Data Mining for Business Analytics

Lecture 5: Model Performance Analytics

Stern School of Business
New York University
Spring 2014

Over-fitting the data

- Finding chance occurrences in data that look like interesting patterns, but which do not generalize, is called over-fitting the data

- We want models to apply not just to the exact training set but to the general population from which the training data came
  - Generalization

Over-fitting

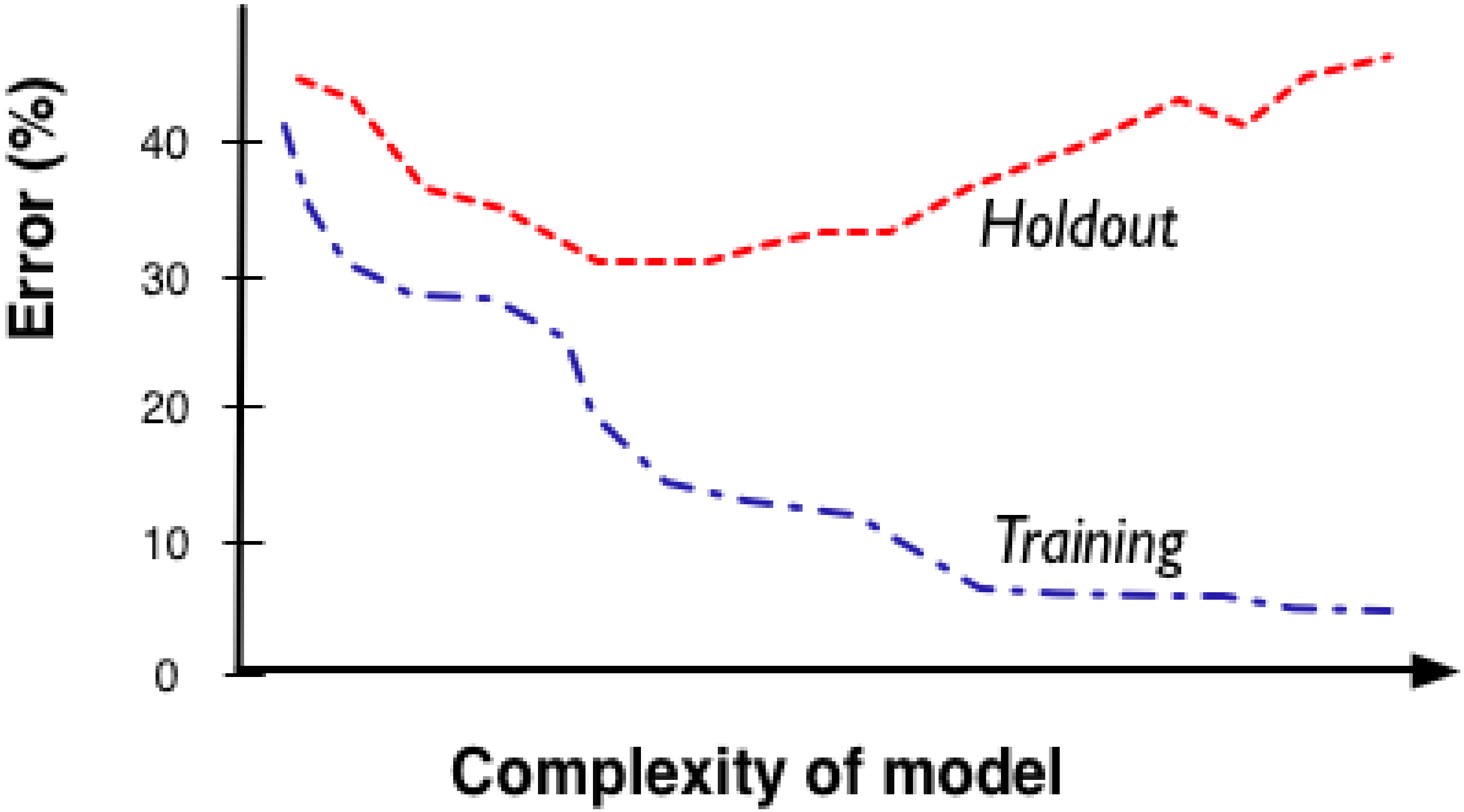
- The tendency of DM procedures to tailor models to the training data, *at the expense of generalization* to previously unseen data points.

- All data mining procedures have the tendency to over-fit to some extent
  - Some more than others.

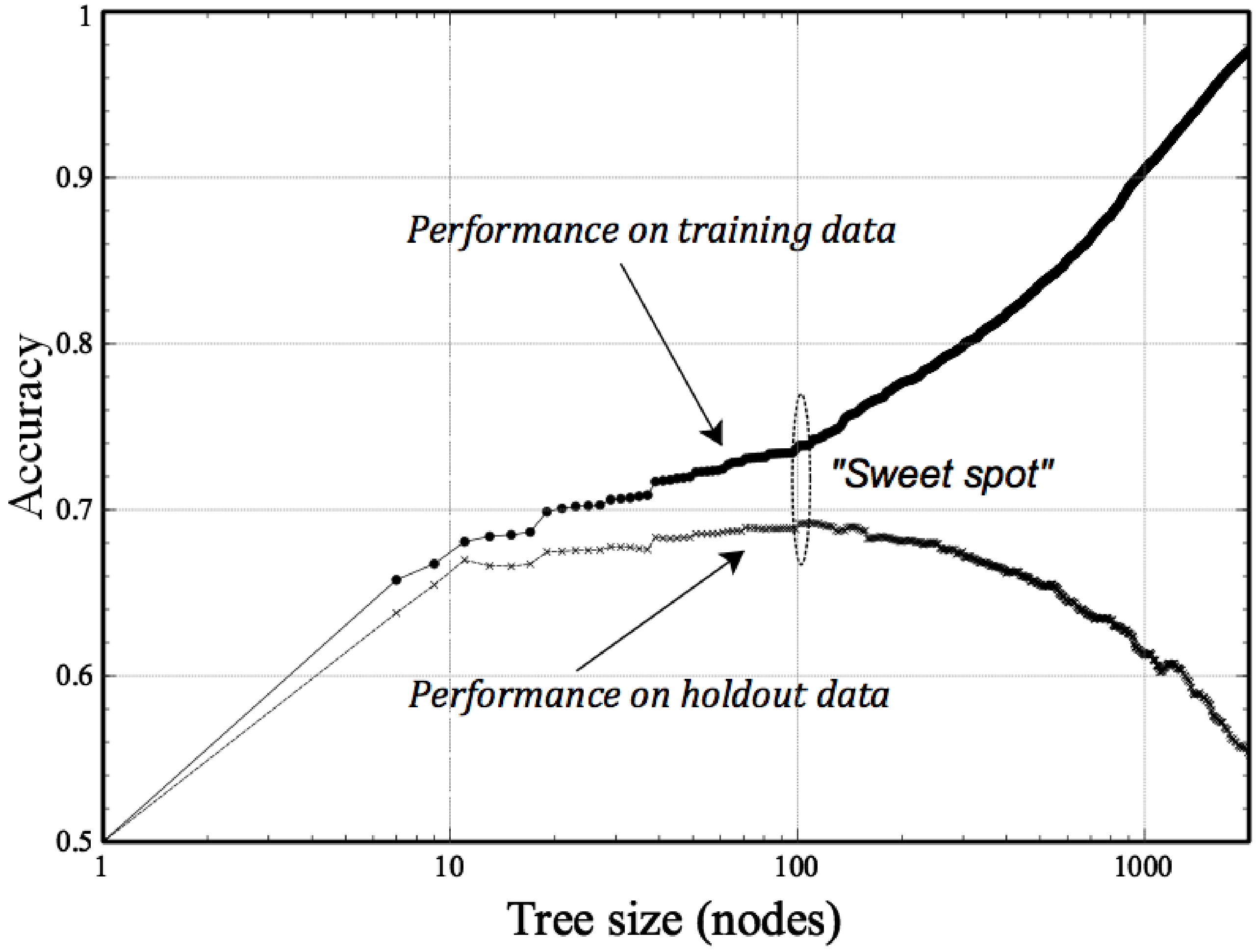
- "If you torture the data long enough, it will confess"

