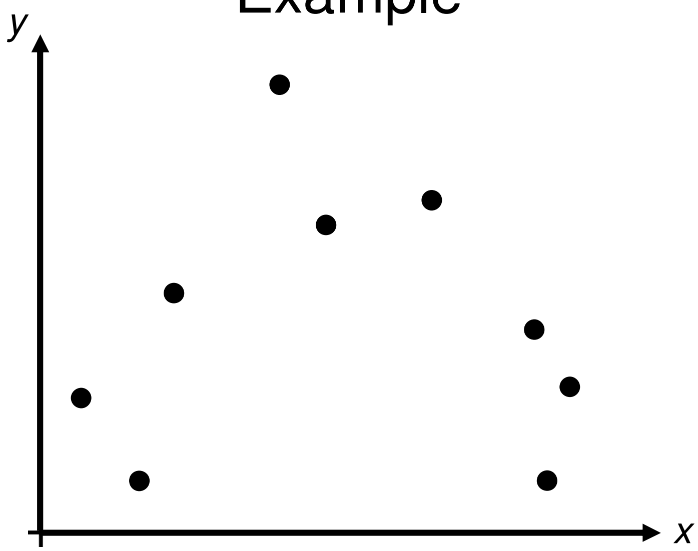
Learning Conclusion: Cross-Validation

Bayes Nets Intro: Representing and Reasoning about Uncertainty

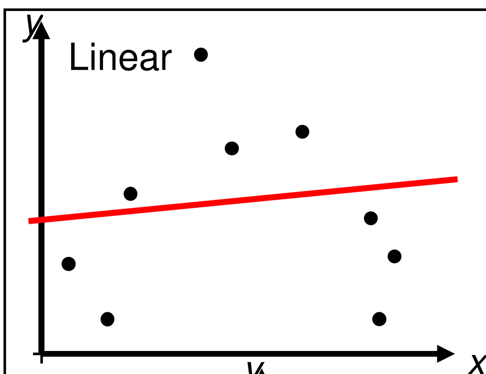
Final Considerations: Avoiding Overfitting

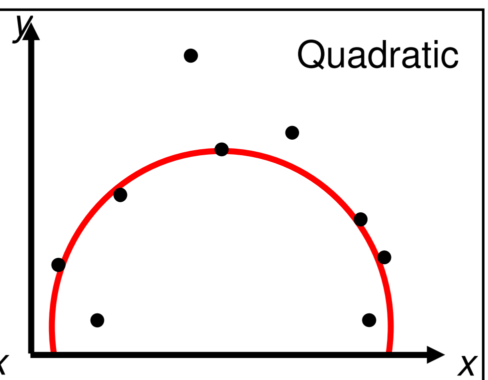
- We have a choice of different techniques:
- Decision trees, Neural Networks, Nearest Neighbors, Bayes Classifier,…
- For each we have different levels of complexity:
  - Depth of trees
  - Number of layers and hidden units
  - Number of neighbors in K-NN
  - …..
- How to choose the right one?
- Overfitting: A complex enough model (e.g., enough units in a neural network, large enough trees,..) will *always* be able to fit the training data well

Example

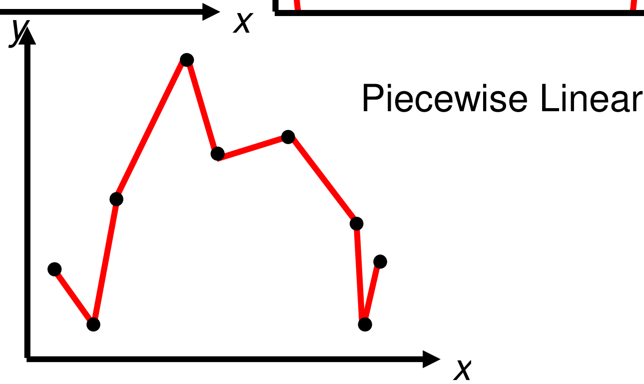


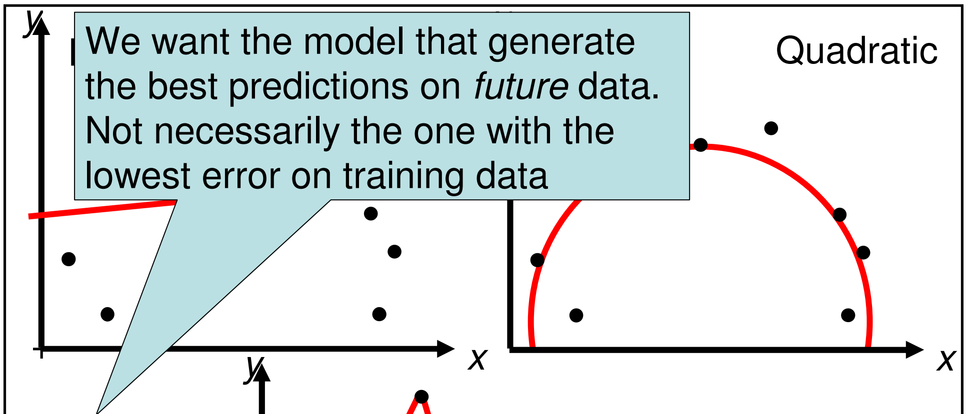
- Construct a predictor of *y* from *x* given this training data



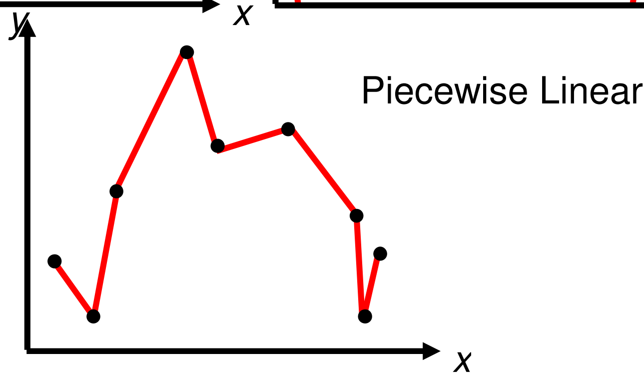


Which model is best for predicting *y* from *x* ????

