Performance of Empirical Models

Ravi Kothari, Ph.D.
ravi.kothari@ashoka.edu.in

"Correlation is not Causation.."

Aspects of Empirical Models

Aspects of Empirical Models

- Model complexity

Aspects of Empirical Models

- Model complexity
- Prediction error

Aspects of Empirical Models

- Model complexity
- Prediction error
    - Average case

Aspects of Empirical Models

- Model complexity
- Prediction error
  - Average case
  - Boosting

Aspects of Empirical Models

- Model complexity
- Prediction error
  - Average case
  - Boosting
  - Minimum Description Length

Aspects of Empirical Models

- Model complexity
- Prediction error
  - Average case
  - Boosting
  - Minimum Description Length
  - Uniform convergence

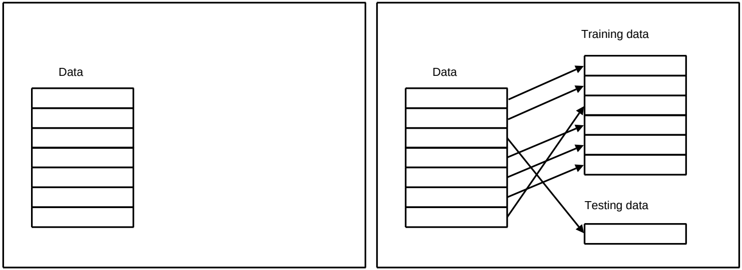
Aspects of Empirical Models

- Model complexity
- Prediction error
  - Average case
  - Boosting
  - Minimum Description Length
  - Uniform convergence
  - Worst case

Estimating the Prediction Error (1/2)

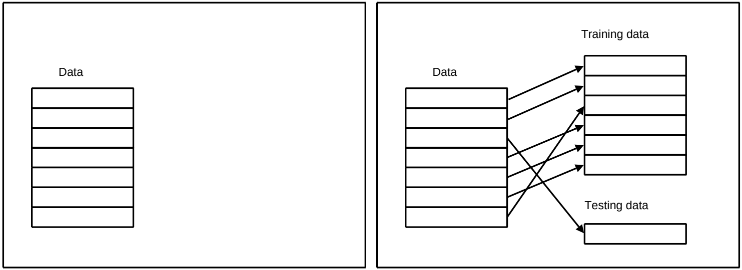
Estimating the Prediction Error (1/2)

- Sampling *without replacement* methods e.g. k-fold cross validation



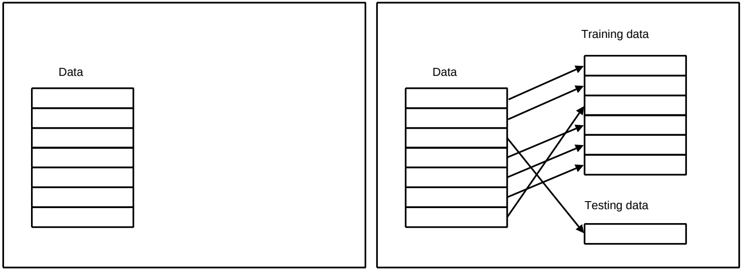
Estimating the Prediction Error (1/2)

- Sampling *without replacement* methods e.g. k-fold cross validation



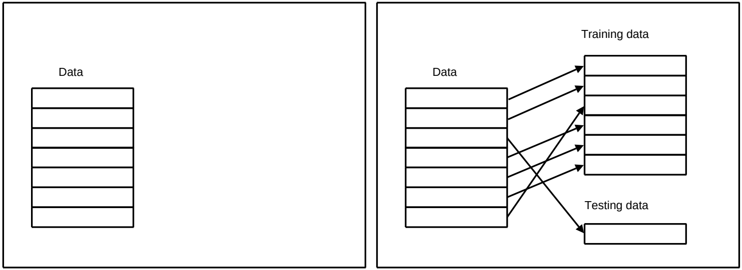
Estimating the Prediction Error (1/2)

