Computational Learning Theory: Agnostic Learning

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

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

So far we have seen…

- The general setting for batch learning

- PAC learning and Occam's Razor
  - How good will a classifier that is ***consistent*** on a training set be?

So far we have seen…

- The general setting for batch learning

- PAC learning and Occam's Razor
  - How good will a classifier that is ***consistent*** on a training set be?

- Assumptions so far:
  1. Training and test examples come from the same distribution
  2. The hypothesis space is finite.
  3. For any concept, there is some function in the hypothesis space that is consistent with the training set

So far we have seen…

- The general setting for batch learning

- PAC learning and Occam's Razor
  - How good will a classifier that is ***consistent*** on a training set be?

- Assumptions so far:
  1. Training and test examples come from the same distribution
  2. The hypothesis space is finite.
  3. For any concept, there is some function in the hypothesis space that is consistent with the training set

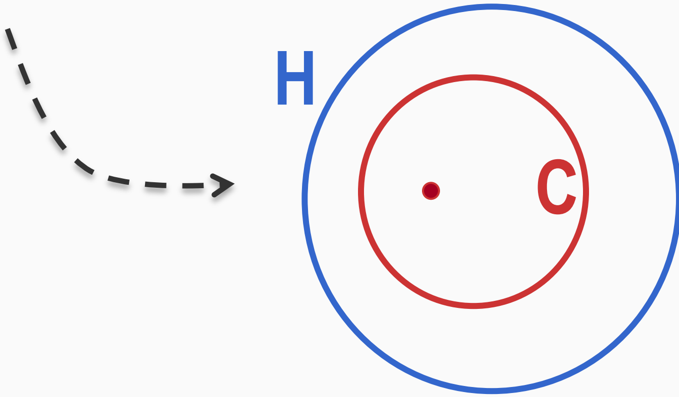
Let's look at the last assumption. Is it reasonable?

What is agnostic learning?

- So far, we have assumed that the learning algorithm could find the true concept

What is agnostic learning?

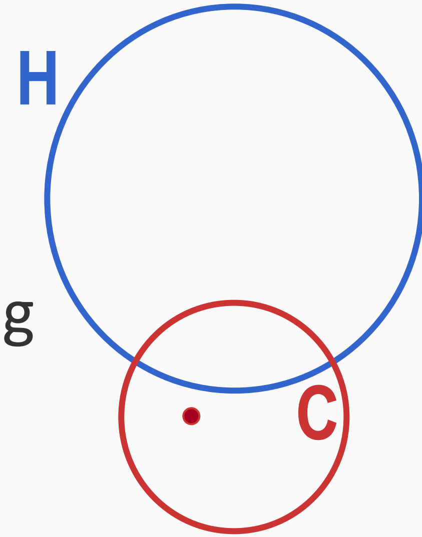
- So far, we have assumed that the learning algorithm could find the true concept



What is agnostic learning?

- So far, we have assumed that the learning algorithm could find the true concept

- What if: We are trying to learn a concept f using hypotheses in H, but f ∉ H
  - That is C is not a subset of H
  - This setting is called *agnostic learning*
  - Can we say something about sample complexity?



More realistic setting than before

Agnostic Learning

Learn a concept f using hypotheses in H, but f $\notin$ H

Are we guaranteed that training error will be zero?

- **No**. There may be no consistent hypothesis in the hypothesis space!

Agnostic Learning

Learn a concept f using hypotheses in H, but f $\notin$ H

Are we guaranteed that training error will be zero?

- **No**. There may be no consistent hypothesis in the hypothesis space!

