GENERAL ASSEMBLY

INTRO TO DATA SCIENCE

LECTURE 3: KNN CLASSIFICATION AND MODEL SELECTION

January 5, 2015

DAT11-SF

CURRICULUM

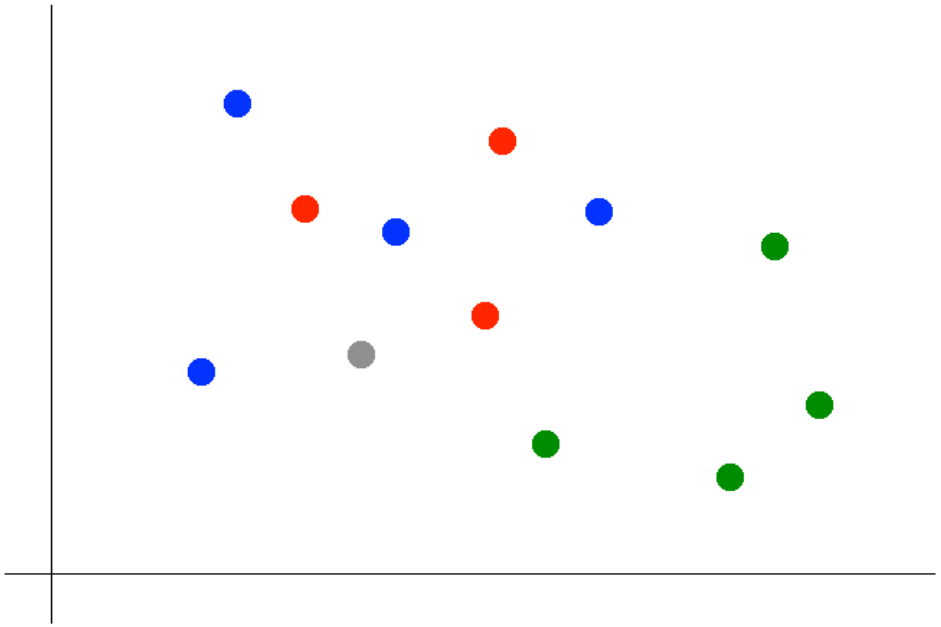
Date	Day	Topic	HW	Project
Nov 19	W	Introduction to data science		
Nov 24	M	Exploratory data analysis		
Dec 1	M	Introduction to machine learning	1	
Dec 3	W	Linear regression and regularization		
Dec 8	M	Logistic regression	2	
Dec 10	W	Support Vector Machines (SVM)		
Dec 15	M	Decision trees		
Dec 17	W	Practice session		
Jan 5	M	kNN and model selection		
Jan 7	W	Ensemble methods: random forest	3	Title
Jan 12	M	Naïve Bayes		
Jan 14	W	K-means and hierarchical clustering	4	Summary
Jan 21	W	Dimensionality reduction: PCA and SVD		
Jan 26	M	Text mining and information retrieval	5	
Jan 28	W	Network analysis		Proposal
Feb 2	M	Recommender systems	6	
Feb 4	W	Relational databases, SQL		
Feb 9	M	Big data storage and retrieval: noSQL, GraphDB		
Feb 11	W	Big data distributed computing: map-reduce, spark rdd		
Feb 18	W	Guest lecture		
Feb 23	M	Final projects presentations		Presentation
Feb 25	W	Final projects presentations		Presentation

I. KNN CLASSIFICATION
II. BIAS AND VARIANCE
III. MODEL SELECTION

LABS:
IV. KNN CLASSIFICATION IN SCIKIT-LEARN

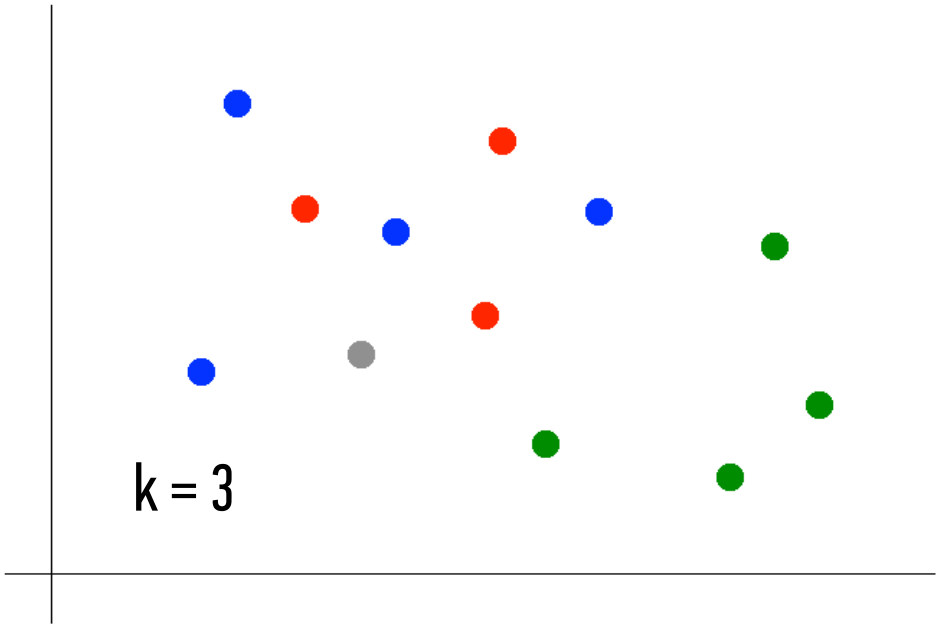
I. KNN CLASSIFICATION

Suppose we want to predict the color of the grey dot.



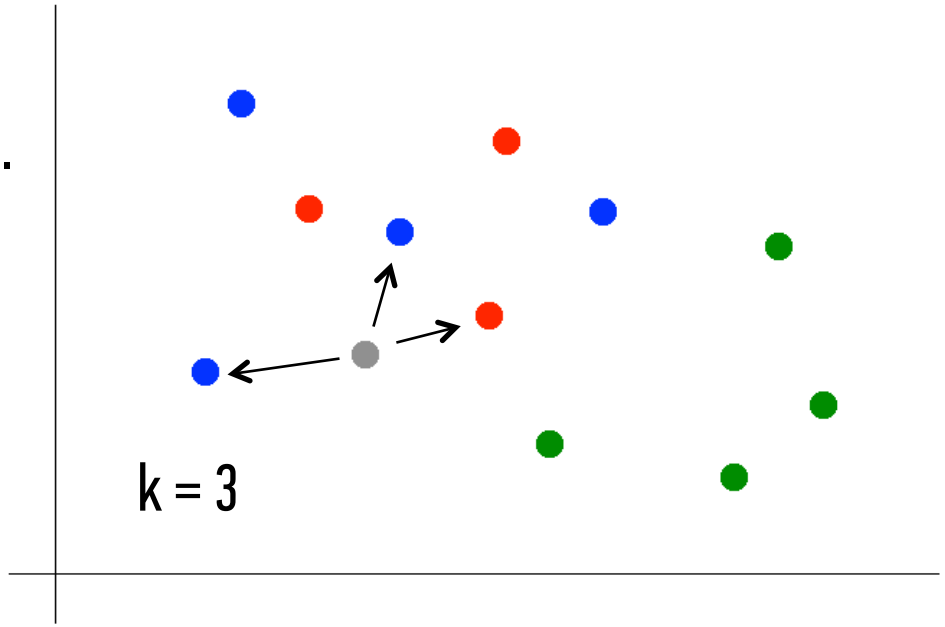
Suppose we want to predict the color of the grey dot.

1) Pick a value for k.



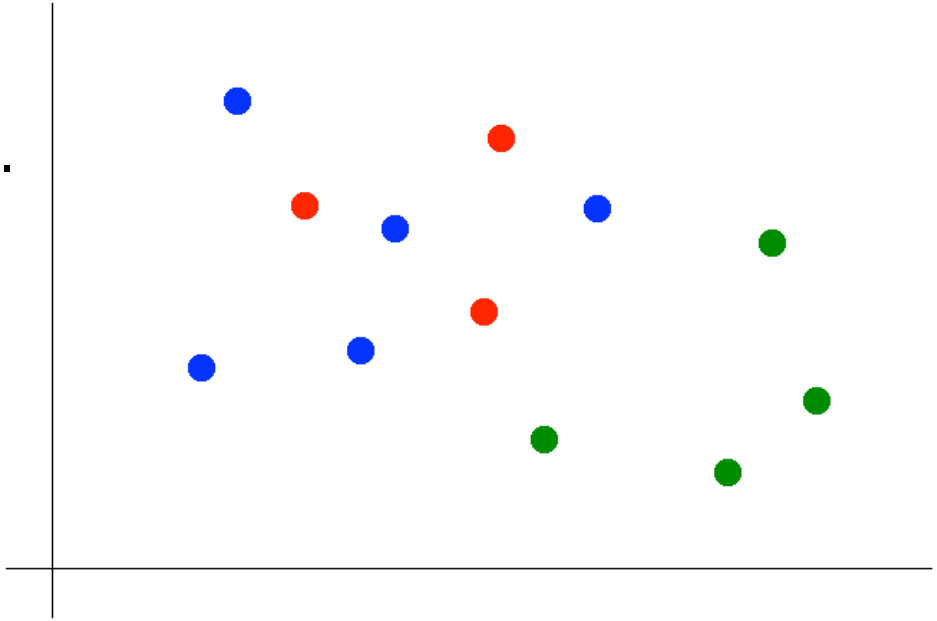
Suppose we want to predict the color of the grey dot.

1) Pick a value for k.
2) Find colors of k nearest neighbors.

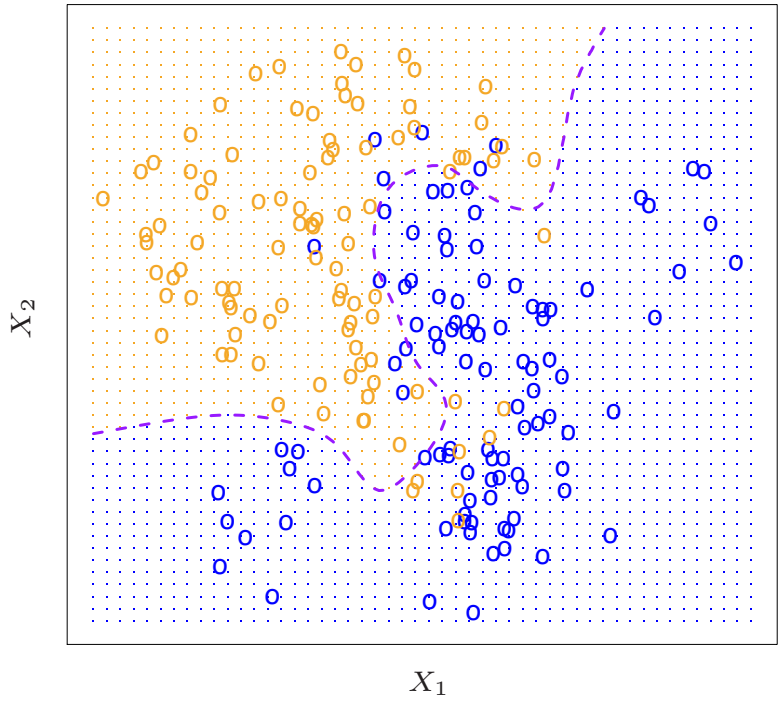


Suppose we want to predict the color of the grey dot.

