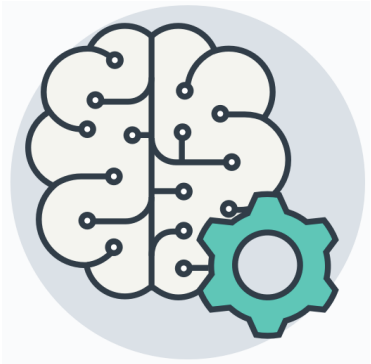
Aprendizado de Máquina

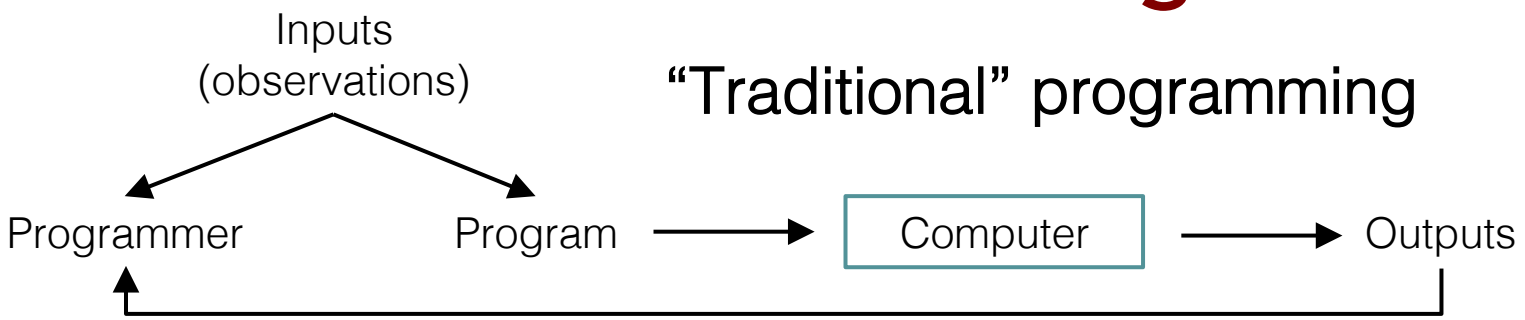
Visão Geral



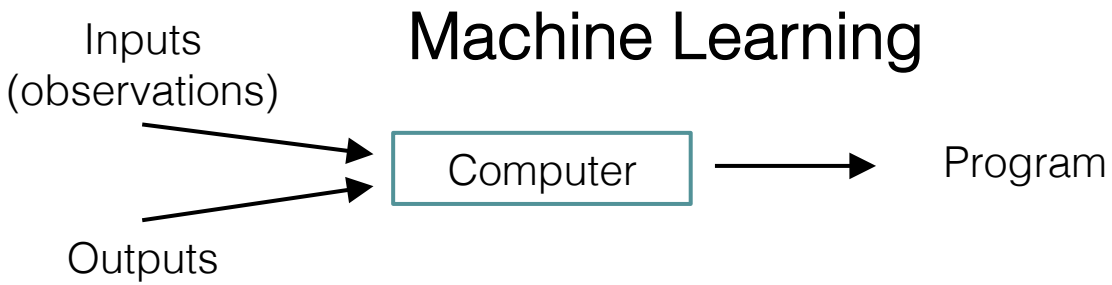
Prof. Regis Pires Magalhães

regismagalhaes@ufc.br - http://bit.ly/ufcregis

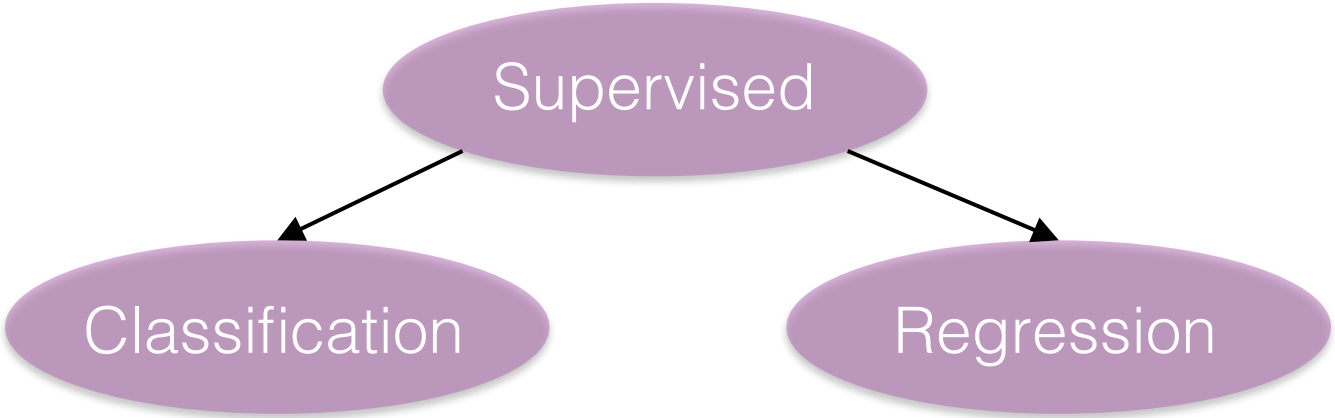
What is Machine Learning?

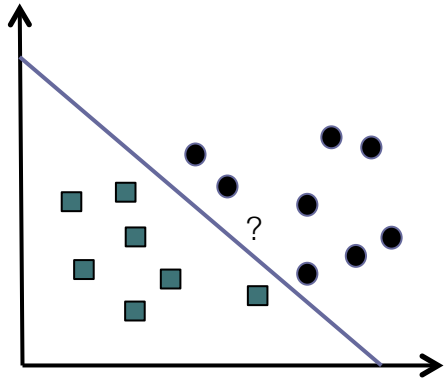


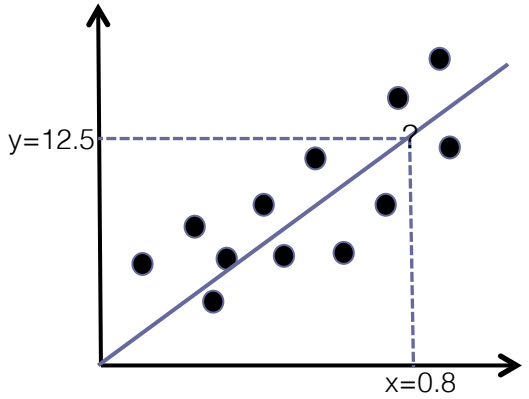
Machine Learning is the field of study that gives computers the ability to learn without being explicitly programmed.
-- Arthur Samuel (1959)



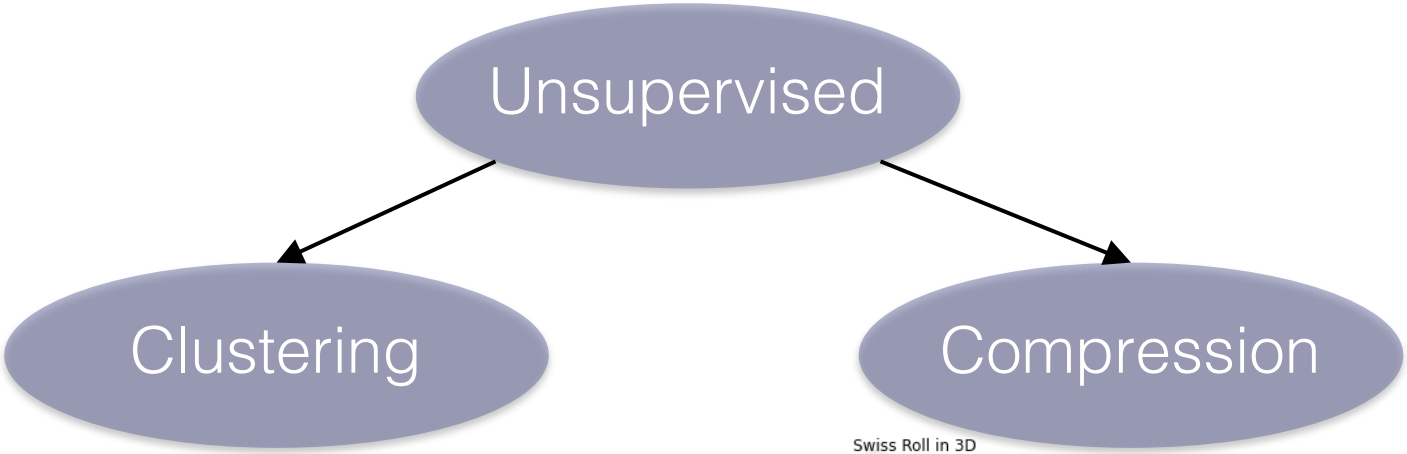
Supervised Learning

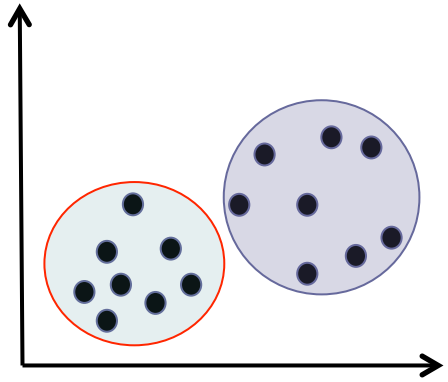


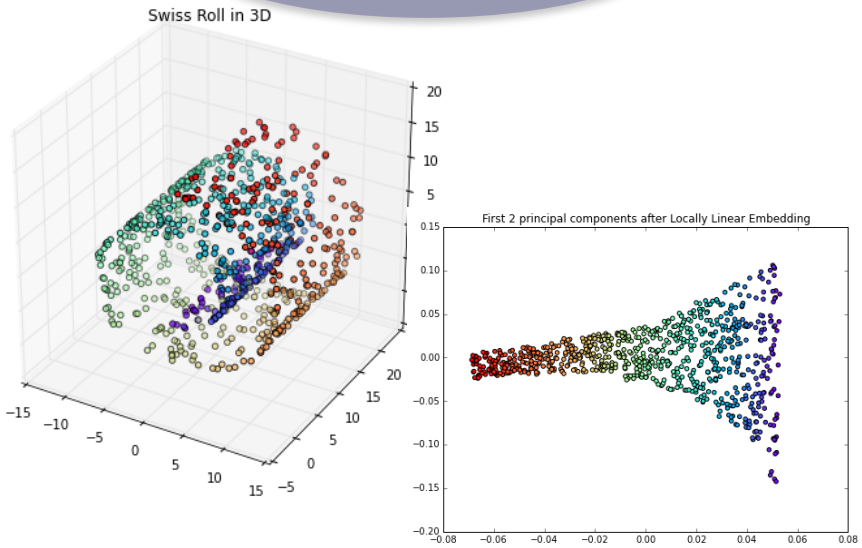




Unsupervised Learning







Statistical Learning Perspective

- Given some input variables (input), what is the predicted output variable (output).

Output = f(Input)

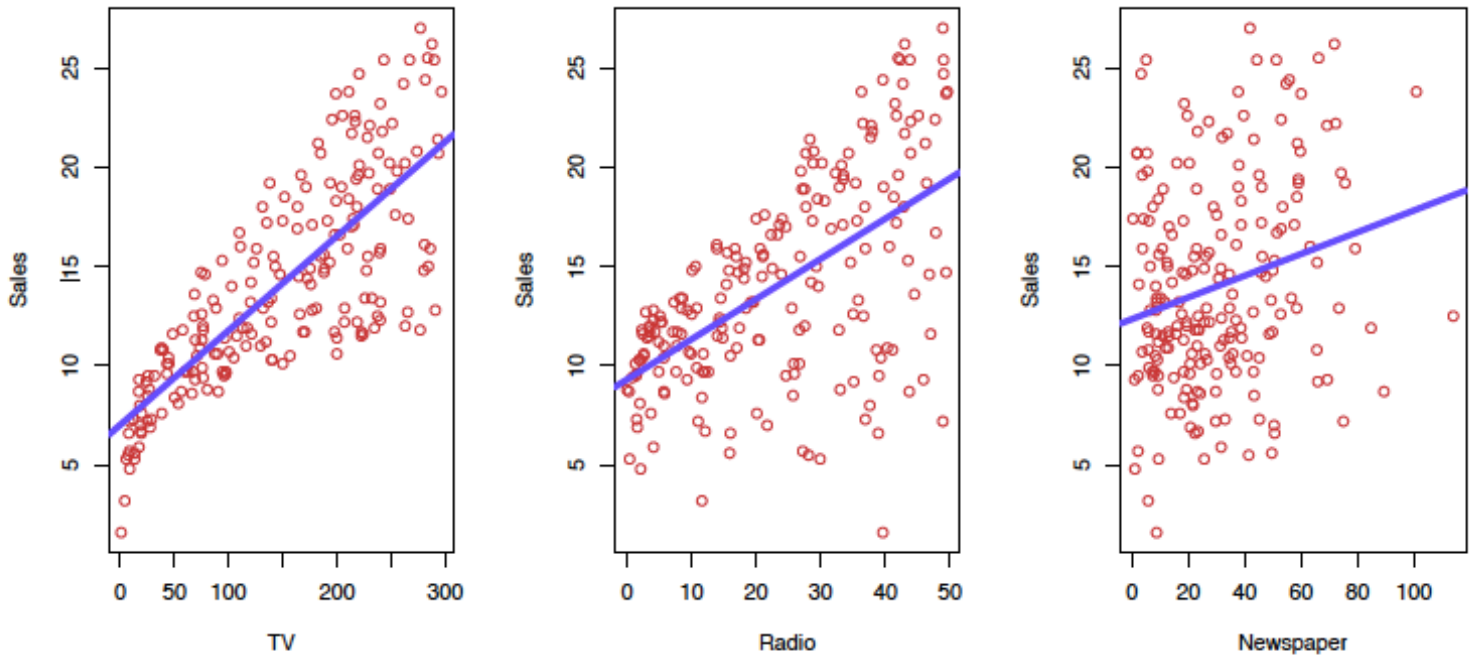
OutputVariable = f(InputVariables)

OutputVariable = f(InputVector)

DependentVariable = f(IndependentVariables)

$Y = f(X)$

Regressão Linear



Shown are `Sales` vs `TV`, `Radio` and `Newspaper`, with a blue linear-regression line fit separately to each.

Can we predict `Sales` using these three?

Perhaps we can do better using a model

$$\texttt{Sales} \approx f(\texttt{TV}, \texttt{Radio}, \texttt{Newspaper})$$

Notação

Here `Sales` is a *response* or *target* that we wish to predict. We generically refer to the response as Y.
`TV` is a *feature*, or *input*, or *predictor*; we name it X_1.
Likewise name `Radio` as X_2, and so on.
We can refer to the *input vector* collectively as

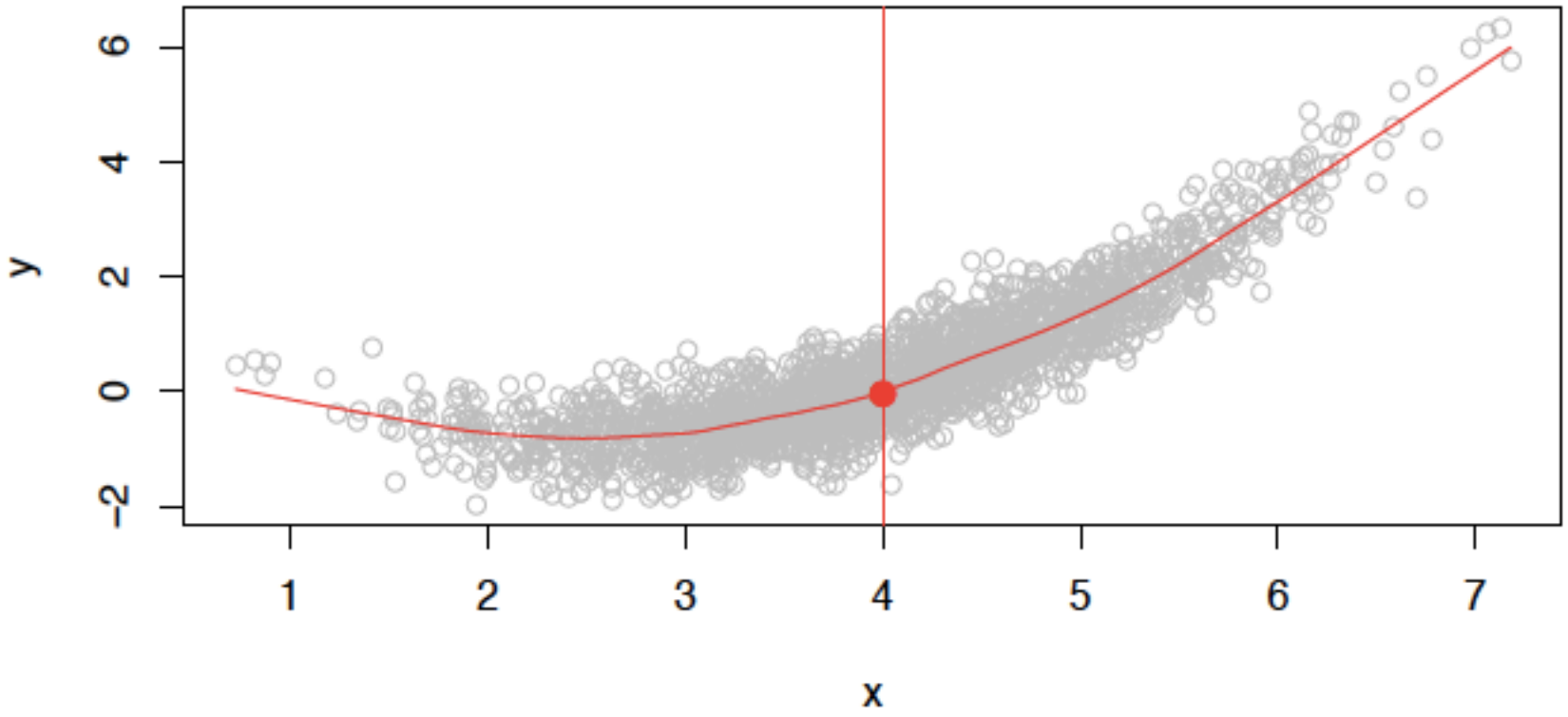
$$X = \begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix}$$

Now we write our model as

$$Y = f(X) + \epsilon$$

where ϵ captures measurement errors and other discrepancies.