- There is no single choice or procedure that will eliminate over-fitting
  - recognize over-fitting and manage complexity in a principled way.

Fitting Graph



Over-fitting in tree induction



Over-fitting in linear discriminants

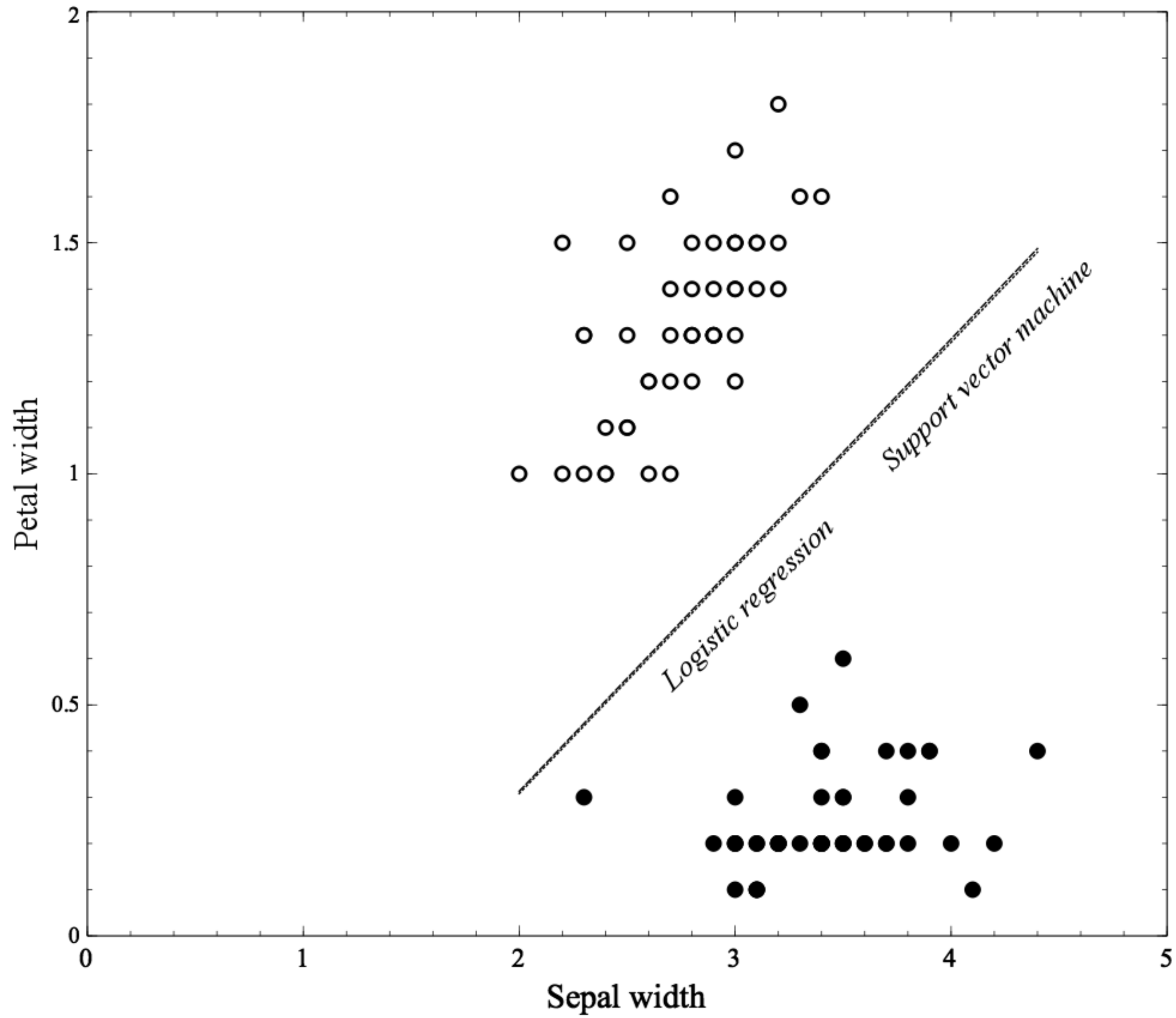
$$f(x) = w_0 + w_1 x_1 + w_2 x_2 + w_3 x_3$$

