Vapnik-Chernovenkis Dimension

Hal Daumé III

CS5350: Machine Learning

06 November 2008

Today's Topics

1. Why VC dimension?
2. Shattering
3. Definition of VC dimension
4. VC dimension and generalization bounds
5. Relationship to margins and SVMs

Why VC dimension

Recall from PAC learning that Occam's bound states:

$$N \geq \frac{1}{b\epsilon}\left[\log|\mathcal{H}| + \log\frac{1}{\delta}\right]$$

Where:

- N is the sample complexity
- b is some constant
- $|\mathcal{H}|$ is the size of the hypothesis space
- ϵ, δ are user-set parameters (error and tolerance)

Big problem: what happens when $|\mathcal{H}|$ is infinite (as in most learners)?

Need a way to measure complexity of infinite hypothesis classes.

Shattering

Idea: measure complexity by the effective number of distinct functions that can be defined.

A hypothesis class is more complex if it can represent lots of possible classification decisions.

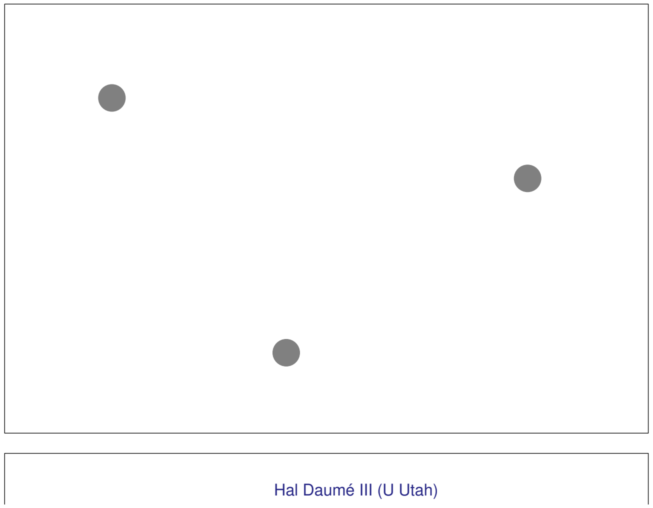
Def: A set of points is *shattered* by $\mathcal{H}$ if for all possible *binary labelings* of the points, there exists $h \in \mathcal{H}$ that can represent this decision.

We will measure complexity by the number of points that can be shattered by $\mathcal{H}$.

Shattering

Def: A set of points is *shattered* by $\mathcal{H}$ if for all possible *binary labelings* of the points, there exists $h \in \mathcal{H}$ that can represent this decision.

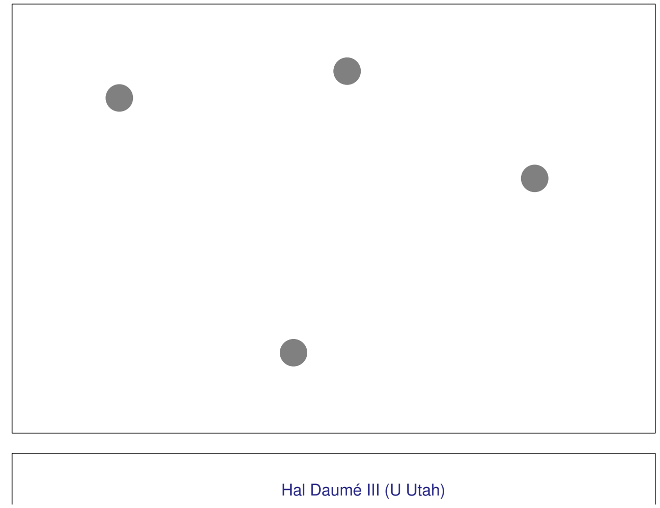
Consider $\mathcal{H}$=linear classifiers in $\mathbb{R}^2$



Shattering

Def: A set of points is *shattered* by $\mathcal{H}$ if for all possible *binary labelings* of the points, there exists $h \in \mathcal{H}$ that can represent this decision.

Consider $\mathcal{H}$=linear classifiers in $\mathbb{R}^2$



VC Dimension: The Game

VC dimension is a shattering game between us and an adversary.