Função de Regressão



$E(Y|X = 4)$ means *expected value* (average) of Y given $X = 4$.

This ideal $f(x) = E(Y|X = x)$ is called the *regression function*.

Modelos paramétricos

A learning model that summarizes data with a set of parameters of fixed size (independent of the number of training examples) is called a parametric model.

Artificial Intelligence: A Modern Approach, page 737

Modelos paramétricos

Modelo de regressão linear

$B0 + B1 \times X1 + B2 \times X2 = 0,$

Where B0, B1 and B2 are the coefficients of the line that control the intercept and slope, and X1 and X2 are two input variables.

We only need to estimate the coefficients of the line equation and we have a predictive model for the problem.

Modelos paramétricos

Exemplos:

- Linear regression
- Logistic Regression
- Linear Discriminant Analysis
- Perceptron

Modelos paramétricos

Vantagens:

- Simpler
  - These methods are easier to understand and interpret results.
- Speed
  - Parametric models are very fast to learn from data.
- Less Data
  - They do not require as much training data and can work well even if the fit to the data is not perfect.

Modelos paramétricos

Desvantagens:

- Constrained
  - By choosing a functional form these methods are highly constrained to the specified form.
- Limited Complexity
  - The methods are more suited to simpler problems.
- Poor Fit
  - In practice the methods are unlikely to match the underlying mapping function.

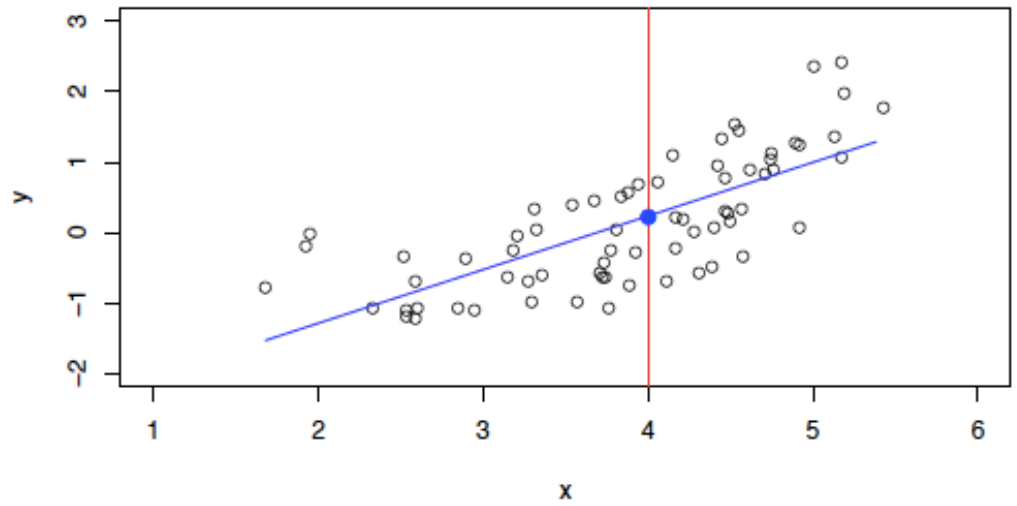
Modelos paramétricos

The *linear* model is an important example of a parametric model:

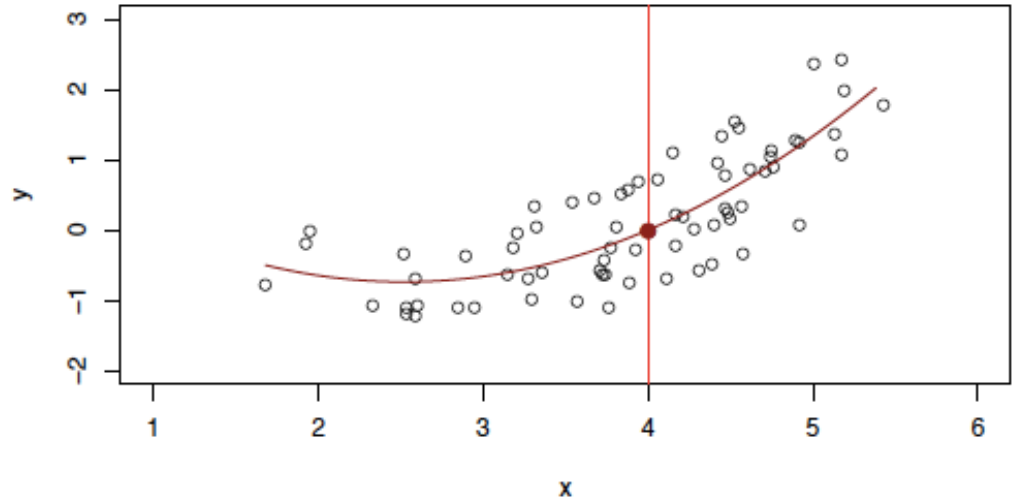
$$f_L(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots \beta_p X_p.$$

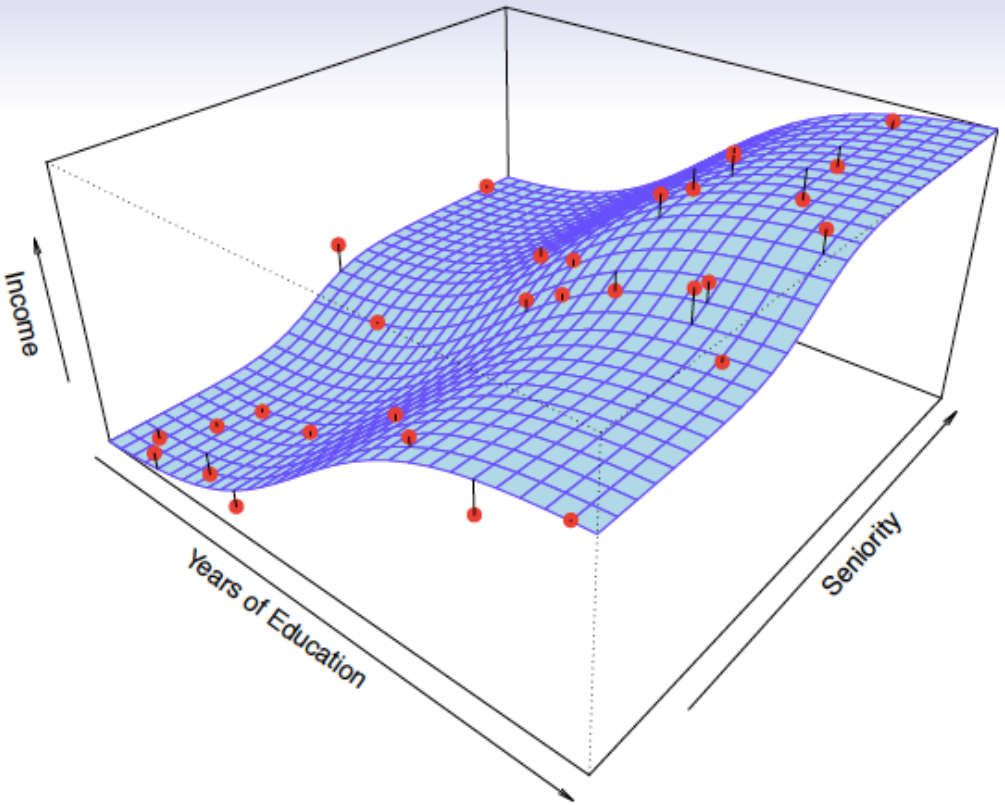
- A linear model is specified in terms of $p + 1$ parameters $\beta_0, \beta_1, \ldots, \beta_p$.
- We estimate the parameters by fitting the model to training data.
- Although it is *almost never correct*, a linear model often serves as a good and interpretable approximation to the unknown true function $f(X)$.

A linear model $\hat{f}_L(X) = \hat{\beta}_0 + \hat{\beta}_1 X$ gives a reasonable fit here



A quadratic model $\hat{f}_Q(X) = \hat{\beta}_0 + \hat{\beta}_1 X + \hat{\beta}_2 X^2$ fits slightly better.

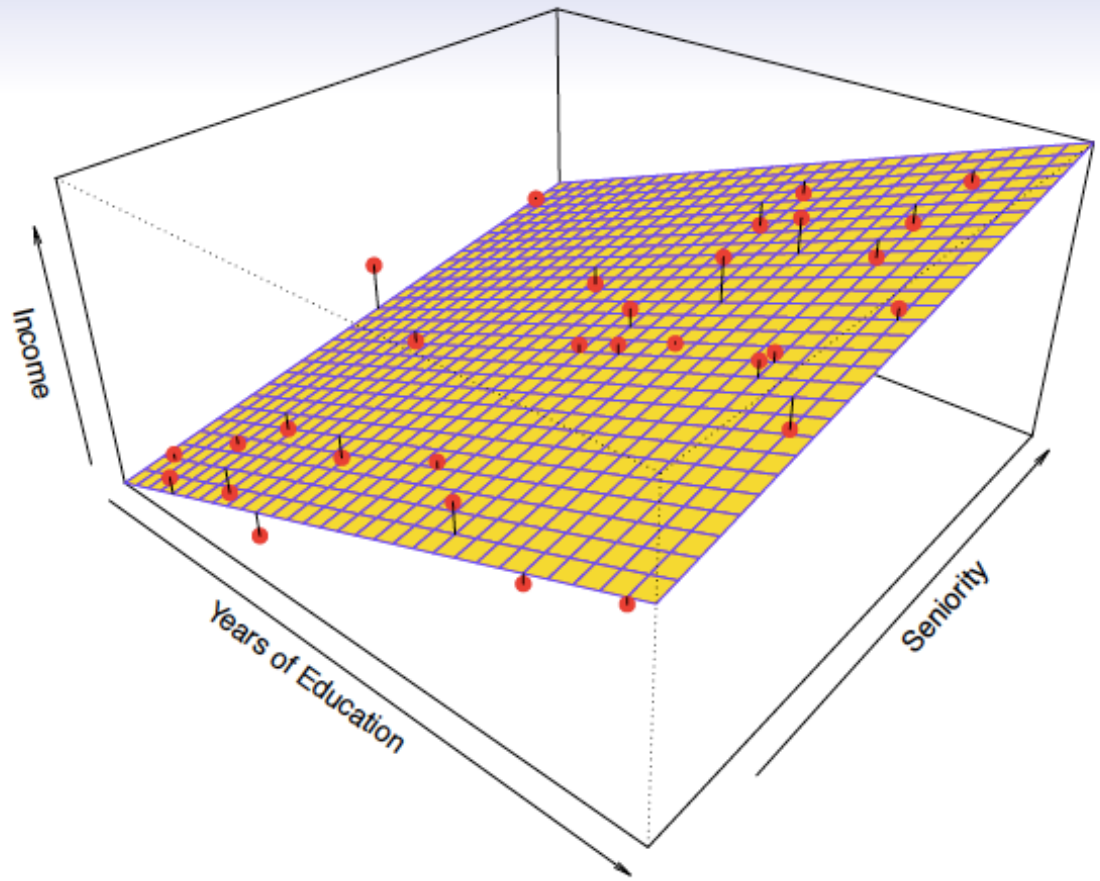




Simulated example. Red points are simulated values for `income` from the model

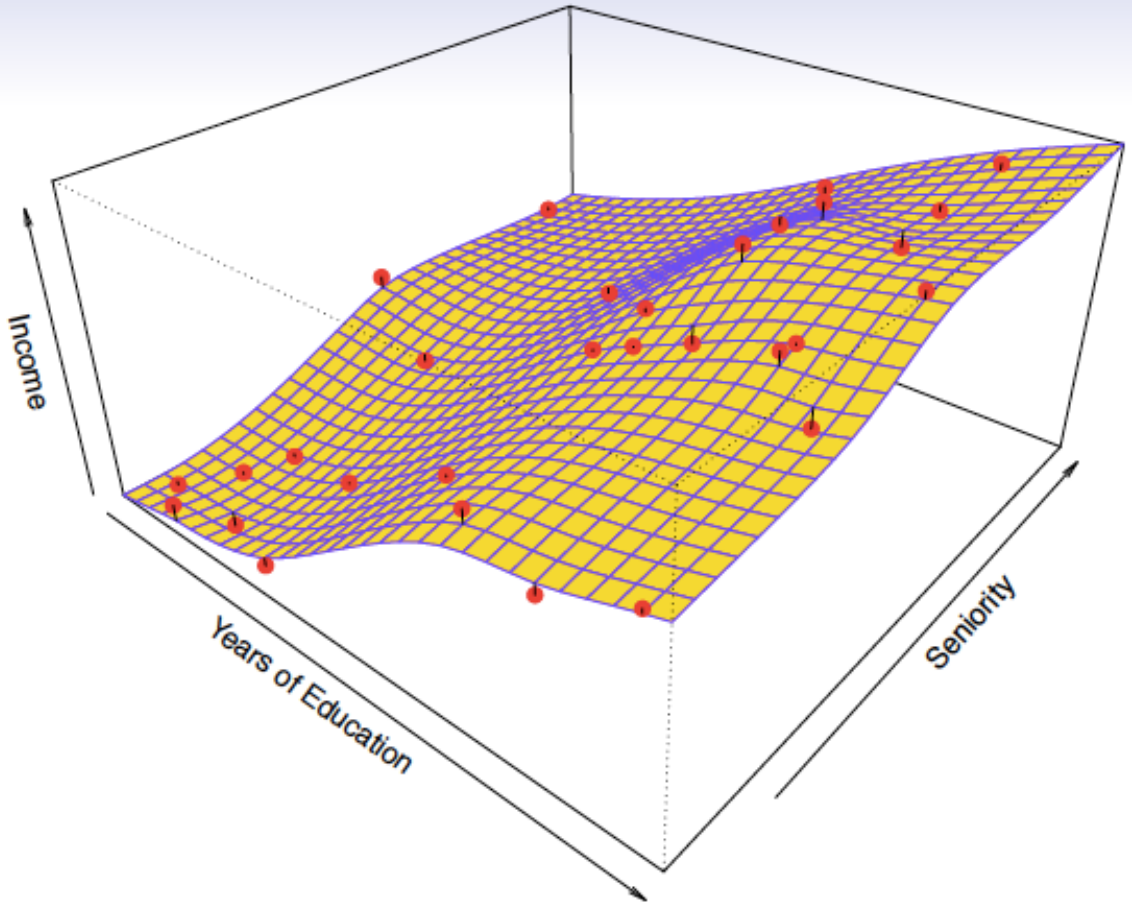
$$\texttt{income} = f(\texttt{education}, \texttt{seniority}) + \epsilon$$

f is the blue surface.

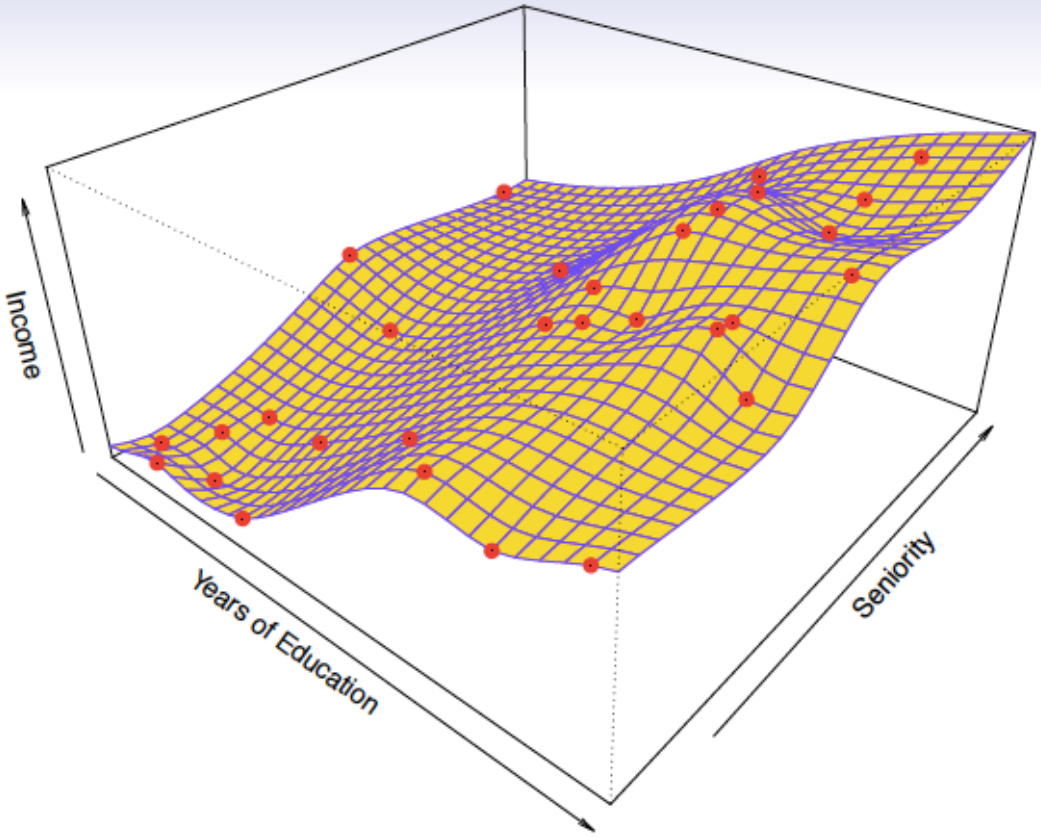


Linear regression model fit to the simulated data.

$$\hat{f}_L(\texttt{education}, \texttt{seniority}) = \hat{\beta}_0 + \hat{\beta}_1 \times \texttt{education} + \hat{\beta}_2 \times \texttt{seniority}$$



More flexible regression model $\hat{f}_S(\texttt{education}, \texttt{seniority})$ fit to the simulated data. Here we use a technique called a *thin-plate spline* to fit a flexible surface. We control the roughness of the fit (chapter 7).



