Computational Learning Theory: Occam's Razor

Machine Learning

Slides based on material from Dan Roth, Avrim Blum, Tom Mitchell and others

This lecture: Computational Learning Theory

- The Theory of Generalization
- Probably Approximately Correct (PAC) learning
- Positive and negative learnability results
- Agnostic Learning
- Shattering and the VC dimension

Where are we?

- The Theory of Generalization
  - When can be trust the learning algorithm?
  - What functions can be learned?
  - Batch Learning
- Probably Approximately Correct (PAC) learning
- Positive and negative learnability results
- Agnostic Learning
- Shattering and the VC dimension

This section

1. Analyze a simple algorithm for learning conjunctions

2. Define the PAC model of learning

3. Make formal connections to the principle of Occam's razor

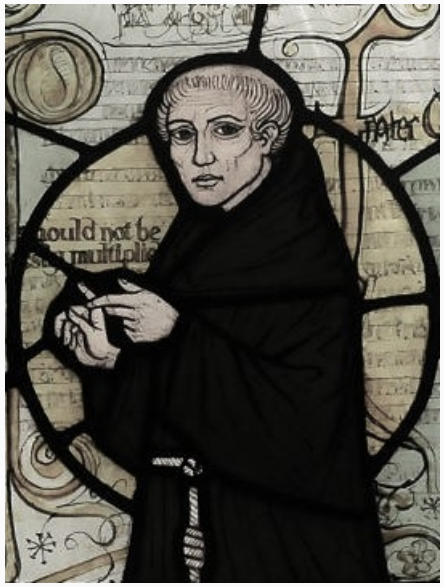
This section

- ✓ Analyze a simple algorithm for learning conjunctions
- ✓ Define the PAC model of learning
3. Make formal connections to the principle of Occam's razor

Occam's Razor

Named after William of Occam

- AD 1300s



Prefer simpler explanations over more complex ones

"Numquam ponenda est pluralitas sine necessitate"

(Never posit plurality without necessity.)

Historically, a widely prevalent idea across different schools of philosophy

Towards formalizing Occam's Razor

Claim: The probability that there is a hypothesis $h \in H$ that:

1. Is Consistent with m examples, and
2. Has $err_D(h) > \epsilon$

is less than $|H| (1 - \epsilon)^m$

Towards formalizing Occam's Razor

(Assuming consistency)

Claim: The probability that there is a hypothesis $h \in H$ that:

1. Is Consistent with m examples, and
2. Has $err_D(h) > \epsilon$

is less than $|H| (1 - \epsilon)^m$

Towards formalizing Occam's Razor

(Assuming consistency)

Claim: The probability that there is a hypothesis $h \in H$ that:

1. Is Consistent with m examples, and
2. Has $err_D(h) > \epsilon$

is less than $|H| (1 - \epsilon)^m$

That is, consistent yet bad

Towards formalizing Occam's Razor

(Assuming consistency)

Claim: The probability that there is a hypothesis $h \in H$ that:

1. Is Consistent with m examples, and
2. Has $err_D(h) > \epsilon$

is less than $|H| (1 - \epsilon)^m$

That is, consistent yet bad

Proof: Let h be such a bad hypothesis that has an error > ϵ

Towards formalizing Occam's Razor

(Assuming consistency)

Claim: The probability that there is a hypothesis $h \in H$ that:

1. Is Consistent with m examples, and
2. Has $err_D(h) > \epsilon$

is less than $|H| (1 - \epsilon)^m$

That is, consistent yet bad

Proof: Let h be such a bad hypothesis that has an error > ϵ

Probability that h is consistent with one example is $Pr[f(x) = h(x)] < 1 - \epsilon$

Towards formalizing Occam's Razor

(Assuming consistency)

Claim: The probability that there is a hypothesis $h \in H$ that:

1. Is Consistent with m examples, and
2. Has $err_D(h) > \epsilon$

is less than $|H| (1 - \epsilon)^m$

That is, consistent yet bad

Proof: Let h be such a bad hypothesis that has an error > ϵ

Probability that h is consistent with one example is $Pr[f(x) = h(x)] < 1 - \epsilon$