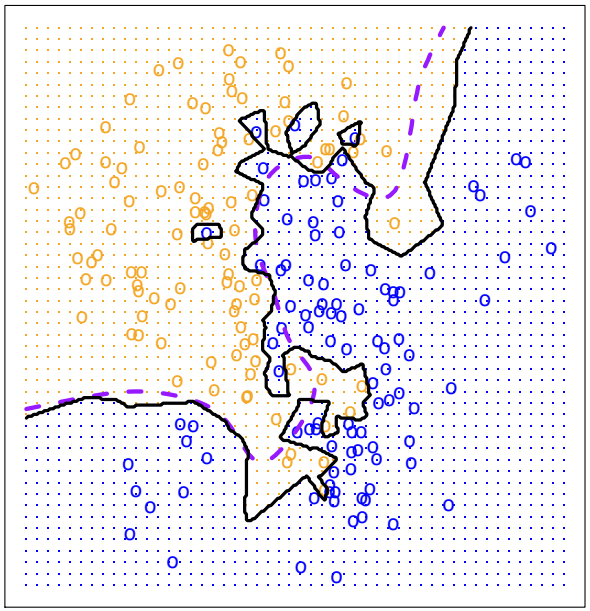
1) Pick a value for k.
2) Find colors of k nearest neighbors.
3) Assign the most common color to the grey dot.



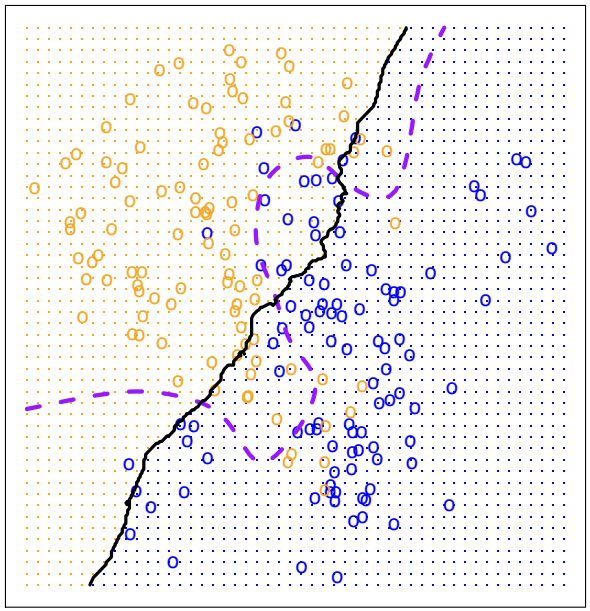
- Similarity/distance based method:
  - Euclidean distance
- Non-parametric method (no model to learn)
- Instance-based learning (in memory learning, no computations until classification)
- Can be used for :
  - classification (majority vote)
  - regression (averaging)