- Sampling *without replacement* methods e.g. k-fold cross validation



Estimating the Prediction Error (1/2)

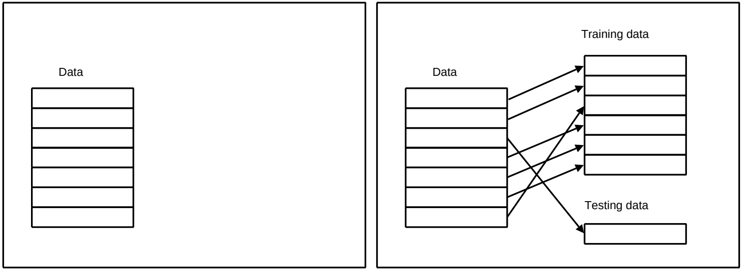
- Sampling *without replacement* methods e.g. k-fold cross validation



- Error estimate is the average of the k-error estimates

Estimating the Prediction Error (1/2)

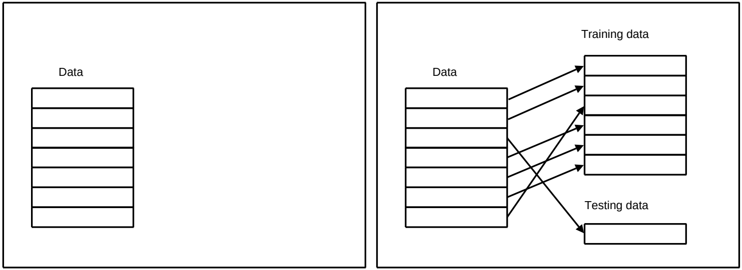
- Sampling *without replacement* methods e.g. k-fold cross validation



- Error estimate is the average of the k-error estimates
- When $k = N$, we get the leave-one-out estimate

Estimating the Prediction Error (1/2)

- Sampling *without replacement* methods e.g. k-fold cross validation



- Error estimate is the average of the k-error estimates
- When $k = N$, we get the leave-one-out estimate
- Time consuming, unbiased estimate with large variance

Estimating the Prediction Error (2/2)

Estimating the Prediction Error (2/2)

- Sampling *with* replacement e.g. bootstrap

Estimating the Prediction Error (2/2)

- Sampling *with* replacement e.g. bootstrap
- Probability that a pattern is chosen is $(1 - (1 - 1/N)^N)$. For large N, this approaches $(1 - e^{-1}) = 0.632$

Estimating the Prediction Error (2/2)

- Sampling *with* replacement e.g. bootstrap
- Probability that a pattern is chosen is $(1 - (1 - 1/N)^N)$. For large N, this approaches $(1 - e^{-1}) = 0.632$
- Patterns not chosen in the i^{th} bootstrap sample become part of the i^{th} test set

Estimating the Prediction Error (2/2)

- Sampling *with* replacement e.g. bootstrap
- Probability that a pattern is chosen is $(1 - (1 - 1/N)^N)$. For large N, this approaches $(1 - e^{-1}) = 0.632$
- Patterns not chosen in the i^{th} bootstrap sample become part of the i^{th} test set
- Induce the model using the i^{th} bootstrap sample. Get the error ϵ_i from the i^{th} test set. Repeat b times

Estimating the Prediction Error (2/2)

- Sampling *with* replacement e.g. bootstrap
- Probability that a pattern is chosen is $(1 - (1 - 1/N)^N)$. For large N, this approaches $(1 - e^{-1}) = 0.632$
- Patterns not chosen in the i^{th} bootstrap sample become part of the i^{th} test set
- Induce the model using the i^{th} bootstrap sample. Get the error ϵ_i from the i^{th} test set. Repeat b times

$$J_{\text{boot}} = \frac{1}{b} \sum_{i=1}^{b} (0.632\epsilon_i + 0.368 J_{\text{total}})$$