Even more flexible spline regression model $\hat{f}_S(\texttt{education}, \texttt{seniority})$ fit to the simulated data. Here the fitted model makes no errors on the training data! Also known as *overfitting*.

Nonparametric models

- Algorithms that do not make strong assumptions about the form of the mapping function.
- They are free to learn any functional form from the training data.
- Nonparametric methods are good when you have a lot of data and no prior knowledge, and when you don't want to worry too much about choosing just the right features (Artificial Intelligence: A Modern Approach, page 757).

Nonparametric models

Exemplos:

- Decision Trees like CART and C4.5
- Naive Bayes
- Support Vector Machines
- Neural Networks

Nonparametric models

Vantagens:

- Flexibility
  - Capable of fitting a large number of functional forms.
- Power
  - No assumptions (or weak assumptions) about the underlying function.
- Performance
  - Can result in higher performance models for prediction.

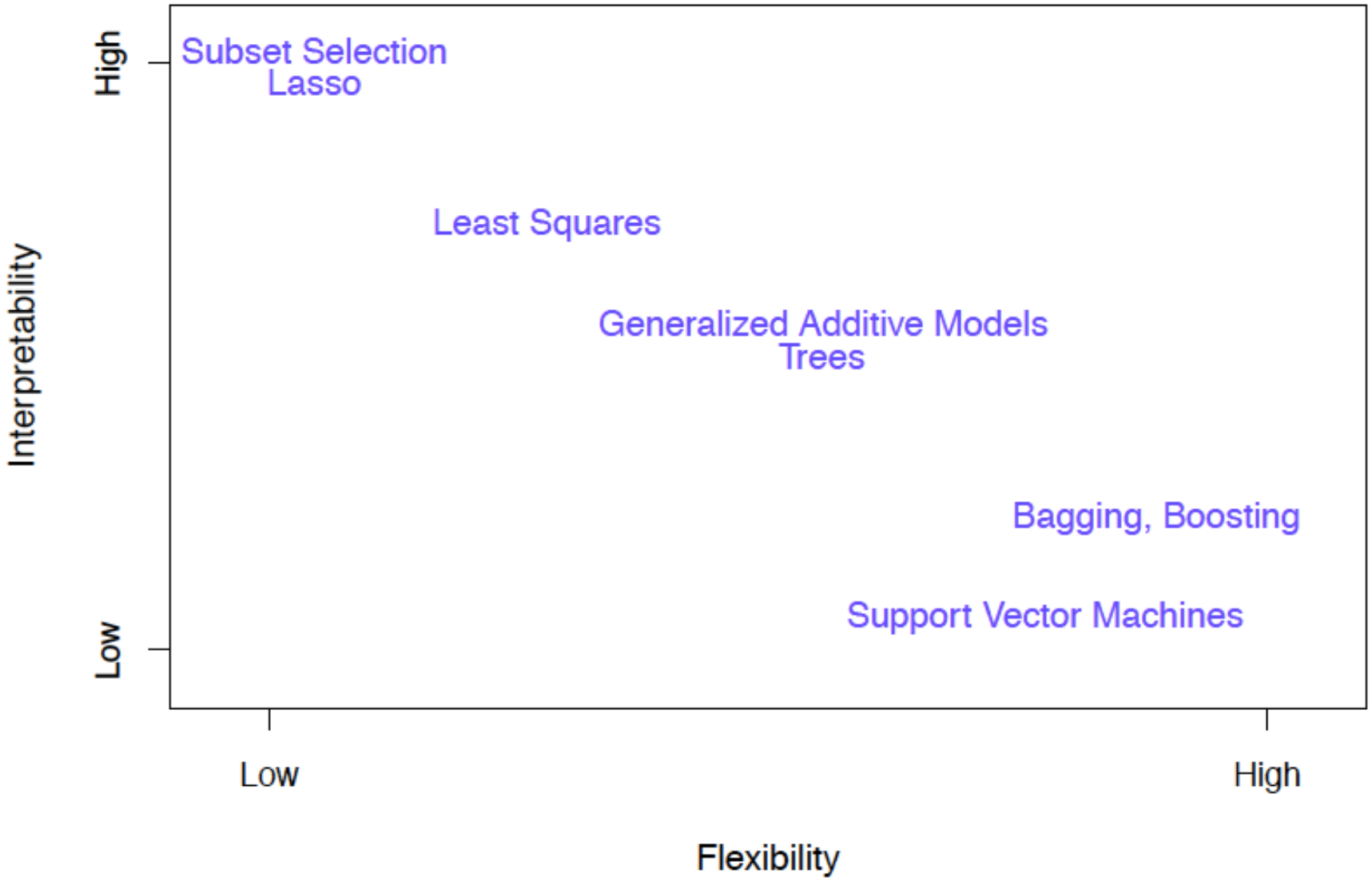
Nonparametric models

Desvantagens:

- More data
  - Require a lot more training data to estimate the mapping function.
- Slower
  - A lot slower to train as they often have far more parameters to train.
- Overfitting
  - More of a risk to overfit the training data and it is harder to explain why specific predictions are made.

Custo-benefício

- Prediction accuracy versus interpretability.
  - Linear models are easy to interpret;
  - thin-plate splines are not.
- Good fit versus over-fit or under-fit.
  - How do we know when the fit is just right?
- Parsimony versus black-box.
  - We often prefer a simpler model involving fewer variables over a black-box predictor involving them all.


Subset Selection
Lasso
Least Squares
Generalized Additive Models
Trees
Bagging, Boosting
Support Vector Machines
High
Low
Interpretability
Low
High
Flexibility

Precisão do modelo

Suppose we fit a model $\hat{f}(x)$ to some training data $\mathsf{Tr} = \{x_i, y_i\}_1^N$, and we wish to see how well it performs.

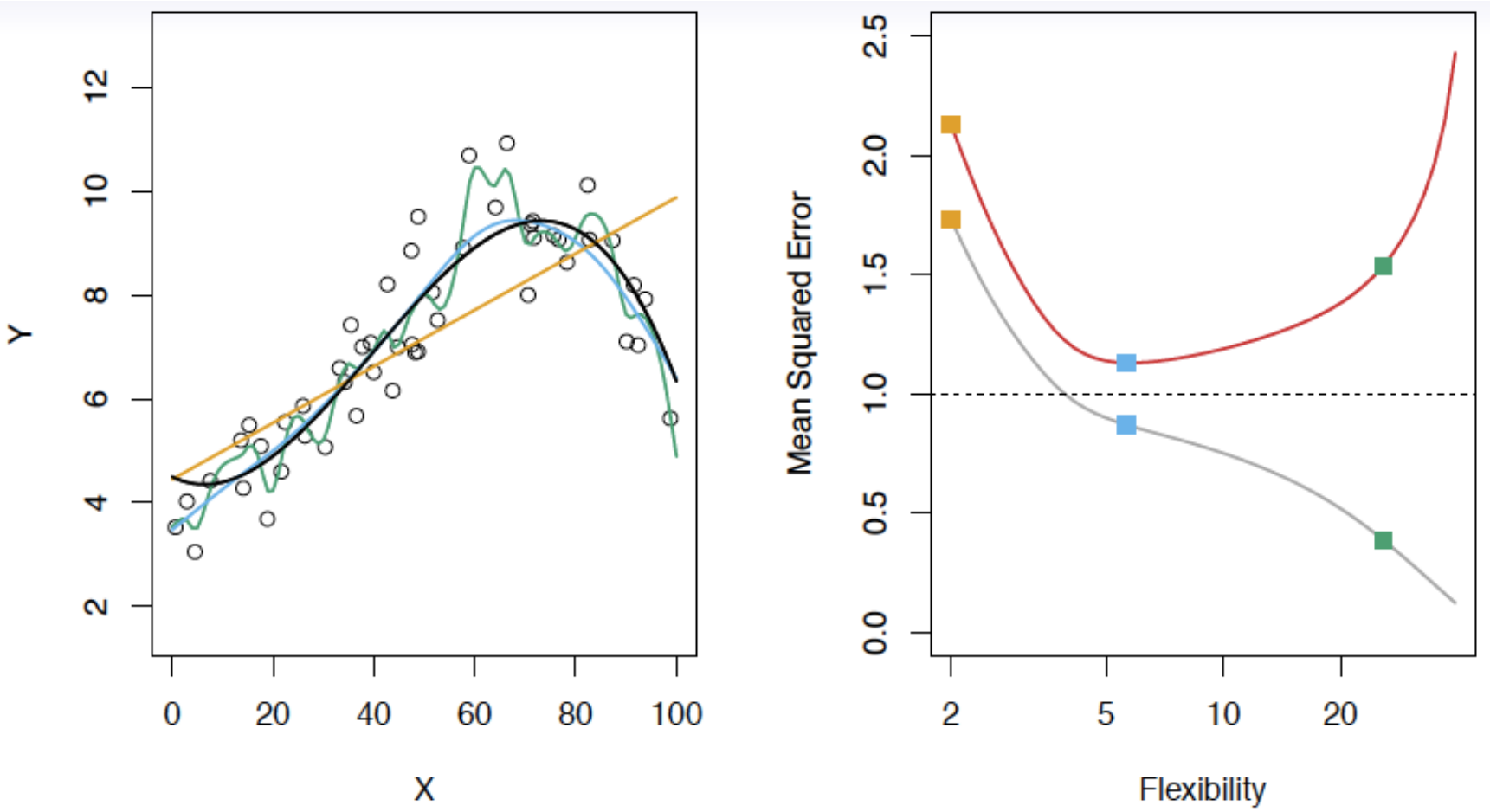
- We could compute the average squared prediction error over Tr:

$$\mathrm{MSE}_{\mathsf{Tr}} = \mathrm{Ave}_{i \in \mathsf{Tr}}[y_i - \hat{f}(x_i)]^2$$

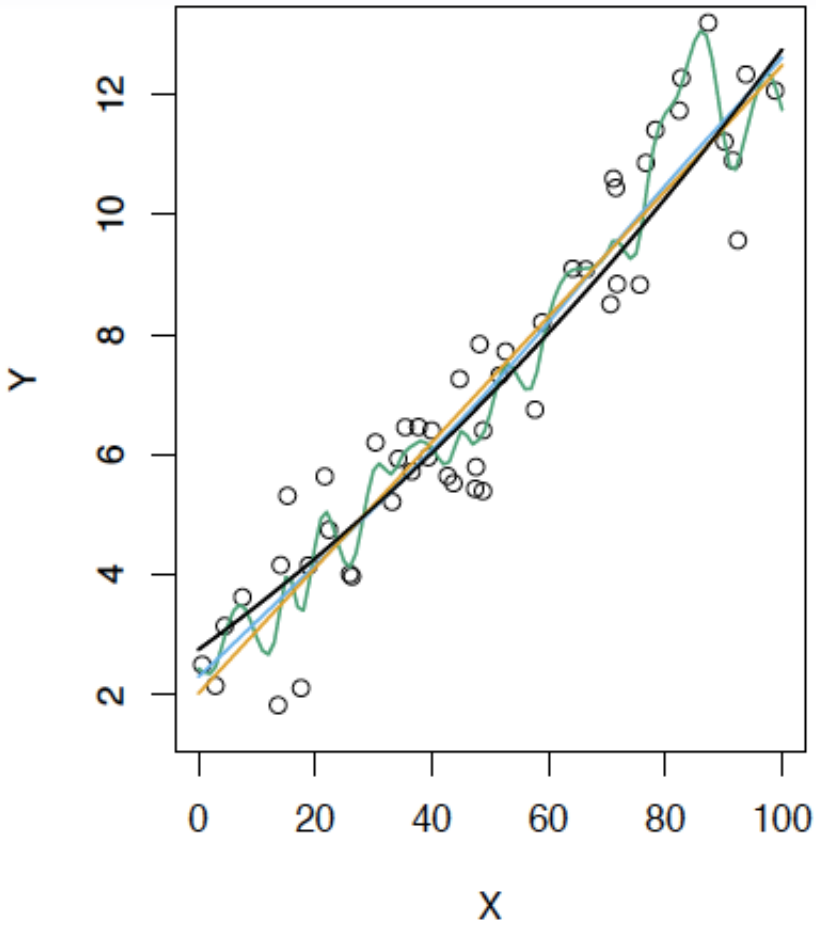
This may be biased toward more overfit models.

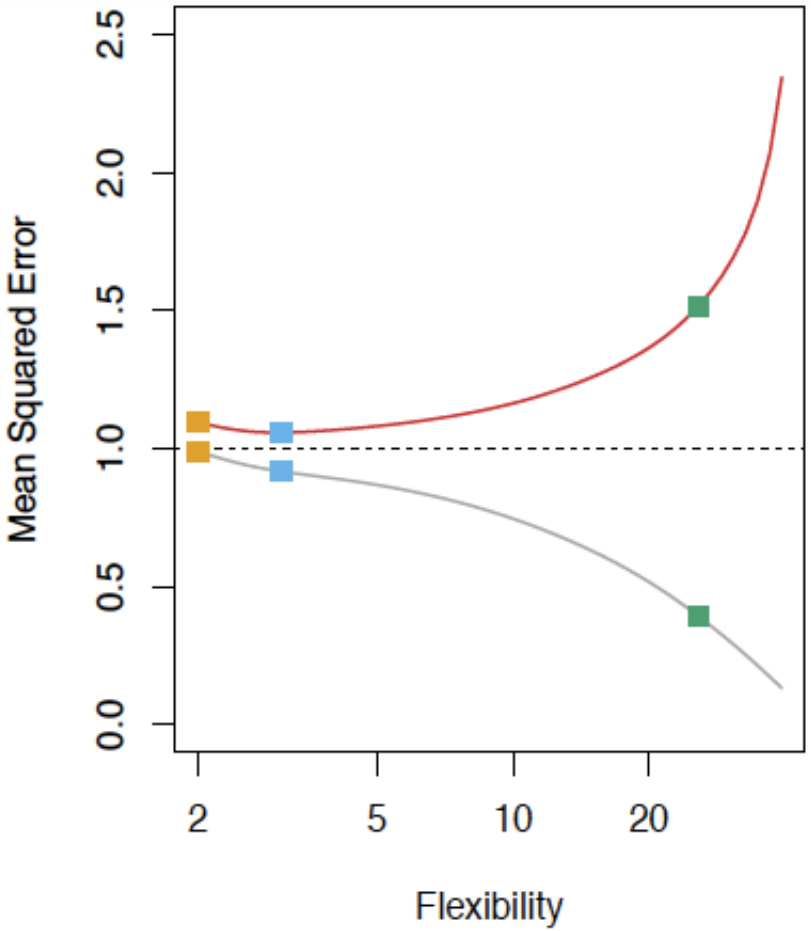
- Instead we should, if possible, compute it using fresh *test* data $\mathsf{Te} = \{x_i, y_i\}_1^M$:

$$\mathrm{MSE}_{\mathsf{Te}} = \mathrm{Ave}_{i \in \mathsf{Te}}[y_i - \hat{f}(x_i)]^2$$