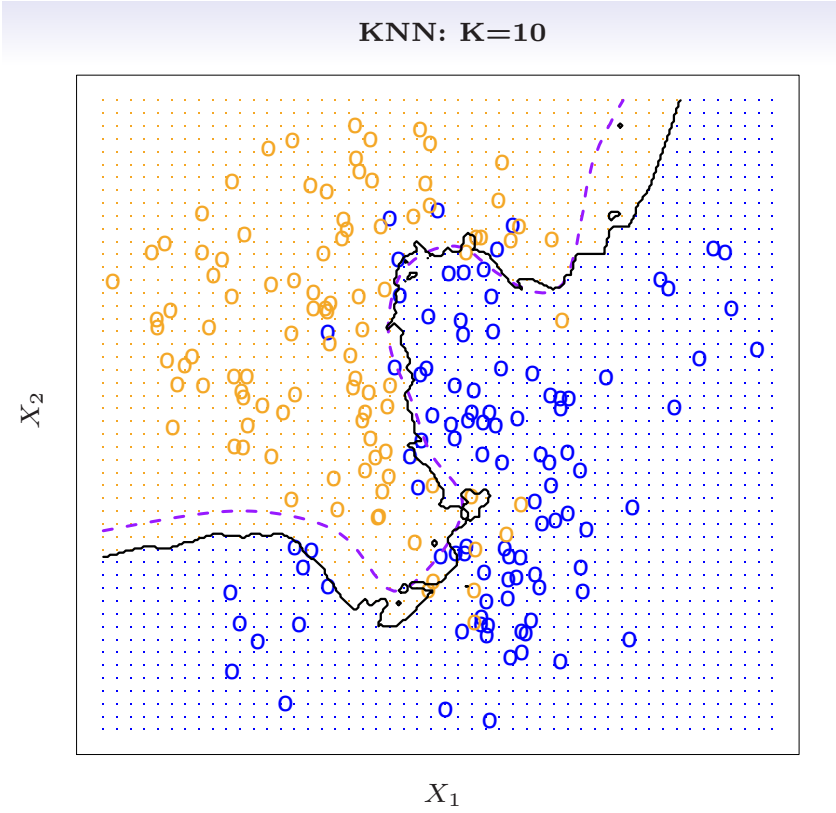

X_2
X_1

KNN: K=1

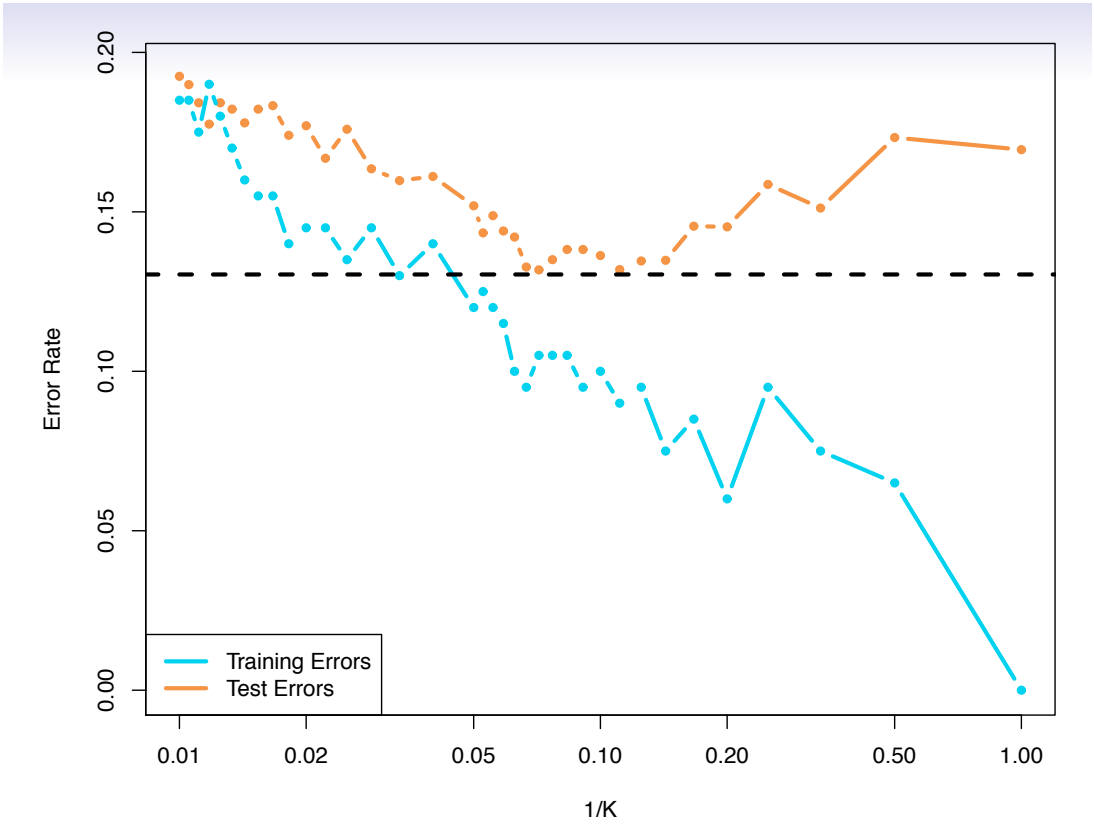


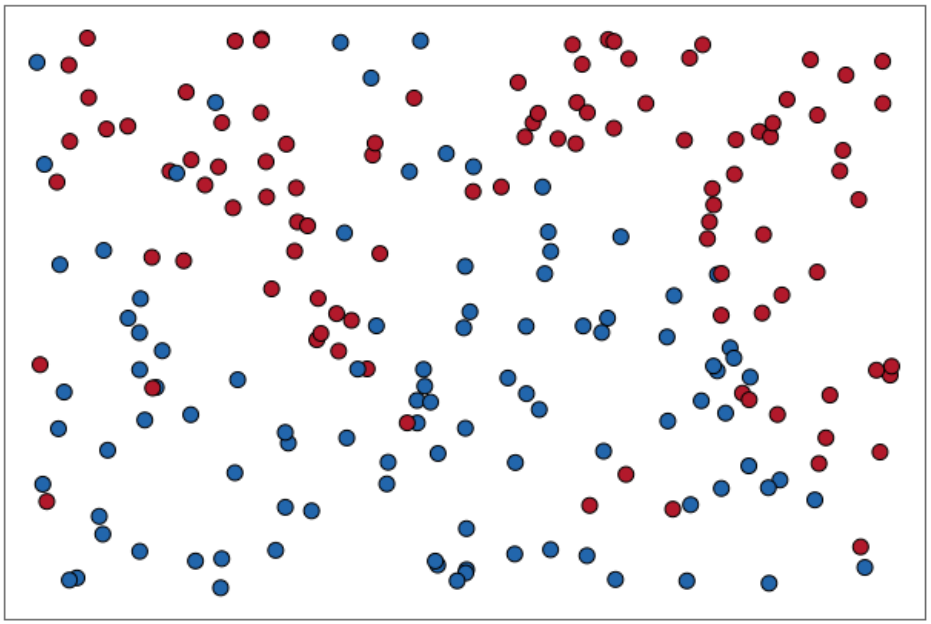
KNN: K=100



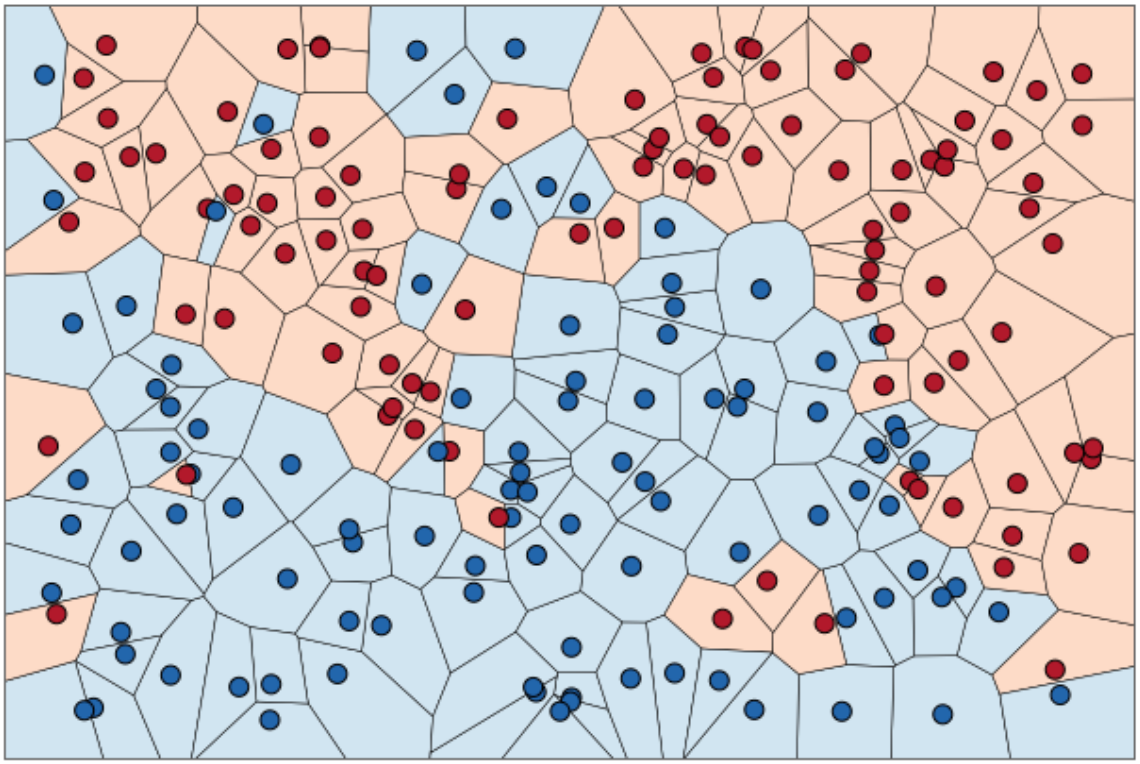

KNN: K=10
X_2
X_1

KNN ERROR


Error Rate
0.00
0.05
0.10
0.15
0.20
0.01
0.02
0.05
0.10
0.20
0.50
1.00
1/K
Training Errors
Test Errors

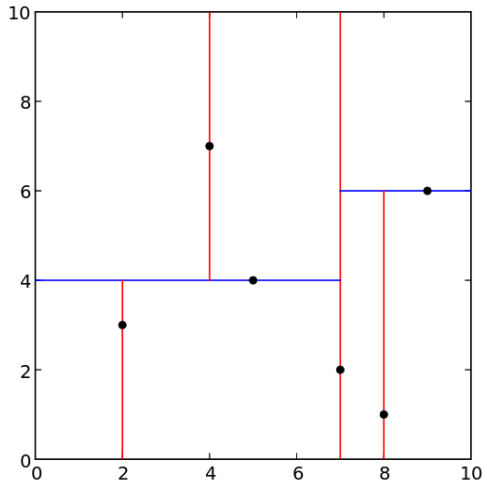


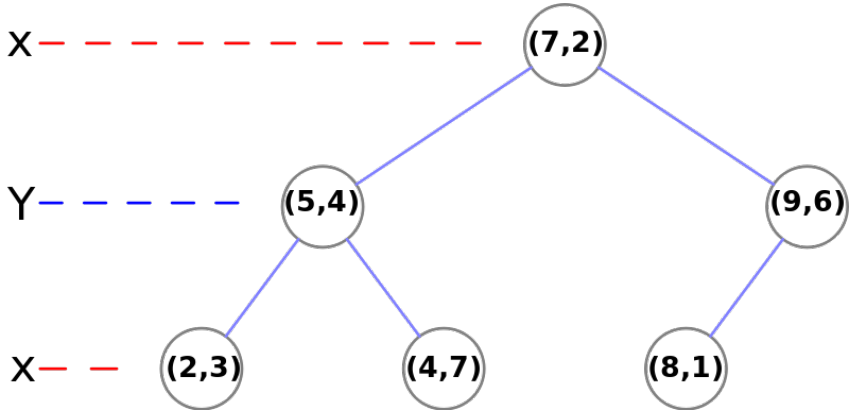
NEAREST NEIGHBOR CLASSIFICATION (KNN WITH K=1)



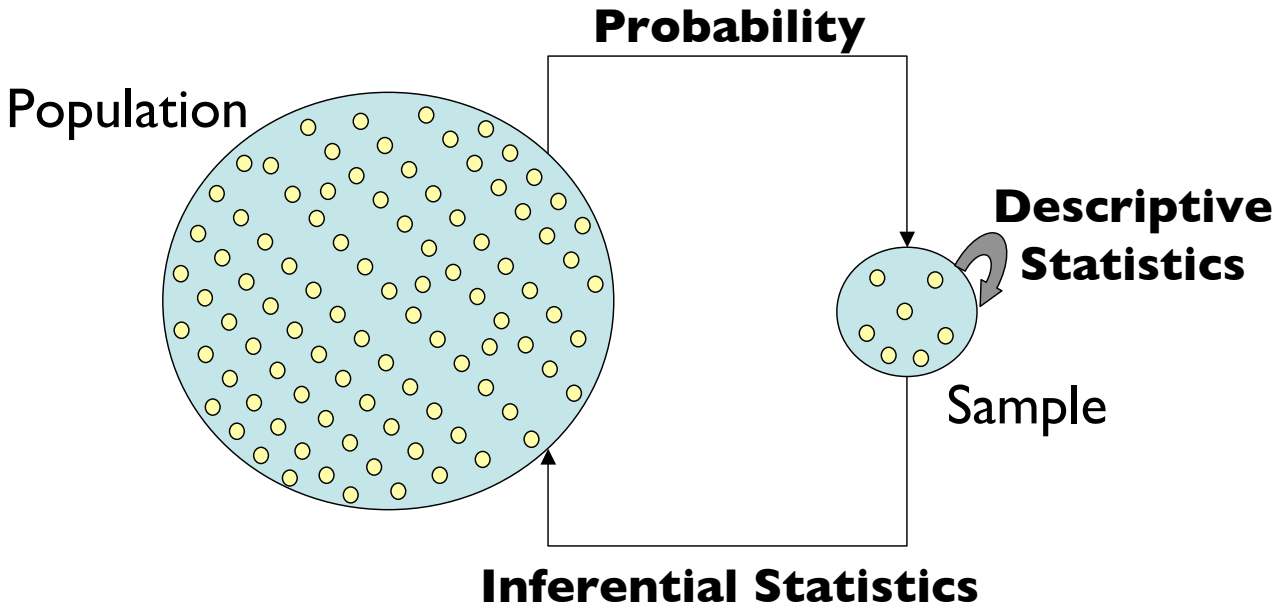
Voronoi diagram / tessellation

Kd-tree (k-dimensional tree) is a space partitioning data structure for organizing points in k-dimensional space by splitting by alternating axis-aligned hyperplanes



