Nastasia Pollas

5/29/24

HW 1

1. What is machine learning?
    a. It is the science of programming computers to learn from data
2. Name 4 types of problems where it shines
    a. In the book's example of spam filtering, machine learning quickly learns what updated subjects/senders could be spam and filter them. In traditional programming, this would require a manual update to keep up with changing information.
    b. Machine learning is best for programs without known algorithms or those too complex for traditional methods, for example speech recognition.
    c. Machine learning is great for changing environments and adapts to new data.
    d. Machine learning can assist people in learning, revealing correlations and trends, and producing greater understanding of problems.
3. What is a labeled training set?
    a. A correctly labeled dataset that can be used to training and to asses a model.
4. Name the four main challenges in Machine Learning
    a. Insufficient quantity of training data
        i. The simplest of problems require a large amount of examples for the algorithms to work properly
    b. Non-presentative training data
        i. Model is unable to make accurate predictions with non-representative data
        ii. If the sample is too small, then that results in sampling noise, and sampling bias in larger datasets.
    c. Poor quality data
        i. Data full of noise, errors, and outliers are hard for the algorithms to detect patterns. Most data requires significant cleaning to ensure proper working algorithm.
    d. Irrelevant features
        i. The system cannot learn if the training data contains too many irrelevant features and not enough relevant ones.
5. Overfitting is when the model performs great in training data but generalizes poorly to new instances. Three possible solutions include
    a. Simplify the model by selecting one with fewer parameters and reducing the number of attributes or constraining the model.
    b. Obtain more training data
    c. Decrease the noise
6. The test set is used after you complete training with the training set, and is used to obtain an estimation of that error.
7. The purpose of the validation set is to allow you to test the best-selected model and hyperparameters, after training, You run a one final test against the test set to get an estimate of the generalization error.

8. If you tune hyperparameters using the test set, then a problem is that you measured the generalization error several times, so the resulting model Is not generalizable to new instances.
9. Cross-validation is when one splits the training set into complementary subsets and then train each model against a different combination of those subsets and validated against the remaining sets. Then, one can train a final model using hyperparameters on full training set. The main advantage is that it avoids wasting too much training data in validation sets.