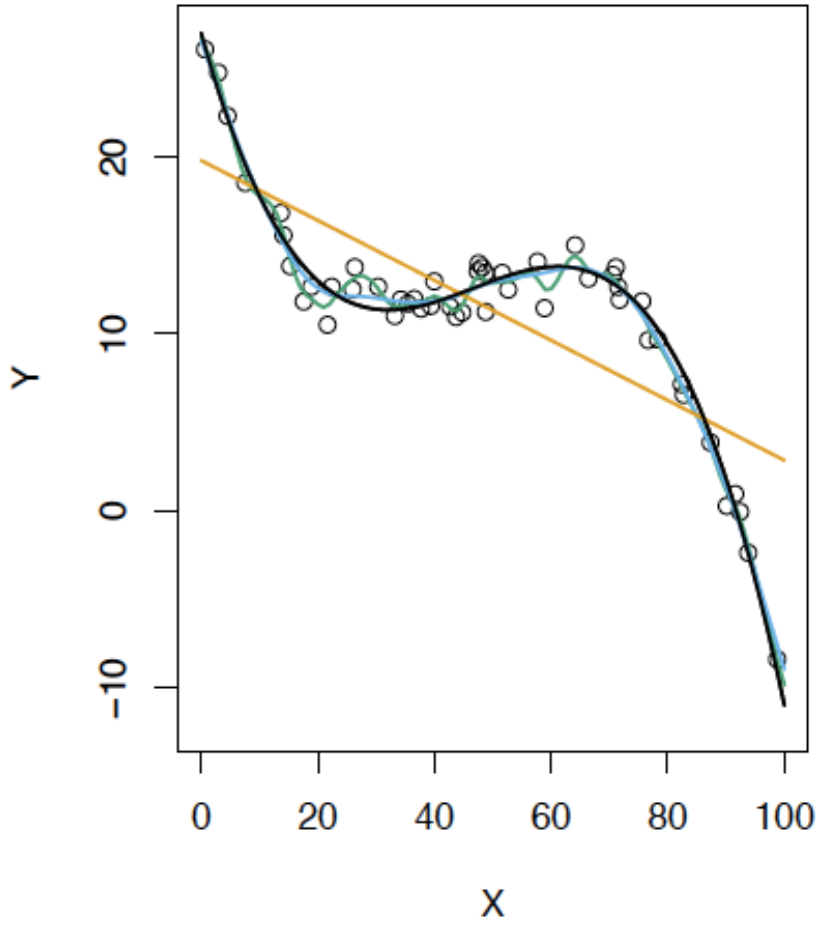


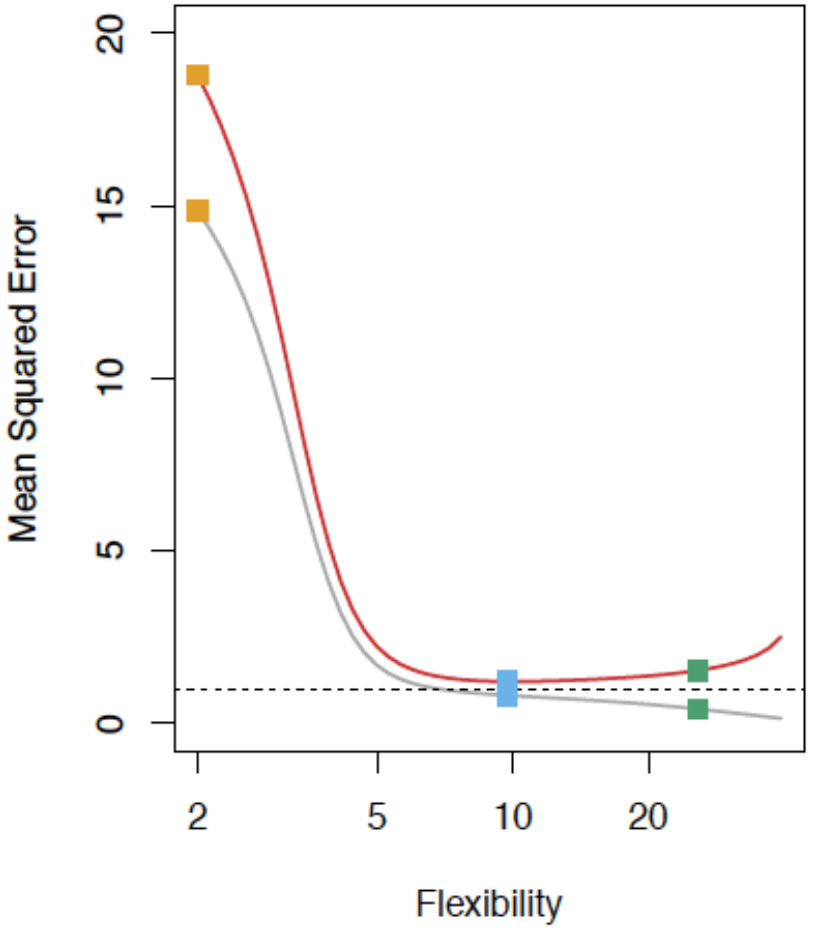
Black curve is truth. Red curve on right is MSE_{Te}, grey curve is MSE_{Tr}. Orange, blue and green curves/squares correspond to fits of different flexibility.





Here the truth is smoother, so the smoother fit and linear model do really well.





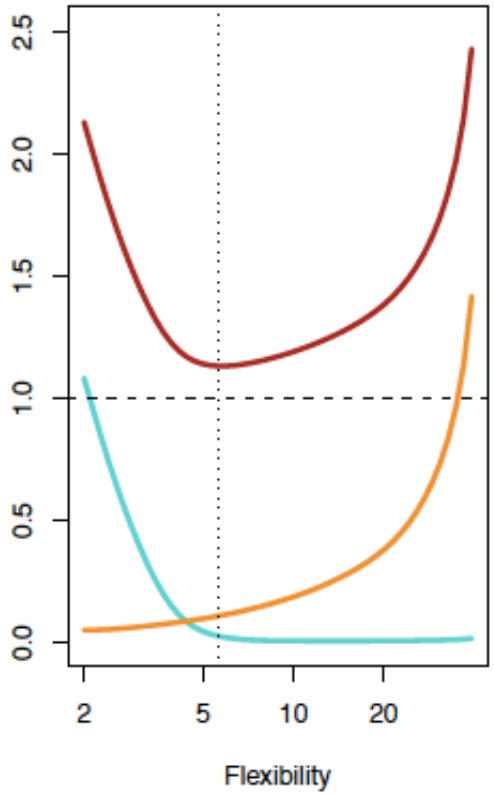
Here the truth is wiggly and the noise is low, so the more flexible fits do the best.

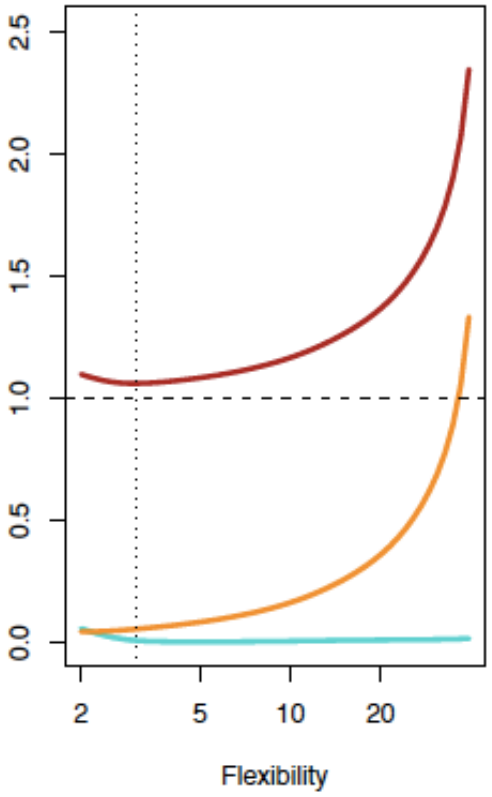
Errors

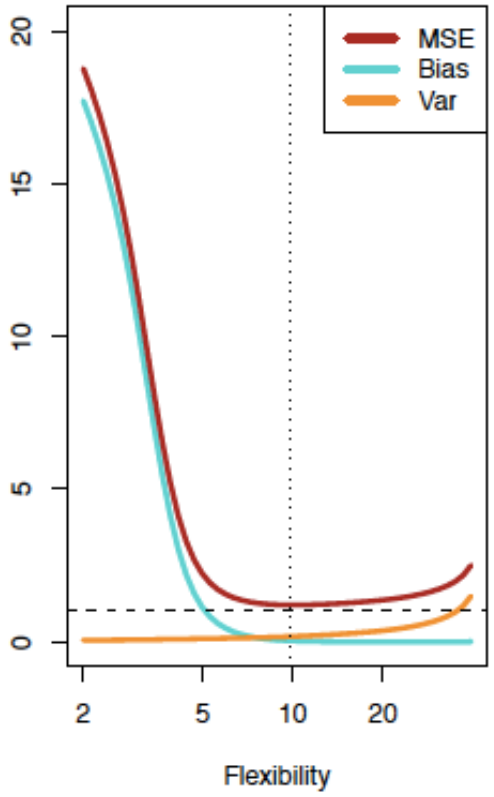
- Bias Error
  - Low Bias
    - Suggests less assumptions about the form of the target function.
  - High-Bias
    - Suggests more assumptions about the form of the target function.
- Variance Error
  - Low Variance
    - Suggests small changes to the estimate of the target function with changes to the training dataset.
  - High Variance
    - Suggests large changes to the estimate of the target function with changes to the training dataset.
- Irreducible Error

Bias-Variance Trade-off

Typically as the *flexibility* of $\hat{f}$ increases, its variance increases, and its bias decreases. So choosing the flexibility based on average test error amounts to a *bias-variance trade-off.*







Overfitting

- In statistics a fit refers to how well you approximate a target function.
- Overfitting refers to a model that models the training data too well.
  - Overfitting happens when a model learns the detail and noise in the training data to the extent that it negatively impacts the performance on the model on new data.

How to limit overfitting

1. Use a resampling technique to estimate model accuracy.
   - The most popular resampling technique is k-fold cross-validation.

2. Hold back a validation dataset.

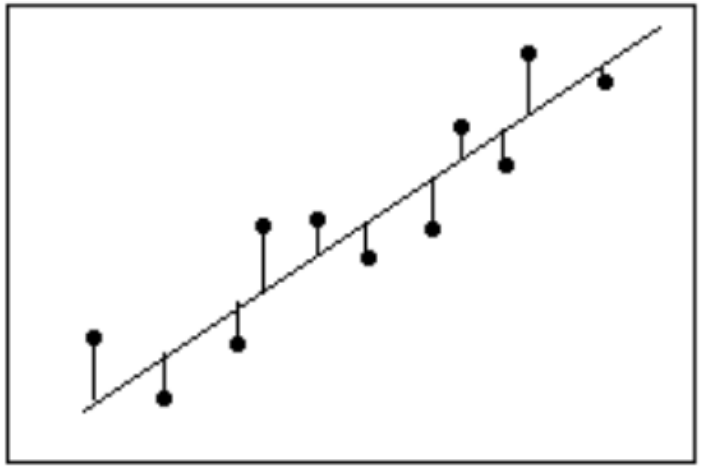
Underfitting

- Underfitting refers to a model that can neither model the training data nor generalize to new data.
- An underfit machine learning model is not a suitable model and it will have poor performance on the training data.
- Underfitting is often not discussed as it is easy to detect given a good performance metric.
- The remedy is to move on and try alternate machine learning algorithms.

Residuals

- Linear regression calculates an equation that minimizes the distance between the fitted line and all of the data points.

- Ordinary least squares (OLS) regression minimizes the sum of the squared residuals.



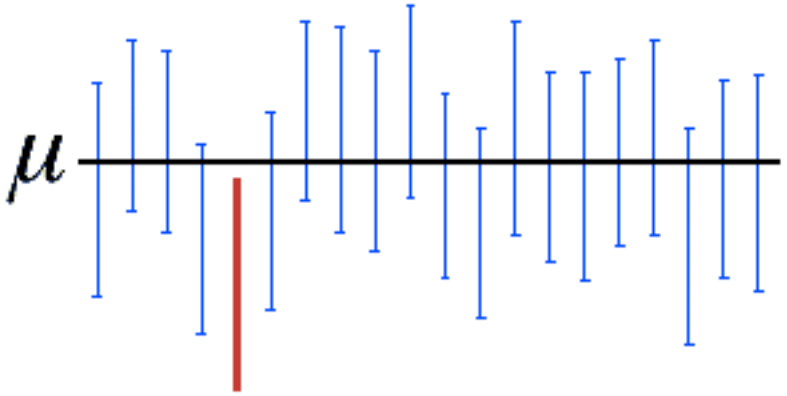
Definition: Residual = Observed value - Fitted value

R-squared (R^2)

- R-squared is a statistical measure of how close the data are to the fitted regression line
- R-squared = Explained variation / Total variation
- Values between 0% and 100%
  - 0% - model explains none of the variability of the response data around its mean.
  - 100% - model explains all the variability of the response data around its mean.
- In general, the higher the R-squared, the better the model fits your data.

Confidence Intervals

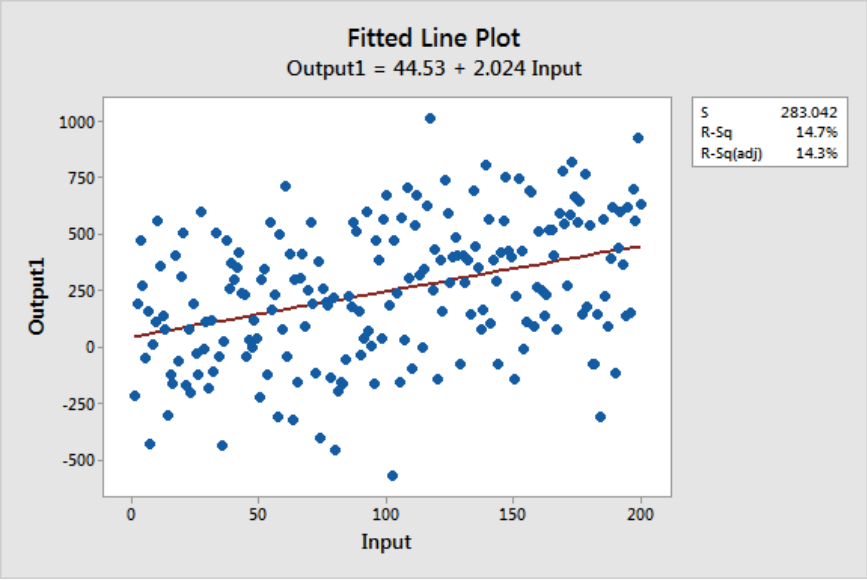
- Confidence interval is a range of values, derived from sample statistics, that is likely to contain the value of an unknown population parameter.

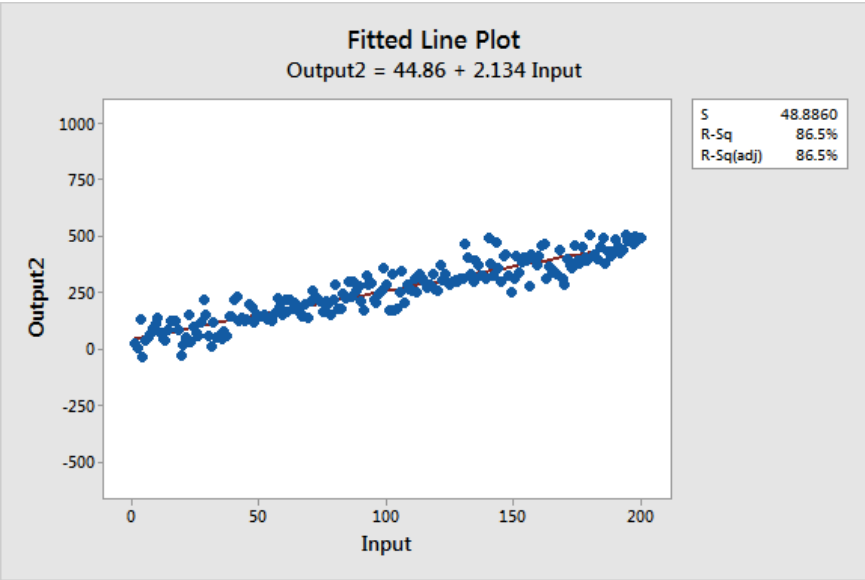


A 95% confidence interval indicates that 19 out of 20 samples (95%) from the same population will produce confidence intervals that contain the population parameter.

- Most frequently, we use confidence intervals to bound the mean or standard deviation, but we can also use regression coefficients, proportions, rates of occurrence (Poisson), and for the differences between populations.

