Model Performance and Visualization

Nov 15th '21

t-test chi-square p-value
dimensionality reduction

overfitting
bias-variance tradeoff
regularization
L1 L2 lasso ridge

true positive true negative
precision recall
sensitivity specificity
expected value

HYPOTHESIS TESTING

Three important tests

- **T-test:** compare two groups, or two interventions on one group.

- **CHI-squared and Fisher's test.** Compare the counts in a "contingency table".

- **ANOVA:** compare outcomes under several discrete interventions.

T-test

Single-sample: Compute the test statistic:

$$t = \frac{\bar{X}}{\bar{\sigma}}$$

where $\bar{X}$ is the sample mean and $\bar{\sigma}$ is the sample standard deviation, which is the square root of the sample variance Var(X).

If X is normally distributed, t is **almost** normally distributed, but not quite because of the presence of $\bar{\sigma}$.

You use the single-sample test for **one group** of individuals in **two conditions**. Just subtract the two measurements for each person, and use the difference for the single sample t-test.

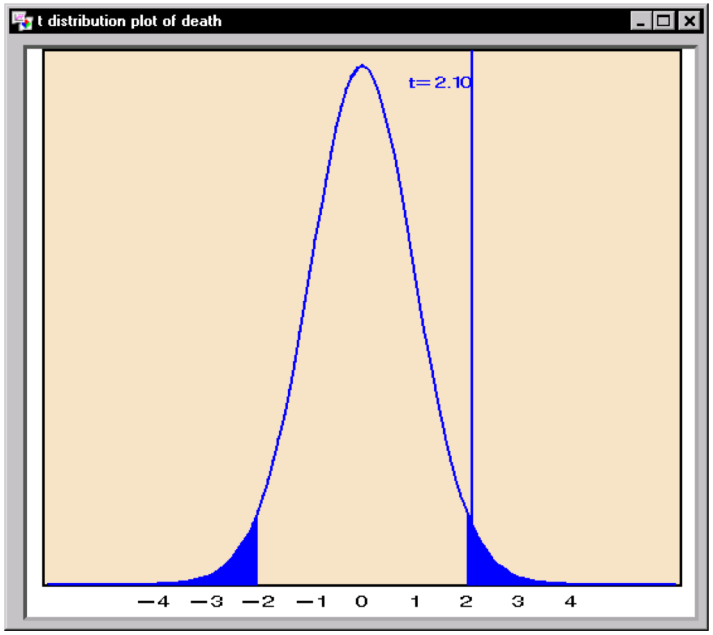
This is called a **within-subjects** design.

T-statistic and T-distribution

- If the underlying population has mean zero, the t-distribution should be distributed like this:

- The area of the tail beyond our measurement tells us how likely it is under the null hypothesis.

- If that probability is low (say < 0.05) we reject the null hypothesis.



Two sample T-test

In this test, there are **two samples** X_1 and X_2. A t statistic is constructed from their sample means and sample standard deviations:

$$t = \frac{\overline{X}_1 - \overline{X}_2}{s_{\overline{X}_1 - \overline{X}_2}}$$

where: $s_{\overline{X}_1 - \overline{X}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}.$

This design is called a **between-subjects** test.

Chi-squared test

Often you will be faced with discrete (count) data. Given a table like this:

	Prob(X)	Count(X)
X=0	0.3	10
X=1	0.7	50

Where Prob(X) is part of a null hypothesis about the data (e.g. that a coin is fair).

The CHI-squared statistic lets you test whether an observation is consistent with the data:

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

O_i is an observed count, and E_i is the expected value of that count. It has a chi-squared distribution, whose p-values you compute to do the test.

ANOVA

In ANOVA we compute a single statistic (an F-statistic) that compares variance between groups with variance within each group.

$$F = \frac{VAR_{between}}{VAR_{within}}$$

The higher the F-value is, the less probable is the null hypothesis that the samples all come from the same population.

We can look up the F-statistic value in a cumulative F-distribution (similar to the other statistics) to get the p-value.

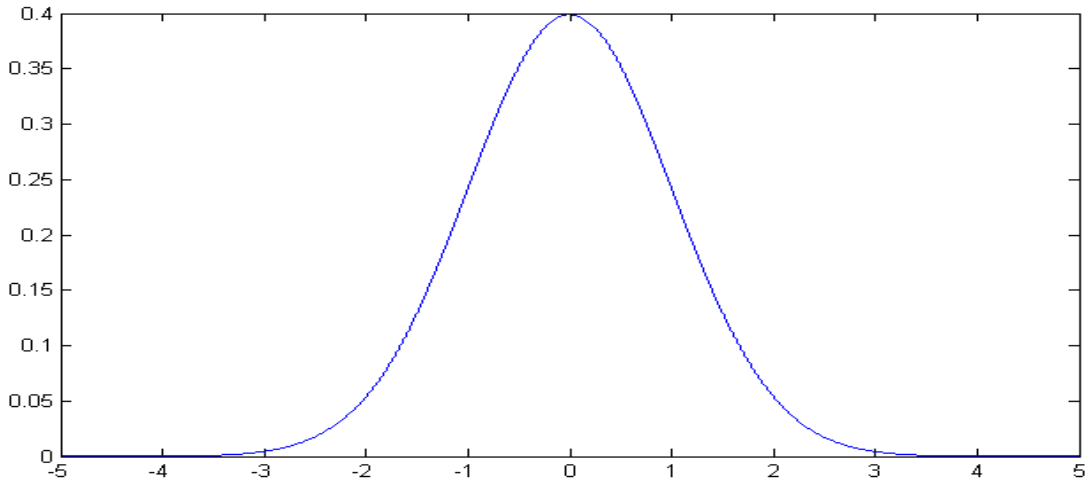
ANOVA tests can be much more complicated, with multiple dependent variables, hierarchies of variables, correlated measurements etc.