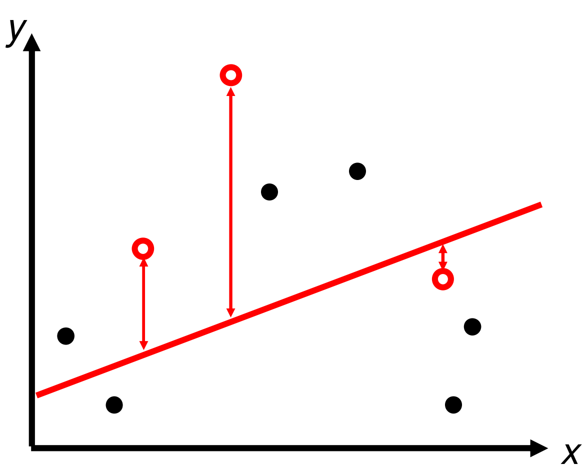




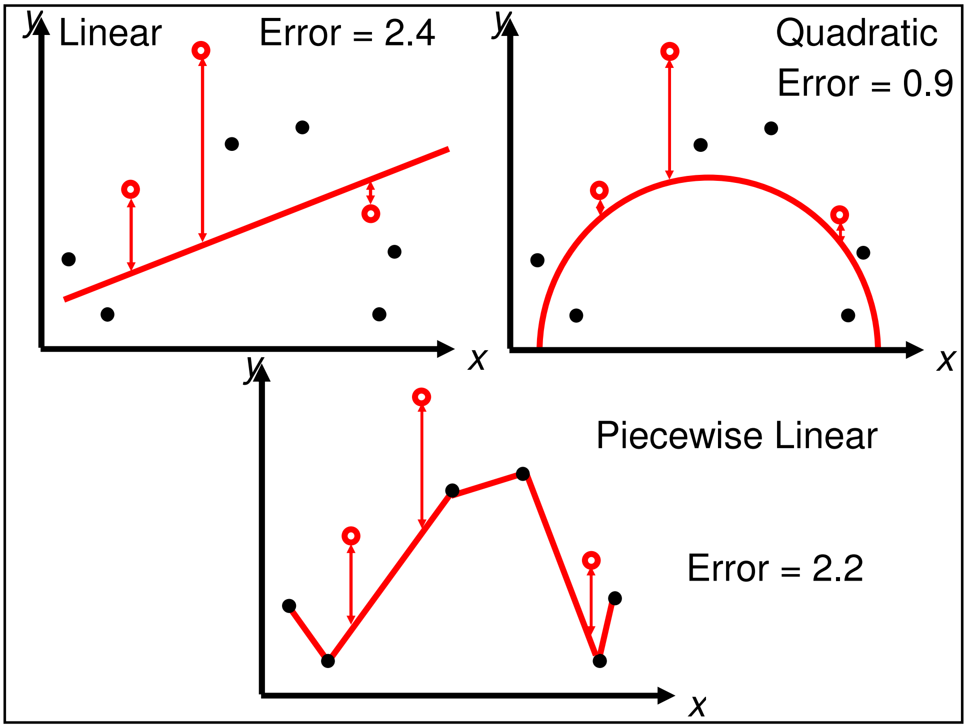
Which model is best for predicting *y* from *x* ????