We can find a classifier $h \in H$ that has low *training* error

$$\mathrm{err}_s(h) = \frac{|\{f(x) \neq h(x) : x \in S\}|}{m}$$

This is the fraction of training examples that are misclassified

Agnostic Learning

Learn a concept f using hypotheses in H, but f ∉ H

We can find a classifier $h \in H$ that has low *training* error

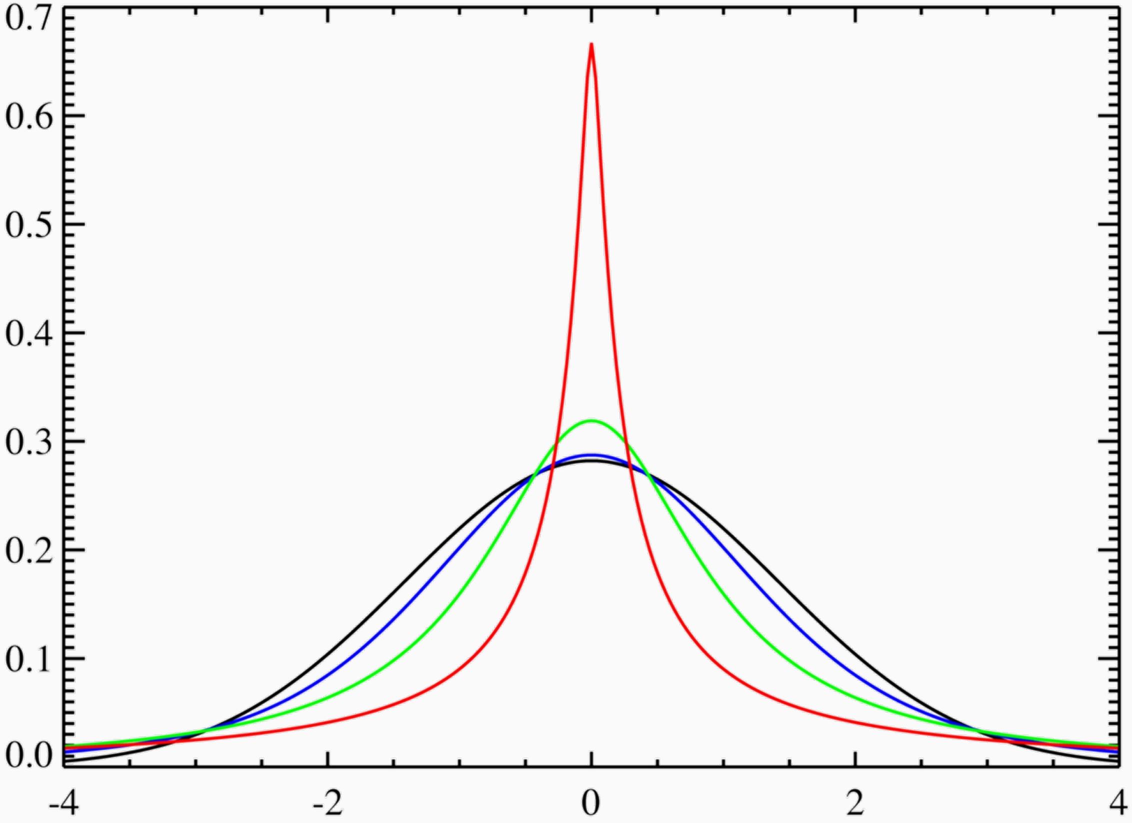
$$\mathrm{err}_S(h) = \frac{|\{f(x) \neq h(x): x \in S\}|}{m}$$

What we want: A guarantee that a hypothesis with small training error will have a good accuracy on unseen examples

