DS 4400

Machine Learning and Data Mining I
Spring 2024

David Liu
Khoury College of Computer Science
Northeastern University

February 2 2024

Outline

- Regularization
  – A geometric understanding
- Classification
  – K Nearest Neighbors (kNN)
- Cross validation
  – K-fold cross validation
  – Leave one out cross validation
- Linear classifiers

Announcements

- Thank you for survey feedback!
  - Continue providing feedback if you haven’t yet via Google Form posted on Piazza.
  - Clear request for more whiteboarding.
    - Starting next Friday
  - Request for examples, demos to understand the equations.
    - Starting next Tuesday
  - Will likely reduce supplemental lectures as a result.

- Homework 2 is posted

Ridge regression

- Linear regression objective function

$$J(\theta) = \underbrace{\frac{1}{2}\sum_{i=1}^{N}[h_\theta(x_i) - y_i]^2}_{\text{model fit to data}} + \underbrace{\frac{\lambda}{2}\sum_{j=1}^{d}\theta_j^2}_{\text{regularization}}$$

  – λ is the regularization parameter ($\lambda \geq 0$)
  – No regularization on θ_0!

    - If λ = 0, we train linear regression
    - If λ is large, the coefficients will shrink close to 0

Lasso Regression

$$J(\theta) = \underbrace{\sum_{i=1}^{N} (h_\theta(x_i) - y_i)^2}_{\text{Squared Residuals}} + \underbrace{\lambda \sum_{j=1}^{d} |\theta_j|}_{\text{Regularization}}$$

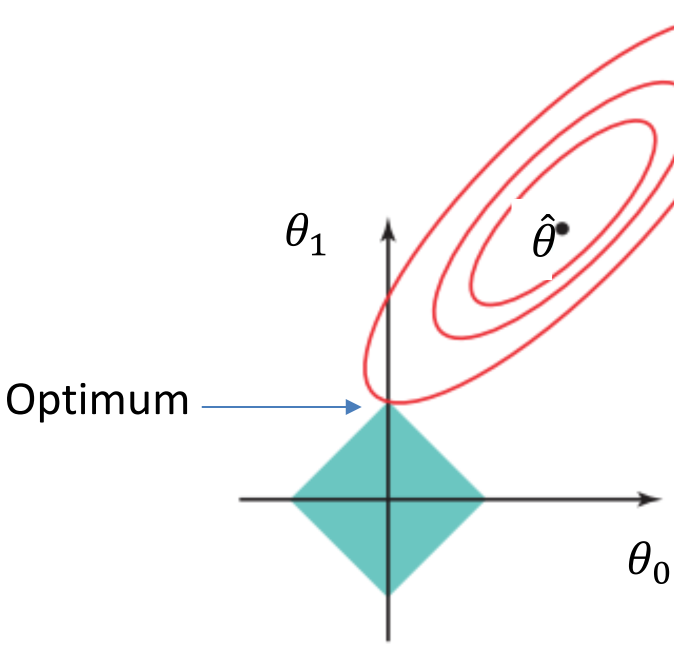
- L1 norm for regularization
- Results in sparse coefficients
- Issue: gradients cannot be computed around 0
- Method of sub-gradient optimization

Alternative Formulations

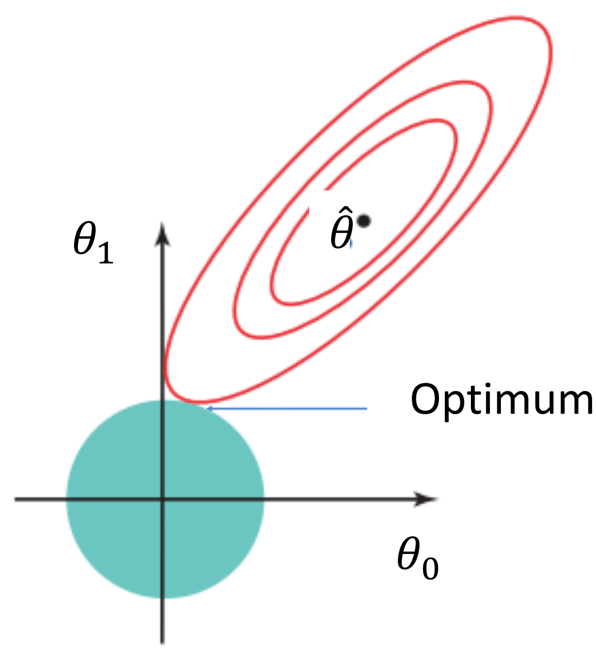
- Ridge
  - L2 Regularization
  - $\min_{\theta} \sum_{i=1}^{N} [h_\theta(x_i) - y_i]^2$ subject to $\sum_{j=1}^{d} |\theta_j|^2 \leq \epsilon$
- Lasso
  - L1 regularization
  - $\min_{\theta} \sum_{i=1}^{N} [h_\theta(x_i) - y_i]^2$ subject to $\sum_{j=1}^{d} |\theta_j| \leq \epsilon$

Lasso vs Ridge

- Ridge shrinks all coefficients
- Lasso sets some coefficients at 0 (sparse solution)
  - Perform feature selection