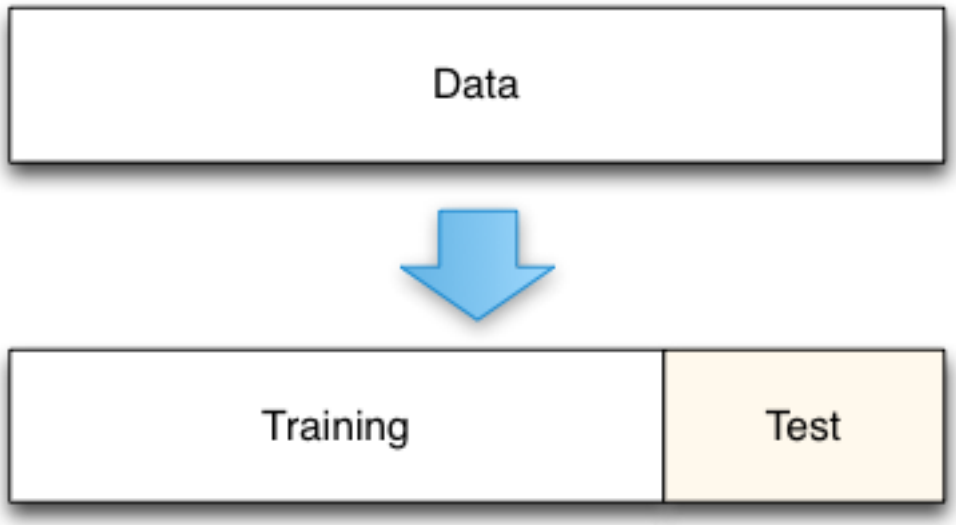


II. BIAS AND VARIANCE



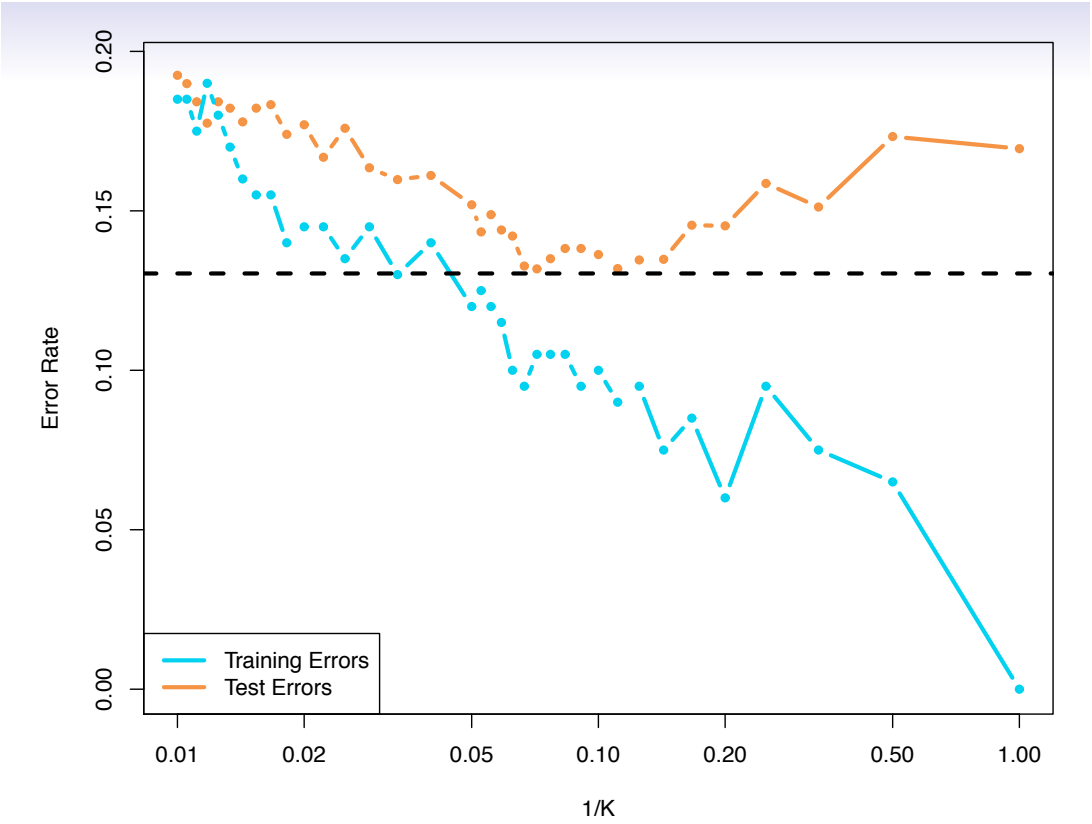
From http://cs109.github.io/2014/

TRAINING SET SAMPLE

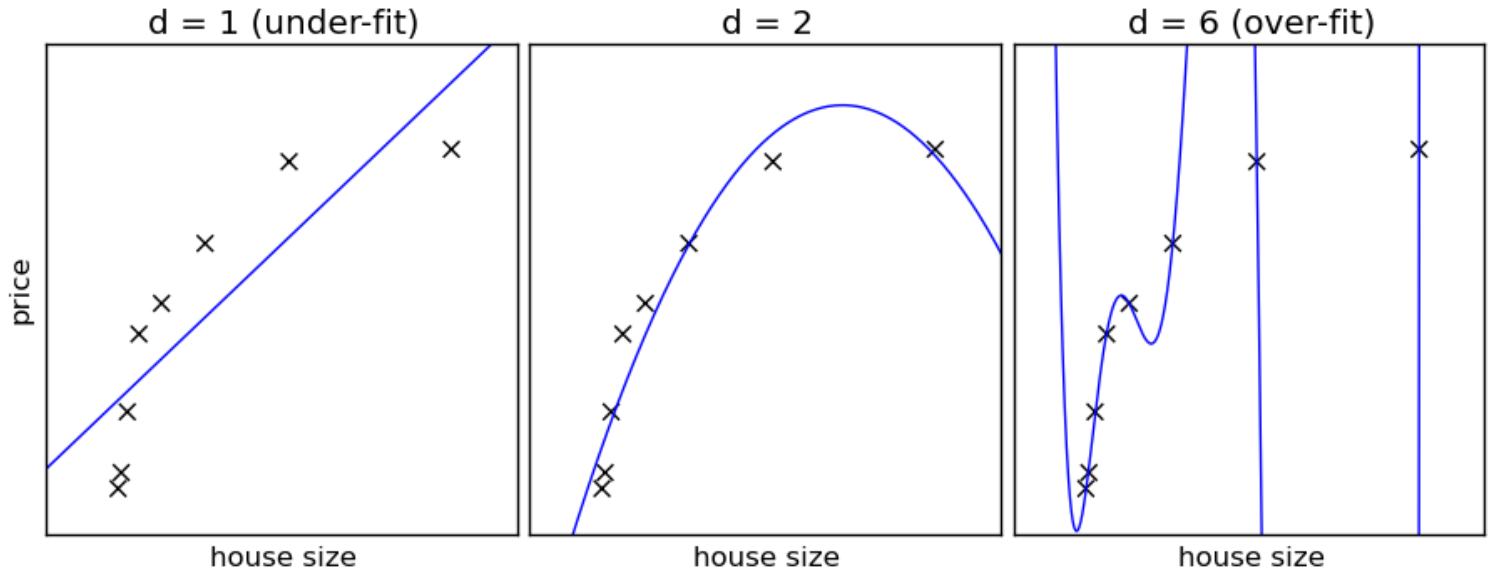


From http://cs109.github.io/2014/

KNN ERROR

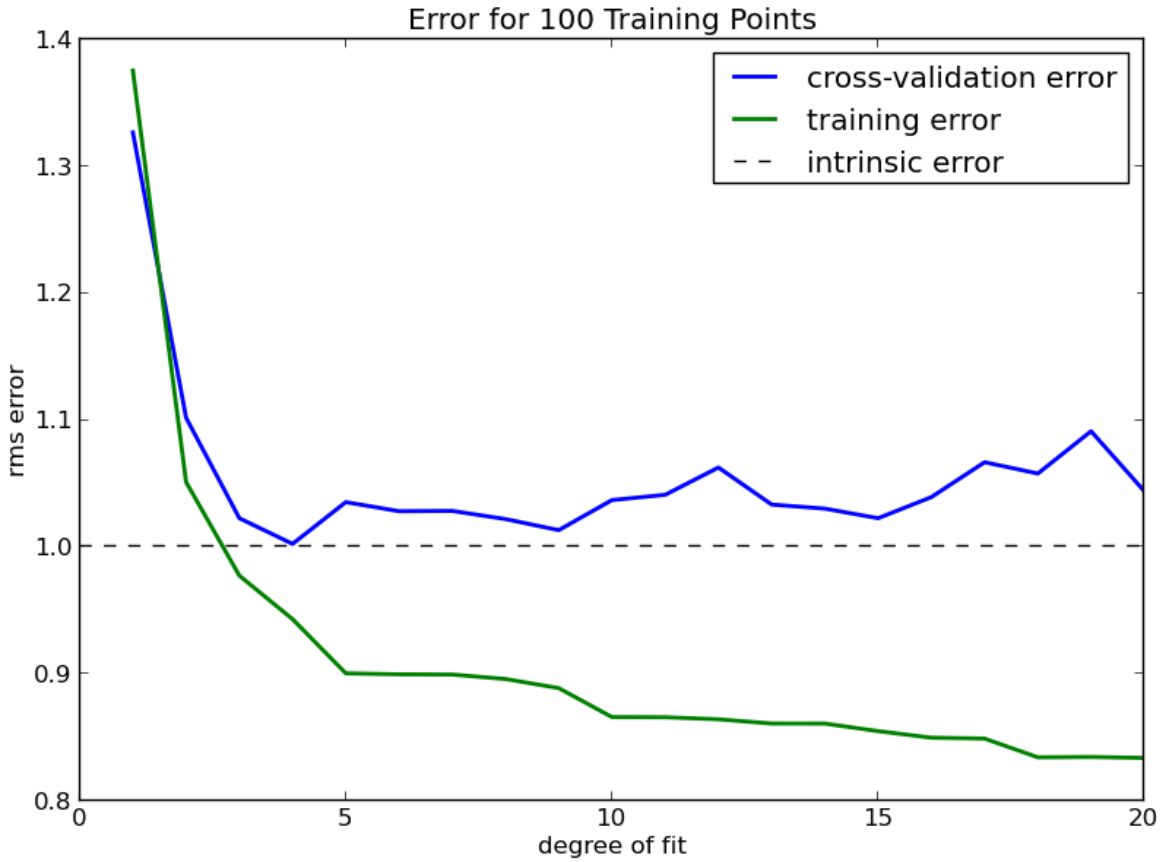

Error Rate
0.00
0.05
0.10
0.15
0.20
0.01
0.02
0.05
0.10
0.20
0.50
1.00
1/K
Training Errors
Test Errors

POLYNOMIAL FIT (LINEAR MODEL)

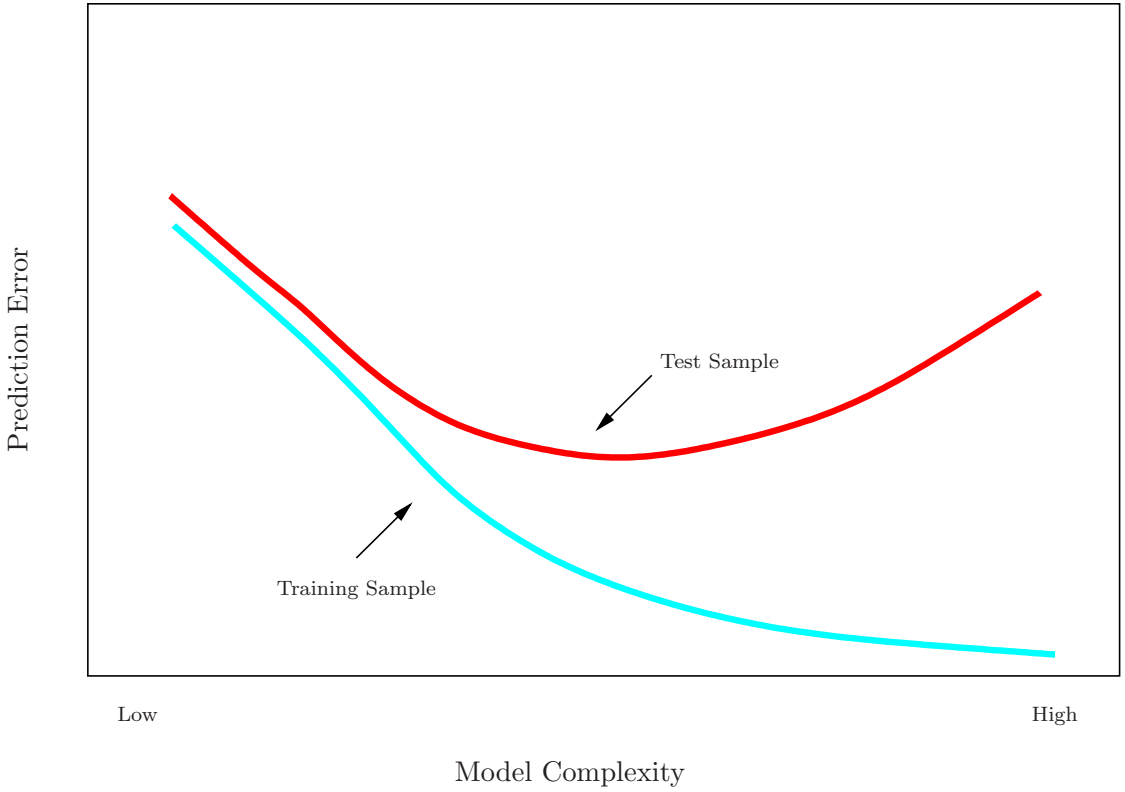


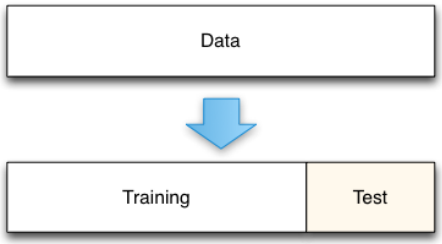
$\theta_0 + \theta_1 x$ $\theta_0 + \theta_1 x + \theta_2 x^2$ $\theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \theta_4 x^4$