Average Case Analysis – The Bias-Variance Decomposition

Optimal Response of an Estimator

Optimal Response of an Estimator

- $y = f(x) + \epsilon$. So, the y obtained corresponding to t repeated observations of x are $y(1), y(2), \ldots, y(t)$

Optimal Response of an Estimator

- $y = f(x) + \epsilon$. So, the y obtained corresponding to t repeated observations of x are $y(1), y(2), \ldots, y(t)$
- What should the *optimal* estimator $\hat{f}^*(x; \theta)$ respond with?

Optimal Response of an Estimator

- $y = f(x) + \epsilon$. So, the y obtained corresponding to t repeated observations of x are $y(1), y(2), \ldots, y(t)$
- What should the *optimal* estimator $\hat{f}^*(x;\theta)$ respond with?
- $$J = \left(\hat{f}^*(x;\theta) - y(1)\right)^2 + \left(\hat{f}^*(x;\theta) - y(2)\right)^2 + \ldots + \left(\hat{f}^*(x;\theta) - y(t)\right)^2$$

Optimal Response of an Estimator

- $y = f(x) + \epsilon$. So, the y obtained corresponding to t repeated observations of x are $y(1), y(2), \ldots, y(t)$
- What should the *optimal* estimator $\hat{f}^*(x;\theta)$ respond with?
- $$J = \left(\hat{f}^*(x;\theta) - y(1)\right)^2 + \left(\hat{f}^*(x;\theta) - y(2)\right)^2 + \ldots + \left(\hat{f}^*(x;\theta) - y(t)\right)^2$$
- Minimum of J is achieved when $\hat{f}^*(x;\theta) = (y(1), y(2), \ldots, y(t))/t$, i.e. $E[y|x]$

The Bias-Variance Decomposition

The Bias-Variance Decomposition

- Average case analysis of the prediction (generalization) error

The Bias-Variance Decomposition

- Average case analysis of the prediction (generalization) error
-

$$
\begin{aligned}
E_{\mathcal{X}}\left[\left(\hat{f}(x;\mathcal{X}) - E[y|x]\right)^2\right] \\
&= E_{\mathcal{X}}\left[\left(\left(\hat{f}(x;\mathcal{X}) - E_{\mathcal{X}}\left[\hat{f}(x;\mathcal{X})\right]\right) + \right.\right. \\
&\qquad \left.\left.\left(E_{\mathcal{X}}\left[\hat{f}(x;\mathcal{X})\right] - E[y|x]\right)\right)^2\right] \\
&= \underbrace{\left(E_{\mathcal{X}}\left[\hat{f}(x;\mathcal{X})\right] - E[y|x]\right)^2}_{\text{Squared} \quad \text{Bias}} + \underbrace{E_{\mathcal{X}}\left[\left(\hat{f}(x;\mathcal{X}) - E_{\mathcal{X}}\left[\hat{f}(x;\mathcal{X})\right]\right)^2\right]}_{\text{Variance}}
\end{aligned}
$$

Bias

Bias

- The Bias term,

$$\underbrace{\left(E_{\mathcal{X}}\left[\hat{f}(x;\mathcal{X})\right] - E[y|x]\right)^2}_{\text{Squared Bias}}$$

Bias

- The Bias term,

$$\underbrace{\left(E_{\mathcal{X}}\left[\hat{f}(x;\mathcal{X})\right]-E[y|x]\right)^2}_{\text{Squared}\quad\text{Bias}}$$

- Measures deviation of the averaged estimator output from the averaged system output

Bias

- The Bias term,

$$\underbrace{\left(E_{\mathcal{X}}\left[\hat{f}(x;\mathcal{X})\right] - E[y|x]\right)^2}_{\text{Squared}\quad\text{Bias}}$$