Central Limit Theorem

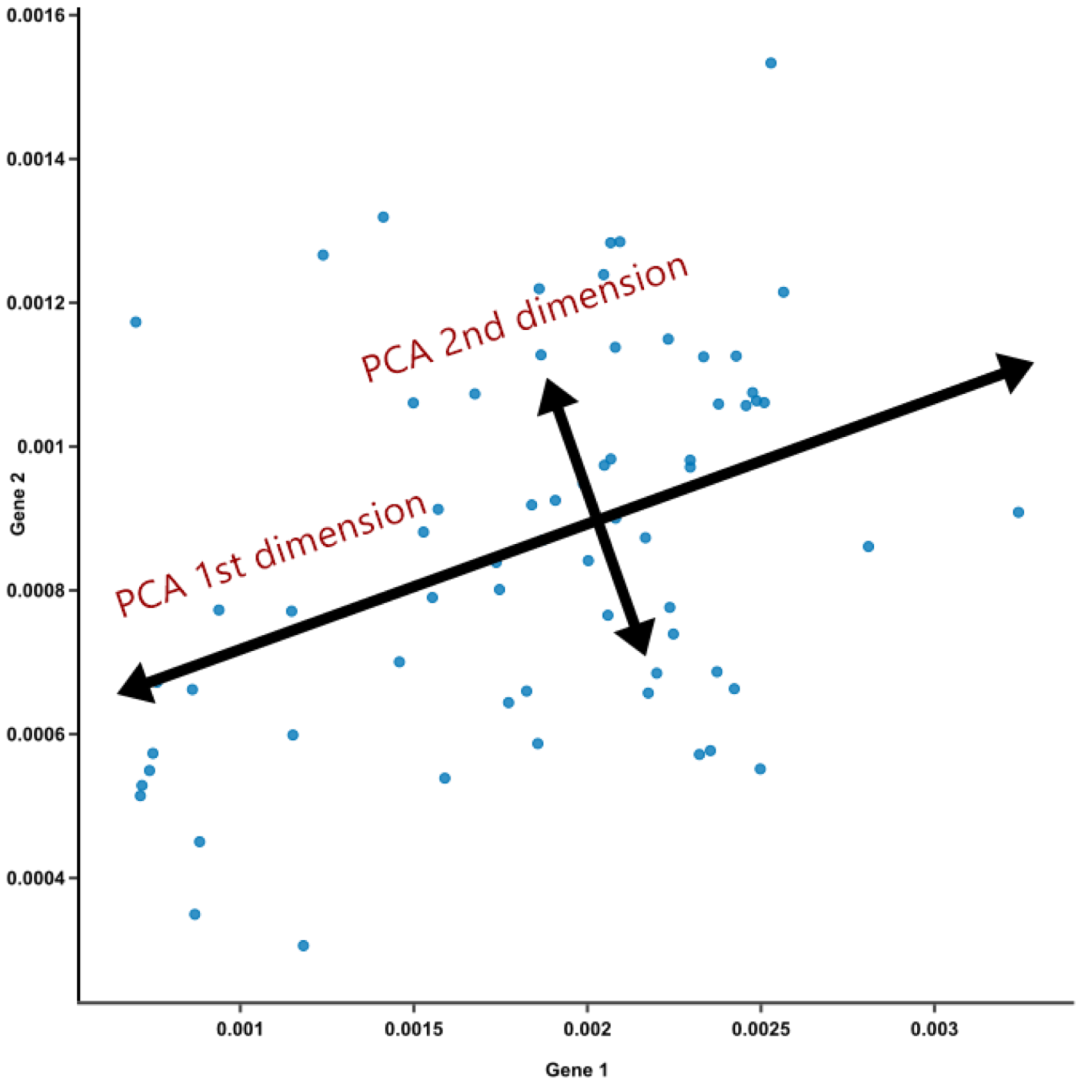
The distribution of the sum (or mean) of a set of n identically-distributed random variables Xi approaches a normal distribution as $n \rightarrow \infty$.

The common parametric statistical tests, like t-test and ANOVA assume normally-distributed data, but depend on sample mean and variance measures of the data.

They typically work reasonably well for data that are not normally distributed as long as the samples are not too small.

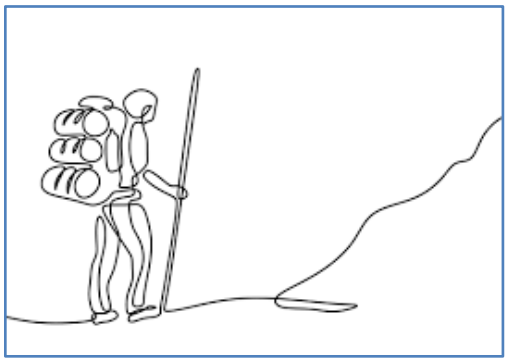


DIMENSIONALITY REDUCTION



https://setosa.io/ev/principal-component-analysis/

MACHINE LEARNING FRAMEWORKS

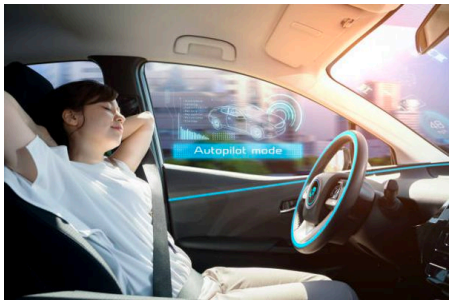


Statistical analysis on paper

Spreadsheets

Statistical analysis with computer tools
R, SPSS, ...

Machine Learning
Weka, SciKit, TensorFlow, PyTorch, ...



***Auto** ML*

Google Cloud AutoML,
AzureML, AWS SageMaker

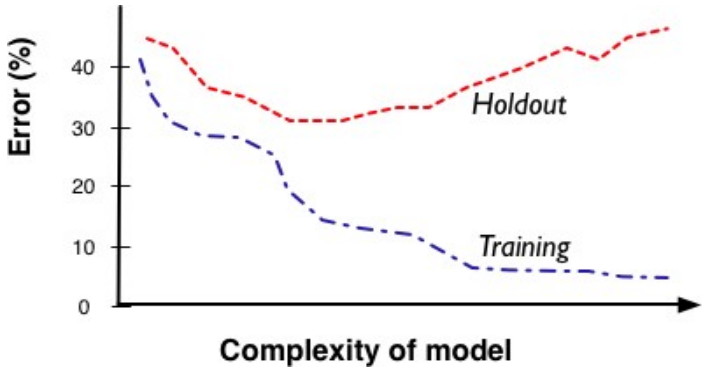
MODEL ASSESSMENT

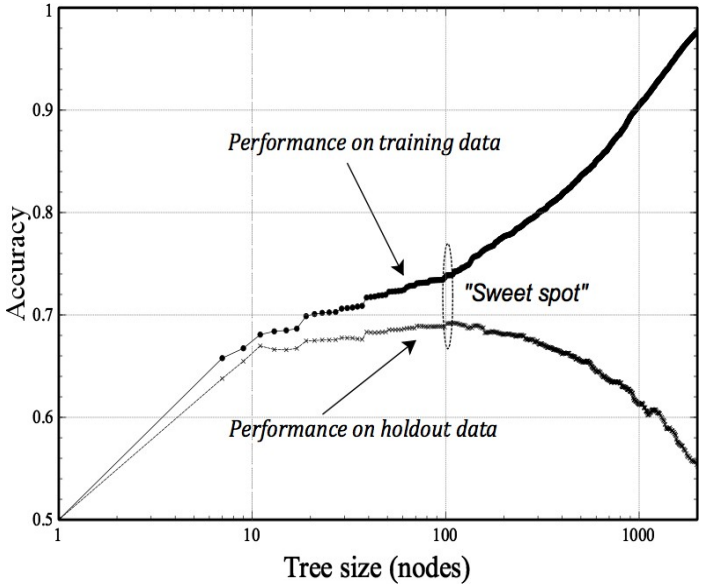
Restaurant	Cuisine			Good for kids	Likes
Art of the Table	Farm fresh, vegetarian			N	
Canlis	American			Y	
Lark	Plant based			Y	
Nue	Vegetarian			N	
Kati Thai	Vegan			Y	