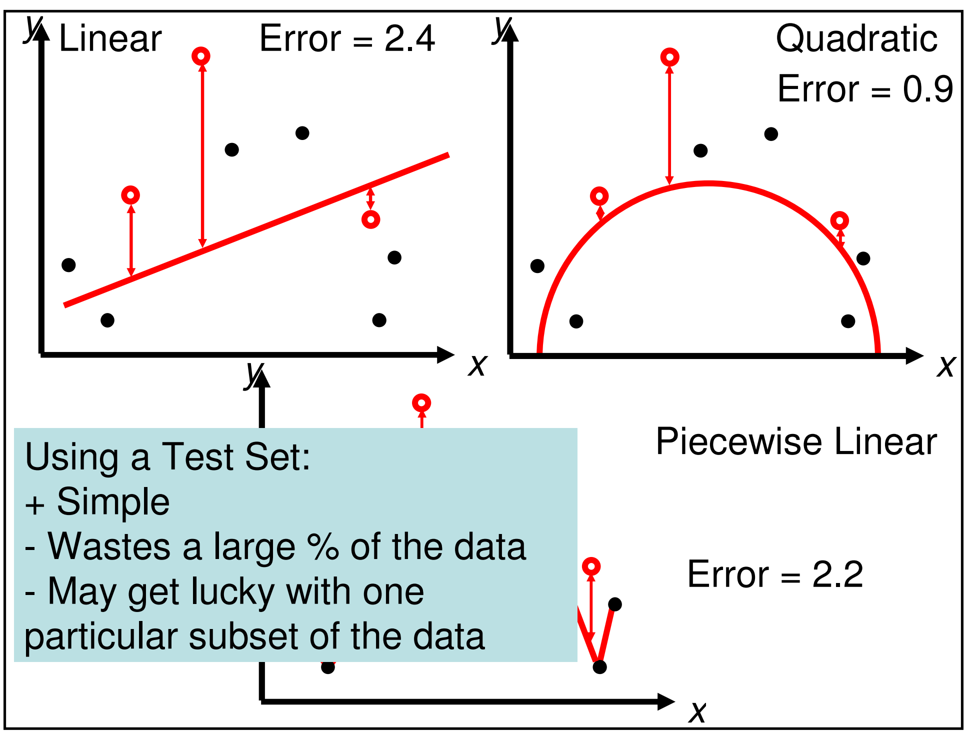


Using a Test Set

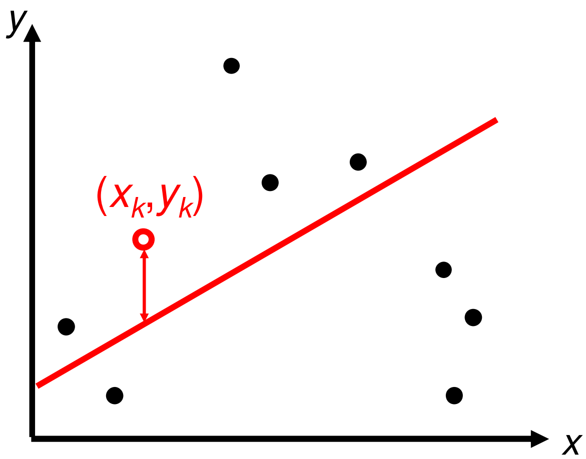


1. Use a portion (e.g., 30%) of the data as *test data*
2. Fit a model to the remaining training data
3. Evaluate the error on the *test data*

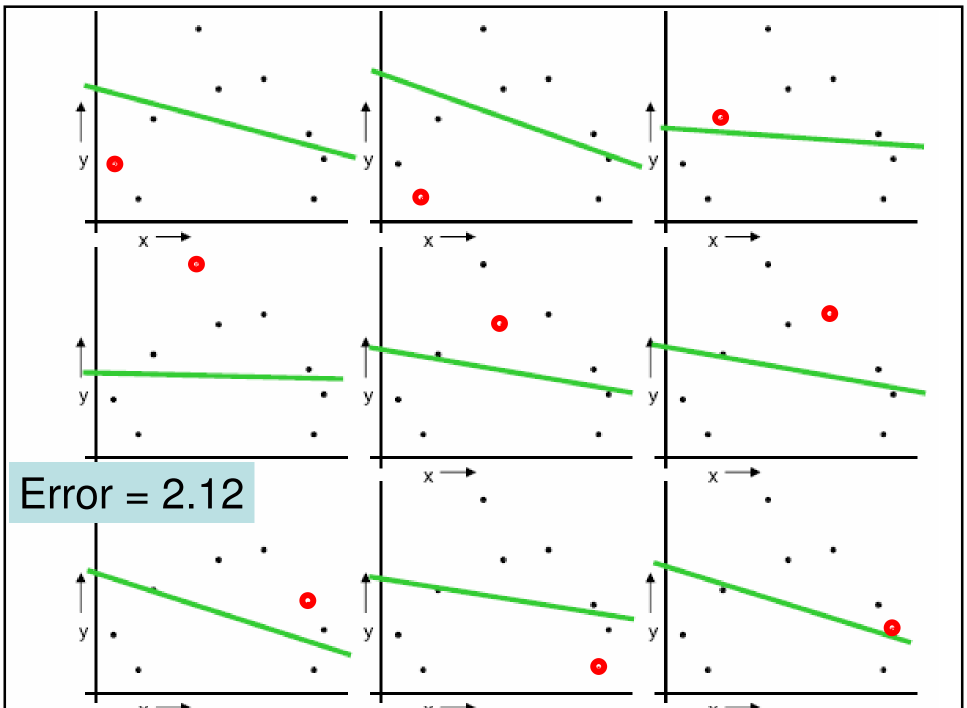

Linear
Error = 2.4
y
x
Quadratic
Error = 0.9
y
x
Piecewise Linear
Error = 2.2
y
x


Linear
Error = 2.4
y
x
Quadratic
Error = 0.9
y
x
Piecewise Linear
Error = 2.2
y
x
Using a Test Set:
+ Simple
- Wastes a large % of the data
- May get lucky with one particular subset of the data