$$\mathrm{err}_D(h) = \Pr_{x \sim D}[f(x) \neq h(x)]$$

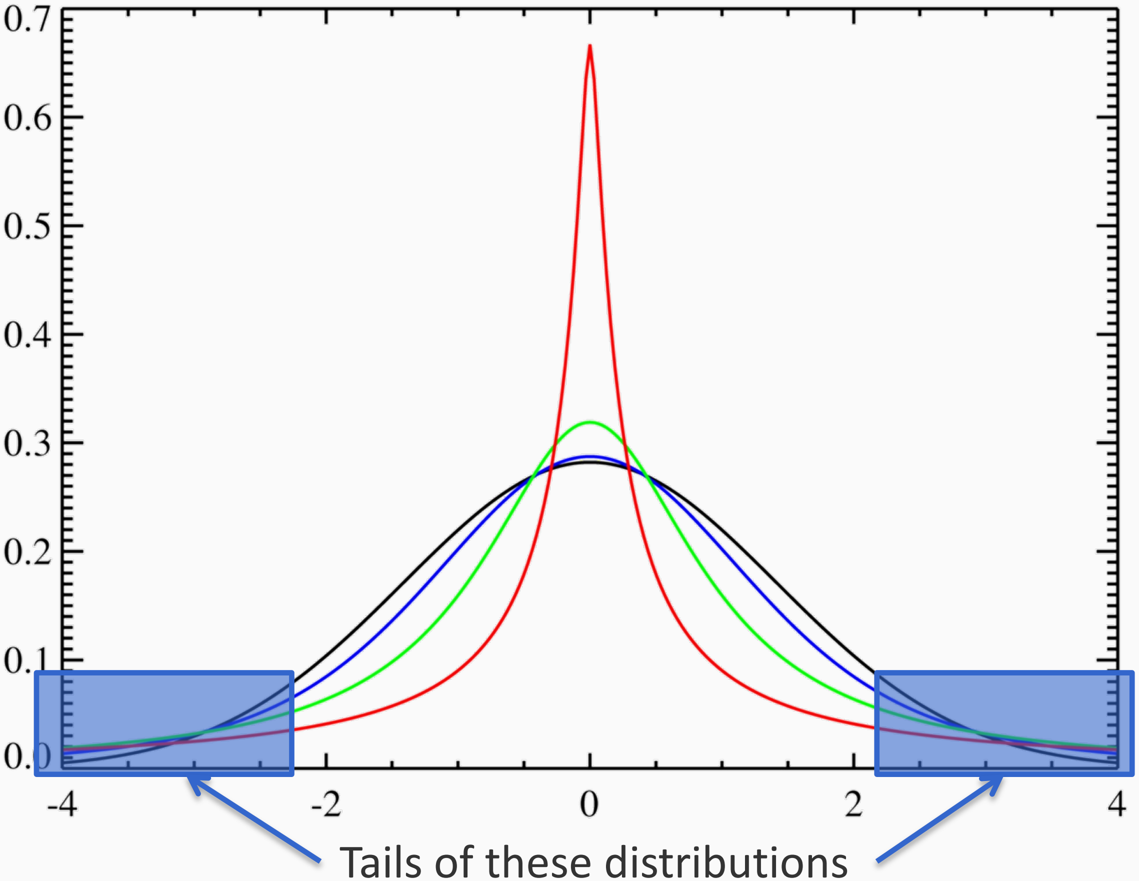
We will use *Tail bounds* for analysis

How far can a random variable get from its mean?



We will use *Tail bounds* for analysis

How far can a random variable get from its mean?



Tails of these distributions

Bounding probabilities

Law of large numbers: As we collect more samples, the empirical average converges to the true expectation

- Suppose we have an unknown coin and we want to estimate its bias (i.e. probability of heads)
- Toss the coin m times

$$\frac{\text{number of heads}}{m} \to \mathrm{P(heads)}$$

As m increases, we get a better estimate of $\mathrm{P(heads)}$

What can we say about the gap between these two terms?

Bounding probabilities

- Markov's inequality: Bounds the probability that a non-negative random variable exceeds a fixed value

$$P[X \geq a] \leq \frac{E[X]}{a}$$

- Chebyshev's inequality: Bounds the probability that a random variable differs from its expected value by more than a fixed number of standard deviations

$$P[|X - \mu| \geq k\sigma] \leq \frac{1}{k^2}$$

What we want: To bound sums of random variables

- Why? Because the training error depends on the number of errors on the training set

