Machine Learning

Pawel Wocjan

University of Central Florida

Spring 2019

Sources for Slides

- I have extensively used the machine learning materials that have been prepared by Google.

  https://developers.google.com/machine-learning/crash-course/

- Google has licensed these materials under the Creative Commons Attribution 3.0 License.

  https://creativecommons.org/licenses/by/3.0/

Outline

Peril of Overfitting

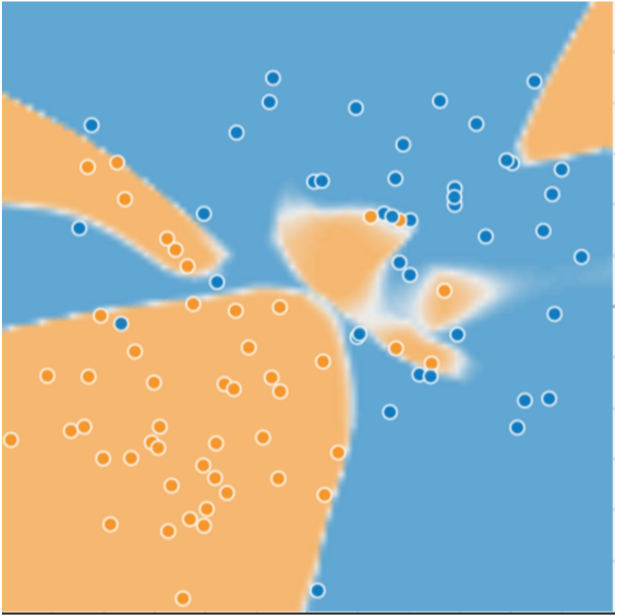
- To gain some intuition about generalization, let's look at the following three figures.
- Assume that each dot in these figures represents a tree's position in a forest. The two colors have the following meanings:
  - The blue dots represent sick trees.
  - The orange dots represent healthy trees.
- Can you imagine a good model for predicting subsequent sick or healthy trees?

Peril of Overfitting

Peril of Overfitting

- Look now at the next figure, which shows how a certain machine learning model separated the sick trees from the healthy trees.
- Note that this model produced a very low loss.

Peril of Overfitting



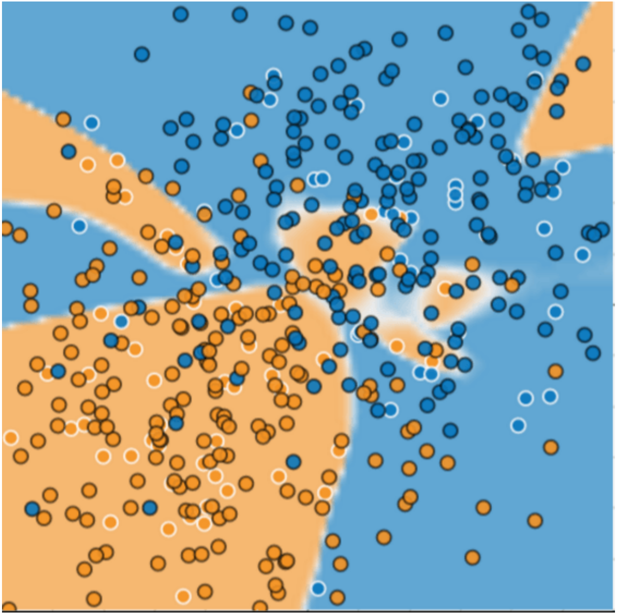
Peril of Overfitting

- At first glance, this model appears to do an excellent job of separating the healthy trees from the sick ones. Or does it?

Peril of Overfitting

- The next figure shows what happened when we added new data to the model.
- It turned out that the model adapted very poorly to the new data.
- Notice that the model miscategorized much of the new data.

Peril of Overfitting



Peril of Overfitting

- The model shown did a bad job predicting new data.
- The model **overfits** the peculiarities of the training data.
- An overfit model has a low loss on the training data, but it has a high loss of the test data (new unseen data).
- Overfitting is caused by making a model more complex than necessary.
- The fundamental tension of machine learning is between fitting our training data well, but also fitting it as simply as possible.

Peril of Overfitting

- Machine learning's goal is to predict well on new data drawn from a (hidden) true probability distribution.
- Unfortunately, the model can't see the whole truth; the model can only sample from a training data set.
- If a model fits the current examples well, how can you trust the model will also make good predictions on never-before-seen examples?

Peril of Overfitting

- William of Ockham, a 14th century friar and philosopher, loved simplicity.
- He believed that scientists should prefer simpler formulas or theories over more complex ones.
- To put **Ockham's razor** in machine learning terms:

  The less complex an ML model, the more likely that a good empirical result is not just due to the peculiarities of the sample.