$$f(x) = w_0 + w_1 x_1 + w_2 x_2 + w_3 x_3 + w_4 x_4 + w_5 x_5$$

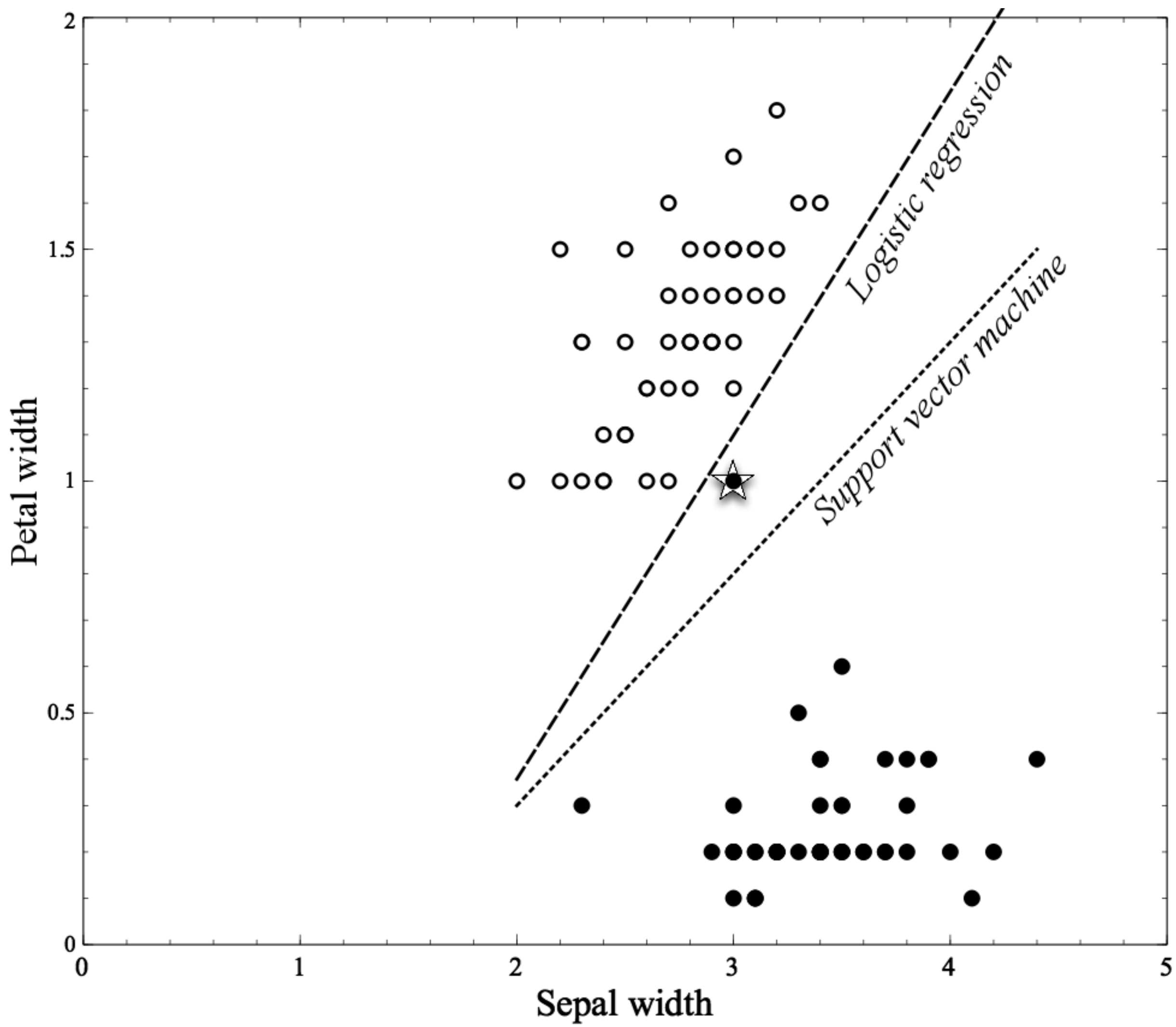
$$f(x) = w_0 + w_1 x_1 + w_2 x_2 + w_3 x_3 + w_4 x_4 + w_5 x_5 + w_6 x_1^2$$

$$f(x) = w_0 + w_1 x_1 + w_2 x_2 + w_3 x_3 + w_4 x_4 + w_5 x_5 + w_6 x_1^2 + w_7 * x_2 / x_3$$