POLYNOMIAL FIT ERROR



TRAINING-SET VS TEST-SET PERFORMANCE

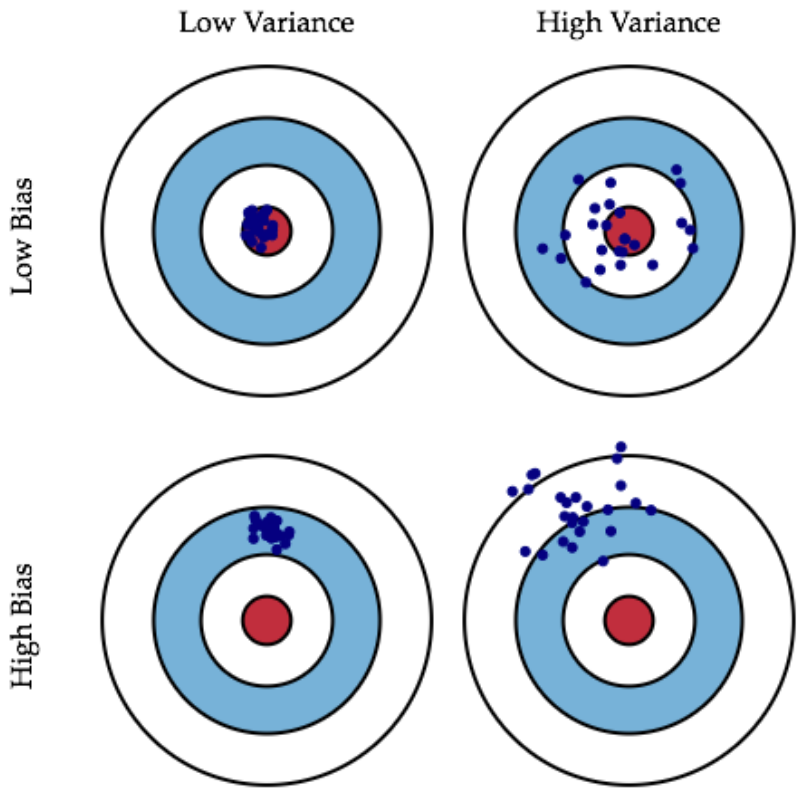




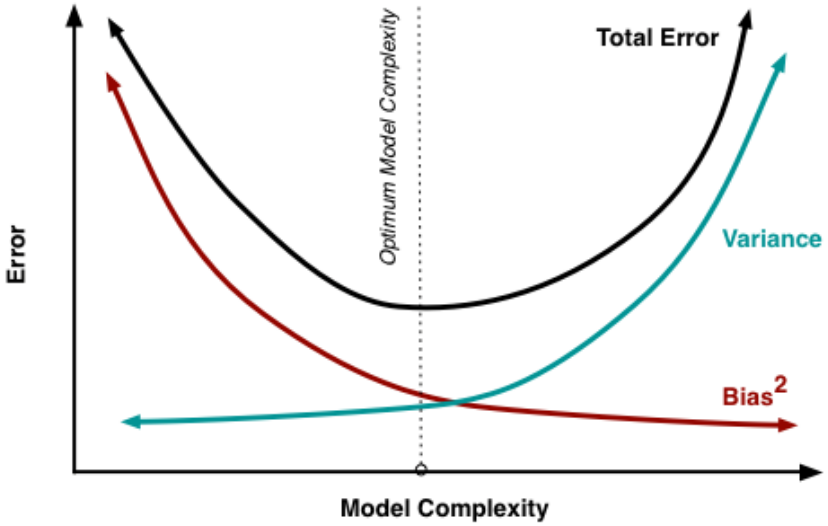
Bias – variance decomposition is a decomposition of the test (out of sample) error into two components:

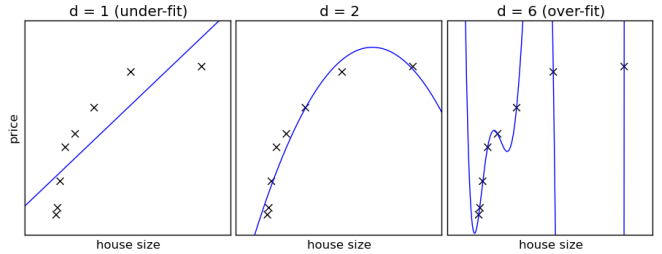
- Bias – difference between average prediction of the model and the true values. Bias is due to systematic erroneous assumptions in the learning algorithm
- Variance - variability of model prediction for a given value. It is due to sensitivity to fluctuations in the training data set

BIAS AND VARIANCE



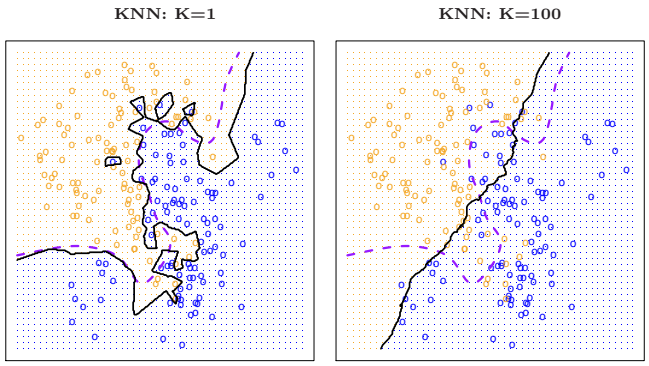
BIAS – VARIANCE TRADEOFF





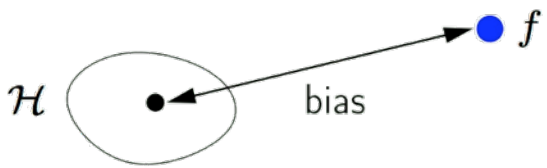
$$Err(x) = \left(E[\hat{f}(x)] - f(x)\right)^2 + E\left[\hat{f}(x) - E[\hat{f}(x)]\right]^2$$

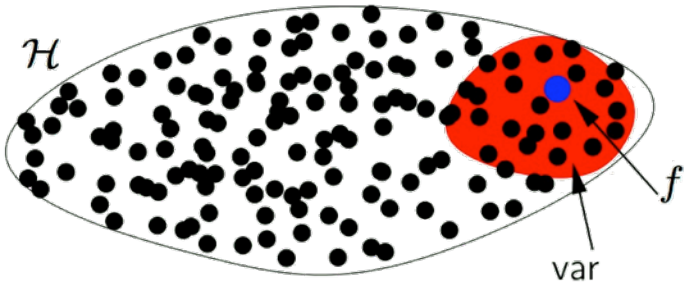
$$Err(x) = \text{Bias}^2 + \text{Variance}$$



BIAS-VARIANCE TRADEOFF

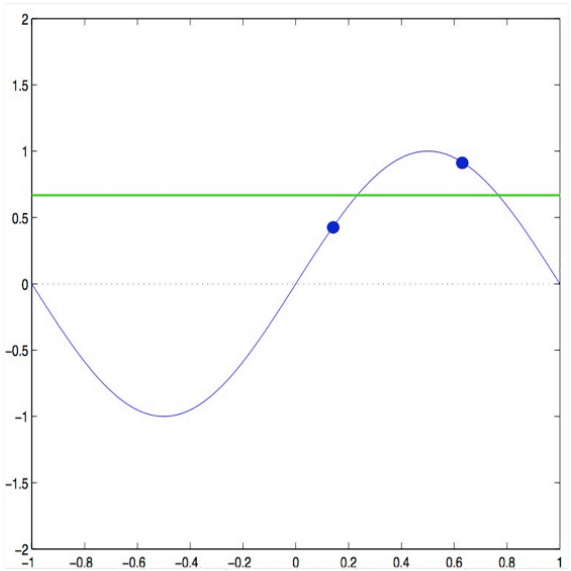
$$\text{bias} = \mathbb{E}_{\mathbf{x}}\left[\left(\bar{g}(\mathbf{x}) - f(\mathbf{x})\right)^2\right] \qquad \text{var} = \mathbb{E}_{\mathbf{x}}\left[\mathbb{E}_{\mathcal{D}}\left[\left(g^{(\mathcal{D})}(\mathbf{x}) - \bar{g}(\mathbf{x})\right)^2\right]\right]$$