The training set consists of m examples drawn independently

So, probability that h is consistent with m examples < $(1 - \epsilon)^m$

Towards formalizing Occam's Razor

(Assuming consistency)

Claim: The probability that there is a hypothesis $h \in H$ that:

1. Is Consistent with m examples, and
2. Has $err_D(h) > \epsilon$

is less than $|H| (1 - \epsilon)^m$

That is, consistent yet bad

Proof: Let h be such a bad hypothesis that has an error > ϵ

Probability that h is consistent with one example is $Pr[f(x) = h(x)] < 1 - \epsilon$

The training set consists of m examples drawn independently

So, probability that h is consistent with m examples < $(1 - \epsilon)^m$

Probability that *some bad hypothesis* in H is consistent with m examples is less than $|H| (1 - \epsilon)^m$

Occam's Razor

The probability that there is a hypothesis $h \in H$ that is

1. Consistent with m examples, and
2. Has $err_D(h) > \epsilon$

is less than $|H| (1 - \epsilon)^m$

Occam's Razor

The probability that there is a hypothesis $h \in H$ that is

1. Consistent with m examples, and
2. Has $err_D(h) > \epsilon$

is less than $|H| (1 - \epsilon)^m$

Just like before, we want to make this probability small, say smaller than δ

Occam's Razor

The probability that there is a hypothesis $h \in H$ that is

1. Consistent with m examples, and
2. Has $err_D(h) > \epsilon$

is less than $|H| (1 - \epsilon)^m$

Just like before, we want to make this probability small, say smaller than δ

$$|H| (1 - \epsilon)^m < \delta$$

Occam's Razor

The probability that there is a hypothesis $h \in H$ that is

1. Consistent with m examples, and
2. Has $err_D(h) > \epsilon$

is less than $|H| (1 - \epsilon)^m$

Just like before, we want to make this probability small, say smaller than δ

$$|H| (1 - \epsilon)^m < \delta$$

$$\ln(|H|) + m \ln(1 - \epsilon) < \ln \delta$$

Occam's Razor

The probability that there is a hypothesis $h \in H$ that is

1. Consistent with m examples, and
2. Has $err_D(h) > \epsilon$

is less than $|H| (1 - \epsilon)^m$

Just like before, we want to make this probability small, say smaller than δ

$$|H| (1 - \epsilon)^m < \delta$$

$$\ln(|H|) + m \ln(1 - \epsilon) < \ln \delta$$

We know that $e^{-x} = 1 - x + \frac{x^2}{2} - \frac{x^3}{6} \cdots > 1 - x$

Let's use $\ln(1 - \epsilon) < -\epsilon$ to get a safer δ

Occam's Razor

The probability that there is a hypothesis $h \in H$ that is

1. Consistent with m examples, and
2. Has $err_D(h) > \epsilon$

is less than $|H| (1 - \epsilon)^m$

Just like before, we want to make this probability small, say smaller than δ

$$|H| (1 - \epsilon)^m < \delta$$

$$\ln(|H|) + m \ln(1 - \epsilon) < \ln \delta$$

We know that $e^{-x} = 1 - x + \frac{x^2}{2} - \frac{x^3}{6} \cdots > 1 - x$

Let's use $\ln(1 - \epsilon) < -\epsilon$ to get a safer δ

That is, if $m > \frac{1}{\epsilon}\left(\ln(|H|) + \ln\frac{1}{\delta}\right)$ then, the probability of getting a bad hypothesis is small

Occam's Razor

Let H be any hypothesis space.

With probability 1 - δ, a hypothesis h $\in$ H that is consistent with a training set of size m will have an error < ϵ on future examples if

$$m > \frac{1}{\epsilon}\left(\ln(|H|) + \ln\frac{1}{\delta}\right)$$

Occam's Razor

Let H be any hypothesis space.

With probability 1 - δ, a hypothesis h $\in$ H that is consistent with a training set of size m will have an error < ϵ on future examples if

$$m > \frac{1}{\epsilon}\left(\ln(|H|) + \ln\frac{1}{\delta}\right)$$

1. Expecting lower error increases sample complexity (i.e more examples needed for the guarantee)

Occam's Razor

Let H be any hypothesis space.