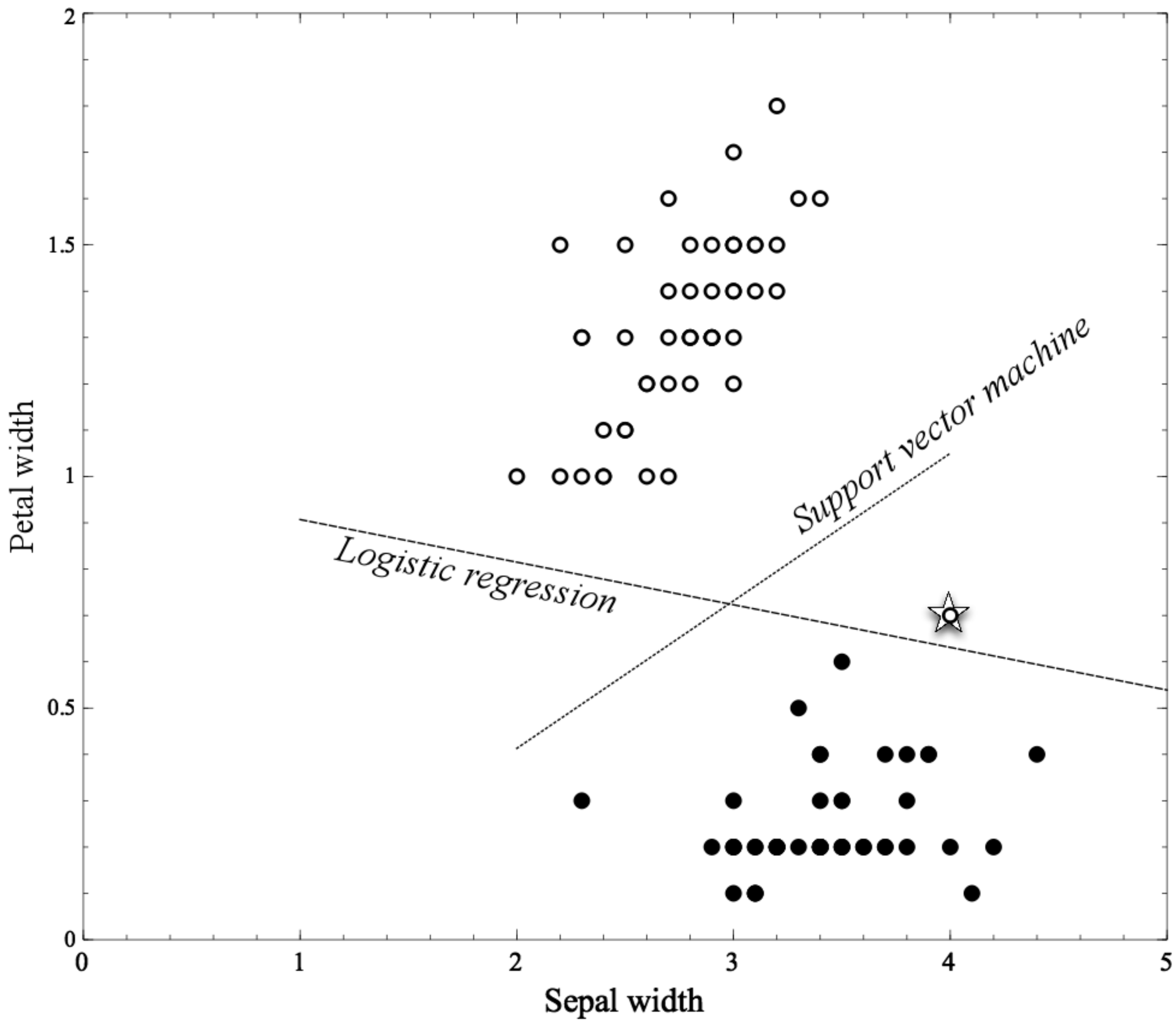
Example: Classifying Flowers



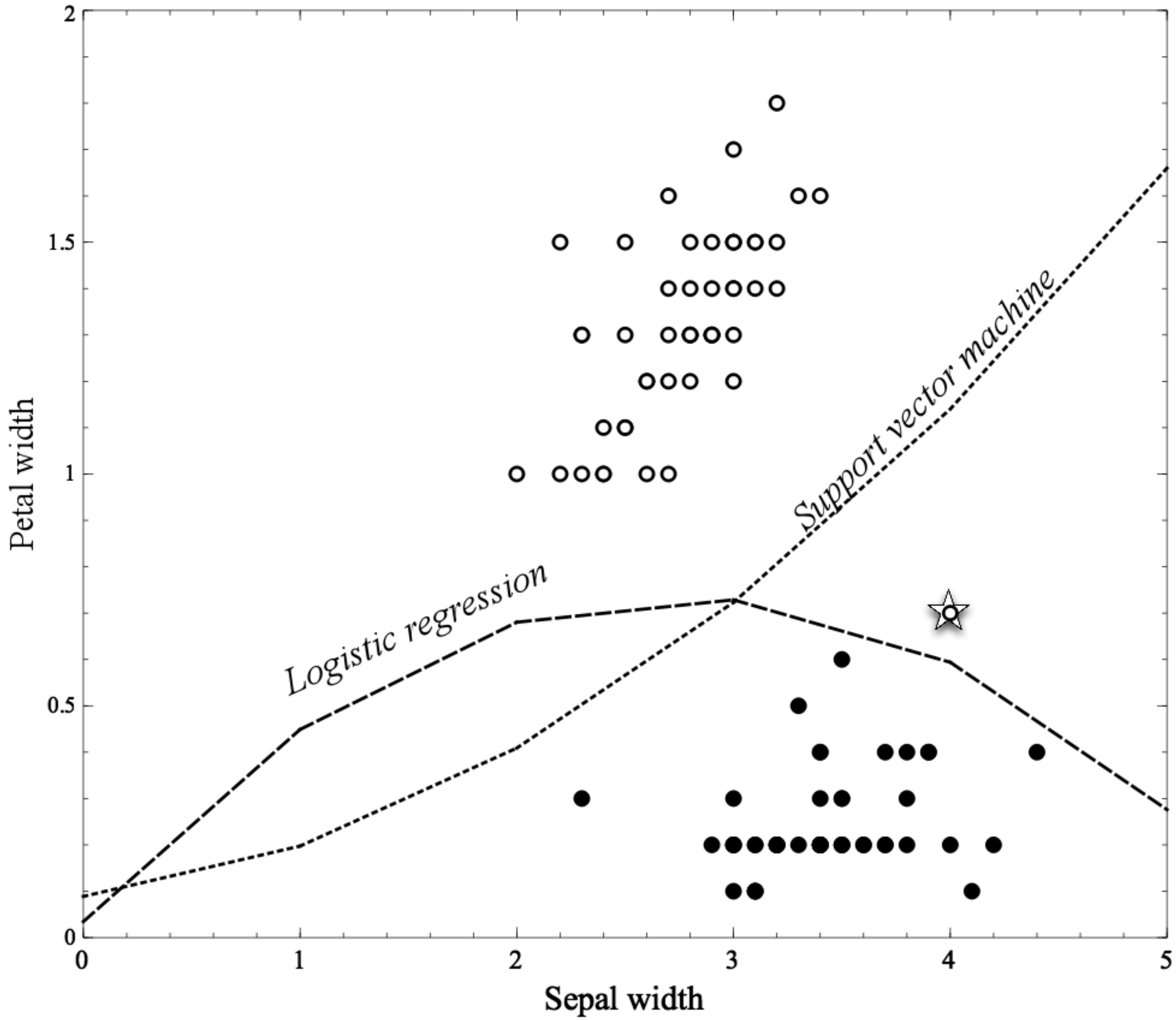
Example: Classifying Flowers



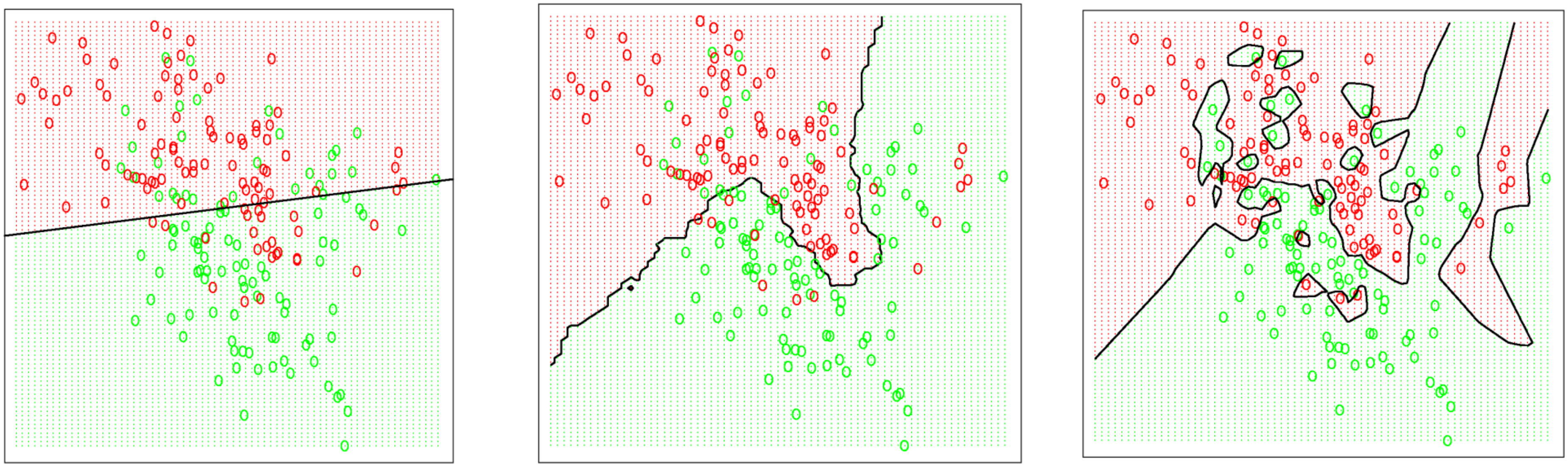
Example: Classifying Flowers



Example: Classifying Flowers



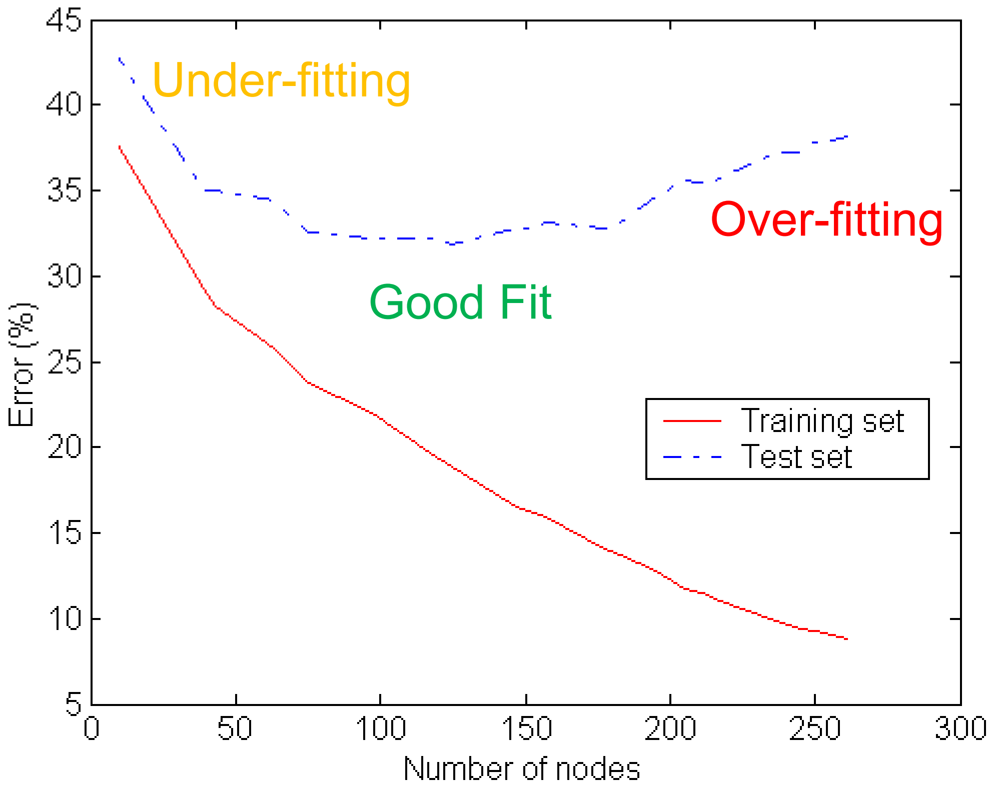
Need for holdout evaluation



Under-fitting | Good | Over-fitting

- In sample evaluation is in favor or “memorizing”
- On the *training data* the right model would be best
- But on *new data* it would be bad

