Foundations of Machine Learning

Brett Bernstein

August 24, 2018

Lab 0: Black box machine learning

Machine learning

Learning Objectives

1. What is machine learning for?
2. What is machine learning?
3. How do we frame a machine learning problem?
4. How do we evaluate machine learning models?
5. What can go wrong?

Concept Check Questions

1. Give 3 examples of ML applications, and name what type of ML problem it is (e.g. binary classification, regression, etc).

   Solution.

   - Spam detection: binary classification
   - Apartment sale price prediction: regression
   - Tree species identification: multiclass classification

2. What is supervised machine learning?

   Solution. The process of producing prediction function from a set of input and output pairs.

3. What is feature extraction?

   Solution. The process of mapping an input example into arrays of numeric values.

4. What are the inputs of a loss function?

   Solution. Inputs are an example and a model's output on that example.

5. Is a small loss or a large loss preferable?

   Solution. A small loss is better, since the loss measures how far off a prediction is from the target output.

6. Write down a loss function for classification and a loss function for regression.

   Solution.

   - Classification loss: 1 if prediction is wrong, 0 is prediction is correct.
   - Regression loss: square loss $(predicted - target)^2$.

7. Describe overfitting in 1 sentence.

   Solution. If a model overfits if training performance is good, but validation/testing performance is poor.

8. What is a hyper parameter? Name two examples of hyperparameters.

   Solution. It's a parameter of the machine learning algorithm itself, e.g. the degree of a polynomial that is being fit, or the number of words that a spam classifier takes into account.

9. What is the difference between a validation set and a test set? Why would the performance on the two sets differ?

   Solution. The validation set is used to find the best (out of many) model architecture for a problem, while the test set is used to evaluate the performance of the final (best) model. Since we are testing many different models on the validation set, it is possible that the model that achieved the best performance on the validation set "was lucky". Hence, it's performance on the validation set may be better than the performance would be on a random fresh sample (e.g. the test set). We follow the principle: if you use a data set to train your model or choose a model, you should not evaluate the final performance on that data set.

10. Briefly explain how to perform k-fold cross-validation.

    Solution. Split the data set into k equally sized subsets $D_1, \ldots, D_k$. For each D_i, train a model on all data points that are not in D_i, and evaluate the model on D_i. Compute the cross-validation performance as the average of the performances computed on the D_i. Note that the cross-validation performance should not be used as a final performance measure!

11. Briefly explain each of the following:

    - Information leakage
    - Sample bias
    - Covariate drift
    - Concept drift

    Solution.

    - Information leakage: information about labels that is unavailable at deployment time is present in the features
    - Sample bias: test and deployment input distributions have different distributions
    - Covariate drift: the input distribution changes over time (as a result, the model "ages")
    - Concept drift: the correct label for a given input changes over time (as a result, the model "ages")