Computational Learning Theory: Occam's Razor

Machine Learning

THE UNIVERSITY OF UTAH

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

This section

1. Define the PAC model of learning

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

This section

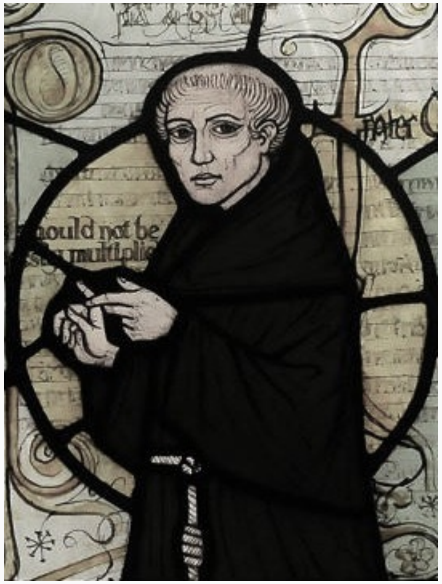
✓ Define the PAC model of learning

2. 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

Why would a *consistent* learner fail?

Consistent learner: Suppose we have a learner that produces a hypothesis that is consistent with a training set...

Why would a *consistent* learner fail?

Consistent learner: Suppose we have a learner that produces a hypothesis that is consistent with a training set...

… but the training set is not a representative sample of the instance space.

Then the hypothesis we learned could be bad even if it is consistent with the entire training set.

Why would a *consistent* learner fail?

Consistent learner: Suppose we have a learner that produces a hypothesis that is consistent with a training set...

… but the training set is not a representative sample of the instance space.

Then the hypothesis we learned could be bad even if it is consistent with the entire training set.

We can try to
1. quantify the probability of such a bad situation occurring and,
2. then, ask: *What will it take for this probability to be low?*

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

Recall that $Err_D(h) = \Pr[f(x) \neq h(x)]$

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 $Err_D(h) > \epsilon$

That is, consistent yet bad

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

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

Union bound

For a set of events, the probability that at least one of them happens < the sum of the probabilities of the individual events

Occam's Razor for consistent hypotheses

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

This situation is a bad one. Let us try to see what we need to do to ensure that this situation is rare.

Occam's Razor for consistent hypotheses

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

This situation is a bad one. Let us try to see what we need to do to ensure that this situation is rare.

We want to make this probability small, say smaller than δ

Occam's Razor for consistent hypotheses

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

This situation is a bad one. Let us try to see what we need to do to ensure that this situation is rare.

We want to make this probability small, say smaller than δ

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

Occam's Razor for consistent hypotheses

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

This situation is a bad one. Let us try to see what we need to do to ensure that this situation is rare.

We want to make this probability small, say smaller than δ

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

Occam's Razor for consistent hypotheses

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

This situation is a bad one. Let us try to see what we need to do to ensure that this situation is rare.

We want to make this probability small, say smaller than δ

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

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

If δ is small, then the probability that there is a consistent, yet bad hypothesis would also be small (because of this inequality)

Occam's Razor for consistent hypotheses

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

We want to make this probability small, say smaller than δ

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

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

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

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

Occam's Razor for consistent hypotheses

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

We want to make this probability small, say smaller than δ

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

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

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

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

Occam's Razor for consistent hypotheses

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

We want to make this probability small, say smaller than δ

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

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

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

Let's use $\log(1-\epsilon) < -\epsilon$ 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 for consistent hypotheses

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

We want to make this probability small, say smaller than δ

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

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

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

Let's use $\log(1-\epsilon) < -\epsilon$ 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

If this is true

Occam's Razor for consistent hypotheses

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

We want to make this probability small, say smaller than δ

Then, this holds

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

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

Let's use $\log(1-\epsilon) < -\epsilon$ 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

If this is true

Occam's Razor for consistent hypotheses

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

Then, this is improbable

We want to make this probability small, say smaller than δ

Then, this holds

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

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

Let's use $\log(1-\epsilon) < -\epsilon$ 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

If this is true

Occam’s Razor for consistent hypotheses

Let H be any hypothesis space.

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

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

Occam's Razor for consistent hypotheses

Let H be any hypothesis space.

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

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

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

Occam's Razor for consistent hypotheses

Let H be any hypothesis space.

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

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

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 for consistent hypotheses

Let H be any hypothesis space.

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

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

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 for consistent hypotheses

Let H be any hypothesis space.

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

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

Occam's Razor for consistent hypotheses

Let H be any hypothesis space.

With probability $1 - \delta$, a hypothesis $h \in H$ that is consistent with a training set of size m will have an error $< \epsilon$ 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 for consistent hypotheses

Let H be any hypothesis space.

With probability $1 - \delta$, a hypothesis $h \in H$ that is consistent with a training set of size m will have an error $< \epsilon$ 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 $< \epsilon$).

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 set of m training 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

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

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

Exercises

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

2. Are decision trees efficiently PAC learnable?

3. Are conjunctions PAC learnable? Can you think of a a PAC algorithm for monotone conjunctions?