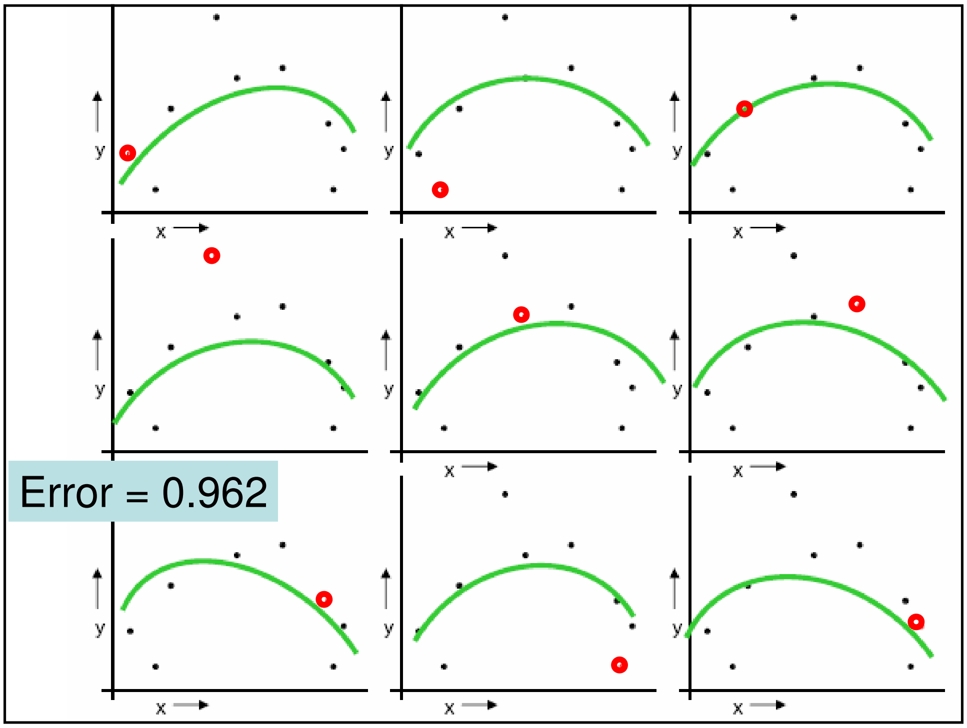
"Leave One Out" Cross-Validation

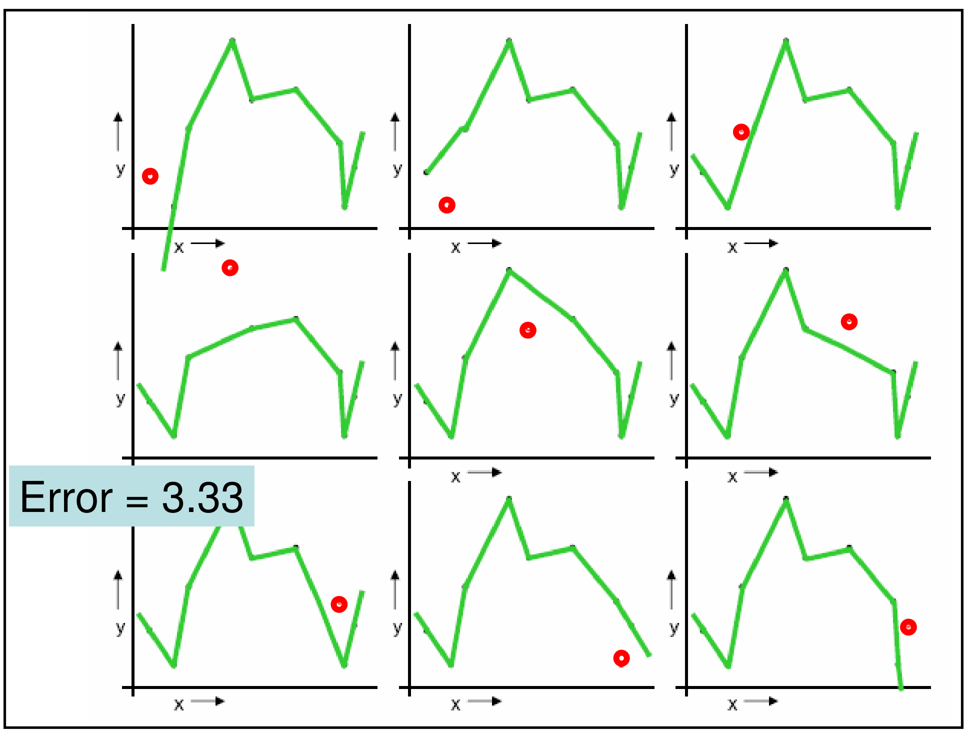


- For k=1 to R
  - Train on all the data leaving out (x_k,y_k)
  - Evaluate error on (x_k,y_k)
- Report the average error after trying *all* the data points



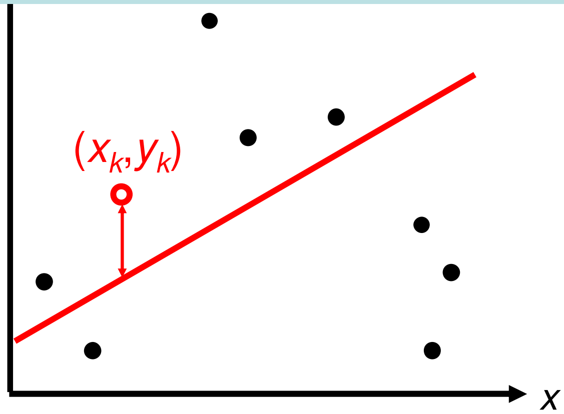
Note: Numerical examples in this and subsequent slides from A. Moore