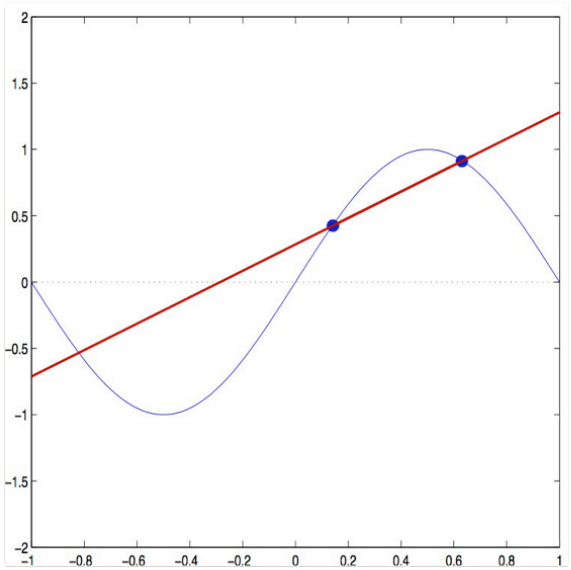


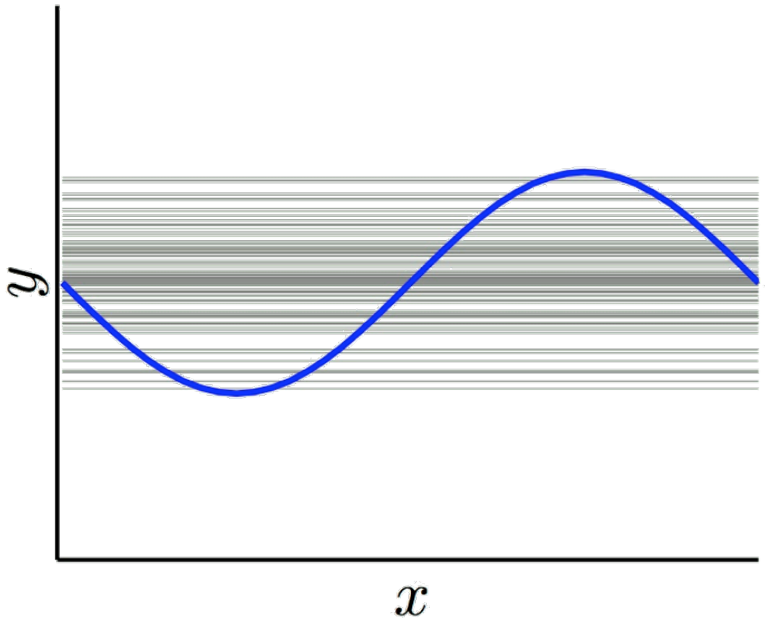
BIAS AND VARIANCE

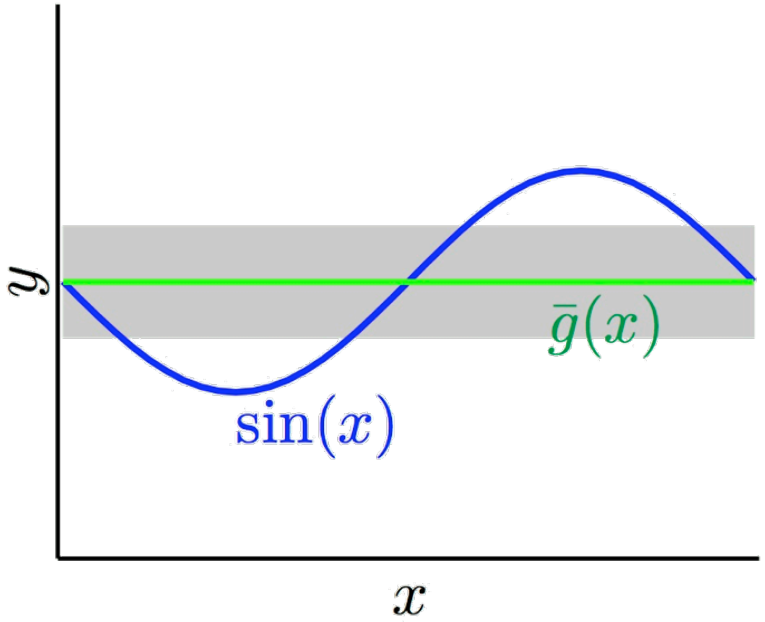
$\mathcal{H}_0$

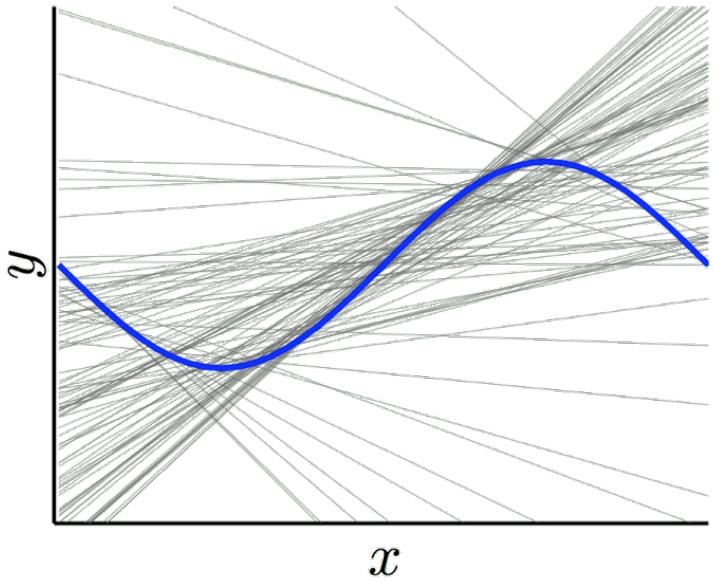


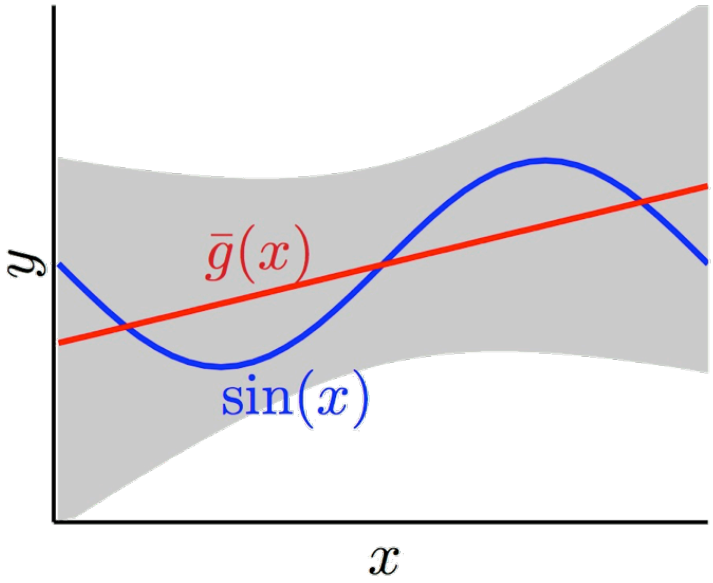
$\mathcal{H}_1$

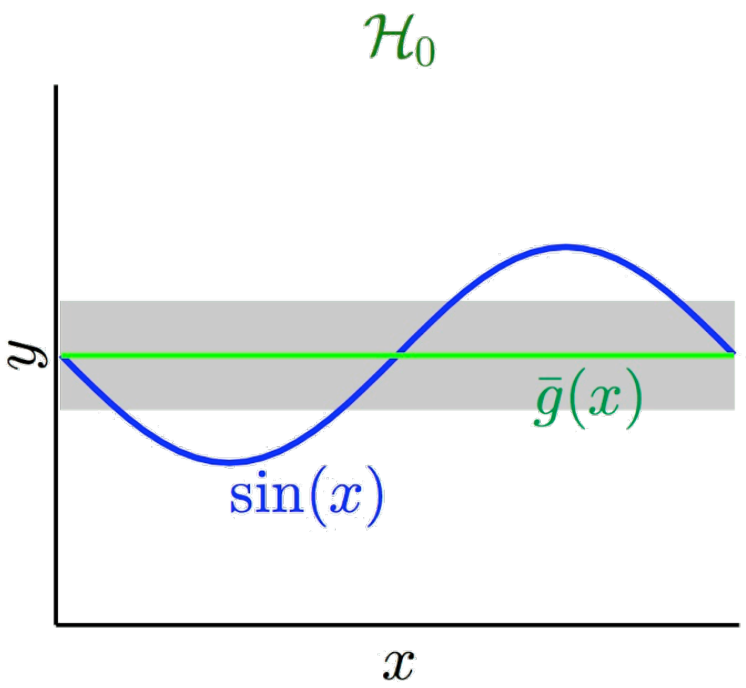



y
x

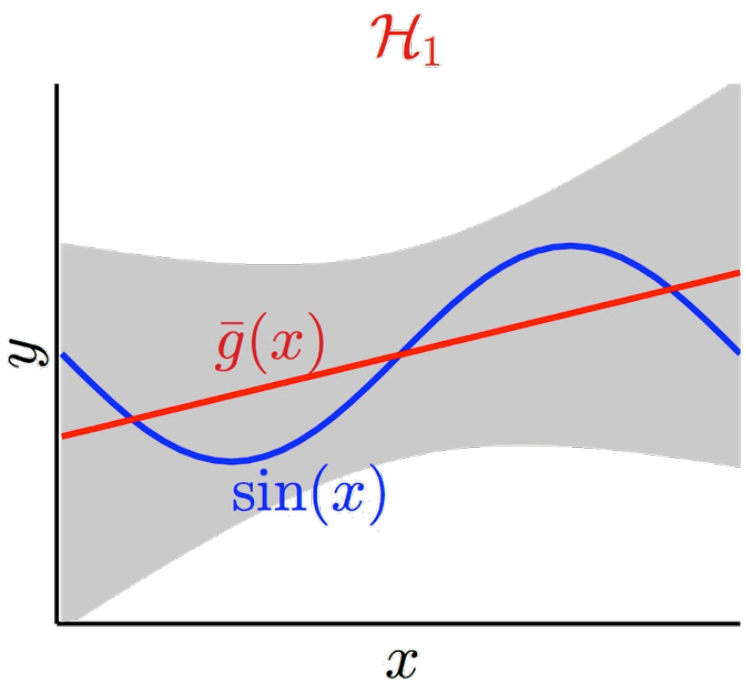

y
$\bar{g}(x)$
$\sin(x)$
x


y
x


$\bar{g}(x)$
$\sin(x)$
y
x



bias $= \mathbf{0.50}$ var $= \mathbf{0.25}$



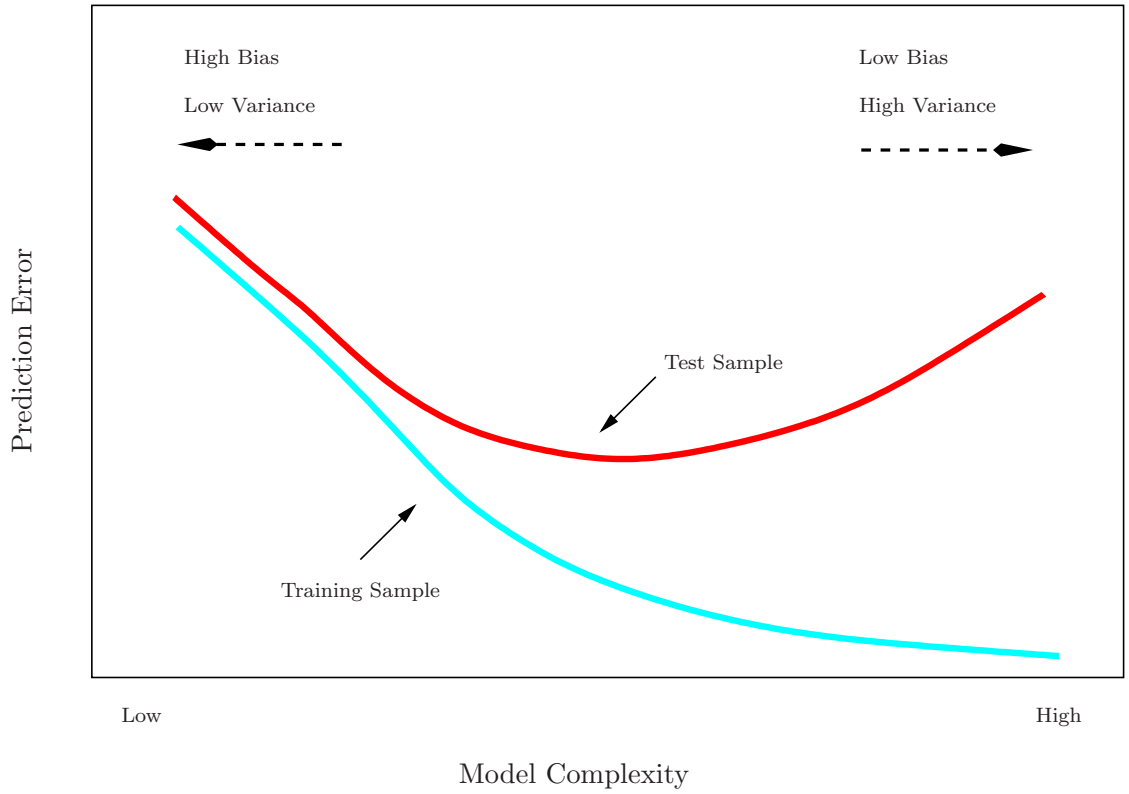
bias $= \mathbf{0.21}$ var $= \mathbf{1.69}$

INTRO TO DATA SCIENCE

III. MODEL SELECTION

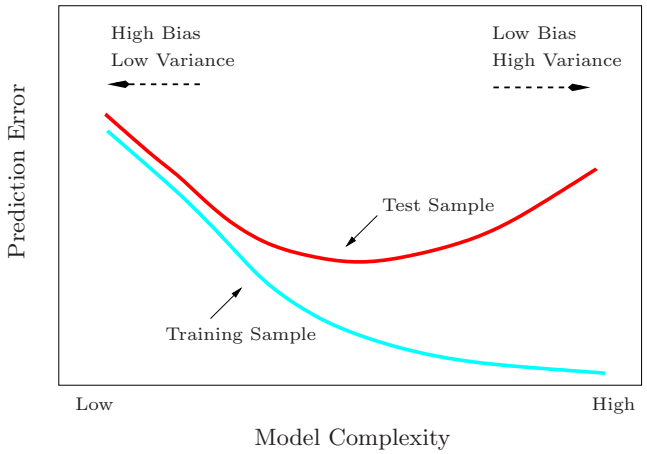
- Model training - fitting to training data
- Model selection – selecting the best model (model parameters)
- Model assessment – estimating prediction error