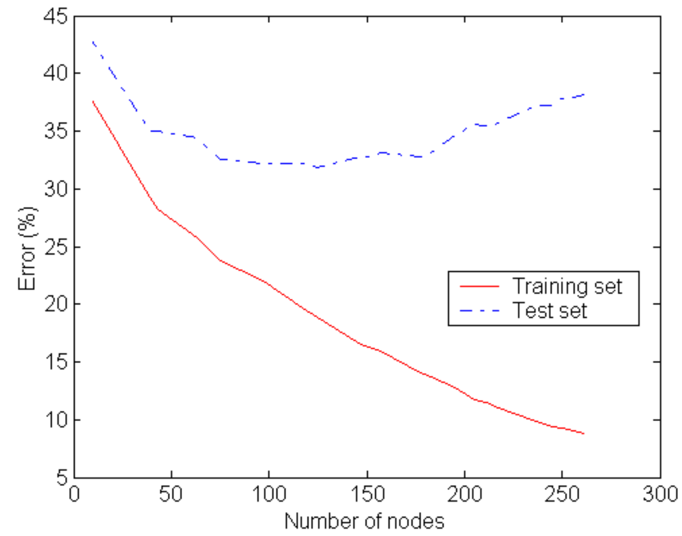
Over-fitting



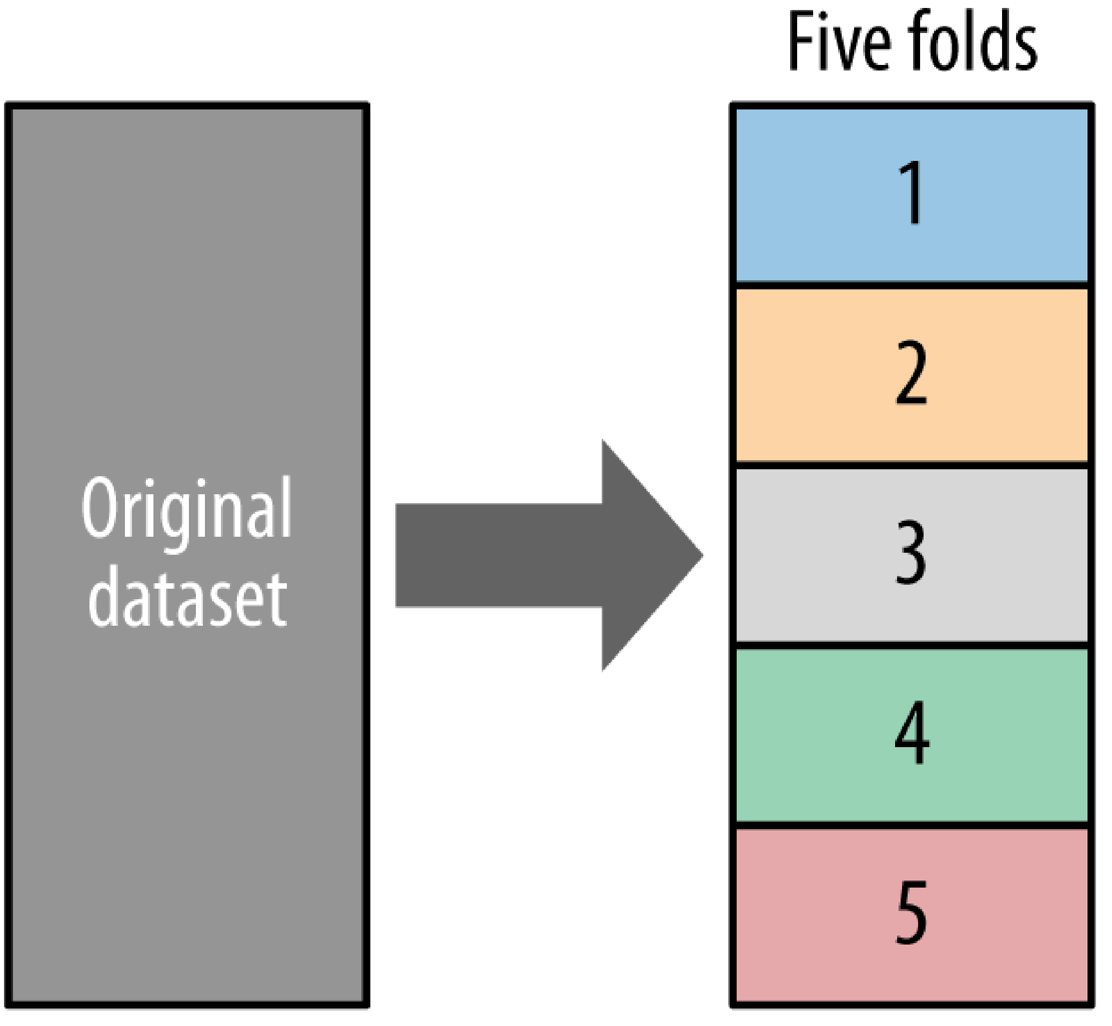
- Over-fitting: Model "memorizes" the properties of the particular training set rather than learning the underlying concept or phenomenon