Error = 0.962
y
x


Error = 3.33
y
x

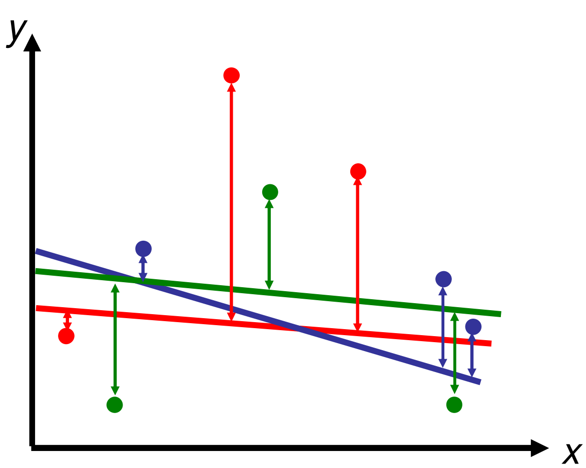
“Leave One Out” Cross-Validation:
+ Does not waste data
+ Average over large number of trials
- Expensive

Validation



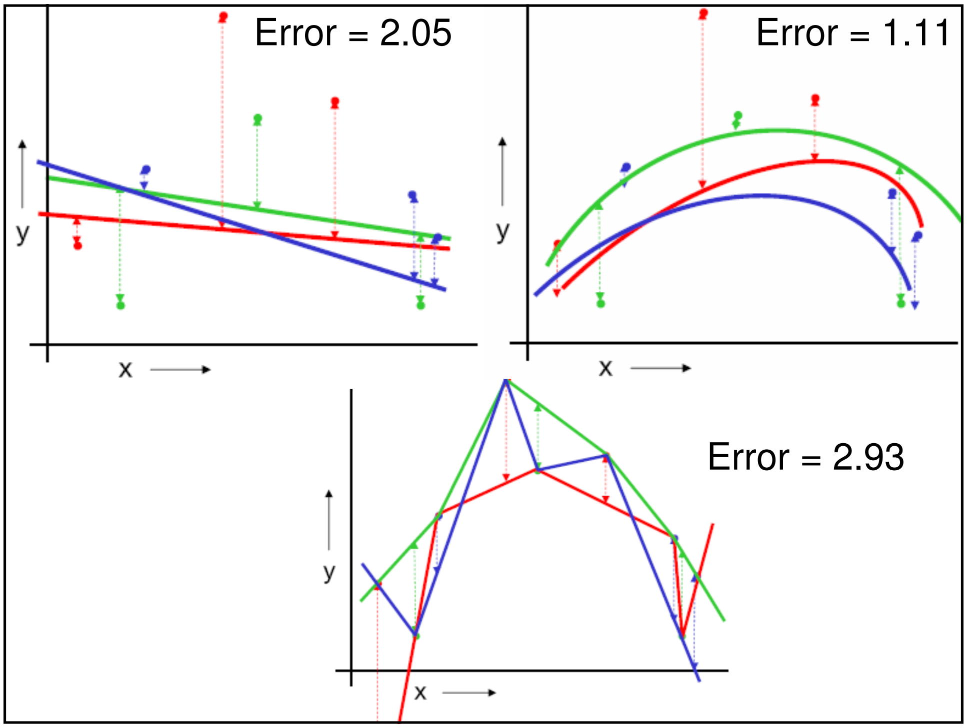
- For k=1 to R
  - Train on all the data leaving out (x_k,y_k)
  - Evaluate error on (x_k,y_k)
- Report the average error after trying *all* the data points

K-Fold Cross-Validation



- Randomly divide the data set into K subsets
- For each subset S:
  - Train on the data *not in S*
  - Test on the data *in S*
- Return the average error over the K subsets

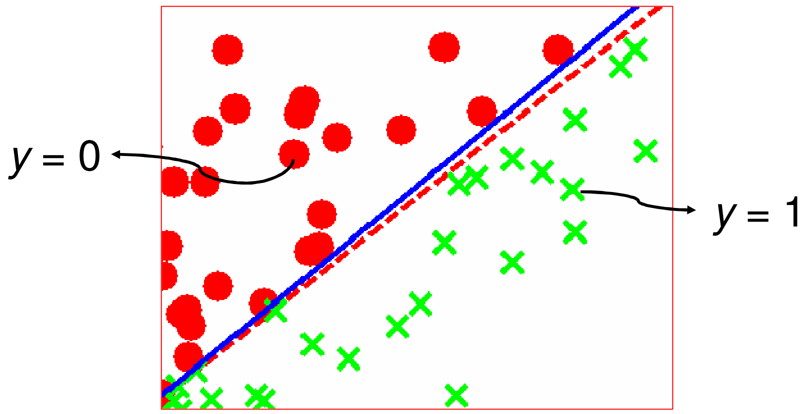
Example: K = 3, each color corresponds to a subset



Cross-Validation Summary

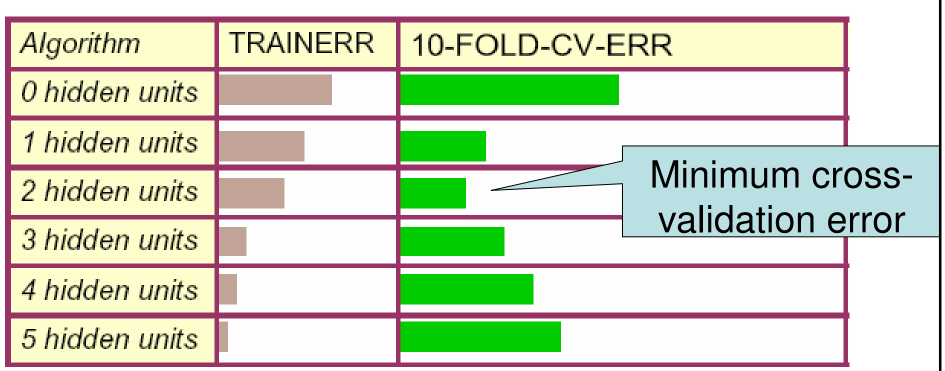
	-	+
Test Set	Wastes a lot of data Poor predictor of future performance	Simple/Efficient
Leave One Out	Inefficient	Does not waste data
K-Fold	Wastes 1/*K* of the data *K* times slower than Test Set	Wastes only 1/*K* of the data! Only *K* times slower than Test Set!