Hoeffding’s inequality

Upper bounds on how much the sum of a set of random variables differs from its expected value

$$P[p > \bar{p} + \epsilon] \leq e^{-2m\epsilon^2}$$

Hoeffding’s inequality

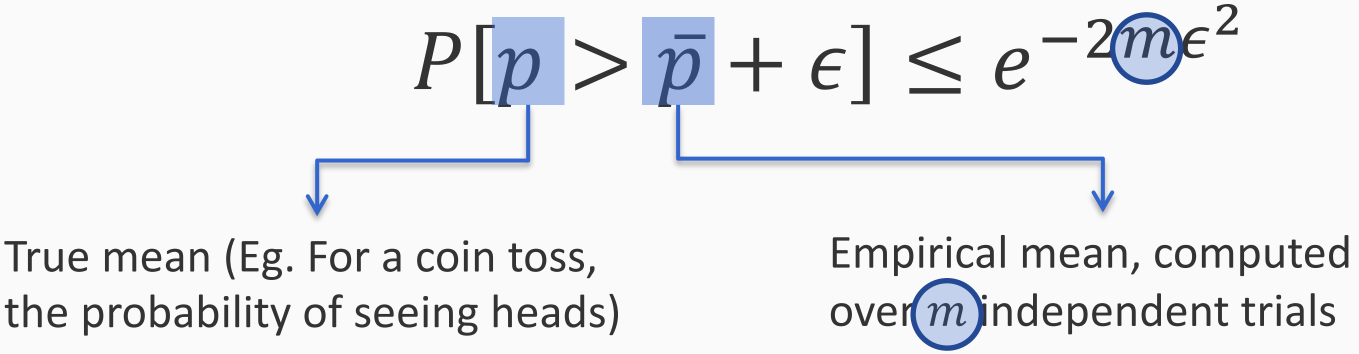
Upper bounds on how much the sum of a set of random variables differs from its expected value

$$P[p > \bar{p} + \epsilon] \leq e^{-2m\epsilon^2}$$

True mean (Eg. For a coin toss, the probability of seeing heads)

Hoeffding’s inequality

Upper bounds on how much the sum of a set of random variables differs from its expected value



$$P[p > \bar{p} + \epsilon] \leq e^{-2m\epsilon^2}$$

True mean (Eg. For a coin toss, the probability of seeing heads)

Empirical mean, computed over m independent trials

Hoeffding’s inequality

Upper bounds on how much the sum of a set of random variables differs from its expected value

$$P[p > \bar{p} + \epsilon] \leq e^{-2m\epsilon^2}$$

True mean (Eg. For a coin toss, the probability of seeing heads)

Empirical mean, computed over m independent trials

The probability that the true mean will be more than ϵ away from the empirical mean, computed over m trials

Hoeffding’s inequality

Upper bounds on how much the sum of a set of random variables differs from its expected value

$$P[p > \bar{p} + \epsilon] \leq e^{-2m\epsilon^2}$$

True mean (Eg. For a coin toss, the probability of seeing heads)

Empirical mean, computed over m independent trials

What this tells us: The empirical mean will not be too far from the expected mean if there are many samples.

And, it quantifies the convergence rate as well.

Back to agnostic learning

Suppose we consider the true error (a.k.a generalization error) $Err_D(h)$ to be a random variable

Back to agnostic learning

Suppose we consider the true error (a.k.a generalization error) $Err_D(h)$ to be a random variable

The training error over m examples $Err_S(h)$ is the empirical estimate of this true error

Back to agnostic learning

Suppose we consider the true error (a.k.a generalization error) $Err_D(h)$ to be a random variable

The training error over m examples $Err_S(h)$ is the empirical estimate of this true error

We can ask: What is the probability that the true error is more than ϵ away from the empirical error?