Holdout validation

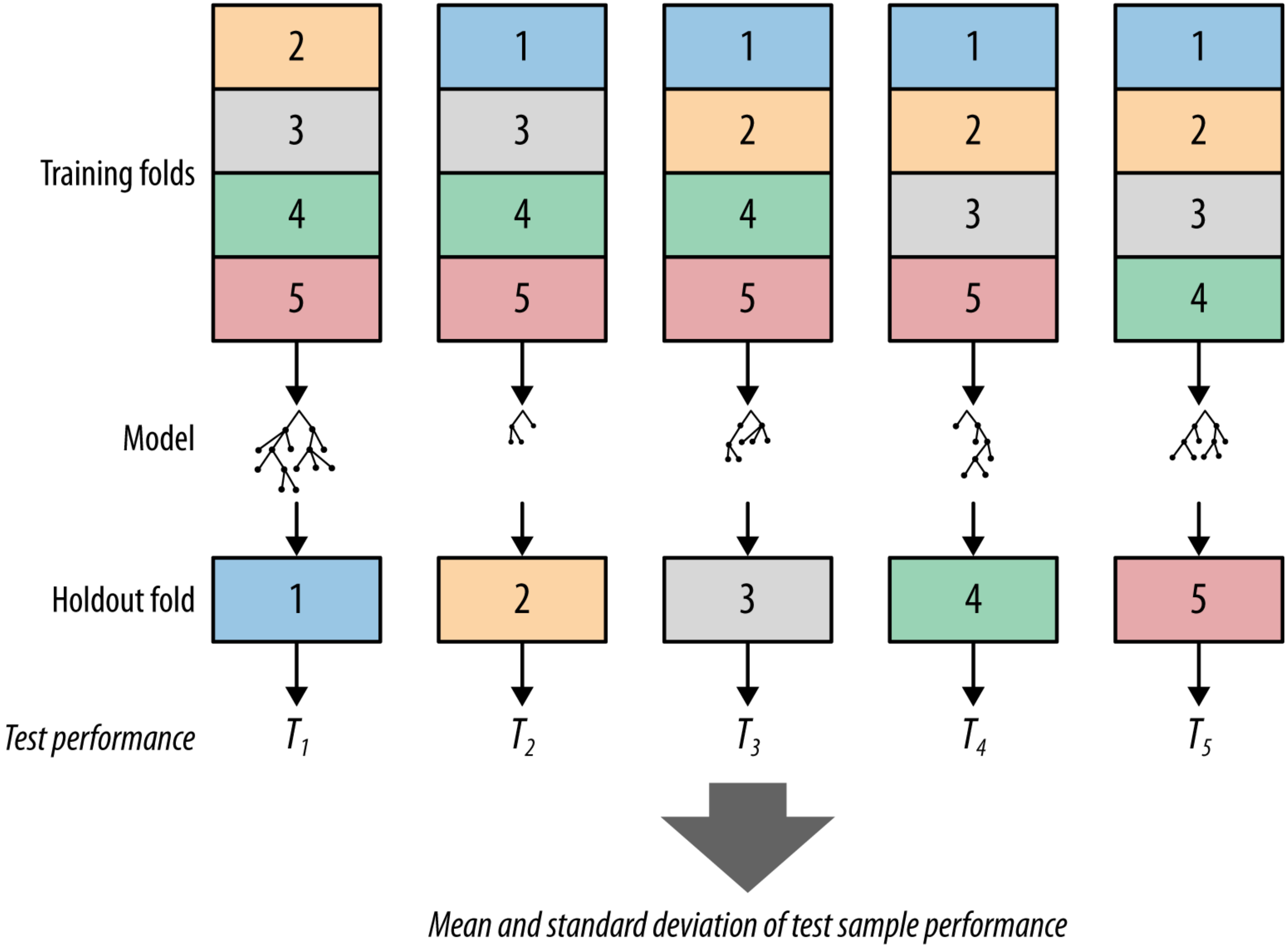
- We are interested in generalization
  - The performance on data not used for training
- Given only one data set, we hold out some data for evaluation
  - Holdout set for final evaluation is called the test set
- Accuracy on training data is sometimes called **“in-sample” accuracy**, vs. **“out-of-sample” accuracy** on test data



