the overall data science purview
60 min	Introduction to new material (lecture/theory)	Present lecture on ML/KNN • What is Machine Learning? • Supervised vs Unsupervised Learning • Categorical vs Continuous data • Show examples of regression, classification, clustering, dimension reduction • Introduce clustering with KNN
10 min	Break	Break

90 min	LAB (We Do)	● Explore the data (iris dataset) using pandas ● Introduce scikit-learn using the iris dataset ● Split the data into test and training dataset ● Train KNN classifier defined function on the train data and test on the test dataset ● Check accuracy of the model and predictions with different numbers of neighbors
10 min	Closure / Q & A	Review, assign homework

MATERIALS

- Students brings laptops with Anaconda Python and scikit-learn installed

HOMEWORK

- Read “Understanding the Bias/Variance trade-off”
- Classification and KNN homework
    - Download the Pima Indians dataset from UCI here: https://archive.ics.uci.edu/ml/datasets/Pima+Indians+Diabetes
    - Describe the content of the dataset in your own words
    - Describe the features and formulate hypothesis on which may be relevant in predicting diabetes
    - Import the dataset to a Pandas dataframe and explore the data:
        - a. Are there any missing data or NULL values? How could they be imputed? Make a choice and impute them or drop them. Justify the choice.
        - b. How many features are there? Are they normalized?
    - Use the KNN classifier from Scikit-learn to predict diabetes occurrence
    - Use Scikit-learn cross-validation routine to evaluate the accuracy of your model with a 5-fold CV.
    - Plot the 5-fold CV accuracy score as a function of K for k up to 50 neighbors
    - Use the Naïve Bayes classifier from Scikit-learn to predict diabetes occurrence
    - Compare the 5-fold CV score for Naïve Bayes and for KNN to find which model is more accurate

Step 4 - Identify how the lesson plan relates to the presentation slides.

Answer the following questions and include the responses at the *bottom of your annotated lesson plan*:

1. How might you have to edit the presentation slides to reflect your lesson plan?

- Slide for hook
- Slide for real world mom Netflix anecdote
- Take out slides 4-5
- Slide for activity explanation

2. Similarly, it helps to think through how you'll reference both the lesson plan and the presentation slides while delivering your lesson. How will you use one or both resources (or not) during class? Why do you think this will help ensure effective lesson delivery?
   I think I'll use the slides as my main guideline, and the lesson plan as Supplemental. Slides will cover main points, lesson plan will contain notes to myself. I think this will ensure main points are covered, but will also give enough flexibility...not just reading slides...how boring would that be.