Back to agnostic learning

Suppose we consider the true error (a.k.a generalization error) $Err_D(h)$ to be a random variable

The training error over m examples $Err_S(h)$ is the empirical estimate of this true error

Let’s apply Hoeffding’s inequality

Back to agnostic learning

Suppose we consider the true error (a.k.a generalization error) $Err_D(h)$ to be a random variable

The training error over m examples $Err_S(h)$ is the empirical estimate of this true error

Let's apply Hoeffding's inequality

$$P[Err_D(h) > Err_S(h) + \epsilon] \leq e^{-2m\epsilon^2}$$

Back to agnostic learning

Suppose we consider the true error (a.k.a generalization error) $Err_D(h)$ to be a random variable

The training error over m examples $Err_S(h)$ is the empirical estimate of this true error

Let's apply Hoeffding's inequality

$$P[Err_D(h) > Err_S(h) + \epsilon] \leq e^{-2m\epsilon^2}$$

$$Err_D(h) = Pr_{x \sim D}[f(x) \neq h(x)]$$

$$Err_S(h) = \frac{|\{f(x) \neq h(x), x \in S\}|}{m}$$

Agnostic learning

The probability that a single hypothesis h has a training error that is more than ϵ away from the true error is bounded above

$$P[Err_D(h) > Err_S(h) + \epsilon] \leq e^{-2m\epsilon^2}$$

Agnostic learning

The probability that a single hypothesis h has a training error that is more than ϵ away from the true error is bounded above

$$P[Err_D(h) > Err_S(h) + \epsilon] \leq e^{-2m\epsilon^2}$$

The learning algorithm looks for the best one of the $|H|$ possible hypotheses

Agnostic learning

The probability that a single hypothesis h has a training error that is more than ϵ away from the true error is bounded above

$$P[Err_D(h) > Err_S(h) + \epsilon] \leq e^{-2m\epsilon^2}$$

The learning algorithm looks for the best one of the $|H|$ possible hypotheses

The probability that there *exists* a hypothesis in H whose training error is ϵ away from the true error is bounded above

$$P\left[\text{for } \mathit{some}\ h \in H,\quad \text{we have } Err_D(h) > Err_S(h) + \epsilon\right] \leq |H|e^{-2m\epsilon^2}$$

Union bound

Agnostic learning

The probability that there *exists* a hypothesis in H whose training error is ϵ away from the true error is bounded above

$$P\left[\text{for } \mathit{some}\ h \in H, \quad \text{we have } Err_D(h) > Err_S(h) + \epsilon\right] \leq |H|e^{-2m\epsilon^2}$$

Agnostic learning

The probability that there *exists* a hypothesis in H whose training error is ϵ away from the true error is bounded above

$$P\left[\text{for } \mathit{some}\ h \in H, \quad \text{we have } Err_D(h) > Err_S(h) + \epsilon\right] \leq |H|e^{-2m\epsilon^2}$$

$$P\left[\begin{array}{l}\textit{Some hypothesis we are considering has generalization error}\\ \textit{that is much worse than the training error.}\end{array}\right] \leq |H|e^{-2m\epsilon^2}$$

Agnostic learning

The probability that there *exists* a hypothesis in H whose training error is ϵ away from the true error is bounded above

$$P\left[\text{for } \mathit{some}\ h \in H, \quad \text{we have } Err_D(h) > Err_S(h) + \epsilon\right] \leq |H|e^{-2m\epsilon^2}$$

$$P\left[\begin{array}{l}\textit{Some hypothesis we are considering has generalization error}\\ \textit{that is much worse than the training error.}\end{array}\right] \leq |H|e^{-2m\epsilon^2}$$

↓

This is an undesirable situation because our learner may end up picking this hypothesis.