Restaurant	Cuisine	Price	Outdoor dining	Good for kids	Front door color	Likes
Art of the Table	Farm fresh, vegetarian	$$$	Y	N	Brown	
Canlis	American	$$$$	Y	Y	Red	
Lark	Plant based	$$$$	N	Y	Grey	
Nue	Vegetarian	$$	N	N	Grey	
Kati Thai	Vegan	$	Y	Y	Black	

Over-fitting the data

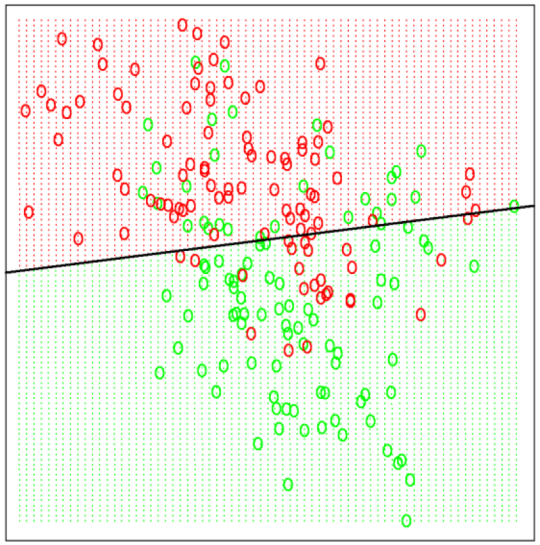
- Finding chance occurrences in data that look like interesting patterns, but which do not generalize, is called over-fitting the data
- We want models to apply not just to the exact training set but to the general population from which the training data came
  - Generalization

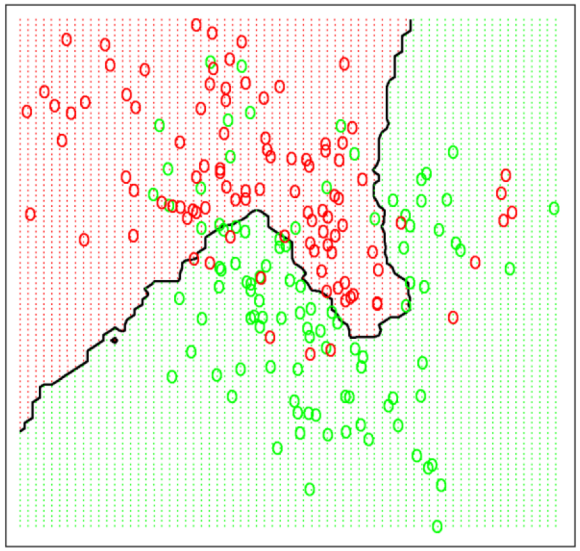


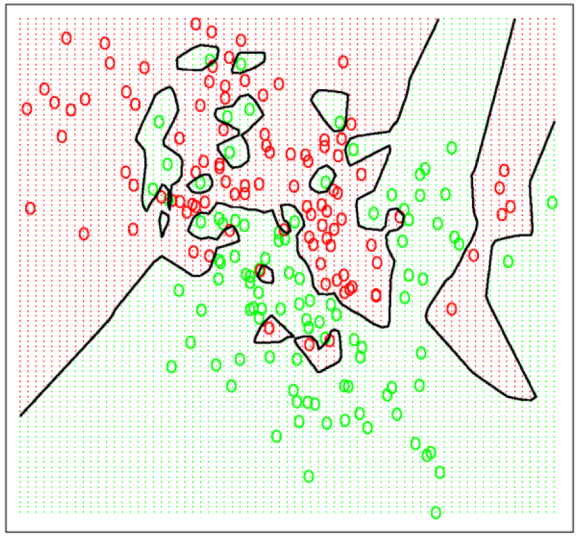


$$f(x) = w_0 + w_1x_1 + w_2x_2 + w_3x_3 + w_4x_4 + w_5x_5 + w_6x^2 + w_7 * x_2/x_3$$