Cross-Validation



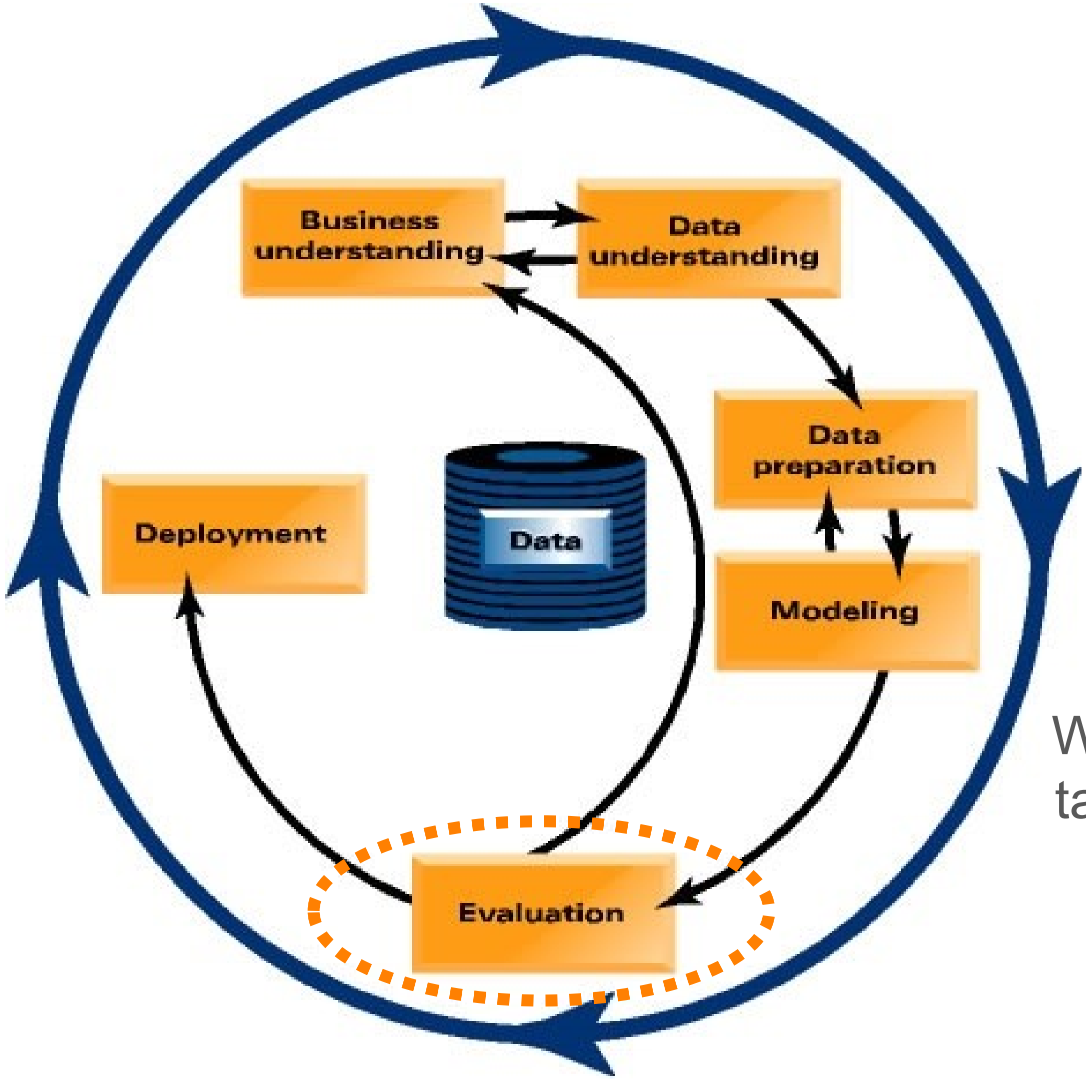
Cross-Validation



From Holdout Evaluation to Cross-Validation

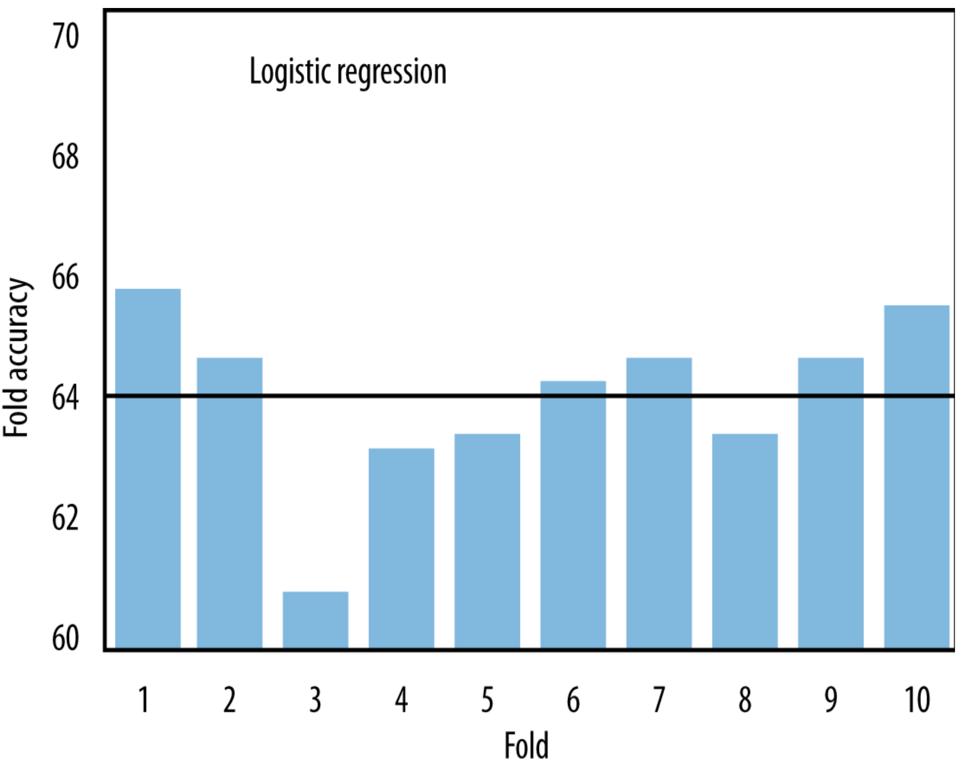
- Not only a simple estimate of the generalization performance, but also some statistics on the estimated performance,
  - such as the mean and variance
- Better use of a limited dataset
  - Cross-validation computes its estimates over *all* the data

Let's focus back in on actually mining the data..

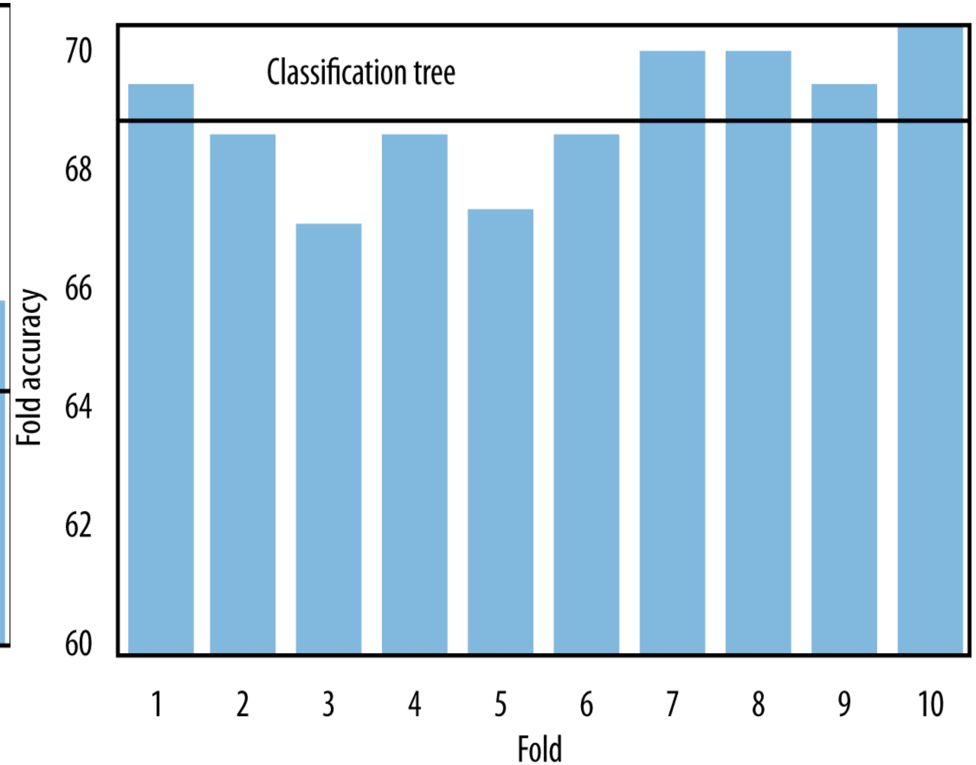


Which customers should TelCo target with a special offer, prior to contract expiration?