Let us see what it takes to make this an improbable situation

Agnostic learning

The probability that there *exists* a hypothesis in H whose training error is ϵ away from the true error is bounded above

$$P\left[\text{for } \mathit{some}\ h \in H, \quad \text{we have } Err_D(h) > Err_S(h) + \epsilon\right] \leq |H|e^{-2m\epsilon^2}$$

Same game as before: We want this probability to be smaller than δ

Agnostic learning

The probability that there *exists* a hypothesis in H whose training error is ϵ away from the true error is bounded above

$$P\left[\text{for } \mathit{some}\ h \in H, \quad \text{we have } Err_D(h) > Err_S(h) + \epsilon\right] \leq |H|e^{-2m\epsilon^2}$$

Same game as before: We want this probability to be smaller than δ

$$|H|e^{-2m\epsilon^2} \leq \delta$$

Agnostic learning

The probability that there *exists* a hypothesis in H whose training error is ϵ away from the true error is bounded above

$$P\left[\text{for } \mathit{some}\ h \in H, \quad \text{we have } Err_D(h) > Err_S(h) + \epsilon\right] \leq |H|e^{-2m\epsilon^2}$$

Same game as before: We want this probability to be smaller than δ

$$|H|e^{-2m\epsilon^2} \leq \delta$$

Rearranging this gives us

$$m \geq \frac{1}{2\epsilon^2}\left[\ln|H| + \ln\left(\frac{1}{\delta}\right)\right]$$

Agnostic learning: Interpretations

1. An agnostic learner makes no commitment to whether f is in H and returns the hypothesis with least training error over at least m examples.

Agnostic learning: Interpretations

1. An agnostic learner makes no commitment to whether f is in H and returns the hypothesis with least training error over at least m examples.

   It can guarantee with probability $1 - \delta$ that the true/generalization error is *not* off by more than ϵ from the training error if

$$m \geq \frac{1}{2\epsilon^2}\left[\ln|H| + \ln\left(\frac{1}{\delta}\right)\right]$$

Agnostic learning: Interpretations

1. An agnostic learner makes no commitment to whether f is in H and returns the hypothesis with least training error over at least m examples.

   It can guarantee with probability $1-\delta$ that the true/generalization error is *not* off by more than ϵ from the training error if

$$m \geq \frac{1}{2\epsilon^2}\left[\ln|H| + \ln\left(\frac{1}{\delta}\right)\right]$$

Difference between generalization and training errors: How much worse will the classifier be in the future than it is at training time?

Agnostic learning: Interpretations

1. An agnostic learner makes no commitment to whether f is in H and returns the hypothesis with least training error over at least m examples.

   It can guarantee with probability $1 - \delta$ that the true/generalization error is *not* off by more than ϵ from the training error if

$$m \geq \frac{1}{2\epsilon^2}\left[\ln|H| + \ln\left(\frac{1}{\delta}\right)\right]$$

Difference between generalization and training errors: How much worse will the classifier be in the future than it is at training time?

Size of the hypothesis class: Again an Occam's razor argument – prefer smaller sets of functions

Agnostic learning: Interpretations

1. An agnostic learner makes no commitment to whether f is in H and returns the hypothesis with least training error over at least m examples.

   It can guarantee with probability $1 - \delta$ that the true/generalization error is *not* off by more than ϵ from the training error if

$$m \geq \frac{1}{2\epsilon^2}\left[\ln|H| + \ln\left(\frac{1}{\delta}\right)\right]$$

2. We have a *generalization bound*: A bound on how much the true error will deviate from the training error. If we have more than m examples, then with high probability (more than $1 - \delta$),

$$err_D(h) - err_S(h) \leq \sqrt{\frac{\ln|H| + \ln(1/\delta)}{2m}}$$

Generalization error   Training error

What we have seen so far

Occam's razor: When the hypothesis space contains the true concept

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

What we have seen so far

Occam’s razor: When the hypothesis space contains the true concept

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

Agnostic learning: When the hypothesis space may not contain the true concept

$$m \geq \frac{1}{2\epsilon^2}\left[\ln|H| + \ln\left(\frac{1}{\delta}\right)\right]$$

What we have seen so far

Occam's razor: When the hypothesis space contains the true concept

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

Agnostic learning: When the hypothesis space may not contain the true concept

$$m \geq \frac{1}{2\epsilon^2}\left[\ln|H| + \ln\left(\frac{1}{\delta}\right)\right]$$

Learnability depends on the log of the size of the hypothesis space

Have we solved everything? Eg: What about linear classifiers?