Comparação entre modelos





Prediction for Output1

Regression Equation

Output1 = 44.5 + 2.024 Input

Variable	Setting
Input	10

Fit	SE Fit	95% CI	95% PI
64.7766	37.2129	(-8.60793, 138.161)	(-498.190, 627.743)

PI – Prediction Interval

Prediction for Output2

Regression Equation

Output2 = 44.86 + 2.1343 Input

Variable	Setting
Input	10

Fit	SE Fit	95% CI	95% PI
66.2076	6.42728	(53.5329, 78.8823)	(-31.0260, 163.441)

Flower Classification

Iris-Setosa

Iris-Setosa

Iris-Versicolor

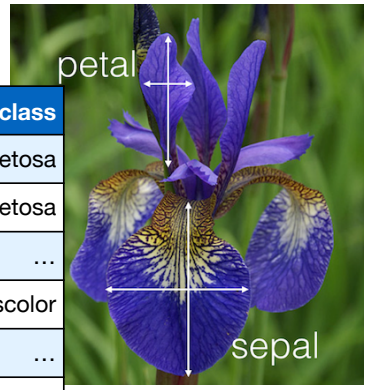
Data Representation

IRIS

https://archive.ics.uci.edu/ml/datasets/Iris

Instances (samples, observations)

	sepal_length	sepal_width	petal_length	petal_width	class
1	5.1	3.5	1.4	0.2	setosa
2	4.9	3.0	1.4	0.2	setosa
…	…	…	…	…	…
50	6.4	3.2	4.5	1.5	veriscolor
…	…	…	…	…	…
150	5.9	3.0	5.1	1.8	virginica



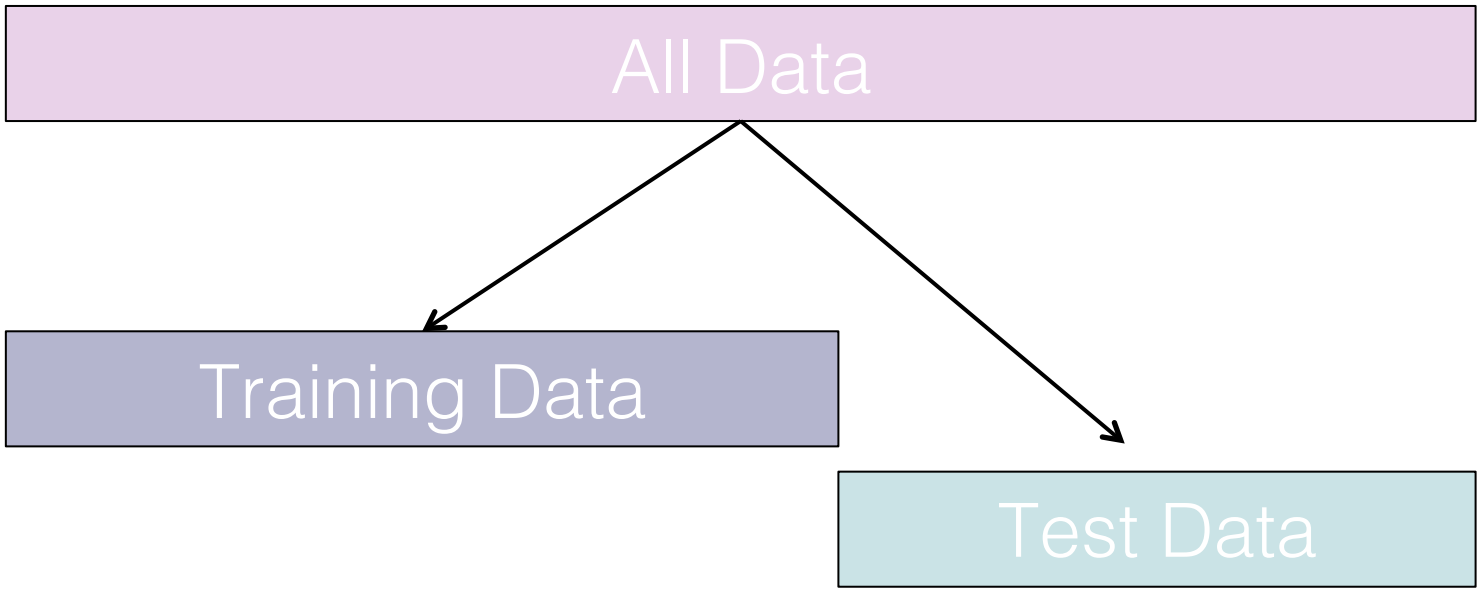
Features (attributes, dimensions)

Classes (targets)

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1D} \\ x_{21} & x_{22} & \cdots & x_{2D} \\ x_{31} & x_{32} & \cdots & x_{3D} \\ \vdots & \vdots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ x_{N1} & x_{N2} & \cdots & x_{ND} \end{bmatrix}$$

$$\mathbf{y} = [y_1, y_2, y_3, \cdots y_N]$$

Training & Test Data

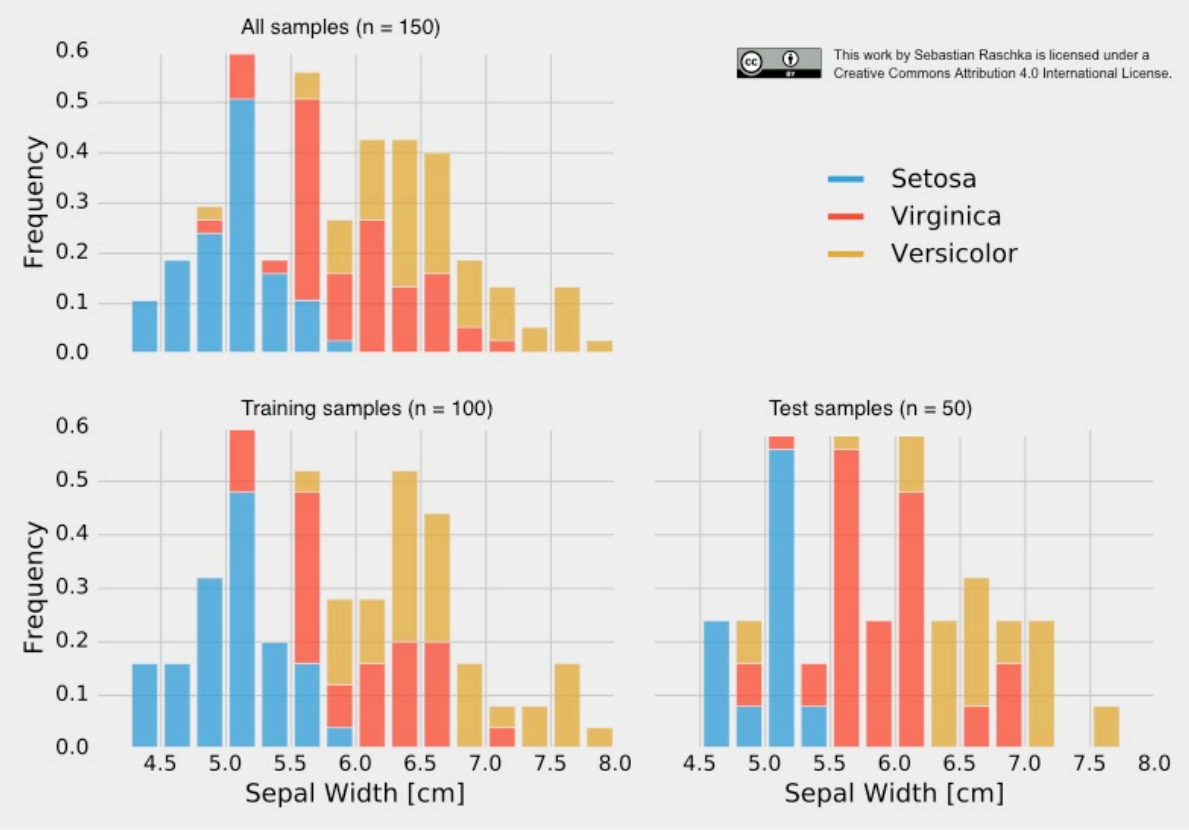


Typically:

- 75% : 25%
- 2/3 : 1/3

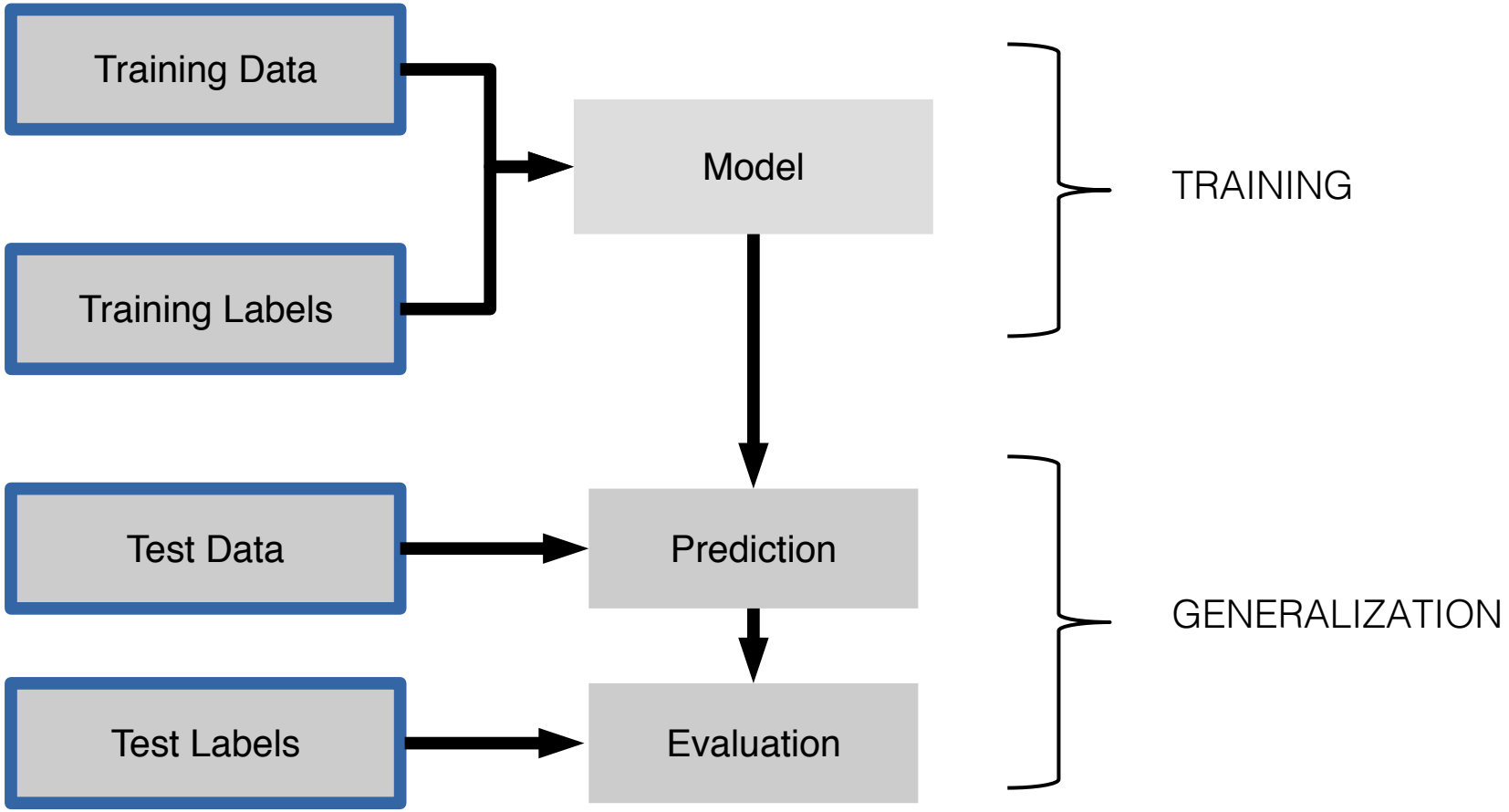
Stratification

Non-stratified split:



- training set → 38 x Setosa, 28 x Versicolor, 34 x Virginica
- test set → 12 x Setosa, 22 x Versicolor, 16 x Virginica

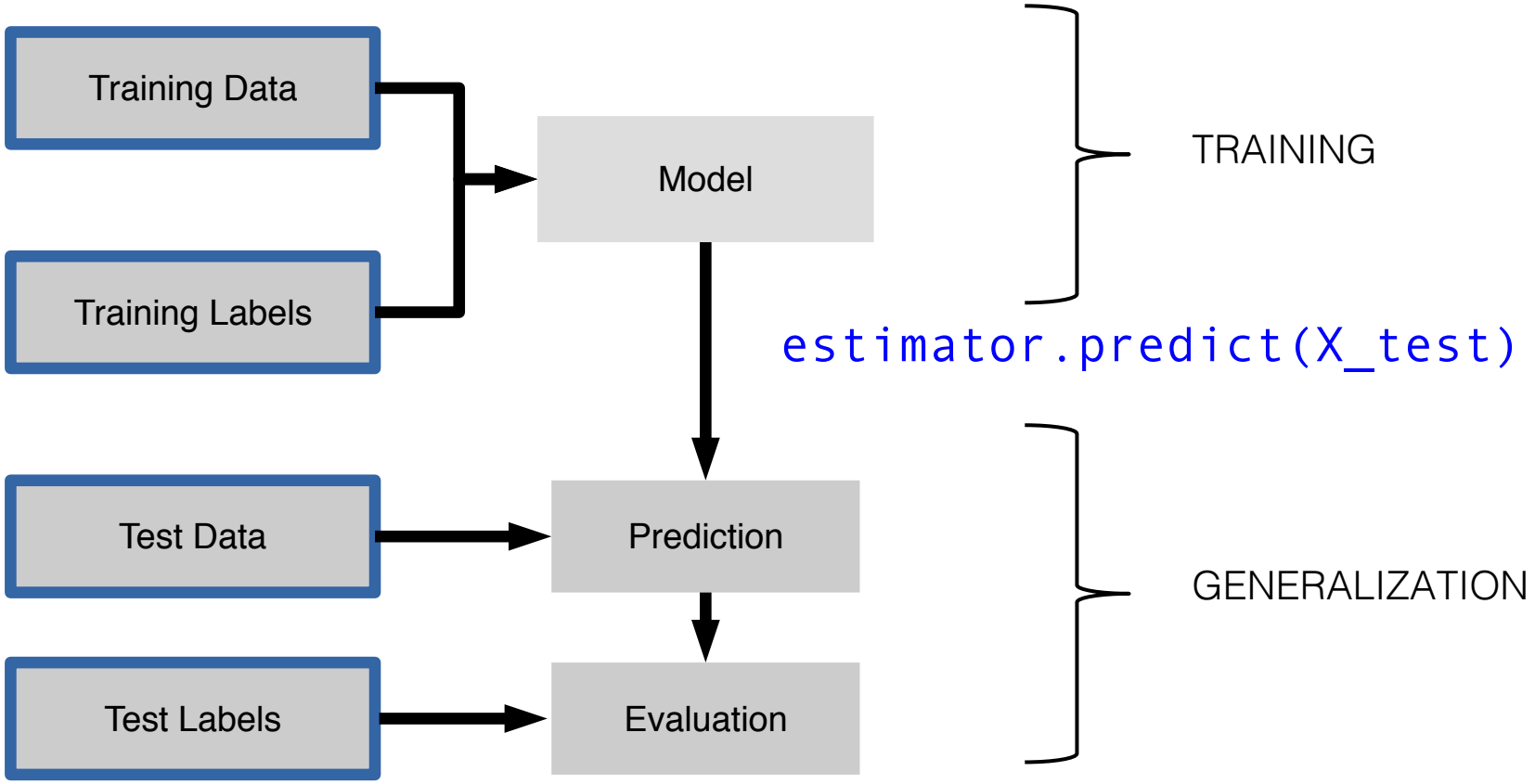
Supervised Workflow



- Fit model on all data after evaluation

Supervised Workflow

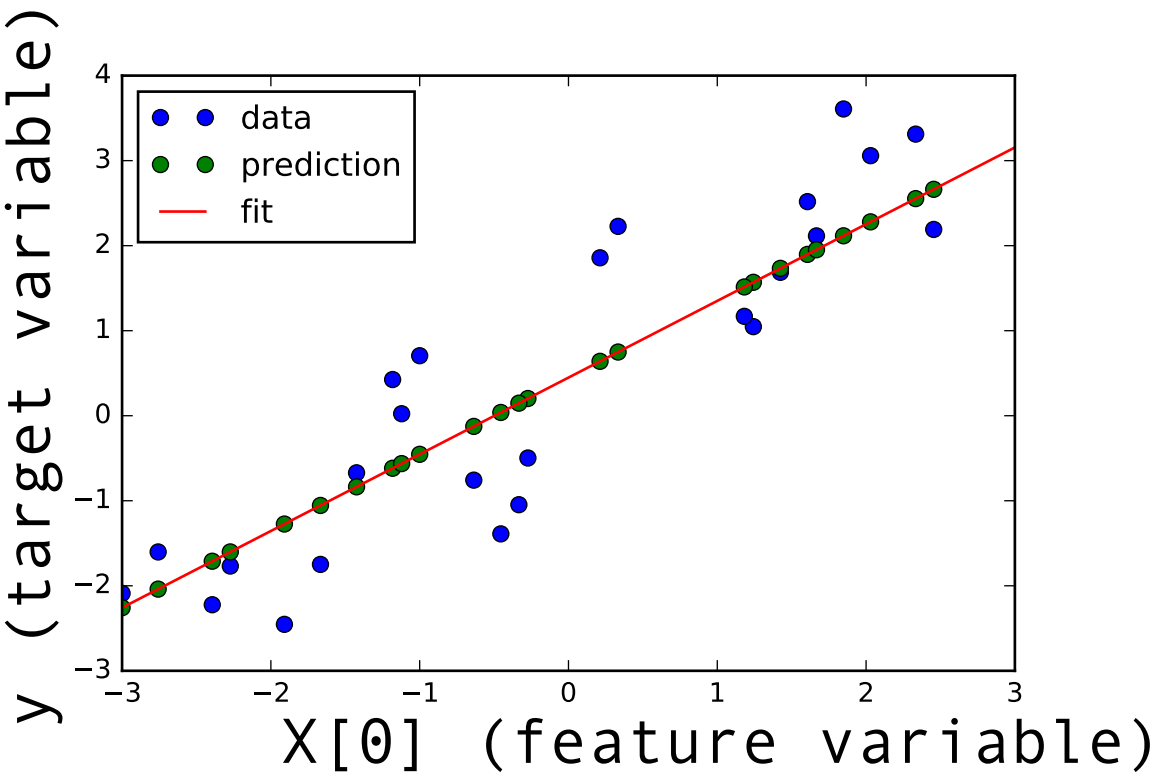
```
estimator.fit(X_train, y_train)
```



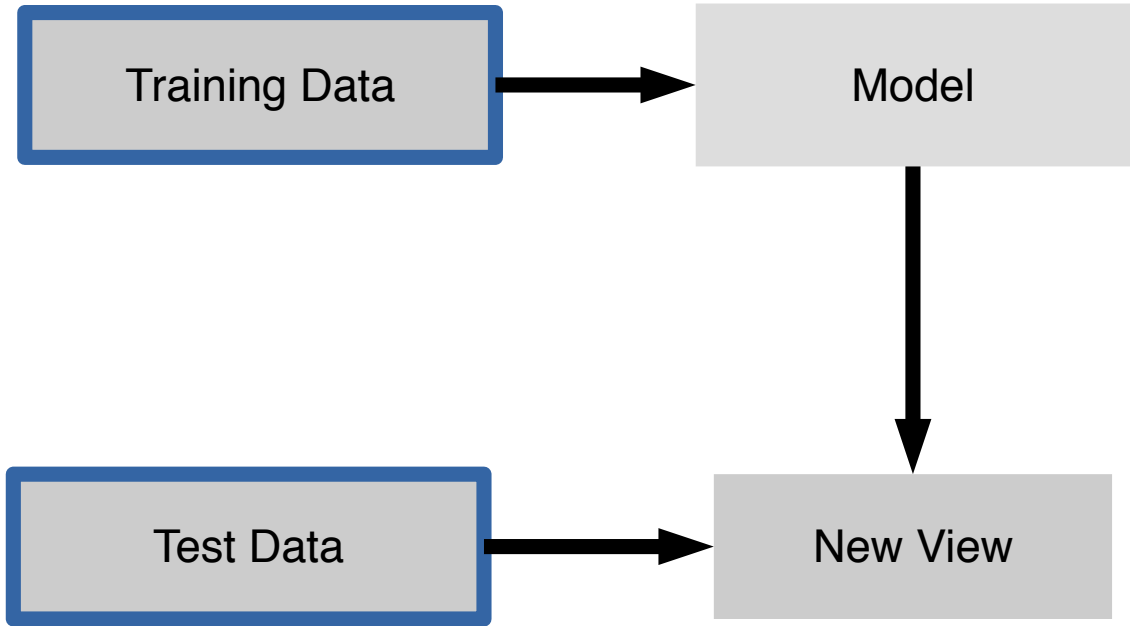

```
estimator.score(X_test, y_test)
```