Need for holdout evaluation







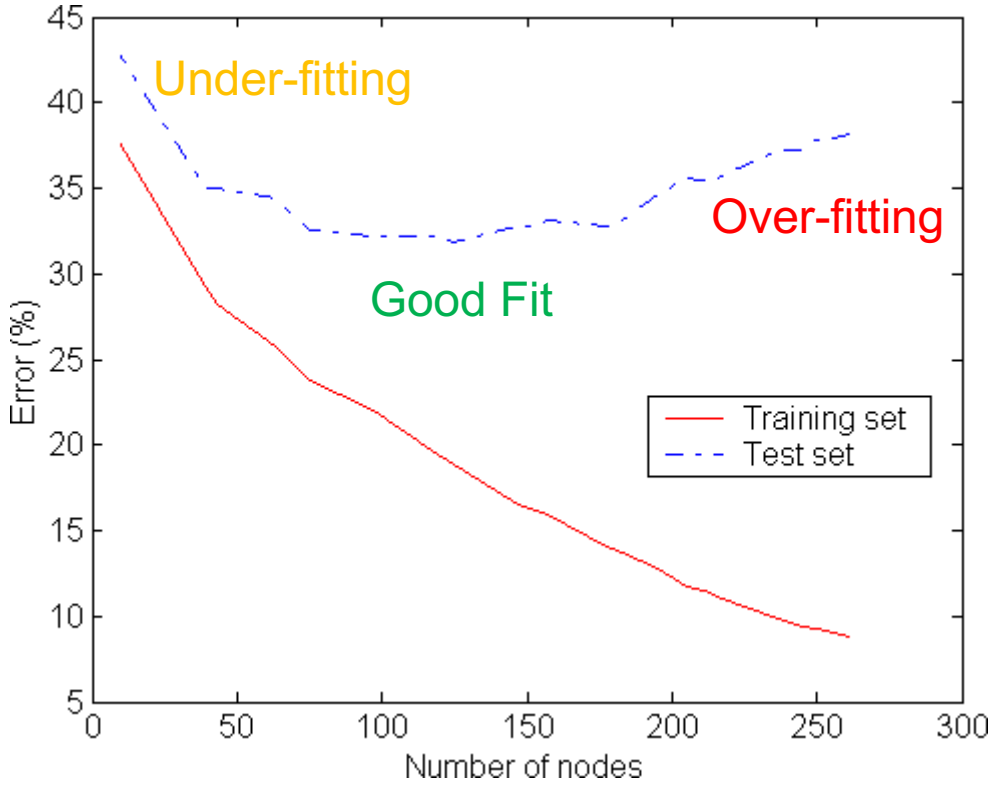
Under-fitting

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

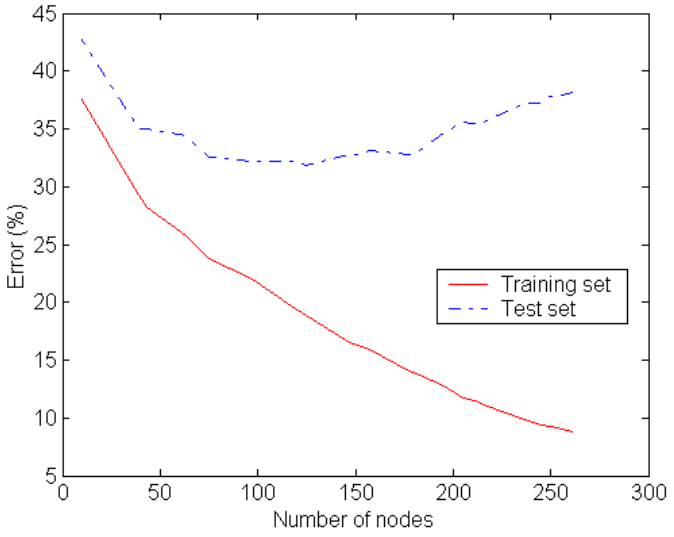
Over-fitting



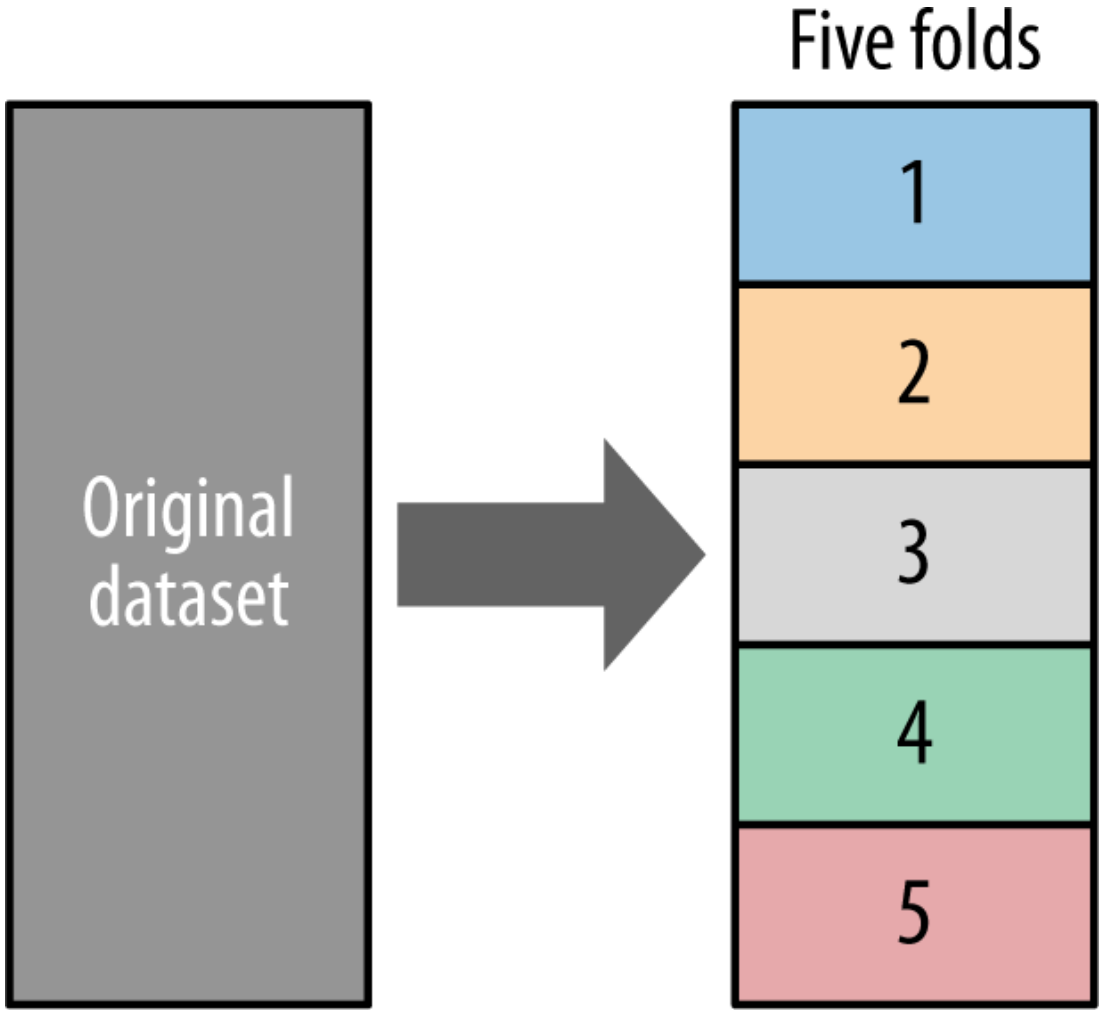
- Over-fitting: Model "memorizes" the properties of the particular training set rather than learning the underlying concept or phenomenon