Peril of Overfitting

- In modern times, we've formalized Ockham's razor into the fields of statistical learning theory and computational learning theory.
- These fields have developed generalization bounds – a statistical description of a model's ability to generalize to new data based on factors such as:
  - the complexity of the model
  - the model's performance on training data
- For instance, take a look at VC-dimension: `https://en.wikipedia.org/wiki/Vapnik-Chervonenkis_dimension`

Peril of Overfitting

- While the theoretical analysis provides formal guarantees under idealized assumptions, they can be difficult to apply in practice.
- In our course, we focus instead on empirical evaluation to judge a model's ability to generalize to new data.

Peril of Overfitting

- A machine learning model aims to make good predictions on new, previously unseen data.
- But if you are building a model from your dataset, how would you get the previously unseen data?
- One way is to divide your data set into two subsets:
  - **training set** – a subset to train a model
  - **test set** – a subset to test the model
- Good performance on the test set is a useful indicator of good performance on the new data in general, assuming that:
  - The test set is large enough.
  - You don't cheat by using the same test set over and over.

Peril of Overfitting

- The following three basic assumptions guide generalization:
  - We draw examples **independently and identically (i.i.d)** at random from the distribution. In other words, examples don't influence each other.
    An alternate explanation: i.i.d. is a way of referring to the randomness of variables.
  - The distribution is **stationary**; that is the distribution doesn't change within the dataset.
  - We draw examples from partitions from the **same distribution**.

Perils of Overfitting

- In practice, these assumptions are sometimes violated. For example:
  - Consider a model that chooses ads to display. The i.i.d. assumption would be violated if the model bases chooses ads, in part, on what ads the user has previously seen.
  - Consider a dataset that contains retail sales information for a year. User's purchases change seasonally, which would violate stationarity.
  - When we know that any of the preceding three basic assumptions are violated, we must pay careful attention to metrics.
- When any of the preceding three basic assumptions are violated, we must pay careful attention to metrics.

Summary

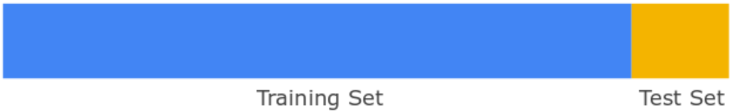
- Overfitting occurs when a model tries to fit the training data so closely that it does not generalize well to new data.
- If the key assumptions of supervised ML are not met, then we lose important theoretical guarantees on our ability to predict on new data.

Key Terms

- generalization
- overfitting
- prediction
- stationarity
- test set
- training set

Splitting Data

- We have introduced the idea of dividing the data set into two subsets:
  - training set—a subset to train a model.
  - test set—a subset to test the trained model.
- You could imagine slicing the single data set as follows:

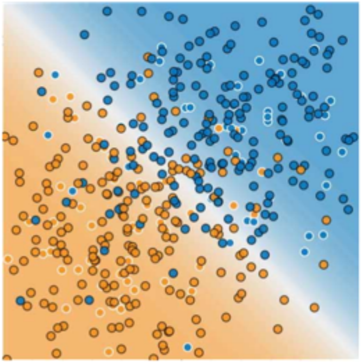


Splitting Data

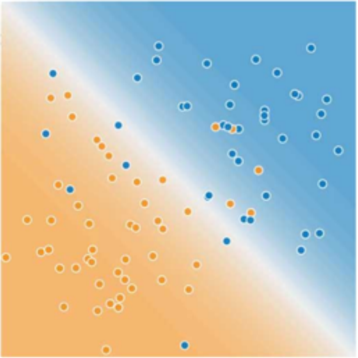
- Make sure that your test set meets the following two conditions:
  - Is large enough to yield statistically meaningful results.
  - Is representative of the data set as a whole. In other words, don't pick a test set with different characteristics than the training set.
- Assuming that your test set meets the preceding two conditions, your goal is to create a model that generalizes well to new data.
- Our test set serves as a proxy for new data.

Splitting Data

- For example, consider the following figure.



Training Data



Test Data

Splitting Data

- Notice that the model learned for the training data is very simple.
- This model doesn't do a perfect job – a few predictions are wrong.
- However, this model does about as well on the test data as it does on the training data.
- In other words, this simple model does not overfit the training data.

Splitting Data

- **Never** train on test data.
- If you see surprisingly good results on your evaluation metrics, it might be a sign that you are accidentally training on the test set.
- For example, high accuracy might indicate that test data has leaked into the training set.

Splitting Data

- For example, consider a model that predicts whether an email is spam, using the subject line, email body, and sender's email address as features.
- We split the data into training and test sets, with an 80-20 split. After training, the model achieves 99% accuracy on both the training set and the test set.
- We'd expect a lower accuracy on the test set, so we take another look at the data and discover that many of the examples in the test set are duplicates of examples in the training set.
- We've inadvertently trained on some of our test data, and as a result, we're no longer accurately measure how well our model generalizes to new data.

Key Terms

- overfitting
- test set
- training sets