Linear Regression

```
y =coef_[0]*X[0] + intercept_
```



Unsupervised Transformers



① `transformer.fit(X_train)`
② `X_train_transf = transformer.transform(X_train)`
③ `X_test_transf  = transformer.transform(X_test)`

Feature Scaling

$$x^{(i)}_{std} = \frac{x^{(i)} - \mu_x}{\sigma_x}$$

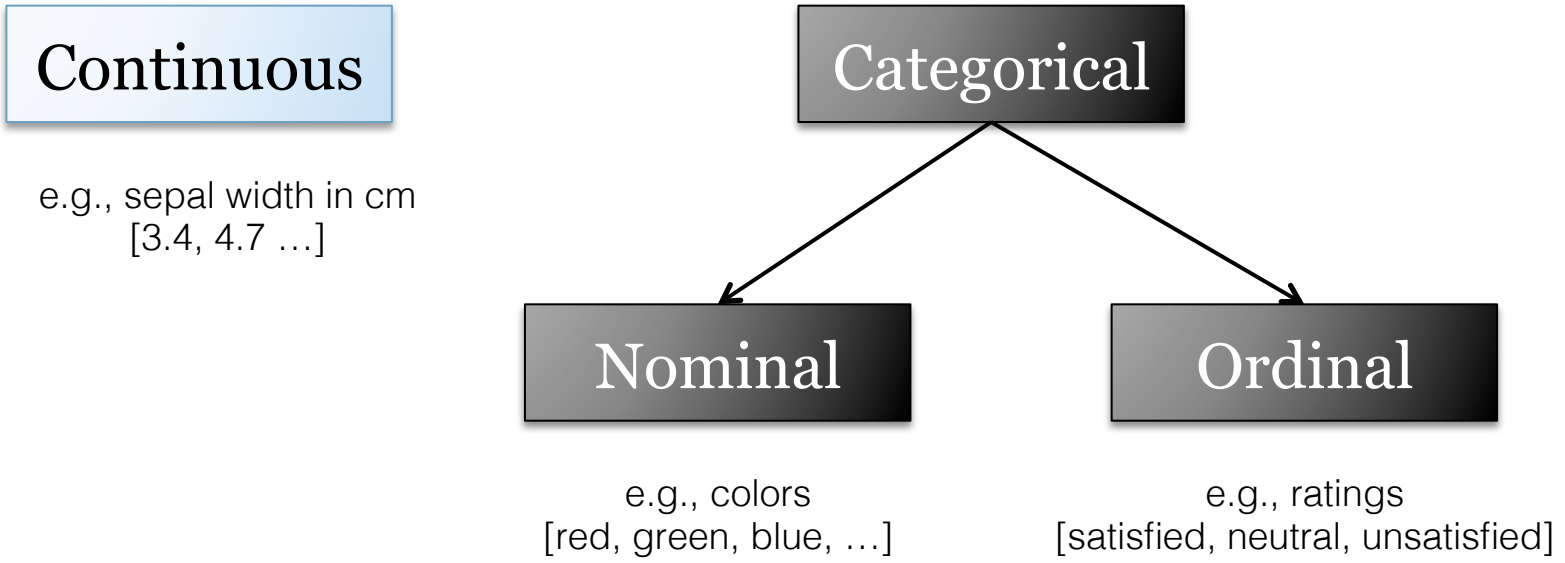
standardization

$$x^{(i)}_{norm} = \frac{x^{(i)} - \mathbf{x}_{min}}{\mathbf{x}_{max} - \mathbf{x}_{min}}$$

min-max scaling ("normalization")

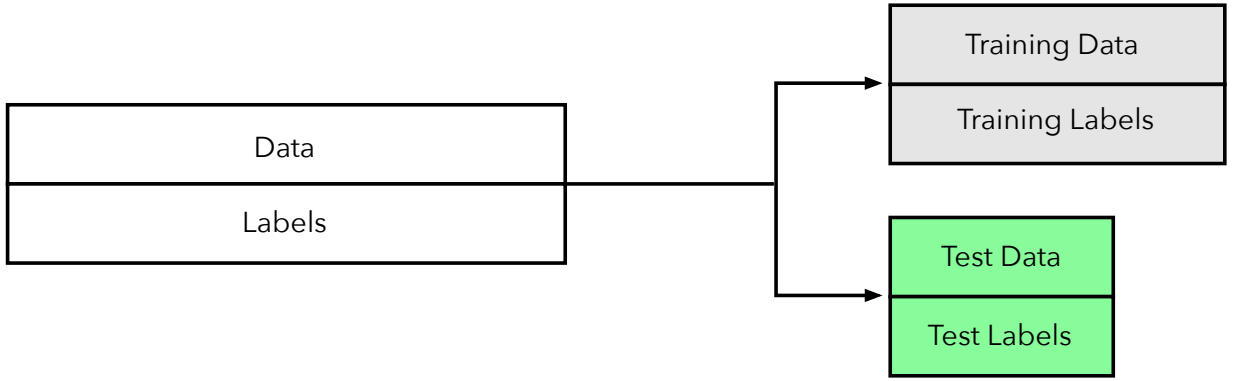
	input	standardized	normalized
0	0	-1.46385	0.0
1	1	-0.87831	0.2
2	2	-0.29277	0.4
3	3	0.29277	0.6
4	4	0.87831	0.8
5	5	1.46385	1.0