MegaTelCo



Logistic regression
Fold accuracy
70
68
66
64
62
60
1
2
3
4
5
6
7
8
9
10
Fold

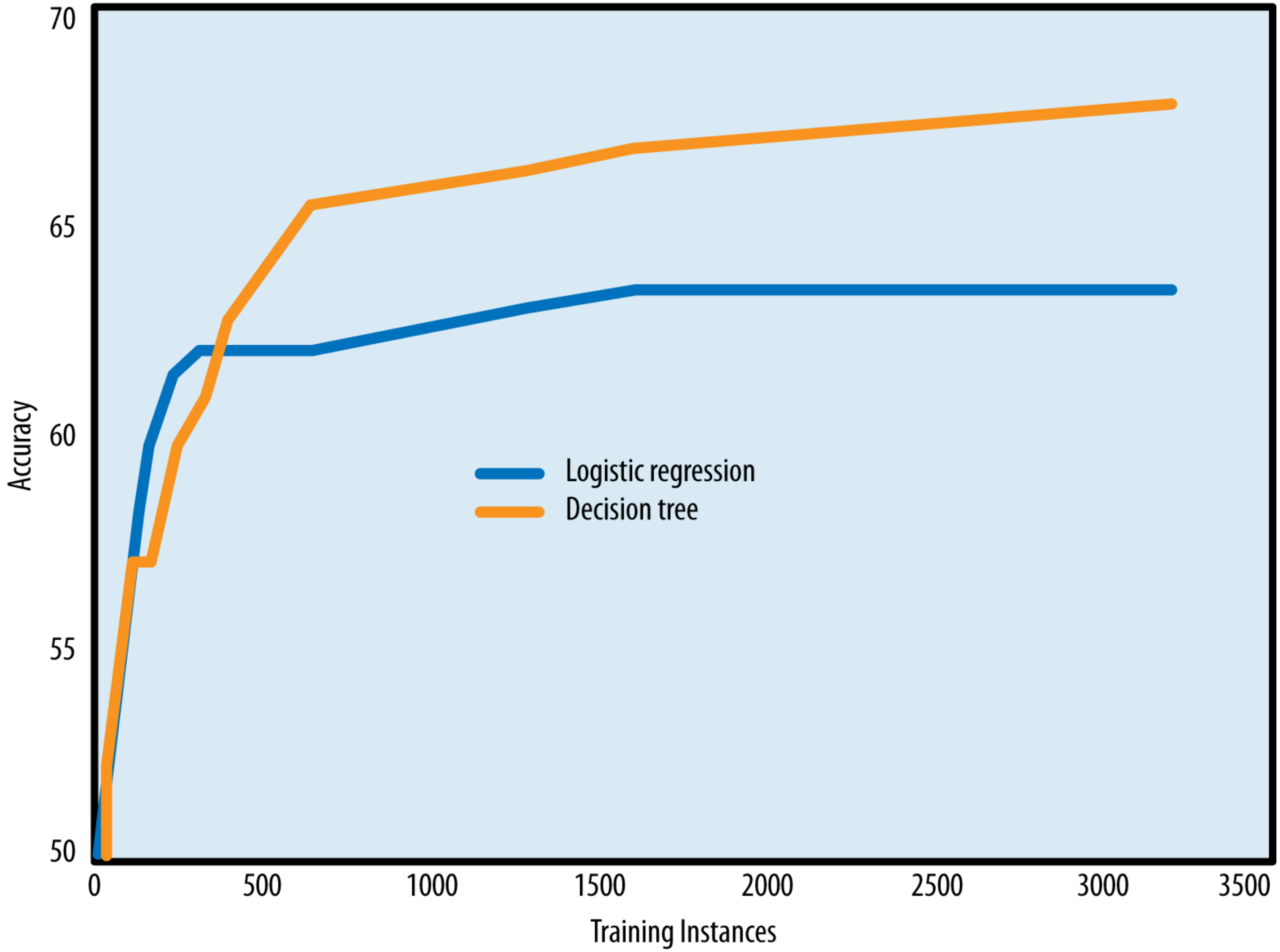


Classification tree
Fold accuracy
70
68
66
64
62
60
1
2
3
4
5
6
7
8
9
10
Fold

Generalization Performance

- Different modeling procedures may have different performance on the same data
- Different training sets may result in different generalization performance
- Different test sets may result in different estimates of the generation performance
- If the training set size changes, you may also expect different generalization performance from the resultant model

Learning Curves



Logistic Regression vs Tree Induction

- For smaller training-set sizes, logistic regression yields better generalization accuracy than tree induction
  - For smaller data, tree induction will tend to over-fit more
- Classification trees are a more flexible model representation than linear logistic regression
- Flexibility of tree induction can be an advantage with larger training sets:
  - Trees can represent substantially nonlinear relationships between the features and the target

Learning curves vs Fitting graphs

- A learning curve shows the generalization performance plotted against the amount of training data used
- A fitting graph shows the generalization performance as well as the performance on the training data, but plotted against model complexity
- Fitting graphs generally are shown for a fixed amount of training data

Avoiding Over-fitting

Tree Induction:

- Post-pruning
  - takes a fully-grown decision tree and discards unreliable parts
- Pre-pruning
  - stops growing a branch when information becomes unreliable

Linear Models:

- Feature Selection
- Regularization
  - Optimize some combination of fit and simplicity

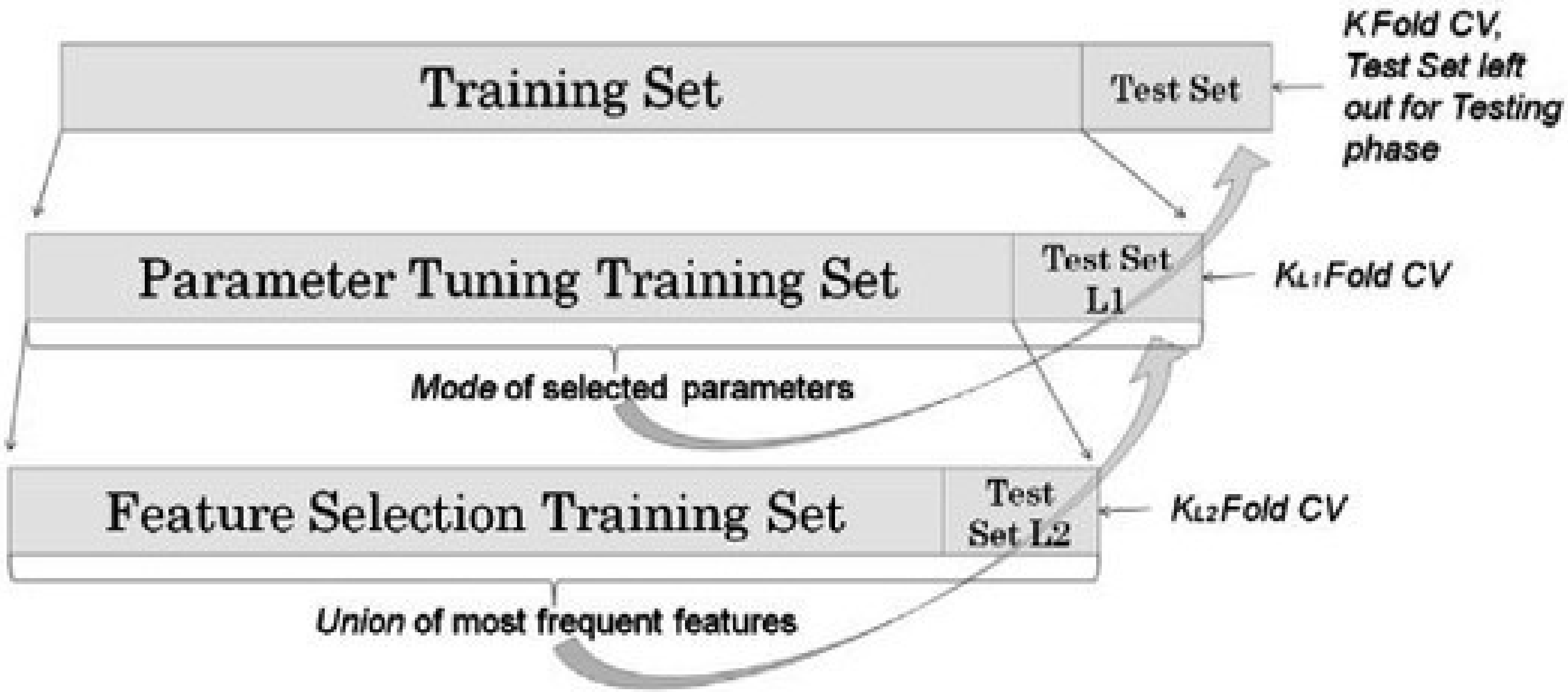
Regularization

Regularized linear model:

$$\underset{\boldsymbol{w}}{\operatorname{argmax}}[\operatorname{fit}(\boldsymbol{x}, \boldsymbol{w}) - \lambda * \operatorname{penalty}(\boldsymbol{w})]$$

- "L2-norm"
  - The sum of the *squares* of the weights
  - L2-norm + standard least-squares linear regression = ridge regression
- "L1-norm"
  - The sum of the *absolute values* of the weights
  - L1-norm + standard least-squares linear regression = lasso
  - Automatic feature selection

Nested Cross-Validation



Thanks!

Questions?