Classification Problems



- The exact same approaches apply for cross-validation except that the error is the number of data points that are misclassified.

Example: Training a Neural Net



- Train neural nets with different numbers of hidden units (more and more complex NNs)
- For each NN, evaluate the error using K-fold Cross-Validation
- Choose the one with the minimum cross-validation error

Summary (R&N Chapter 20)

- Learning Algorithms:
  - Naïve Bayes
  - Decision Trees
  - Nearest Neighbors
  - Neural Networks
- Validation:
  - Error on training set should never be used directly for evaluate learning algorithm on a data set
  - Validation on test set
  - Cross-validation to avoid wasting data
    - Leave one out
    - K-fold
  - Used for:
    - Finding best configuration of learned model (complexity of neural network, K-NN, etc.)
    - Deciding between different learning algorithms (neural networks, nearest neighbors, decision trees,…)

Bayes Nets
Representing and Reasoning about Uncertainty

Bayes Nets

- Material covered in Russell & Norvig, Chapter 14
- Not covered in lectures: Networks with continuous variables
- Not covered in chapter: d-separation

Reasoning with Uncertainty

- Most real-world problems deal with uncertain information
  - Diagnosis: Likely disease given observed symptoms
  - Equipment repair: Likely component failure given sensor reading
  - Help desk: Likely operation based on past operations

Reasoning with Uncertainty

- We saw how to use probability to represent uncertainty and to perform queries such as inference
  - Diagnosis: Prob (disease | observed symptoms)
  - Equipment repair: Prob (component | sensor readings)
  - Help desk: Prob (Likely operation | past operations)
- We saw that representing probability distributions can be inefficient (or intractable) for large problems.

Reasoning with Uncertainty

- We saw how to use probability to represent uncertainty and to perform queries such as inference
  - Diagnosis: Prob (disease | observed symptoms)
  - Equipment repair: Prob (component | sensor readings)
  - Help desk: Prob (Likely operation | past operations)
- We saw that representing probability distribution can be inefficient (or intractable) for large problems.
- Today: Bayes Nets provide a powerful tool for making reasoning with uncertainty manageable by taking advantage of dependence relations between variables
- For example: Knowing that the hand brake is operational does not help diagnose why the engine does not start!
- We'll start by reviewing our key probability tools.

Probability Reminder

- Conditional probability for 2 events A and B:

$$P(A|B) = \frac{P(A,B)}{P(B)}$$

- Chain rule:

$$P(A,B) = P(A|B)\ P(B)$$

Probability Reminder

- Conditional probability for 2 variables X and Y:

$$P(X=x \mid Y=y) = \frac{P(X=x, Y=y)}{P(Y=y)}$$

- Chain rule:

$$P(X=x,Y=y) = P(X=x|Y=y)\ P(Y=y)$$

- For any values x,y

The Joint Distribution

- Joint distribution = collection of all the probabilities $P(X = x, Y = y, Z = z\ldots)$ for all possible combinations of values.
- For m binary variables, size is 2^m
- Any query can be computed from the joint distribution

X	Y	Z	Prob
T	T	T	0.1
T	T	F	0.22
T	F	T	0.2
T	F	F	0.08
F	T	T	0.1
F	T	F	0.15
F	F	T	0.07
F	F	F	0.08

The Joint Distribution

- Any query can be computed from the joint distribution
- Marginal distribution

  $P(X = \text{True}), P(X = \text{False})$

- Conditional distribution:

  $P(X = \text{True} \mid Y = \text{True}) = P(X = \text{True}, Y = \text{True})/P(Y = \text{True})$

- In general:

  $P(E_1 \mid E_2) = P(E_1, E_2)/P(E_2)$

  $$P(E_2) = \sum_{\text{Entries that match } E_2} P(\text{Joint Entries})$$

X	Y	Z	Prob
T	T	T	0.1
T	T	F	0.22
T	F	T	0.2
T	F	F	0.08
F	T	T	0.1
F	T	F	0.15
F	F	T	0.07
F	F	F	0.08

The Joint Distribution

- Any query can be computed from the joint distribution
- Marginal distribution

  P(Y = True), P(Y = False)

- Conditional distribution:

  P(X = True | Y = True) =

  P (X = True,Y = True)/P(Y =

- In general:

  $P(E_1 \mid E_2) = P(E_1,E_2)/P(E_2)$

  $P(E_2) = \sum_{\text{Entries that match } E_2} P(\text{Joint Entries})$

X	Y	Z	Prob
T	T	T	0.1
T	T	F	0.22
T	F	T	0.2

E_1 and E_2 are assignments of values to subsets of variables. E_2 = evidence, observed variables,…

The Joint Distribution

- Joint distribution = collection of all the probabilities

X	Y	Z	Prob
T	T	T	0.1

Minor point about our notations and examples:

- We'll use "^" or "," to mean "and" in the joint probabilities. It's the same thing.
- Sometimes P(X = True) is abbreviated to P(X) and P(X=False) to P(¬X).