Holdout validation

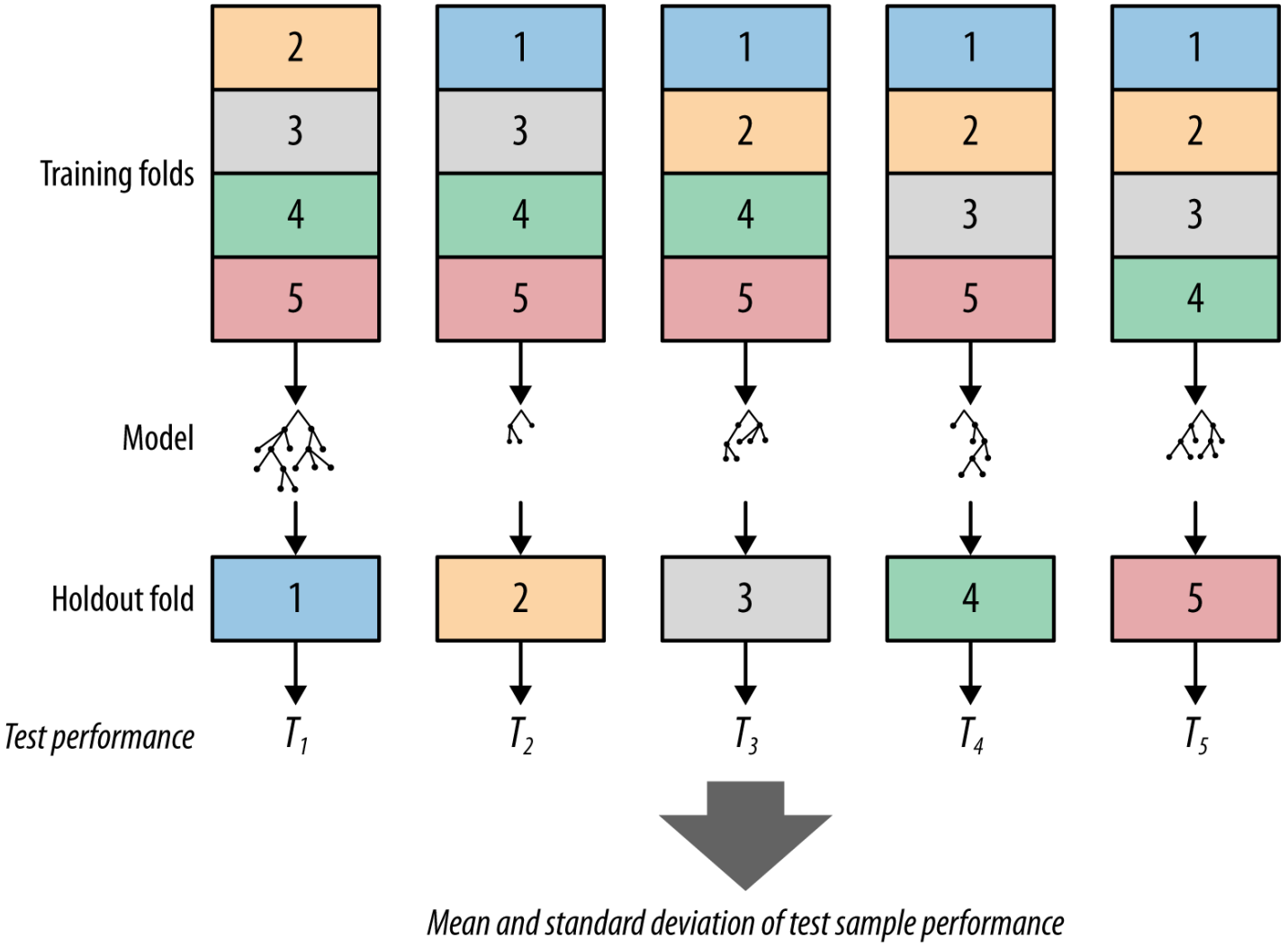
- We are interested in generalization
  - The performance on data not used for training
- Given only one data set, we hold out some data for evaluation
  - Holdout set for final evaluation is called the test set
- Accuracy on training data is sometimes called **"in-sample" accuracy**, vs. **"out-of-sample"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