- Measures deviation of the averaged estimator output from the averaged system output
- Bias is 0 even when a particular estimator has a large error which is canceled out by an opposite error generated by another model

Variance

Variance

- The Variance term,

$$\underbrace{E_{\mathcal{X}}\left[\left(\hat{f}(x;\mathcal{X}) - E_{\mathcal{X}}\left[\hat{f}(x;\mathcal{X})\right]\right)^2\right]}_{\text{Variance}}$$

Variance

- The Variance term,

$$\underbrace{E_{\mathcal{X}}\left[\left(\hat{f}(x;\mathcal{X}) - E_{\mathcal{X}}\left[\hat{f}(x;\mathcal{X})\right]\right)^2\right]}_{\text{Variance}}$$

- Measures the sensitivity of the estimator

Variance

- The Variance term,

$$\underbrace{E_{\mathcal{X}}\left[\left(\hat{f}(x;\mathcal{X}) - E_{\mathcal{X}}\left[\hat{f}(x;\mathcal{X})\right]\right)^2\right]}_{\text{Variance}}$$

- Measures the sensitivity of the estimator
- It is independent of the underlying system $f(x)$

Understanding Bias and Variance

Understanding Bias and Variance

- Let the estimator be k-nearest neighbor

Understanding Bias and Variance

- Let the estimator be k-nearest neighbor
- When $k = N$, the output is simply the average of the training set output i.e. $(1/N)\sum_{i=1}^{N} y^{(i)}$
  - Estimate is likely to be unchanged from one training data set to another. Bias is high but variance is low

Understanding Bias and Variance

- Let the estimator be k-nearest neighbor
- When $k = N$, the output is simply the average of the training set output i.e. $(1/N) \sum_{i=1}^{N} y^{(i)}$
  - Estimate is likely to be unchanged from one training data set to another. Bias is high but variance is low
- If $k = 1$, the output follows local changes (as opposed to the population behavior)
  - Indeed, when $N \to \infty$ then Bias $\to 0$. Bias is low but variance will be high

Understanding Bias and Variance

- Let the estimator be k-nearest neighbor
- When $k = N$, the output is simply the average of the training set output i.e. $(1/N)\sum_{i=1}^{N} y^{(i)}$
  - Estimate is likely to be unchanged from one training data set to another. Bias is high but variance is low
- If $k = 1$, the output follows local changes (as opposed to the population behavior)
  - Indeed, when $N \to \infty$ then Bias $\to 0$. Bias is low but variance will be high
- The best solution is usually some intermediate k

The Bias/Variance Trade-Off

The Bias/Variance Trade-Off

- Bias and Variance are complementary, i.e. reducing one (most often) increases the other

The Bias/Variance Trade-Off

- Bias and Variance are complementary, i.e. reducing one (most often) increases the other
- The trick is to allow one to increase if the other decreases more than the increase

The Bias/Variance Trade-Off

- Bias and Variance are complementary, i.e. reducing one (most often) increases the other
- The trick is to allow one to increase if the other decreases more than the increase
- Approaches like weight decay, pruning, growing, early stopping are based on the above premise (avoid overfitting, i.e. tolerate increased bias in the *hope* that variance reduces more)

The Bias/Variance Trade-Off

- Bias and Variance are complementary, i.e. reducing one (most often) increases the other
- The trick is to allow one to increase if the other decreases more than the increase
- Approaches like weight decay, pruning, growing, early stopping are based on the above premise (avoid overfitting, i.e. tolerate increased bias in the *hope* that variance reduces more)
- Other approaches are based on aggregation (e.g. bagging – bootstrap aggregating, boosting)

Boosting (1/2)

Boosting (1/2)

- AdaBoost for a two-class classification setting,

Boosting (1/2)

- AdaBoost for a two-class classification setting,
  - Create T models each trained with a different distribution on the training set,

$$F(x) = \text{sign}\left(\sum_{t=1}^{T} \alpha_t \hat{f}_t(x; \theta_t)\right)$$