- Most of the examples use binary (True/False) variables. This is for convenience only, everything works with variables with arbitrary domains.

We'll consider only discrete variables. Everything can be extended to continuous variables.

Avoiding Using the Full Joint

- Consider two events:
  - My house is being burglarized → Binary variable B = {True, False}
  - There is an earthquake → Binary variable E = {True, False}
- We can model the joint distribution with four numbers
- Can we model it with fewer numbers?
- Can we use only P(B) and P(E)?

Independence

- The fact that an earthquake occurs does not depend on whether or not a burglary is in progress.

$$P(E=e|B=b) = P(E=e)$$

- The knowledge of B does not add anything to our estimate of how likely E is.

Independence

- In general, if two sets of random variables S_1 and S_2 are independent:
- P(any assignment to S_1| any assignment to S_2) = P(any assignment to S_1)
- P(any assignment to S_1 ^ any assignment to S_2) = P(any assignment to S_1) x P(any assignment to S_2)

Independence

- P(E = True) = 0.002
- P(B = True) = 0.001
- E and B independent
- From these assumptions, we can derive the joint distribution
- From the joint distribution, we can answer any query

E	B	Prob
T	T	
T	F	
F	T	
F	F	

A More Complicated Case

- The house is equipped with an alarm system that can be triggered by a burglar or by an earthquake
- Model this with a new binary variable A = {True, False}
- To answer queries, we now need a joint table with variables E,B,A

A More Complicated Case

- The house is equipped with an alarm system that can be triggered by a burglar or by an earthquake
- Model this with a new binary variable A = {True, False}
- To answer queries, we now need a joint table with variables E,B,A

This is what we know so far:
We know the distributions P(E) and P(B)
We know that E and B are independent P(E|B) = P(E)
The Alarm is *NOT* independent of B and is *NOT* independent of E

A More Complicated Case

- The house is equipped with an alarm system that can be triggered by a burglar or by an earthquake
- Model this with a new binary variable A = {True, False}
- To answer queries, we now need a joint table with variables E,B,A

This is what we know so far:
We know the distributions P(E) and P(B)
We know that E and B are independent P(E|B) = P(E)
The Alarm is *NOT* independent of B and is *NOT* independent of E

We know the joint of E and B, so all we need is:
P(A | E=e,B=b) for all 4 combinations of e and b in {True, False}

A More Complicated Case

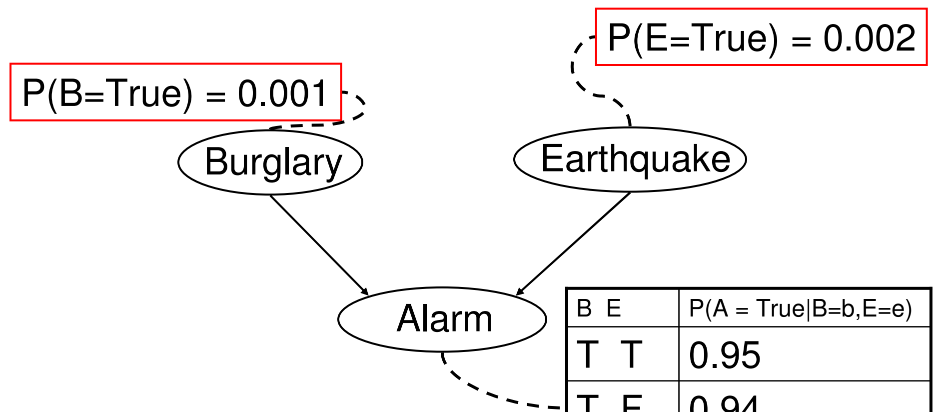
- The house is equipped with an alarm system that can be triggered by a burglar or by an earthquake
- Model this with a new binary variable A = {True, False}

P(E = True) = 0.002
P(B = True) = 0.001
E and B independent

P(A = True | B = True, E = True) = 0.95
P(A = True | B = True, E = False) = 0.94
P(A = True | B = False, E = True) = 0.29
P(A = True | B = False, E = False) = 0.001

We can specify the entire distribution by 6 numbers. How can you compute P(A = a , B = b , E = e) for any value of a,b,e??