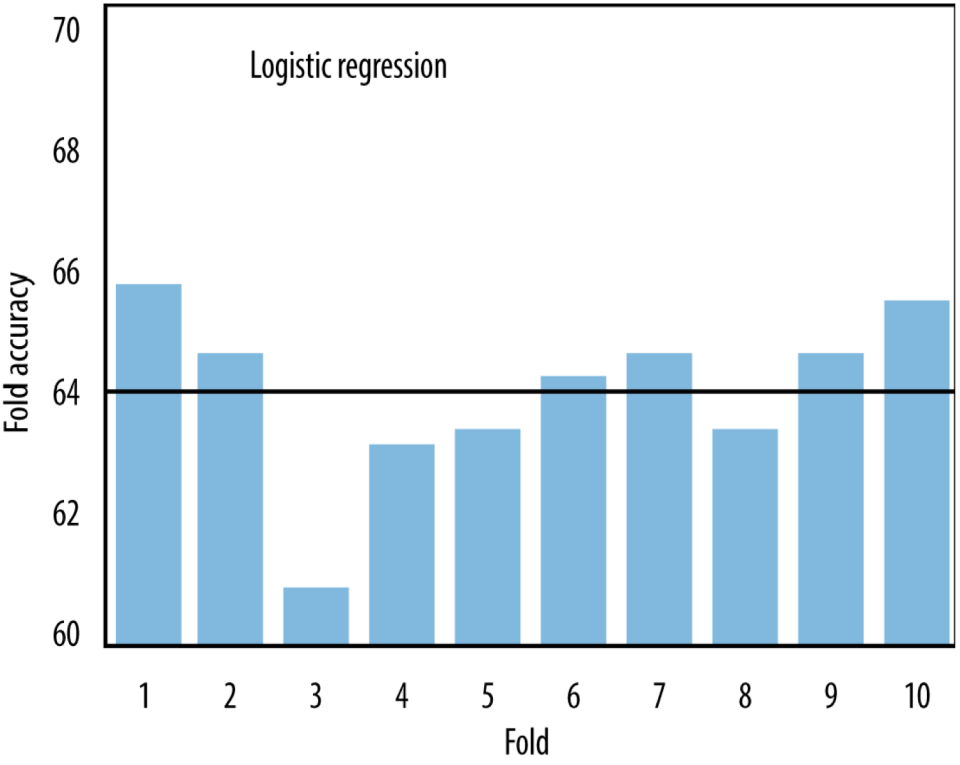
Cross-Validation

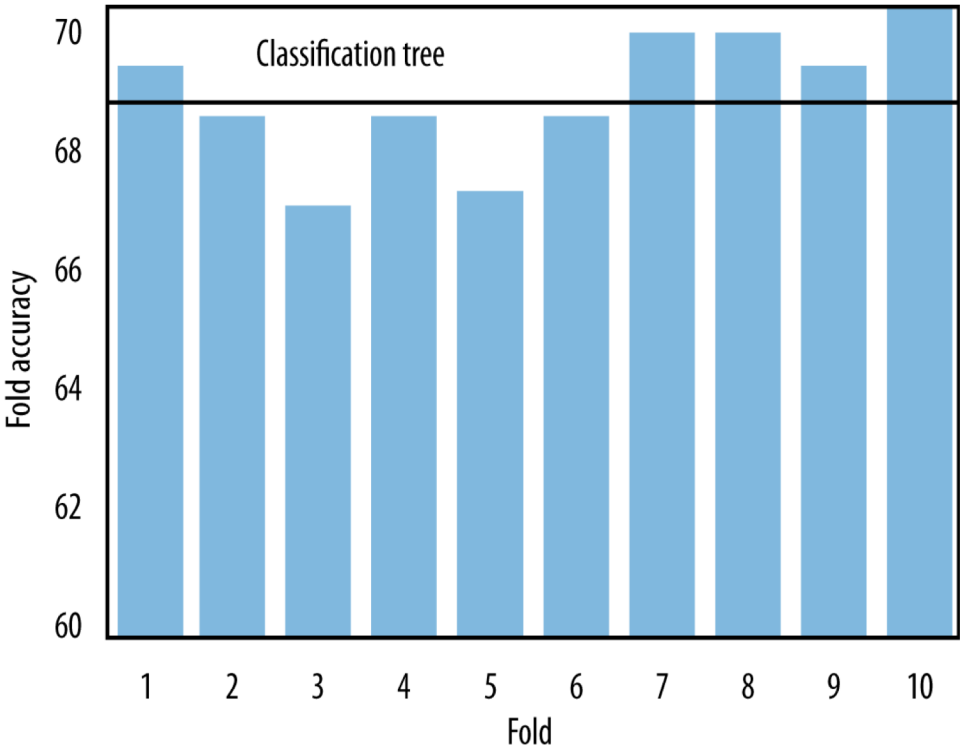


Cross-Validation



MegaTelCo

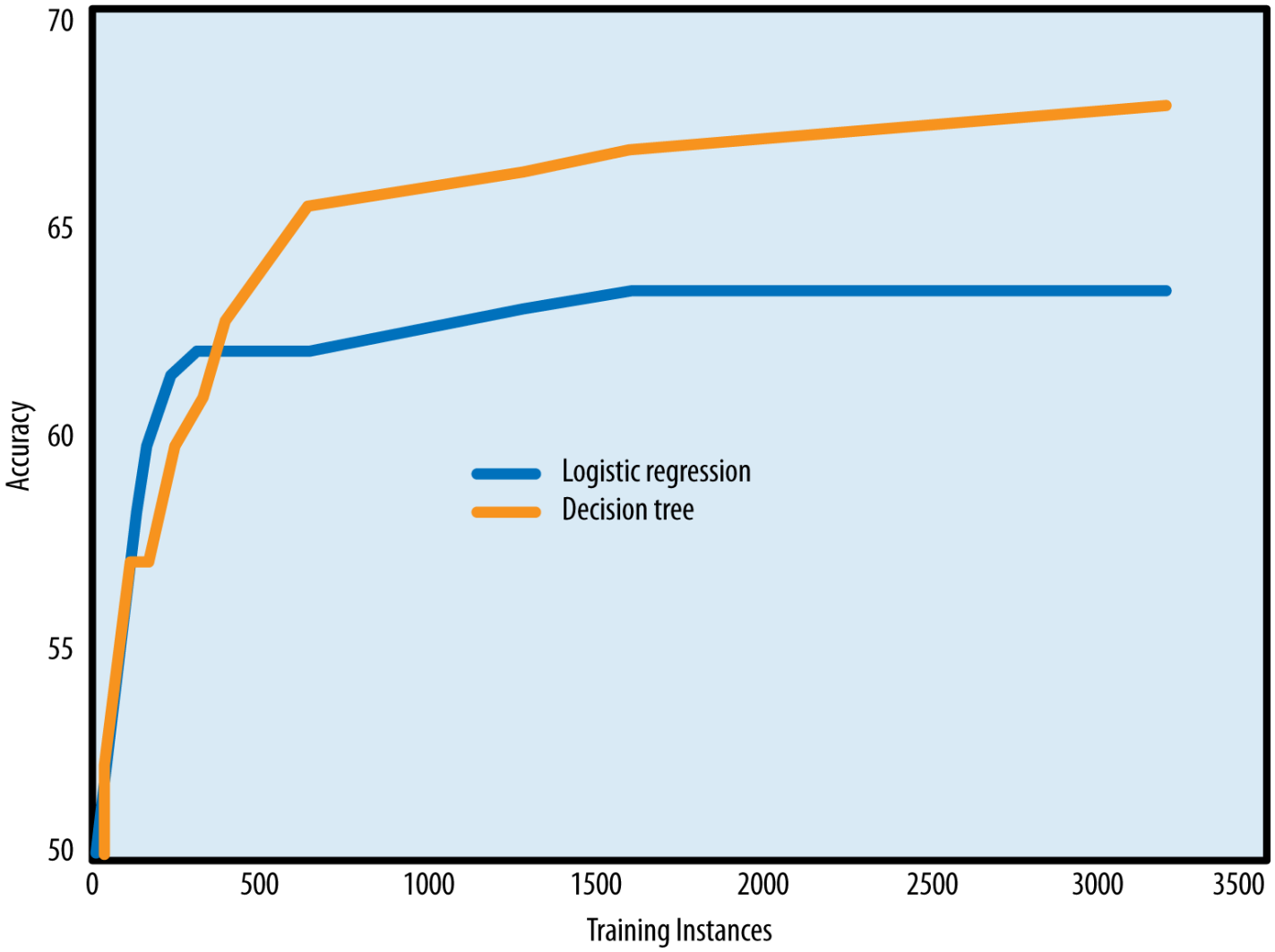

Logistic regression
Fold accuracy
70
68
66
64
62
60
1
2
3
4
5
6
7
8
9
10
Fold


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



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:

$$\arg\max_{\mathbf{w}} \; [\text{fit}(\mathbf{x}, \mathbf{w}) - \lambda \cdot \text{penalty}(\mathbf{w})]$$

- "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

Evaluating Classifiers: Plain Accuracy

$$\text{Accuracy} = \frac{\text{Number of correct decisions made}}{\text{Total number of decisions made}}$$

$$= 1 - \text{error rate}$$

- *Too* simplistic..

Evaluating Classifiers: The Confusion Matrix

- A confusion matrix for a problem involving $n$classes is an $n\times n$ matrix,
  - with the columns labeled with actual classes and the rows labeled with predicted classes
- It separates out the decisions made by the classifier,
  - making explicit how one class is being confused for another

	p	n
Y	True Positives	False Positives
N	False Negatives	True Negatives

- The errors of the classifier are the false positives and false negatives

Building a Confusion Matrix

Default Truth	Model Prediction
0	0
1	1
0	1
0	1
0	0
1	1
0	0
0	0
1	1
1	0

Predicted class Actual class	Default	No Default	Total
Default	3	1	4
No Default	2	4	6
Total	5	5	10

Other Evaluation Metrics

- $\text{Precision} = \frac{TP}{TP + FP}$
- $\text{Recall} = \frac{TP}{TP + FN}$
- $\text{F-measure} = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$

A Key Analytical Framework: Expected Value

- The expected value computation provides a framework that is useful in organizing thinking about data-analytic problems
- It decomposes data-analytic thinking into:
  - the structure of the problem,
  - the elements of the analysis that can be extracted from the data, and
  - the elements of the analysis that need to be acquired from other sources
- The general form of an expected value calculation:

$$EV = p(o_1) \cdot v(o_1) + p(o_2) \cdot v(o_2) + p(o_3) \cdot v(o_3) \ldots$$

$$\text{Expected benefit of targeting} = p_R(\mathbf{x}) \cdot v_R + [1 - p_R(\mathbf{x})] \cdot v_{NR}$$

Expected Value Framework in Use Phase

- Online marketing:

- Expected benefit of targeting = p_R

$$\text{Expected benefit of targeting} = p_R(\mathbf{x}) \cdot v_R + [1 - p_R(\mathbf{x})] \cdot v_{NR}$$