Boosting (1/2)

- AdaBoost for a two-class classification setting,
  - Create T models each trained with a different distribution on the training set,

$$F(x) = \text{sign}\left(\sum_{t=1}^{T} \alpha_t \hat{f}_t(x; \theta_t)\right)$$

  - Initially, weight $D_1^{(i)}$ on each pattern is $1/N$

Boosting (1/2)

- AdaBoost for a two-class classification setting,
  - Create T models each trained with a different distribution on the training set,
  $$F(x) = \text{sign}\ \left(\sum_{t=1}^{T} \alpha_t \hat{f}_t(x; \theta_t) \right)$$
  - Initially, weight $D_1^{(i)}$ on each pattern is $1/N$
  - Train using the distribution D_t

Boosting (2/2)

Boosting (2/2)

- Get error rate ϵ_t as,

$$\epsilon_t = \frac{1}{N} \sum_{i=1}^{N} I \left[\hat{f}_i(x^{(i)}; \Theta) \neq y^{(i)} \right]$$

$$\alpha_t = \log \left(\frac{1 - \epsilon_t}{\epsilon_t} \right)$$

- Reassign weightage on each pattern as,

$$D_{t+1}(i) = \frac{D_t}{Z_t} \times \begin{cases} e^{-\alpha_t} & \text{if } \hat{f}_i(x^{(i)}; \theta_i) = y^{(i)} \\ e^{\alpha_t} & \text{if } \hat{f}_i(x^{(i)}; \theta_i) \neq y^{(i)} \end{cases}$$

Minimum Description Length

Minimum Description Length

$$\underbrace{L(D)}_{\text{Description Length}} = \underbrace{L(D|H)}_{\text{Perturbation}} + \underbrace{L(H)}_{\text{Complexity (Nominal Model)}} \tag{1}$$

Minimum Description Length

$$\underbrace{L(D)}_{\text{Description Length}} = \underbrace{L(D|H)}_{\text{Perturbation}} + \underbrace{L(H)}_{\text{Complexity (Nominal Model)}} \tag{1}$$

- In ffnn, we could say the weights are $\mathcal{N}(0, \sigma_1)$ i.e. the model is $exp(- \parallel w \parallel^2)/(2{\sigma_1}^2)$

Minimum Description Length

$$\underbrace{L(D)}_{\text{Description Length}} = \underbrace{L(D|H)}_{\text{Perturbation}} + \underbrace{L(H)}_{\text{Complexity (Nominal Model)}} \tag{1}$$

- In ffnn, we could say the weights are $\mathcal{N}(0, \sigma_1)$ i.e. the model is $exp(- \| w \|^2)/(2{\sigma_1}^2)$
- Say, perturbation based on the data is $\mathcal{N}(0, \sigma_2)$ i.e. $exp(- \| \epsilon \|^2)/(2{\sigma_2}^2)$

Minimum Description Length

$$\underbrace{L(D)}_{\text{Description Length}} = \underbrace{L(D|H)}_{\text{Perturbation}} + \underbrace{L(H)}_{\text{Complexity (Nominal Model)}} \tag{1}$$

- In ffnn, we could say the weights are $\mathcal{N}(0, \sigma_1)$ i.e. the model is $exp(- \| w \|^2)/(2{\sigma_1}^2)$
- Say, perturbation based on the data is $\mathcal{N}(0, \sigma_2)$ i.e. $exp(- \| \epsilon \|^2)/(2{\sigma_2}^2)$
- The total description length is,

$$\begin{aligned} L(D) &= -\log(D|H) - \log(H) \\ &= \frac{1}{2{\sigma_2}^2}(y^{(i)} - \hat{f}(x^{(i)};\theta))^2 + \frac{1}{2{\sigma_1}^2} \| w \|^2 \end{aligned}$$