Continuous & Categorical Features

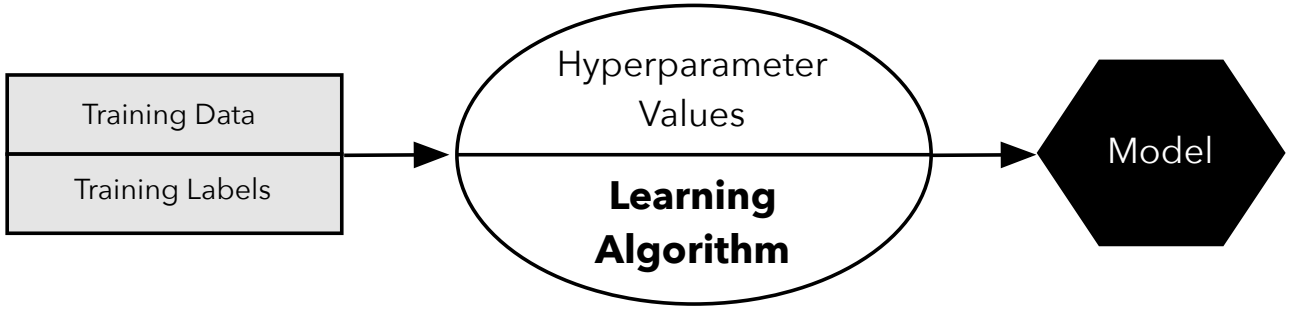


Holdout Evaluation I

1

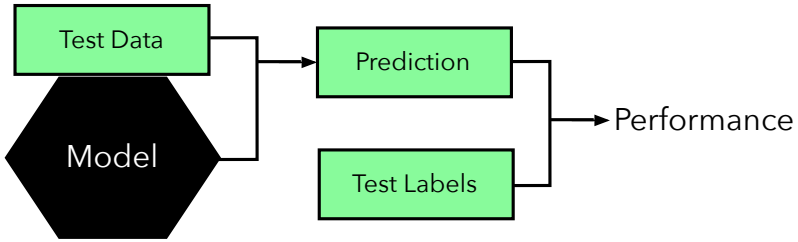


2

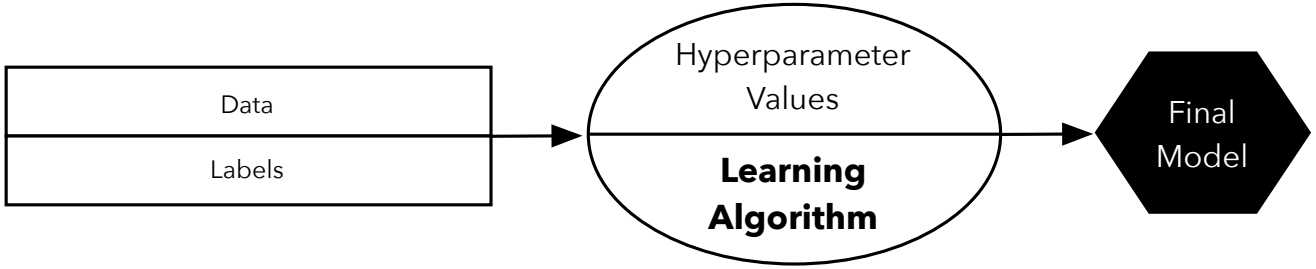


Holdout Evaluation II

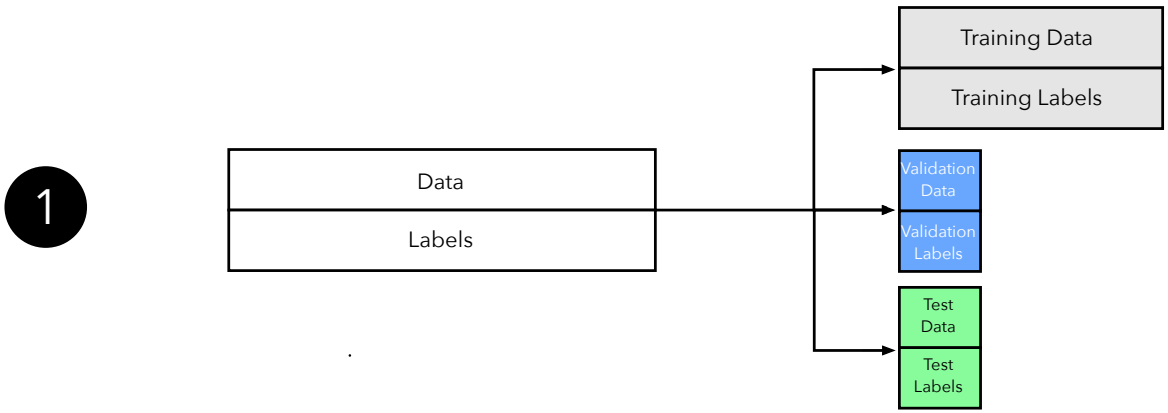
3

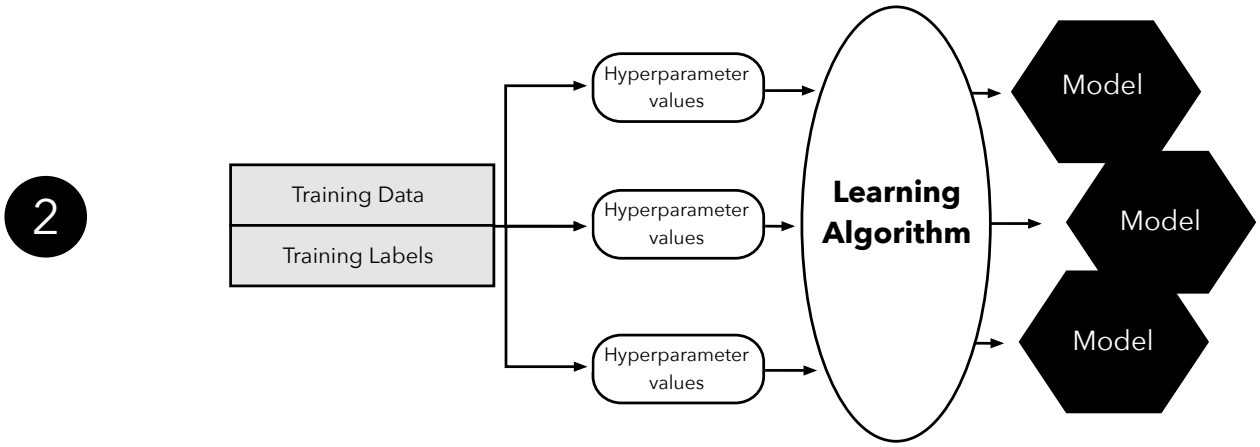


4



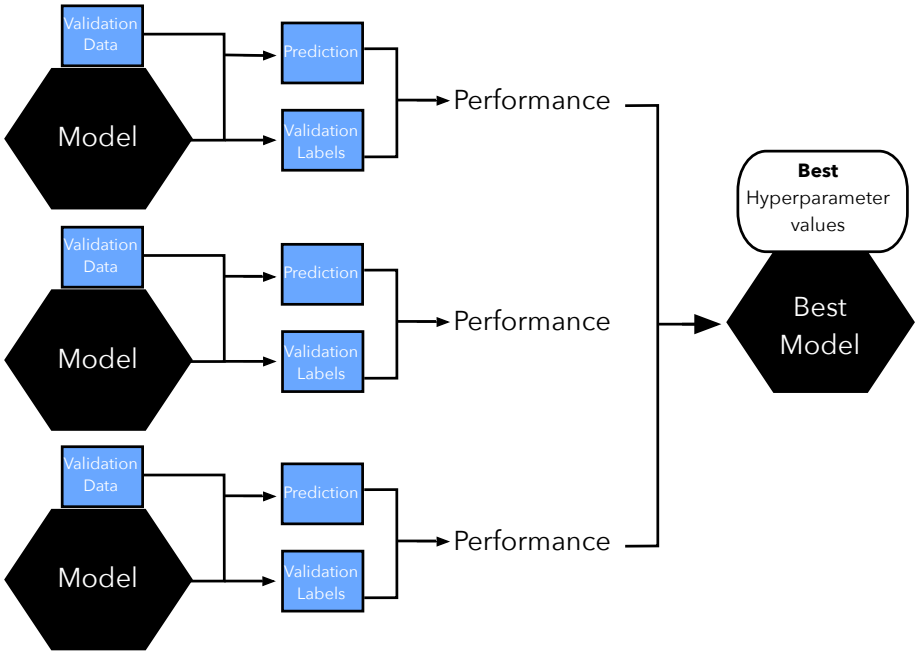
Holdout Validation I



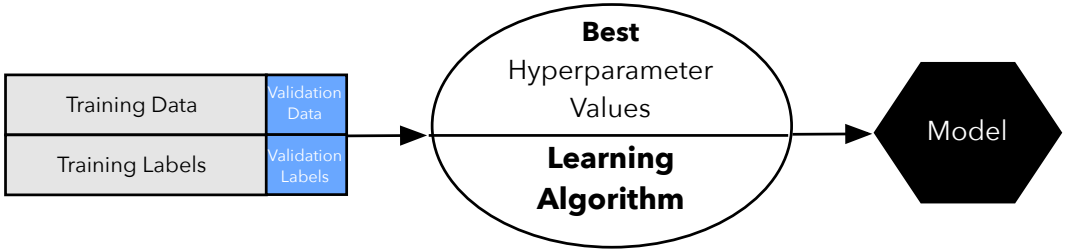


Holdout Validation II

3

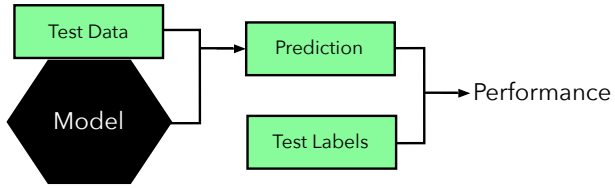


4

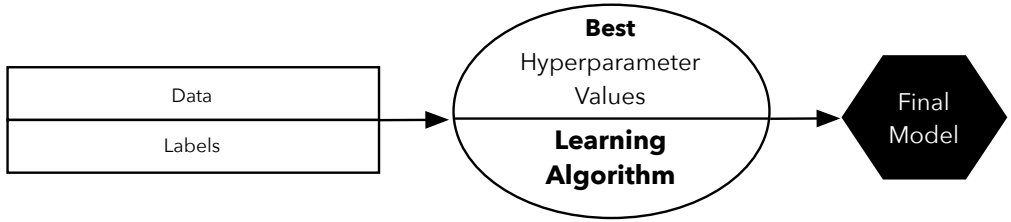


Holdout Validation III

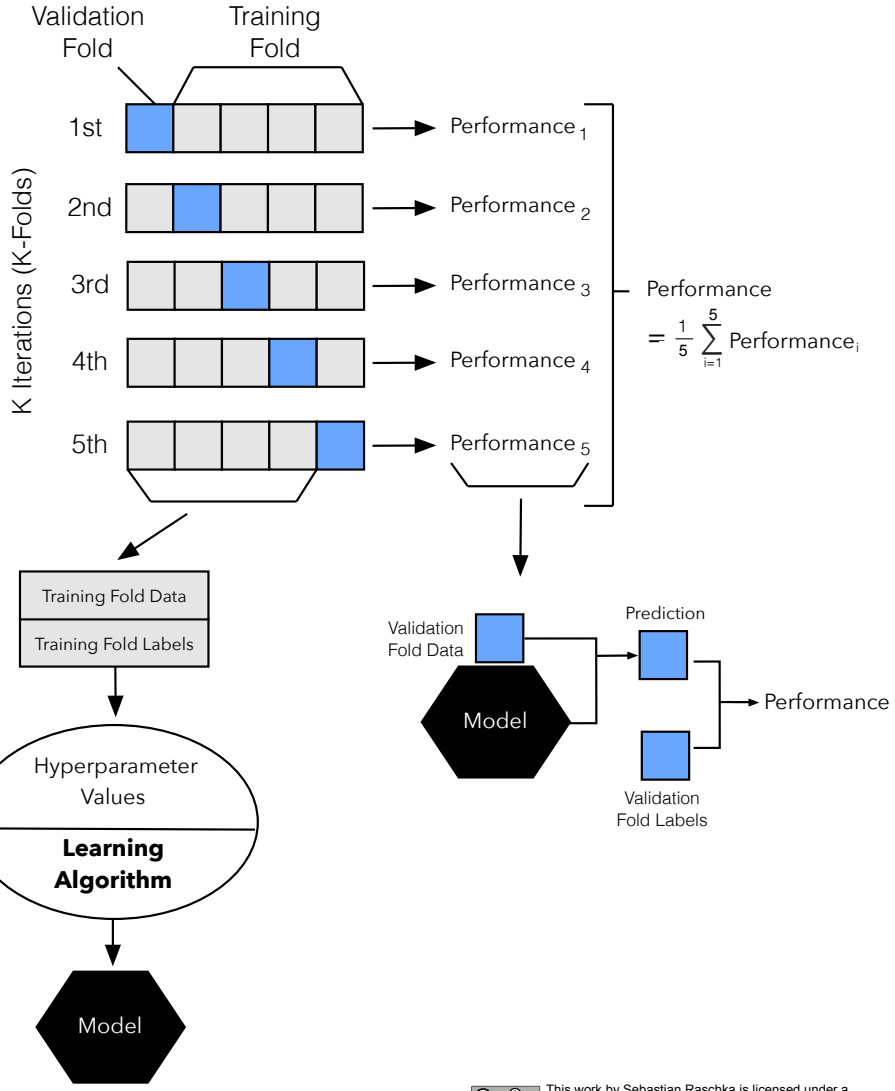
5



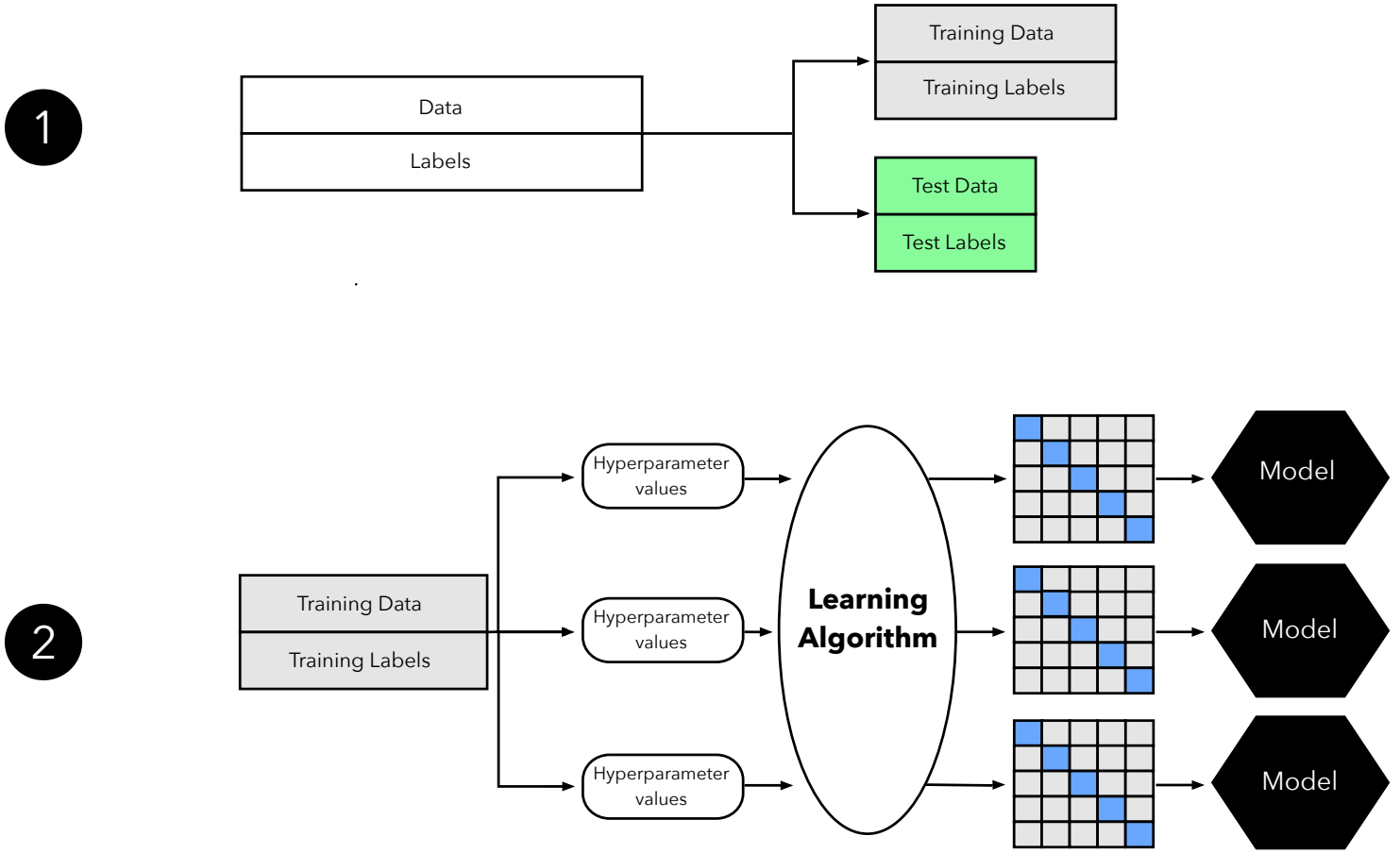
6



K-fold Cross-Validation

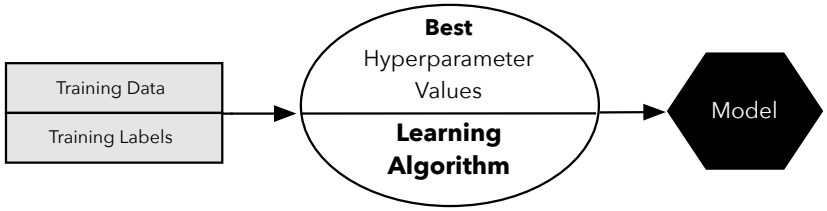


K-fold Cross-Validation Pipeline I

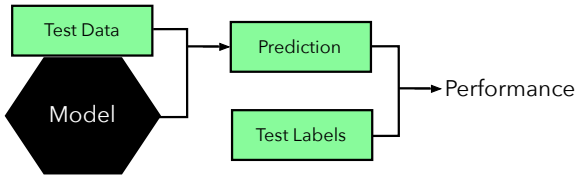


K-fold Cross-Validation Pipeline II

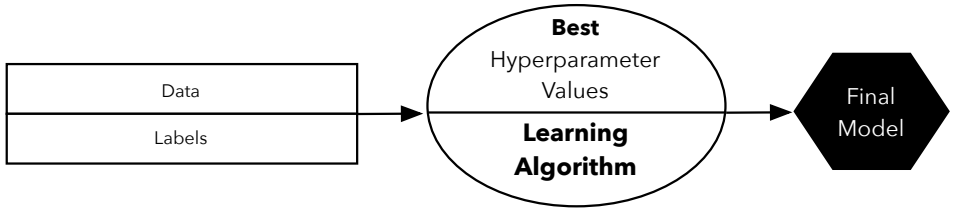
3



4



5



Learning Curves

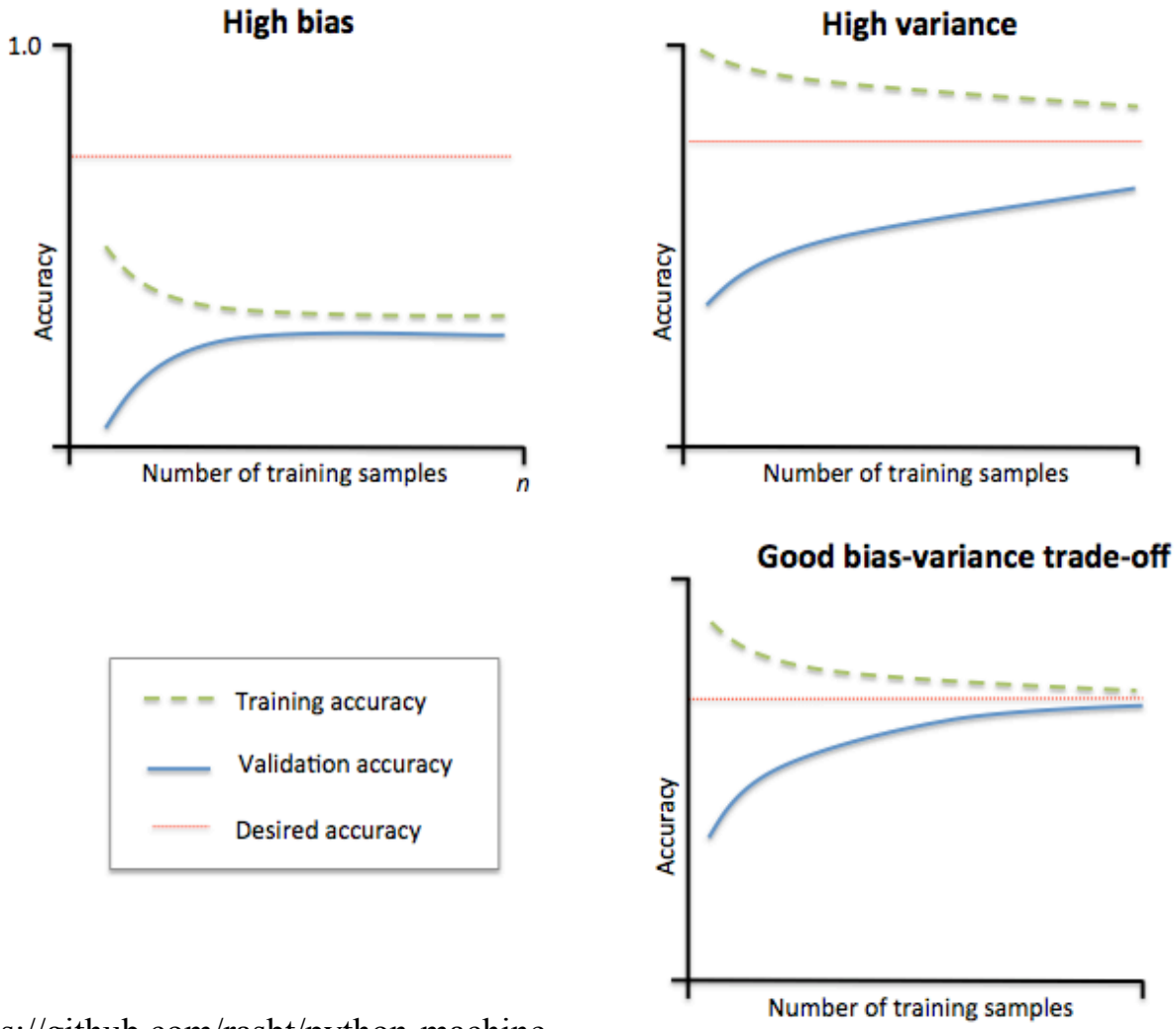
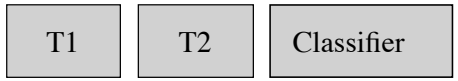


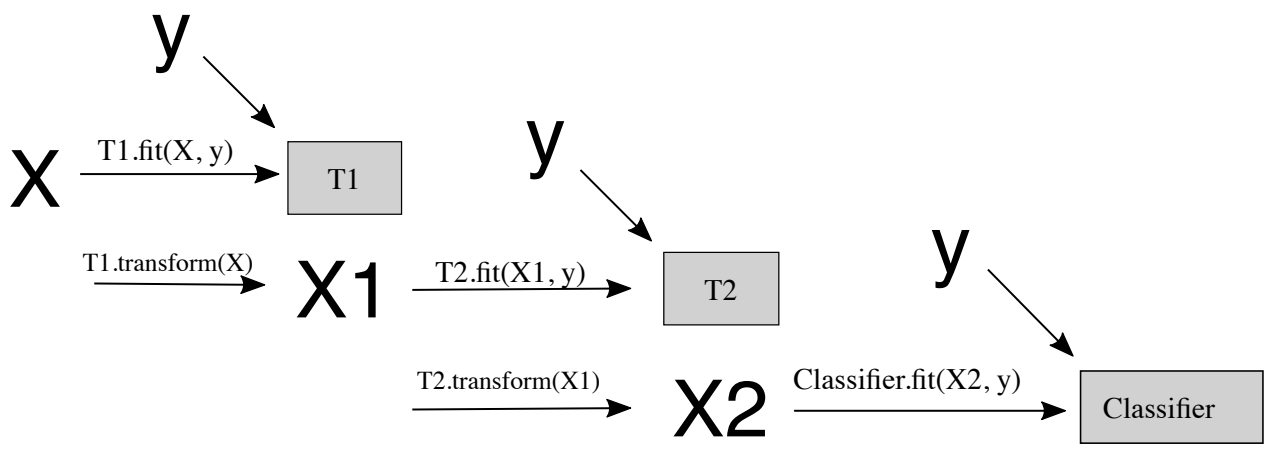
Image source: https://github.com/rasbt/python-machine-learning-book/blob/master/code/ch06/images/06_04.png

Pipelines

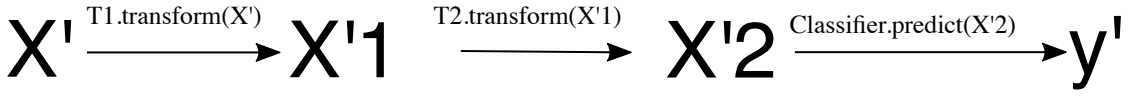
```
pipe = make_pipeline(T1(), T2(), Classifier())
```



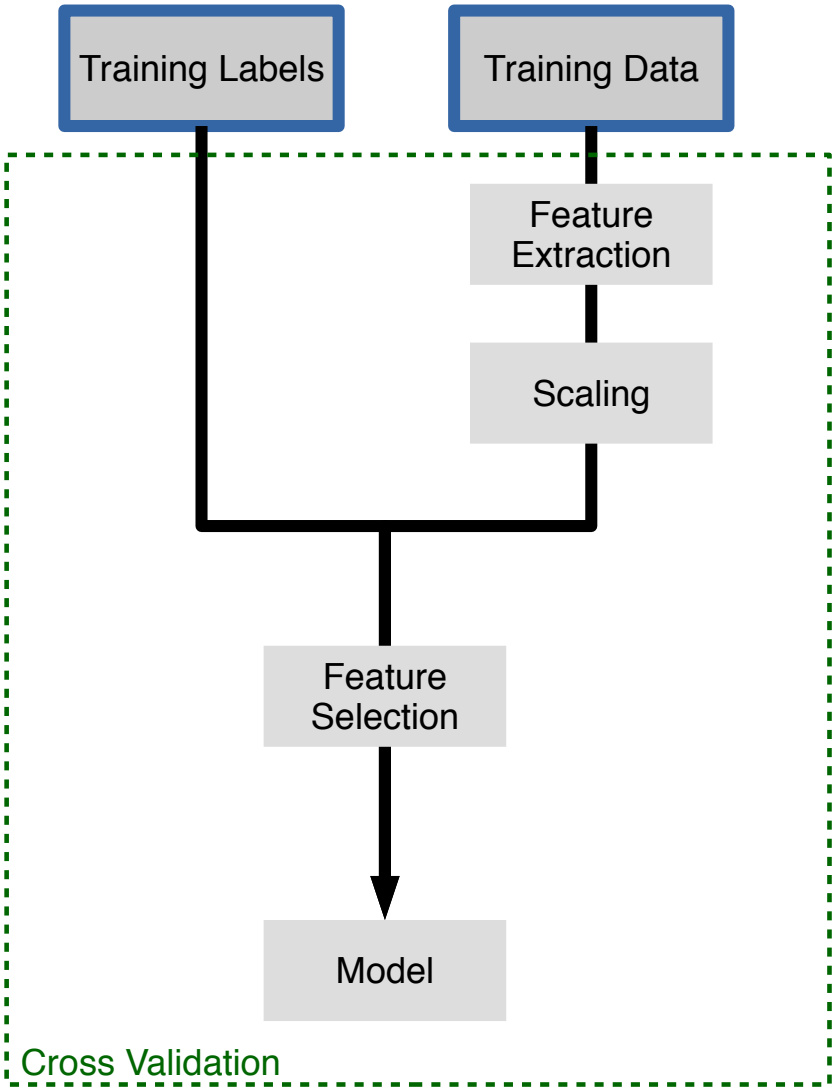

```
pipe.fit(X, y)
```




```
pipe.predict(X')
```