Dataset:

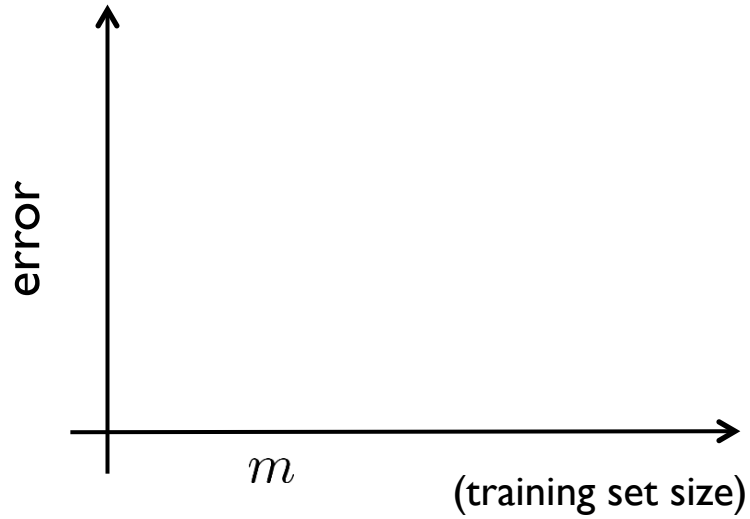

High Bias
Low Variance
Low Bias
High Variance
Test Sample
Training Sample
Prediction Error
Low
High
Model Complexity

- Selecting the best model complexity:
  - Hyperparameter tuning
  - Regularization level

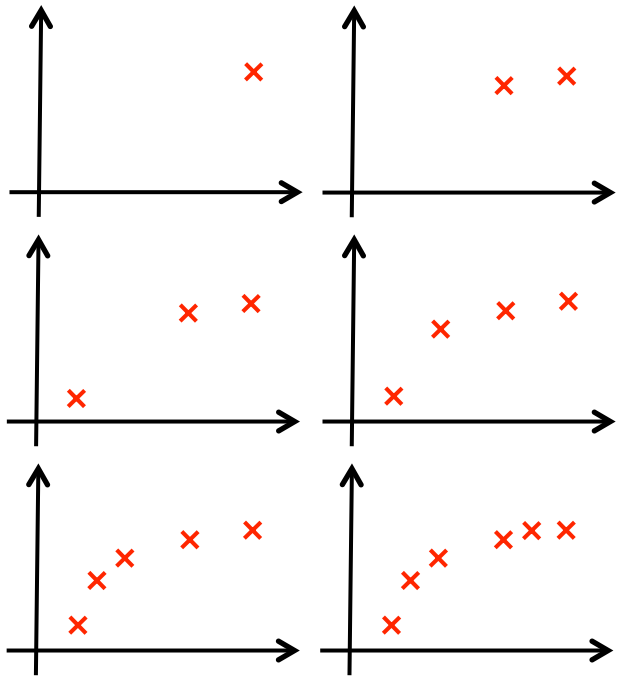
- Parameter search:
  - Exhaustive grid search
  - Randomized optimization



Learning curves

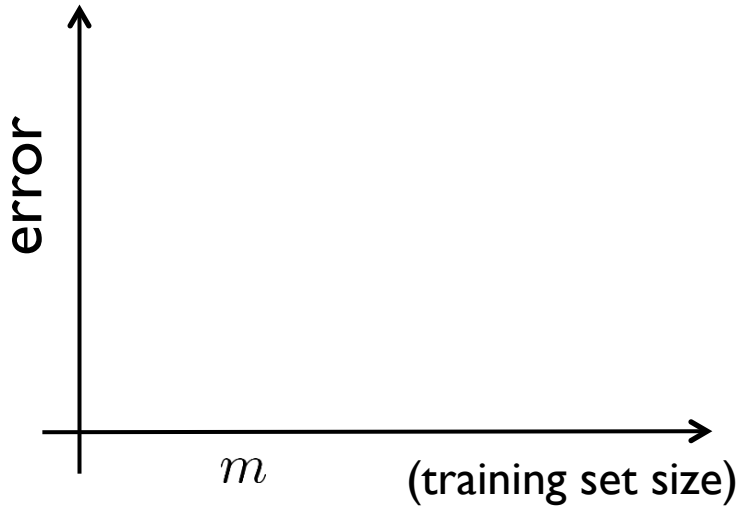


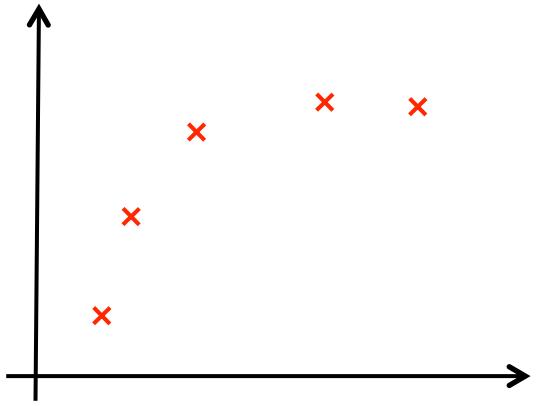
$$h_\theta(x) = \theta_0 + \theta_1 x + \theta_2 x^2$$

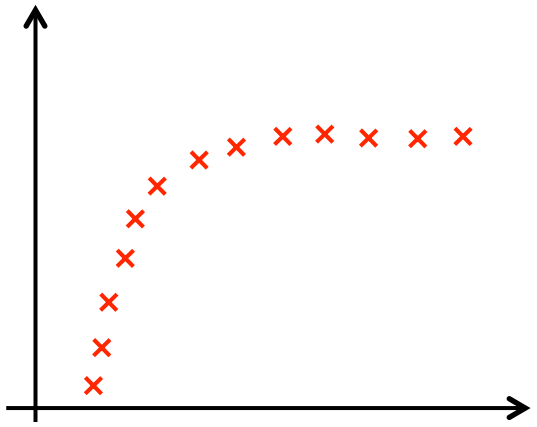


Learning curves

$h_\theta(x) = \theta_0 + \theta_1 x$



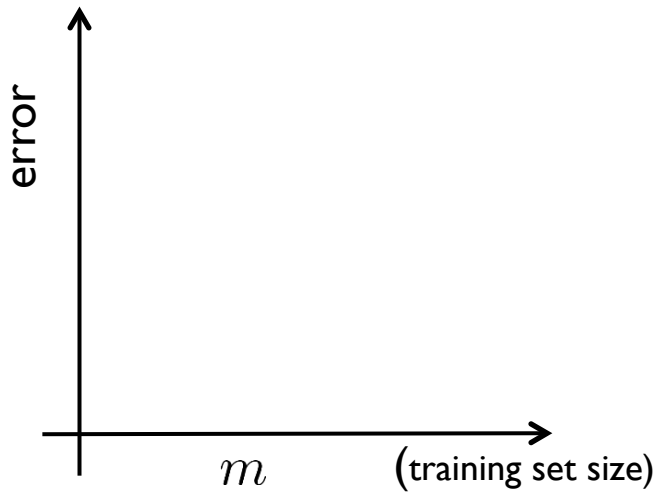


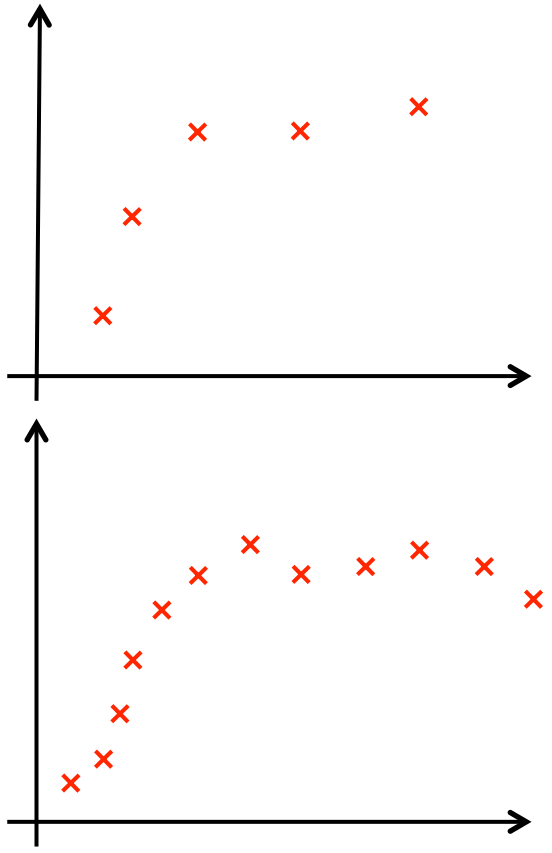


Simple model (high bias)