Worst Case Analysis – The VC Dimension

Empirical and True Risk

Empirical and True Risk

- Let $x \in \{-1, +1\}^n$. Then,

$$J_{\mathrm{emp}}(\theta) = \frac{1}{N} \sum_{i=1}^{N} \left(y^{(i)} - \hat{f}(x^{(i)}; \theta) \right)^2$$

$$J(\theta) = E \left[\left(y - \hat{f}(x; \mathcal{X}) \right)^2 \right]$$

Empirical and True Risk

- Let $x \in \{-1, +1\}^n$. Then,

$$J_{\mathrm{emp}}(\theta) = \frac{1}{N} \sum_{i=1}^{N} \left(y^{(i)} - \hat{f}(x^{(i)}; \theta) \right)^2$$

$$J(\theta) = E\left[\left(y - \hat{f}(x; \mathcal{X}) \right)^2 \right]$$

- $J(\theta) \gg J_{\mathrm{emp}}(\theta)$. From the law of large numbers,

$$\Pr\left[|J(\theta) - J_{\mathrm{emp}}(\theta)| > \epsilon \right] \rightarrow 0, \quad \text{as } N \rightarrow \infty$$

Uniform Convergence

Uniform Convergence

- However the stronger result also holds,

$$\Pr\left[\sup_{\theta}|J(\theta) - J_{\mathrm{emp}}(\theta)| > \epsilon\right] \to 0, \quad \text{as } N \to \infty$$

Uniform Convergence

- However the stronger result also holds,

$$\Pr\left[\sup_{\theta} |J(\theta) - J_{\mathrm{emp}}(\theta)| > \epsilon\right] \to 0, \quad \text{as } N \to \infty$$

- This is the uniform convergence of the empirical error rate to the true error rate

Uniform Convergence

- However the stronger result also holds,

$$\Pr\left[\sup_{\theta}|J(\theta) - J_{\mathrm{emp}}(\theta)| > \epsilon\right] \to 0, \quad \text{as } N \to \infty$$

- This is the uniform convergence of the empirical error rate to the true error rate
- What is the rate of uniform convergence?

Rate of Uniform Convergence

Rate of Uniform Convergence

- However the stronger result also holds,

$$\Pr\left[\sup_{\theta}|J(\theta) - J_{\text{emp}}(\theta)| > \epsilon\right] < 4\ \Delta(2N)\ e^{-\frac{\epsilon^2 N}{8}}$$

Rate of Uniform Convergence

- However the stronger result also holds,

$$\Pr\left[\sup_{\theta} |J(\theta) - J_{\mathrm{emp}}(\theta)| > \epsilon\right] < 4\ \Delta(2N)\ e^{-\frac{\epsilon^2 N}{8}}$$

All realizations Growth function

Rate of Uniform Convergence

- However the stronger result also holds,

$$\Pr\left[\sup_{\theta} |J(\theta) - J_{\mathrm{emp}}(\theta)| > \epsilon\right] < 4\ \Delta(2N)\ e^{-\frac{\epsilon^2 N}{8}}$$

All realizations Growth function

- This is the uniform convergence of the empirical error rate to the true error rate

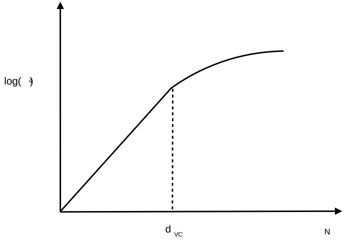
VC Dimension

VC Dimension

- $\Delta(2N)$ is identically equal to 2^N or bounded above by $\Delta(N) \leq N^{d_{\mathrm{vc}}} + 1$ i.e. the machine can shatter upto d_{vc} points in a general position (realize all possible dichotomies)

