UNC CHARLOTTE

The WILLIAM STATES LEE COLLEGE *of* ENGINEERING

Introduction to ML
Lecture 6: Generalization

Hamed Tabkhi
Department of Electrical and Computer Engineering,
University of North Carolina Charlotte (UNCC)
htabkhiv@uncc.edu

Generalization

- As machine learning scientists/engineers, our goal is to discover *patterns*. But how can we be sure that we have truly discovered a *general* pattern and not simply memorized our data?
  - We don't want to predict yesterday's stock prices, but tomorrow's.
  - We don't need to recognize already diagnosed diseases for previously seen patients, but rather previously undiagnosed ailments in previously unseen patients.
- This problem—how to discover patterns that *generalize*—is the fundamental problem of machine learning, and arguably of all of statistics.

- We might cast this problem as just one slice of a far grander question that engulfs all of science:

When are we ever justified in making the leap from particular observations to more general statements?

Generalization

- In real life, we must fit out models using a finite collection of data.
  - For many important medical problem, we can only access a few thousand data points. When studying rare diseases, we might be lucky to access hundreds.
- However, even at this extreme scale, the number of available data points remains infinitesimally small compared to the space of all possible images at 1 megapixel resolution.
- Whenever we work with finite samples, we must keep in mind the risk that we might fit our training data, only to discover that we failed to discover a generalizable pattern.
- The phenomenon of fitting closer to our training data than to the underlying distribution is called ***overfitting***, and techniques for combatting overfitting are often called *regularization* methods.
- While there is no substitute for a proper introduction to statistical learning theory, we will give you just enough intuition to get going.
- **Generalizations are group of techniques to combat overfitting!**

Generalization: Training Error and Generalization Error

- In the standard supervised learning setting, we assume that the training data and the test data are drawn *independently* from *identical* distributions.
  - This is commonly called the *IID assumption*.
  - It's worth noting that absent any such assumption. We would be dead in the water.

  - **In practice, we must *estimate* the generalization error by applying our model to an independent test set constituted of a random selection of examples $\mathbf{X}'$ and labels $\mathbf{y}'$ that were withheld from our training set.**

  - This consists of applying the same formula as for calculating the empirical training error but to a test set $\mathbf{X}'$; $\mathbf{y}'$.

Generalization: Model Complexity

- When we have simple models and abundant data, the training and generalization errors tend to be close.
- However, when we work with more complex models and/or fewer examples, we expect the training error to go down but the generalization gap to grow.

This should not be surprising!

- In general, we cannot conclude based on fitting the training data alone that our model has discovered any generalizable pattern
- On the other hand, if our model class was not capable of fitting arbitrary labels, then it must have discovered a pattern.

Karl Popper, an influential philosopher of science formalized the criterion of falsifiability:

"a theory that can explain any and all observations is not a scientific theory at all!"

Generalization: Model Complexity

- Does a complexity matter?
  - Often, models with more parameters are able to fit a greater number of arbitrarily assigned labels.
  - However, a model which is capable of fitting arbitrary labels, low training error does not necessarily imply low generalization error. *However, it does not necessarily imply high generalization error either!*

- All we can say confidently is that low training error alone is not enough to certify low generalization error!
- Deep neural networks turn out to be just such models:
  - While they generalize well in practice, they are too powerful to allow us to conclude much on the basis of training error alone.
  - In these cases, we must rely more heavily on our holdout data to certify generalization after the fact.

Error on the holdout data, i.e., validation set, is called the *validation error*

.

Generalization: Underfitting or Overfitting?

- When we compare the training and validation errors, we want to be mindful of two common situations:

- If the model is unable to reduce the training error, that could mean that our model is too simple (i.e., insufficiently expressive) to capture the pattern that we are trying to model.
- This phenomenon is known as ***underfitting***.

- On the other hand, as we discussed above, we want to watch out for the cases when our training error is significantly lower than our validation error, **severe *overfitting***.
  - Ultimately, we usually care about driving the generalization error lower, and only care about the gap insofar as it becomes an obstacle to that end!

Generalization: Dataset Size

- As the above bound already indicates, another big consideration to bear in mind is dataset size.
- Fixing our model, the fewer samples we have in the training dataset, the more likely (and more severely) we are to encounter overfitting.
- As we increase the amount of training data, the generalization error typically decreases.
- For a fixed task and data distribution, model complexity should not increase more rapidly than the amount of data.
- Given more data, we might attempt to fit a more complex model.
- Absent sufficient data, simpler models may be more difficult to beat.
- For many tasks, deep learning only outperforms linear models when many thousands of training examples are available.
- In part, the current success of deep learning owes considerably to the abundance of massive datasets arising from Internet companies, cheap storage, connected devices, and the broad digitization of the economy.