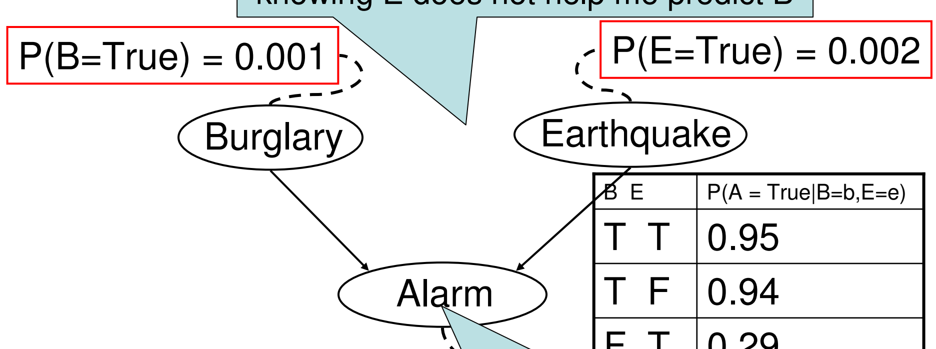
Graphical Representation



B	E	P(A = True|B=b,E=e)
T	T	0.95
T	F	0.94
F	T	0.29
F	F	0.001

Graphical Representation

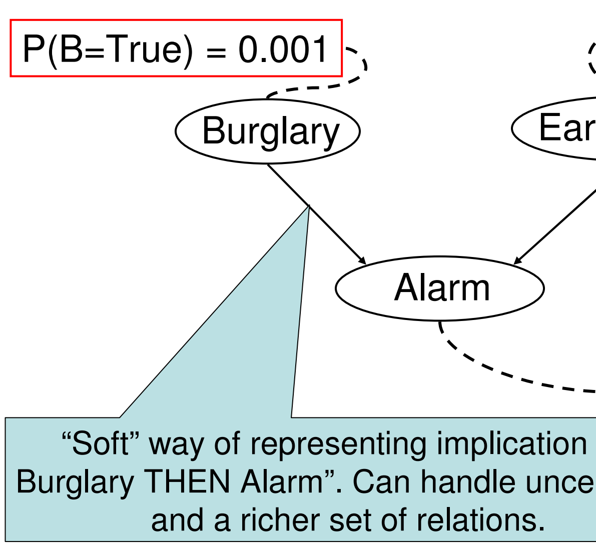
No arrow between B and E means that knowing E does not help me predict B



B	E	P(A = True|B=b,E=e)
T	T	0.95
T	F	0.94
F	T	0.29
F	F	0.001

The two arrows coming into A mean that to know the value of A, if helps to know the values of B and E

Graphical Representation



B E	P(A = True|B=b,E=e)
T T	0.95
T F	0.94
F T	0.29
F F	0.001

Another Type of Independence

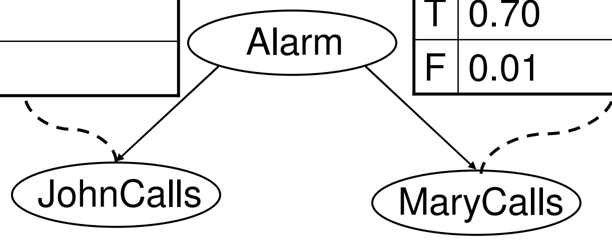
- A: Alarm goes off
- J: Neighbor John calls
- M: Neighbor Mary calls

- New kind of independence:
- Once we know that the alarm went off, we know if John will call, irrespective of what Mary does
- P (J | A=a, M=m) = P(J | A=a) for any values of a and m
- J and M are *conditionally independent* given A

Graphical Representation

A	P(J = True|A=a)
T	0.90
F	0.05

A	P(M = True|A= a)
T	0.70
F	0.01



Given knowledge of A, knowing anything else in the diagram won't help with J and M

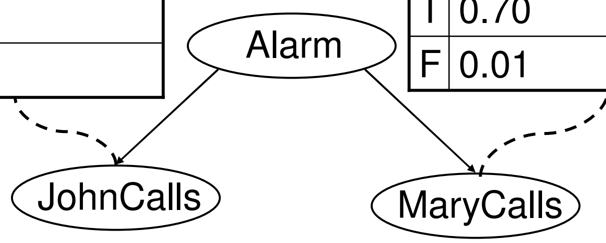
Conditional Independence

- In general, if two sets of random variables S_1 and S_2 are conditionally independent given S_3:
- P(any assignments to S_1\| any assignments to S_2 ,any assignments to S_3) = P(assignment to S_1 \| assignments to S_3)

Graphical Representation

A	P(J = True|A=a)
T	0.90
F	0.05

A	P(M = True|A= a)
T	0.70
F	0.01



If we know the distribution P(A), we can compute the joint (and therefore we can answer any query with 5 numbers)

Summary

Conditional probability to represent relation between variables:
$P(X = x \mid Y = y) = P(X = x , Y = y)/P(Y = y)$ for all x,y
"How probable is it for X to take value x, given that we know that the value of Y is y?"

Independence of variables:
$P(X = x \mid Y = y) = P(X = x)$ for all x,y
"knowledge of Y does not affect knowledge of X"

Conditional independence:
$P(X = x \mid Y = y , Z = z) = P(X = x \mid Z = z)$ for all x,y,z
"Given knowledge of Z, knowledge of Y does add anything to our knowledge of X"

A set of variables is represented by the collection of values
$P(X = x , Y = y , Z = z , W = w)$, the joint distribution.
For m binary variables the joint distribution requires 2^m entries.
Any query can be answered from the joint distribution.

Graphical representation: Directed graph in which nodes are the variables, arcs represent conditional dependencies

Conditional probability to represent relation between variables:

$P(X = x \mid Y = y) = P(X = x , Y = y)/P(Y = y)$ for all x,y

"How probable is it for X to take value x, given that we know that the value of Y is y?"

Independence of variables:

$P(X = x \mid Y = y) = P(X = x)$ for all x,y

"knowledge of Y does not affect knowledge of X"

Conditional independence:

$P(X = x \mid Y = y , Z = z) = P(X = x \mid Z = z)$ for all x,y,z

"Given knowledge of Z, knowledge of Y does add anything to our knowledge of X"

A set of variables is represented by the collection of values
$P(X = x , Y = y , Z = z , W = w)$, the joint distribution.
For m binary variables the joint distribution requires 2^m entries.
Any query can be answered from the joint distribution.

Key insight: The need for enumerating and storing entries can be drastically reduced by exploiting (conditional) independence relations between variables