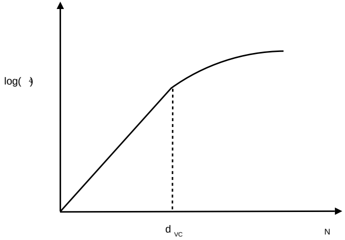
VC Dimension

- $\Delta(2N)$ is identically equal to 2^N or bounded above by $\Delta(N) \leq N^{d_{\mathrm{vc}}} + 1$ i.e. the machine can shatter upto d_{vc} points in a general position (realize all possible dichotomies)



VC Dimension

- $\Delta(2N)$ is identically equal to 2^N or bounded above by $\Delta(N) \leq N^{d_{\mathrm{vc}}} + 1$ i.e. the machine can shatter upto d_{vc} points in a general position (realize all possible dichotomies)



- The estimator can generalize beyond d_{vc} – the VC dimension

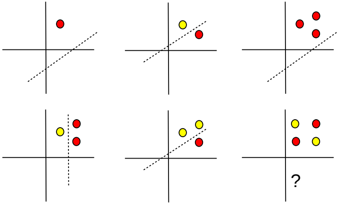
VC Dimension – Example

VC Dimension – Example

- d_{VC} for a linear classifier in n dimensions is $(n+1)$

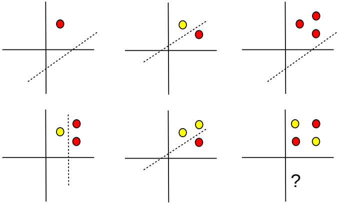
VC Dimension – Example

- d_{VC} for a linear classifier in n dimensions is $(n+1)$



VC Dimension – Example

- d_{VC} for a linear classifier in n dimensions is $(n+1)$



- The estimator can generalize beyond d_{VC} – the VC dimension

Linear Classifier

Linear Classifier

$$
\begin{aligned}
\Pr\left[\sup_{\theta} |J(\theta) - J_{\mathrm{emp}}(\theta)| > \epsilon\right] &< 4\, \Delta(2N)\, e^{-\frac{\epsilon^2 N}{8}} \\
&< 4\, \left[(2N)^{d_{\mathrm{vc}}} + 1\right]\, e^{-\frac{\epsilon^2 N}{8}} \\
&< 4\, \left[(2N)^{n+1} + 1\right]\, e^{-\frac{\epsilon^2 N}{8}}
\end{aligned}
$$

Linear Classifier

$$\begin{aligned}
\Pr\left[\sup_{\theta}|J(\theta) - J_{\text{emp}}(\theta)| > \epsilon\right] &< 4\,\Delta(2N)\,e^{-\frac{\epsilon^2 N}{8}} \\
&< 4\,\left[(2N)^{d_{\text{vc}}} + 1\right]\,e^{-\frac{\epsilon^2 N}{8}} \\
&< 4\,\left[(2N)^{n+1} + 1\right]\,e^{-\frac{\epsilon^2 N}{8}}
\end{aligned}$$

- For right side to be small, approximately,

$$N > \frac{8n\log n}{\epsilon^2}$$

Some VC Dimensions

Some VC Dimensions

- Intervals in $\mathcal{R}$: 2

Some VC Dimensions

- Intervals in $\mathcal{R}$: 2
- Axis parallel rectangles in $\mathcal{R}^2$: 4

Some VC Dimensions

- Intervals in $\mathcal{R}$: 2
- Axis parallel rectangles in $\mathcal{R}^2$: 4
- Circle in $\mathcal{R}^2$: 3

Some VC Dimensions

- Intervals in $\mathcal{R}$: 2
- Axis parallel rectangles in $\mathcal{R}^2$: 4
- Circle in $\mathcal{R}^2$: 3
- Triangle in $\mathcal{R}^2$: 7

Conclusion

Conclusion

- Empirical models need to be carefully designed and used to realize their advantages

Conclusion

- Empirical models need to be carefully designed and used to realize their advantages
- B^2V, VC-Dimension or MDL are good points for designing/analyzing new algorithms