To show that $\mathcal{H}$ has VC dimension $\geq d$:

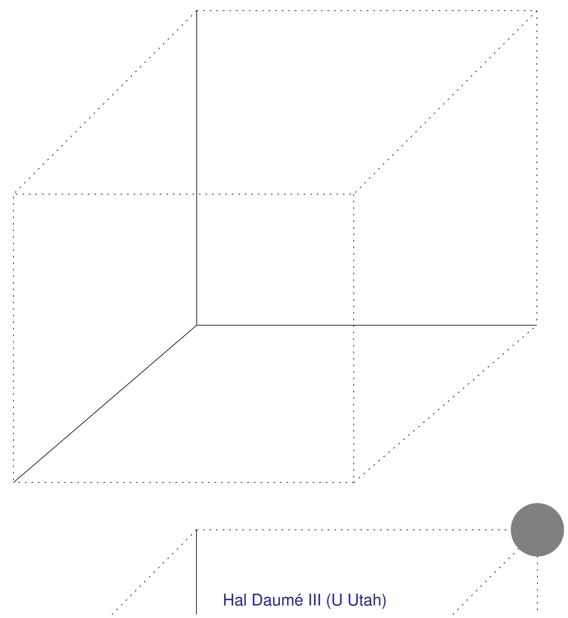
1. I choose d points positioned however I want
2. Adversary chooses a labeling of these d points
3. I choose $h \in \mathcal{H}$ that labels the points properly

The VC dimension of $\mathcal{H}$ is the maximum d I can choose so that I can always succeed in the game.

We just (informally) showed that the VC dimension of linear classifiers in $\mathbb{R}^2$ is . . . 3

VC Dimension in $\mathbb{R}^3$

What about linear classifiers in $\mathbb{R}^3$?



VC Dimension in $\mathbb{R}^D$

What about linear classifiers in $\mathbb{R}^3$?
$VC = 4$ seems like a reasonable guess...

What about $\mathbb{R}^D$? Yup, $VC = D + 1$

What about the VC dimension of a 1-nearest neighbor classifier?

...or an SVM with an RBF kernel?

VC Dimension and Generalization Bounds

Recall from PAC learning that Occam's bound states:

$$\text{ExpectedLoss}(h) \leq \text{TrainLoss}(h) + \sqrt{\frac{\log |\mathcal{H}| + \log \frac{1}{\delta}}{2N}}$$

If $\mathcal{H}$ is infinite, but has finite VC-dimension d, then:

$$\text{ExpectedLoss}(h) \leq \text{TrainLoss}(h) + \sqrt{\frac{d\left(\log \frac{2N}{d} + 1\right) + \log \frac{4}{\delta}}{N}}$$

Again, for linear classifiers, having fewer features is better!

Relationship to SVMs...

Thm (Vapnik, 1982)

Given:

- N data points $X = \{x_1, \dots x_N\}$,
- Living in $\mathbb{R}^D$,
- All with $||x_n|| \leq R$
- $\mathcal{H}_\gamma$ = set of linear classifiers in $\mathbb{R}^D$ with margin γ on X

Then, the VC dimension of $\mathcal{H}_\gamma$ is bounded by:

$$VC(\mathcal{H}_\gamma) \leq \min\left\{D, \left\lceil \frac{4R^2}{\gamma^2} \right\rceil\right\}$$

Relationship to SVMs...

Vapnik tells us:

$$VC(\mathcal{H}_\gamma) \leq \min\left\{D, \left\lceil \frac{4R^2}{\gamma^2} \right\rceil\right\}$$

This means that hypothesis spaces with large margin have small VC dimension.

From before, we know that:

$$\text{ExpectedLoss}(h) \leq \text{TrainLoss}(h) + \sqrt{\frac{VC(\mathcal{H}_\gamma)\left(\log \frac{2N}{VC(\mathcal{H}_\gamma)} + 1\right) + \log \frac{4}{\delta}}{N}}$$

Together, this means that large margin implies good generalization! Hooray!

Take home messages

- Standard PAC bounds only apply to finite hypothesis classes
- VC dimension is a measure of complexity of infinite hypothesis classes
- Based on the idea of shattering: how many points can we always correctly classify?
- Generalization now scales in terms of VC dimension
- Large margins imply small VC dimension