- Product Price: $200

- Product Cost: $100

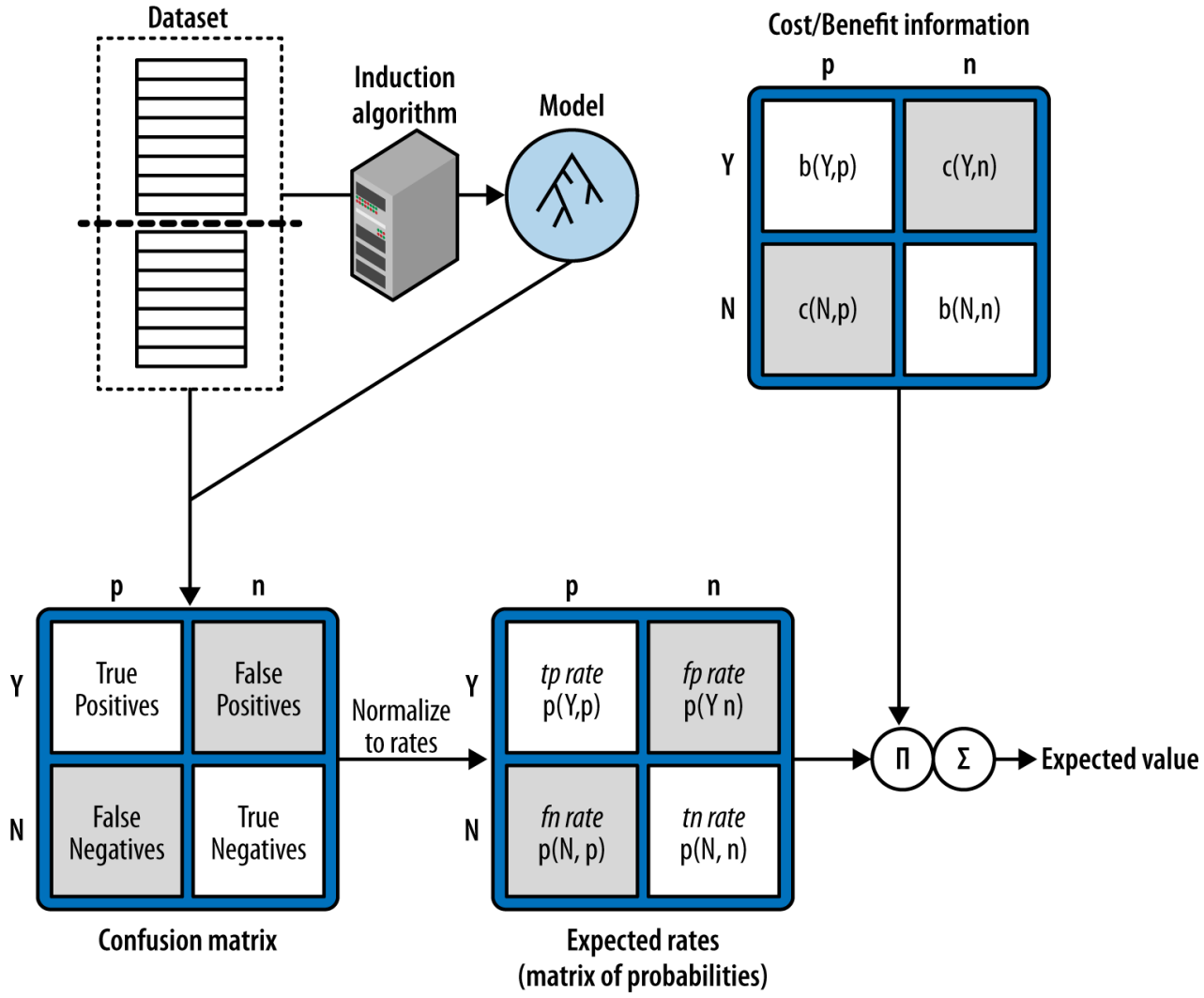
- Targeting Cost: $1

$$p_R(\mathbf{x}) \cdot \$99 - [1 - p_R(\mathbf{x})] \cdot \$1 > 0$$

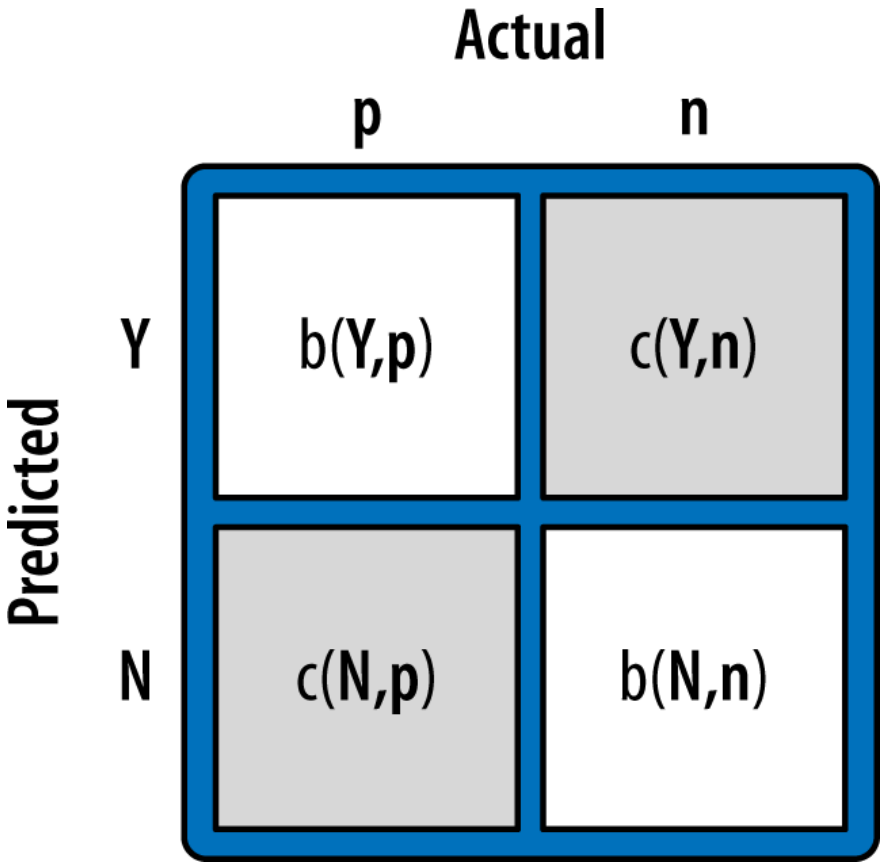
$$p_R(\mathbf{x}) \cdot \$99 > [1 - p_R(\mathbf{x})] \cdot \$1$$

$$p_R(\mathbf{x}) > 0.01$$

Using Expected Value to Frame Classifier Evaluation



A cost-benefit matrix



A cost-benefit matrix for the marketing example

		Actual	
		p	**n**
Predicted	**Y**	99	-1
	N	0	0

Using Expected Value to Frame Classifier Evaluation

$$p(x, y) = p(y) \cdot p(x \mid y)$$

$$\textit{Expected profit} = p(\mathbf{Y}, \mathbf{p}) \cdot b(\mathbf{Y}, \mathbf{p}) + p(\mathbf{N}, \mathbf{p}) \cdot b(\mathbf{N}, \mathbf{p}) + p(\mathbf{N}, \mathbf{n}) \cdot b(\mathbf{N}, \mathbf{n}) + p(\mathbf{Y}, \mathbf{n}) \cdot b(\mathbf{Y}, \mathbf{n})$$