Learning curves

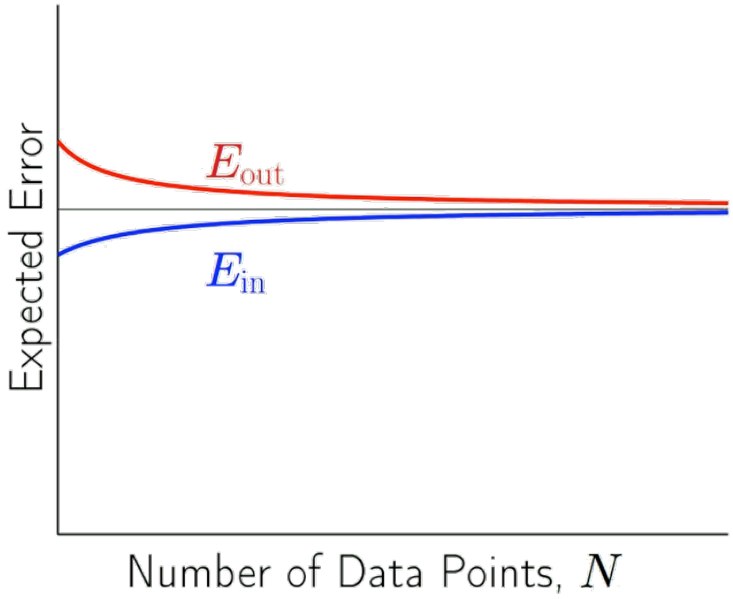
$$h_\theta(x) = \theta_0 + \theta_1 x + \cdots + \theta_{100} x^{100}$$



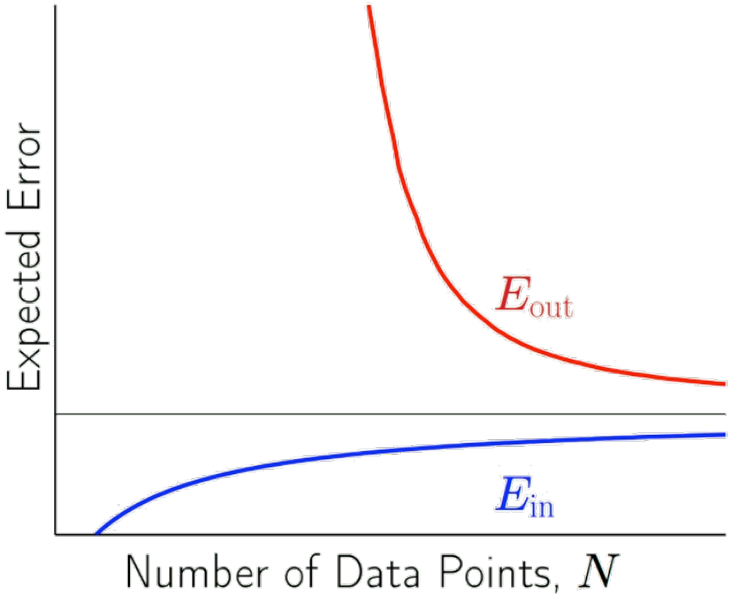


Complex model (high variance)

LEARNING CURVE

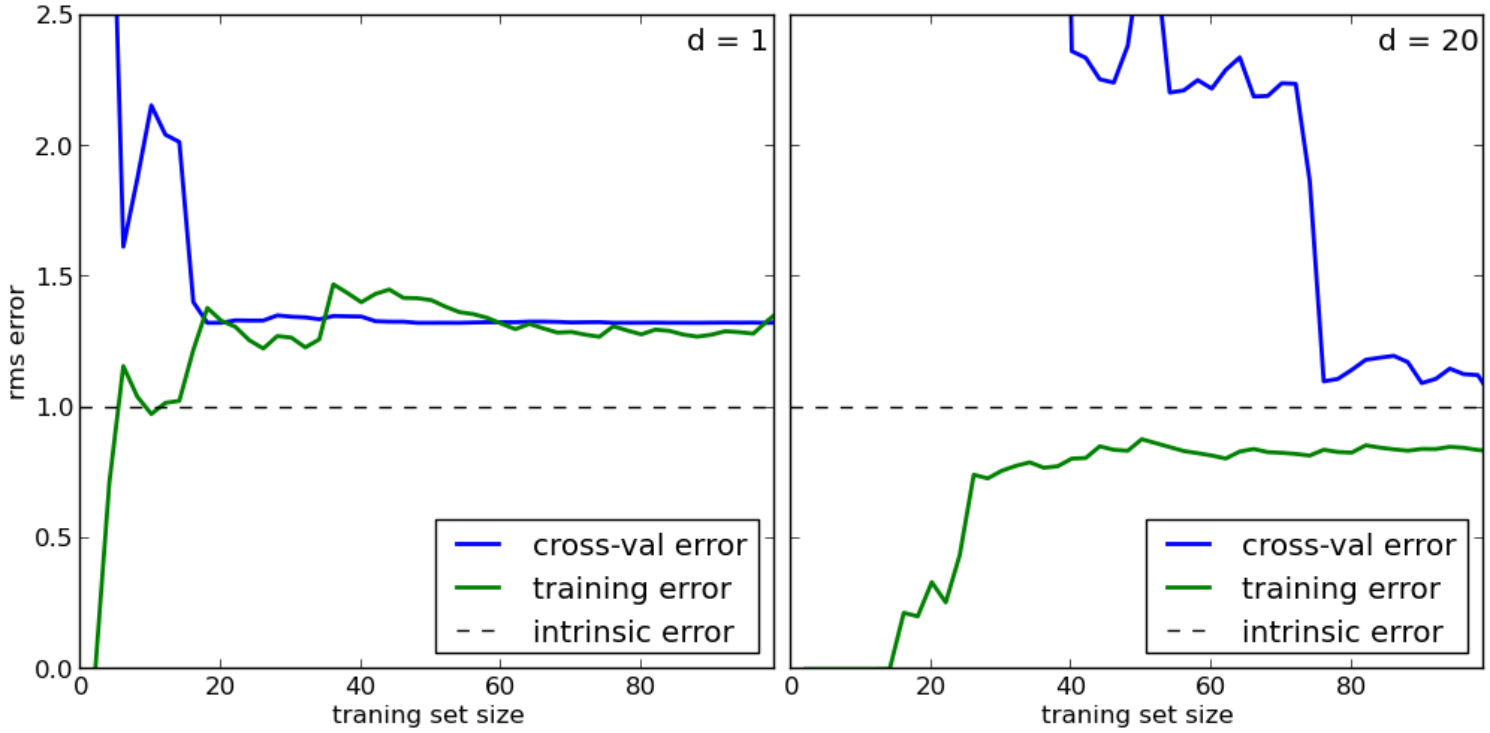


Simple Model

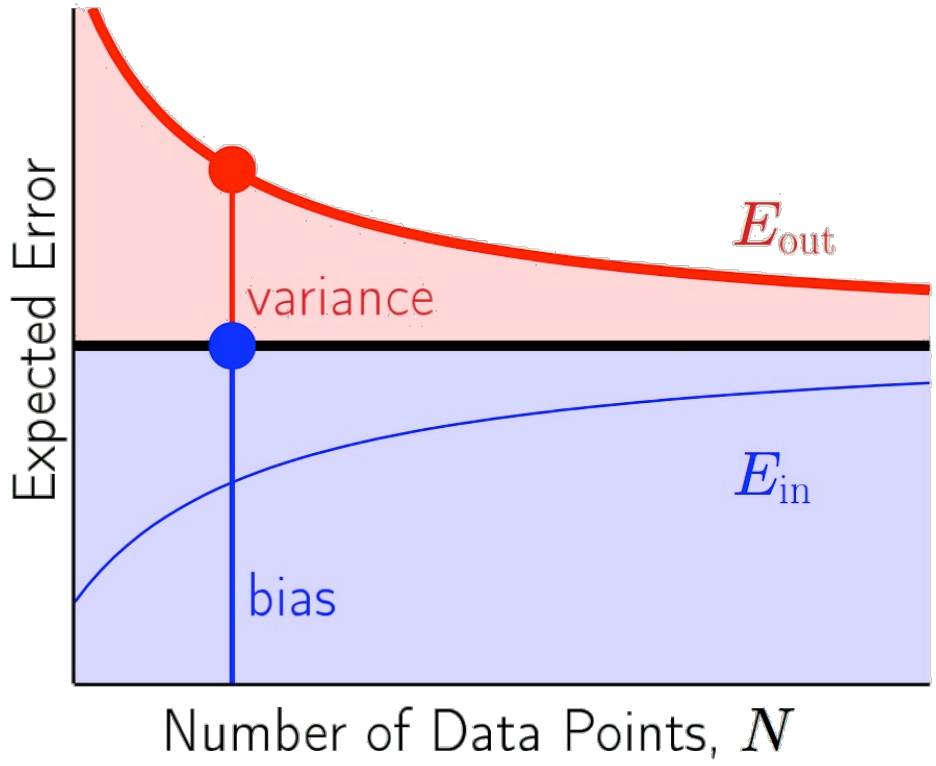


Complex Model

LEARNING CURVE



LEARNING CURVE



INTRO TO DATA SCIENCE

IV LAB: KNN CLASSIFICATION IN SCIKIT-LEARN