With probability 1 - δ, a hypothesis h $\in$ H that is consistent with a training set of size m will have an error < ϵ on future examples if

$$m > \frac{1}{\epsilon}\left(\ln(|H|) + \ln\frac{1}{\delta}\right)$$

1. Expecting lower error increases sample complexity (i.e more examples needed for the guarantee)

2. If we have a larger hypothesis space, then we will make learning harder (i.e higher sample complexity)

Occam's Razor

Let H be any hypothesis space.

With probability 1 - δ, a hypothesis h $\in$ H that is consistent with a training set of size m will have an error < ϵ on future examples if

$$m > \frac{1}{\epsilon}\left(\ln(|H|) + \ln\frac{1}{\delta}\right)$$

1. Expecting lower error increases sample complexity (i.e more examples needed for the guarantee)

2. If we have a larger hypothesis space, then we will make learning harder (i.e higher sample complexity)

3. If we want a higher confidence in the classifier we will produce, sample complexity will be higher.

Occam's Razor

Let H be any hypothesis space.

With probability 1 - δ, a hypothesis h $\in$ H that is consistent with a training set of size m will have an error < ϵ on future examples if

$$m > \frac{1}{\epsilon}\left(\ln(|H|) + \ln\frac{1}{\delta}\right)$$

Occam's Razor

Let H be any hypothesis space.

With probability 1 - δ, a hypothesis h $\in$ H that is consistent with a training set of size m will have an error < ϵ on future examples if

$$m > \frac{1}{\epsilon}\left(\ln(|H|) + \ln\frac{1}{\delta}\right)$$

This is called the Occam's Razor because it expresses a preference towards smaller hypothesis spaces.

Occam's Razor

Let H be any hypothesis space.

With probability 1 - δ, a hypothesis h $\in$ H that is consistent with a training set of size m will have an error < ϵ on future examples if

$$m > \frac{1}{\epsilon}\left(\ln(|H|) + \ln\frac{1}{\delta}\right)$$

This is called the Occam's Razor because it expresses a preference towards smaller hypothesis spaces.

Shows when a m-consistent hypothesis generalizes well (i.e error < ϵ).

Occam's Razor

Let H be any hypothesis space.

With probability 1 - δ, a hypothesis h $\in$ H that is consistent with a training set of size m will have an error < ϵ on future examples if

$$m > \frac{1}{\epsilon}\left(\ln(|H|) + \ln\frac{1}{\delta}\right)$$

This is called the Occam's Razor because it expresses a preference towards smaller hypothesis spaces.

Shows when a m-consistent hypothesis generalizes well (i.e error < ϵ).

Complicated/larger hypothesis spaces are not necessarily bad. But simpler ones are unlikely to fool us by being consistent with many examples!

Consistent Learners and Occam's Razor

From the definition, we get the following general scheme for PAC learning

Given a sample D of m examples

- Find some $h \in H$ that is consistent with ***all*** m examples
  - If m is large enough, a consistent hypothesis must be *close enough* to f
  - Check that m does not have to be too large (i.e polynomial in the relevant parameters): we showed that the "closeness" guarantee requires that

$$m > 1/\epsilon \ (\ln |H| + \ln 1/\delta)$$

- Show that the consistent hypothesis $h \in H$ can be computed efficiently

Consistent Learners and Occam's Razor

From the definition, we get the following general scheme for PAC learning

Given a sample D of m examples

- Find some $h \in H$ that is consistent with ***all*** m examples
  - If m is large enough, a consistent hypothesis must be *close enough* to f
  - Check that m does not have to be too large (i.e polynomial in the relevant parameters): we showed that the "closeness" guarantee requires that
    $$m > 1/\epsilon \, (\ln |H| + \ln 1/\delta)$$
- Show that the consistent hypothesis $h \in H$ can be computed efficiently

We worked out the details for conjunctions

- The Elimination algorithm to find a hypothesis h that is consistent with the training set (easy to compute)
- We showed directly that if we have sufficiently many examples (polynomial in the parameters), than h is close to the target function.

Exercises

We have seen the decision tree learning algorithm. Suppose our problem has n binary features. What is the size of the hypothesis space?

Are decision trees efficiently PAC learnable?