$$\text{Expected profit} = p(\mathbf{Y} \mid \mathbf{p}) \cdot p(\mathbf{p}) \cdot b(\mathbf{Y}, \mathbf{p}) + p(\mathbf{N} \mid \mathbf{p}) \cdot p(\mathbf{p}) \cdot b(\mathbf{N}, \mathbf{p}) + p(\mathbf{N} \mid \mathbf{n}) \cdot p(\mathbf{n}) \cdot b(\mathbf{N}, \mathbf{n}) + p(\mathbf{Y} \mid \mathbf{n}) \cdot p(\mathbf{n}) \cdot b(\mathbf{Y}, \mathbf{n})$$

	p	n
Y	56	7
N	5	42

T = 110

P = 61 N = 49

$p(\mathbf{p}) = 0.55$ $p(\mathbf{n}) = 0.45$

tp rate = 56/61 = 0.92 *fp rate* = 7/49 = 0.14

fn rate = 5/61 = 0.08 *tn rate* = 42/49 = 0.86

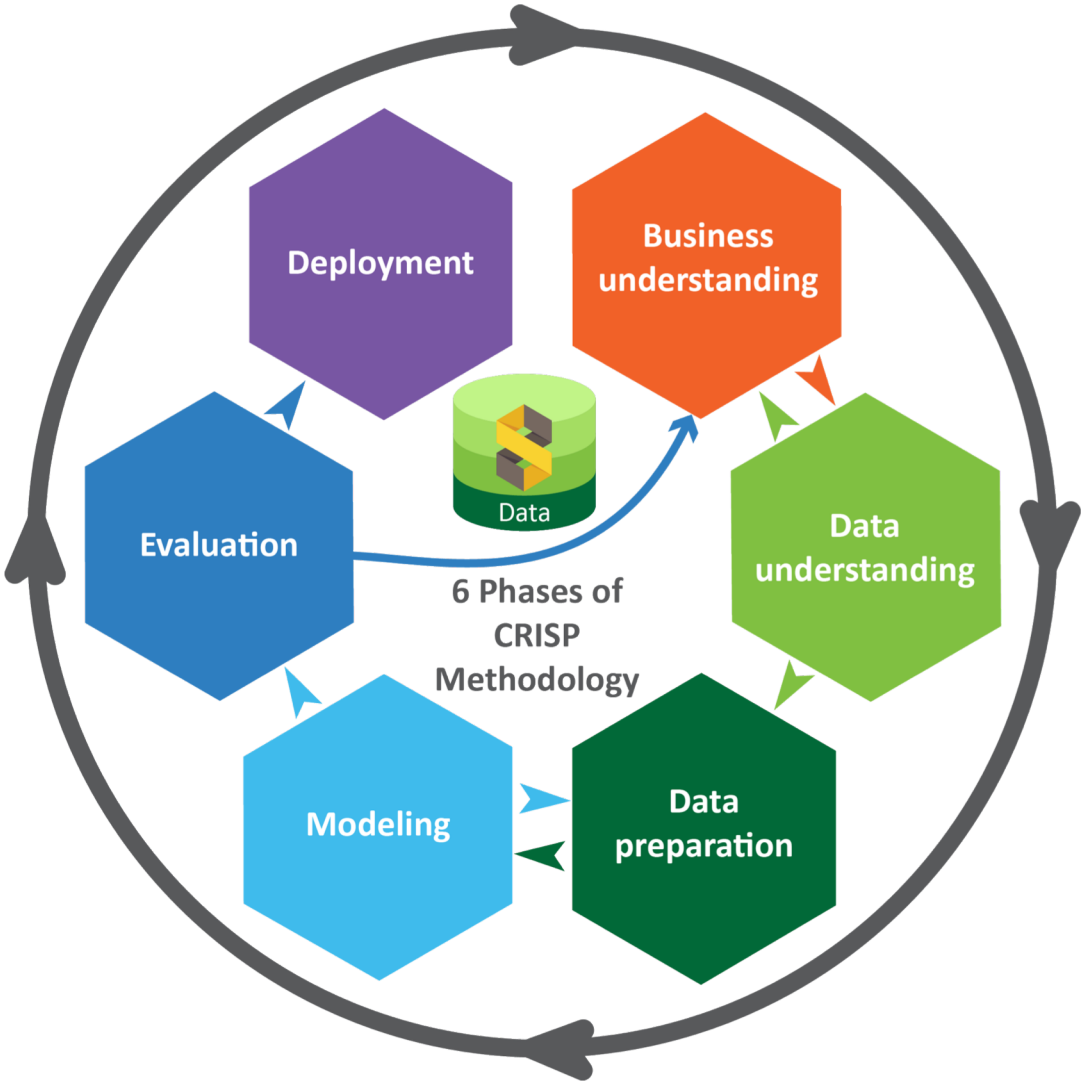
$$
\begin{aligned}
\text{expected profit} &= p(\mathbf{p}) \cdot [\, p(\mathbf{Y} \mid \mathbf{p}) \cdot b(\mathbf{Y}, \mathbf{p}) + p(\mathbf{N} \mid \mathbf{p}) \cdot c(\mathbf{N}, \mathbf{p})] + \\
&\quad\; p(\mathbf{n}) \cdot [\, p(\mathbf{N} \mid \mathbf{n}) \cdot b(\mathbf{N}, \mathbf{n}) + p(\mathbf{Y} \mid \mathbf{p}) \cdot c(\mathbf{Y}, \mathbf{n})] \\
&= 0.55 \cdot [0.92 \cdot b(\mathbf{Y}, \mathbf{p}) + 0.08 \cdot b(\mathbf{N}, \mathbf{p})] + \\
&\quad\; 0.45 \cdot [0.86 \cdot b(\mathbf{N}, \mathbf{n}) + 0.14 \cdot p(\mathbf{Y}, \mathbf{n})] \\
&= 0.55 \cdot [0.92 \cdot 99 + 0.08 \cdot 0] + \\
&\quad\; 0.45 \cdot [0.86 \cdot 0 + 0.14 \cdot \text{ - } 1] \\
&= 50.1 \text{ - } 0.063 \\
&\approx \mathbf{\$50.04}
\end{aligned}
$$

Other terms

Precision = TP / (TP+FP)
Recall = TP / (TP + FN)

Sensitivity = $TN / (TN + FP)$ = True negative rate = 1 - False positive rate
Specificity = $TP / (TP + FN)$ = True positive rate

PROJECT !


Deployment
Business understanding
Data understanding
Data preparation
Modeling
Evaluation
Data
6 Phases of CRISP Methodology

Project Aspects, Criteria

- Due date
- Components
- Approach
- Evaluation