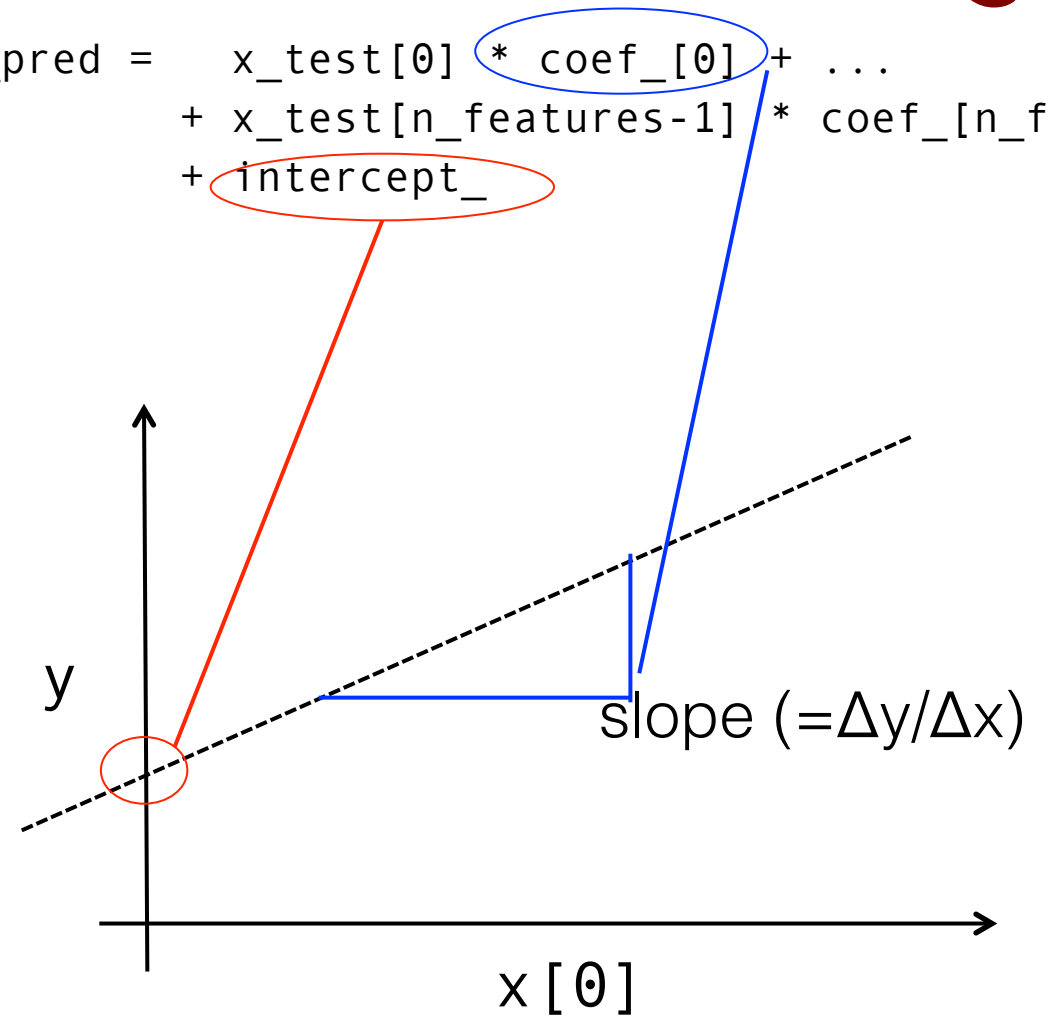


Pipelines & Cross Validation

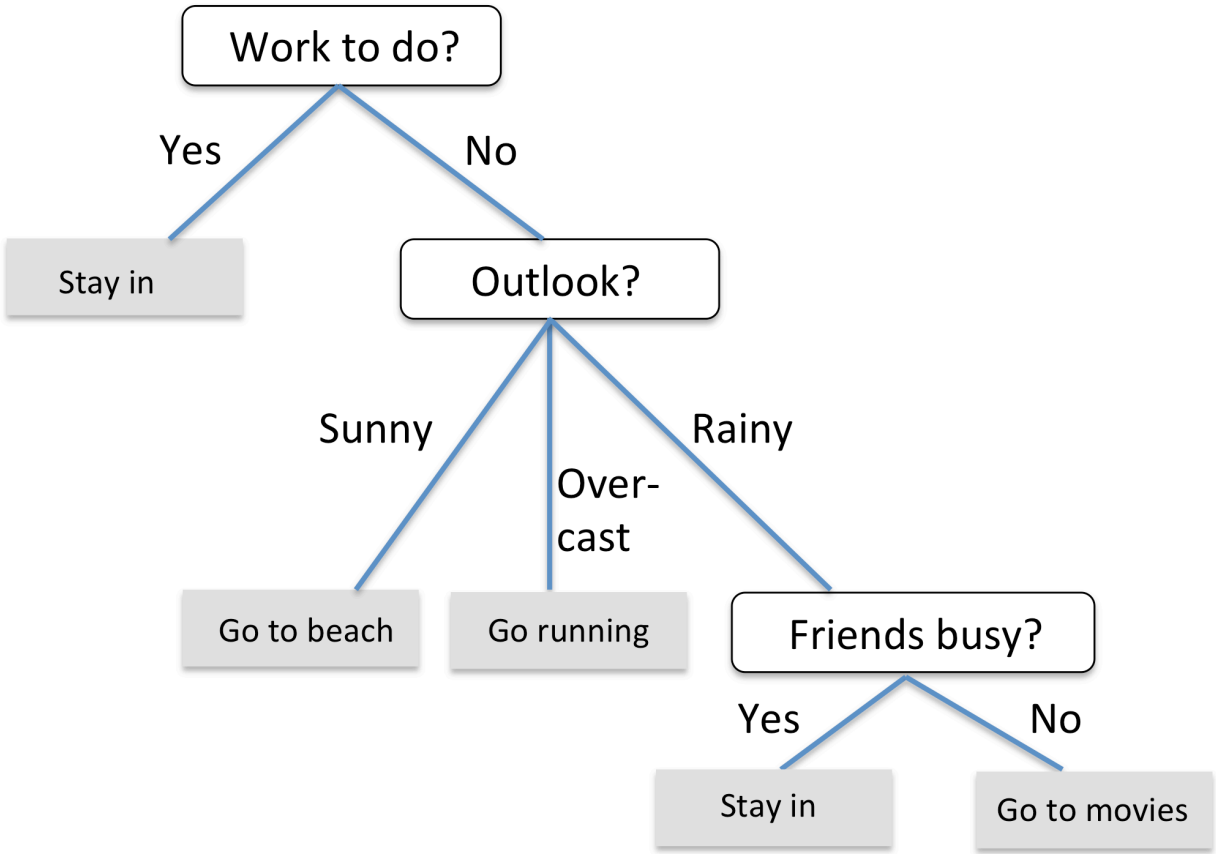


Linear models for regression

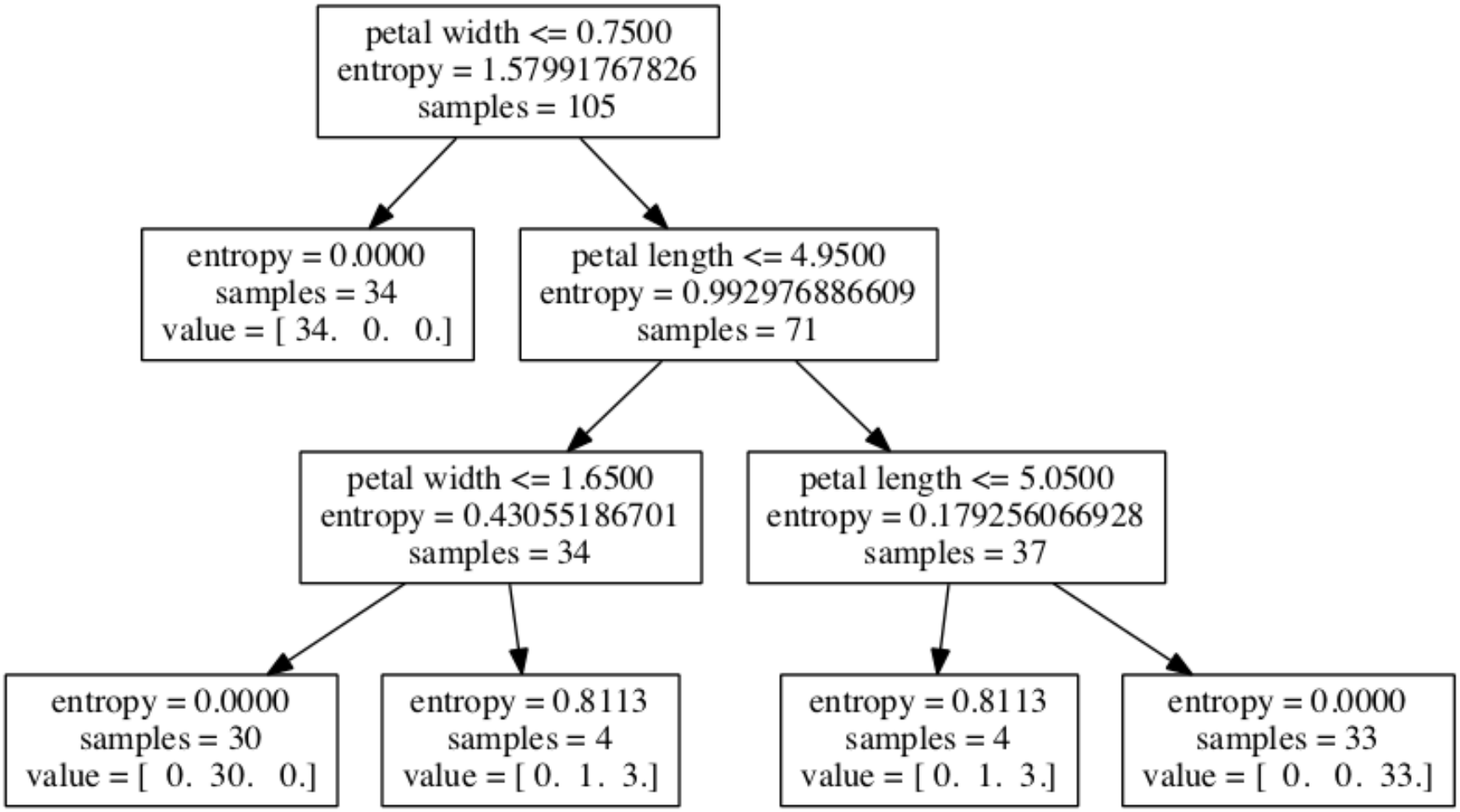
```
y_pred =   x_test[0] * coef_[0] + ...
         + x_test[n_features-1] * coef_[n_features-1]
         + intercept_
```



Decision Trees



Classification w. Continuous Features

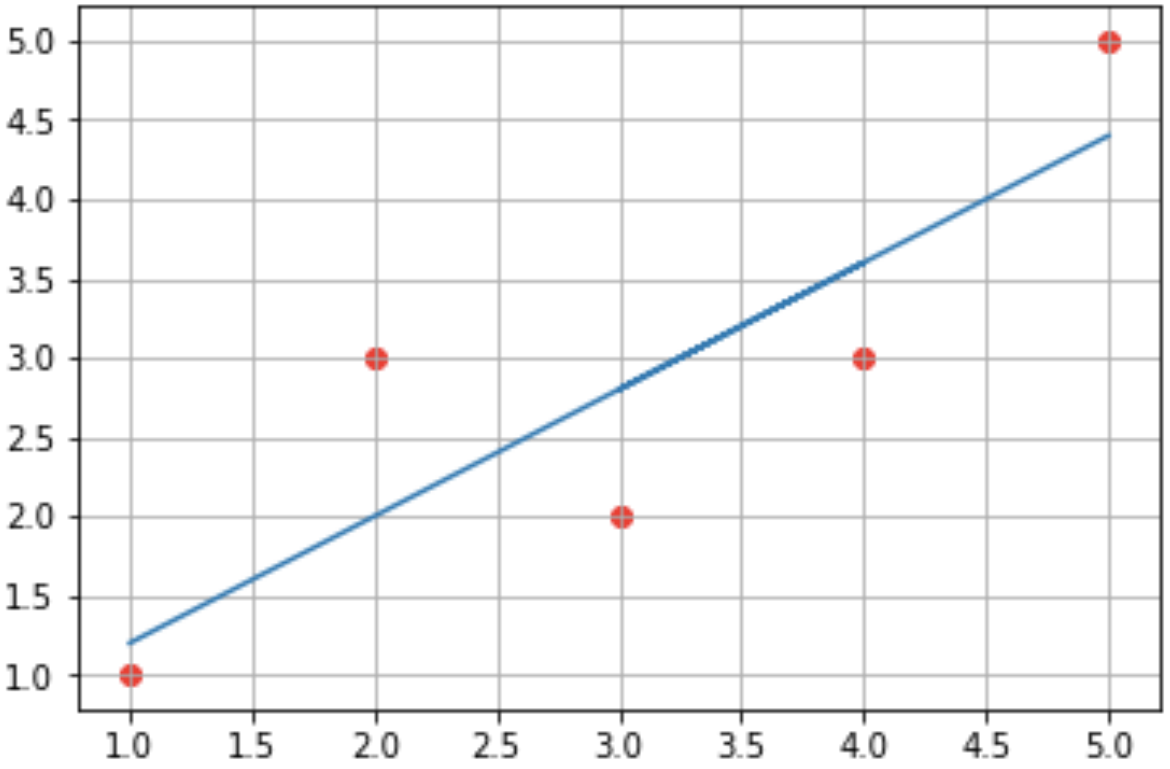


scikit-learn datasets

datasets.load_boston ([return_X_y])	Load and return the boston house-prices dataset (regression).
datasets.load_breast_cancer ([return_X_y])	Load and return the breast cancer wisconsin dataset (classification).
datasets.load_diabetes ([return_X_y])	Load and return the diabetes dataset (regression).
datasets.load_digits ([n_class, return_X_y])	Load and return the digits dataset (classification).
datasets.load_files (container_path[, ...])	Load text files with categories as subfolder names.
datasets.load_iris ([return_X_y])	Load and return the iris dataset (classification).
datasets.load_linnerud ([return_X_y])	Load and return the linnerud dataset (multivariate regression).
datasets.load_mlcomp (name_or_id[, set_, ...])	DEPRECATED: since the http://mlcomp.org/ website will shut down in March 2017, the load_mlcomp function was deprecated in version 0.19 and will be removed in 0.21.
datasets.load_sample_image (image_name)	Load the numpy array of a single sample image
datasets.load_sample_images ()	Load sample images for image manipulation.
datasets.load_svmlight_file (f[, n_features, ...])	Load datasets in the svmlight / libsvm format into sparse CSR matrix
datasets.load_svmlight_files (files[, ...])	Load dataset from multiple files in SVMlight format
datasets.load_wine ([return_X_y])	Load and return the wine dataset (classification).

Tutorial

x	y
1	1
2	3
4	3
3	2
5	5



Simple Linear Regression

- $y = B0 + B1 \times x$

$$B1 = \frac{\sum_{i=1}^{n}(x_i - mean(x)) \times (y_i - mean(y))}{\sum_{i=1}^{n}(x_i - mean(x))^2}$$

$$B0 = mean(y) - B1 \times mean(x)$$

$$mean(x) = 3$$
$$mean(y) = 2.8$$

$$B1 = \frac{8}{10}$$
$$B1 = 0.8$$

$$B0 = mean(y) - B1 \times mean(x)$$
$$B0 = 2.8 - 0.8 \times 3$$
$$B0 = 0.4$$

$$y = B0 + B1 \times x$$
$$y = 0.4 + 0.8 \times x$$

RMSE = 0.692820323

Simple Linear Regression

- Atalho

$$B1 = corr(x, y) \times \frac{stdev(y)}{stdev(x)}$$

$$B1 = 0.852802865 \times \frac{1.483239697}{1.58113883}$$

$$B1 = 0.8$$

Simple Linear Regression with Stochastic Gradient Descent

$$B0(t+1) = B0(t) - alpha \times error$$

$$B1(t+1) = B1(t) - alpha \times error \times x$$

RMSE = 0.720626401

alpha → learning rate
error → error we calculate for the training instance.

Simple Linear Regression with Stochastic Gradient Descent

B0	B1
0.01	0.01
0.0397	0.0694
0.066527	0.176708
0.08056049	0.21880847
0.118814462	0.410078328
0.123525534	0.4147894
0.14399449	0.455727313
0.154325453	0.497051164
0.157870663	0.507686795
0.180907617	0.622871563
0.182869825	0.624833772
0.198544452	0.656183024
0.200311686	0.663251962
0.19841101	0.657549935
0.213549404	0.733241901
0.21408149	0.733773988
0.227265196	0.760141398
0.224586888	0.749428167
0.219858174	0.735242025
0.230897491	0.79043861

Obrigado!

Dúvidas, comentários, sugestões?

Regis Pires Magalhães
regismagalhaes@ufc.br

UFC