Lasso



Ridge

Ridge vs Lasso

- Both methods can be applied to any loss function (regression or classification)

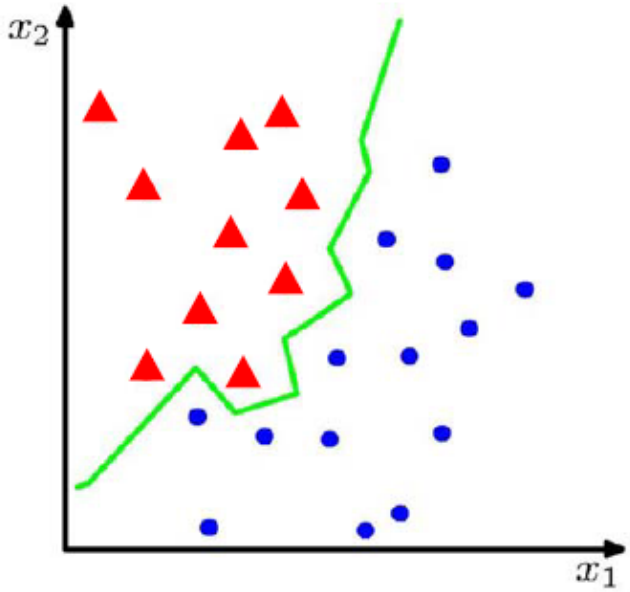
• Ridge	• Lasso

Ridge vs Lasso

- Both methods can be applied to any loss function (regression or classification)
- In both methods, value of regularization parameter λ needs to be adjusted
- Both reduce model complexity

Ridge	**Lasso**
+ Differentiable objective	- Gradient descent needs to be adapted
+ Gradient descent converges to global optimum	+ Results in sparse model
- Shrinks all coefficients	+ Can be used for feature selection in large dimensions

Classification



Binary or discrete

- Suppose we are given a training set of N observations

$$\{x_1, \dots, x_N\} \text{ and } \{y_1, \dots, y_N\}, x_i \in R^d, y_i \in \{0, 1\}$$

- Classification problem is to estimate f(x) from this data such that

$$f(x_i) = y_i$$

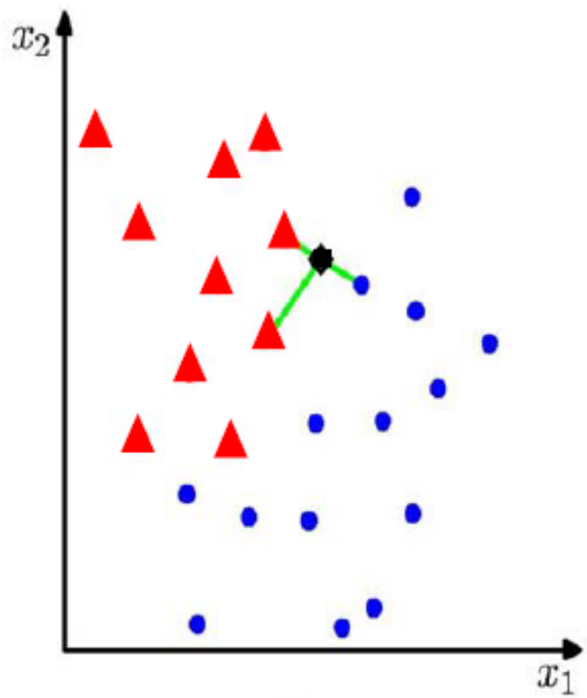
K Nearest Neighbour (K-NN) Classifier

Algorithm

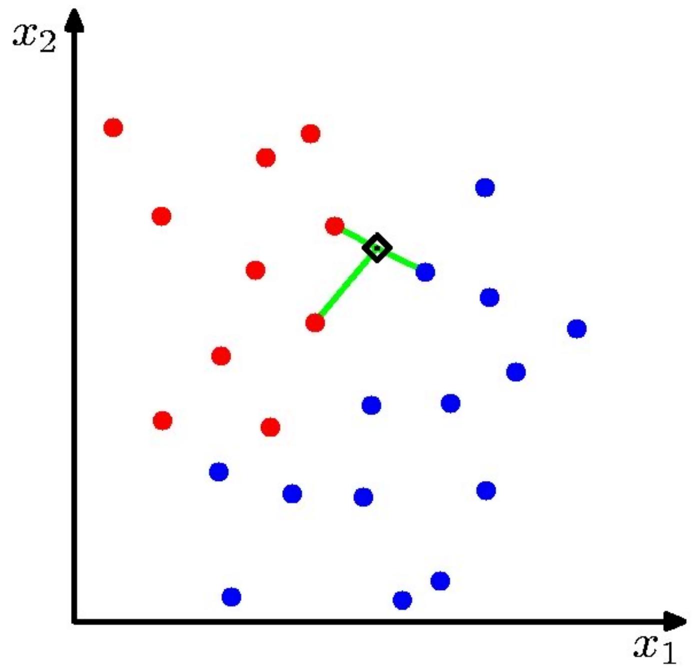
- For each test point, x, to be classified, find the K nearest samples in the training data
- Classify the point, x, according to the majority vote of their class labels

e.g. K = 3

• applicable to multi-class case

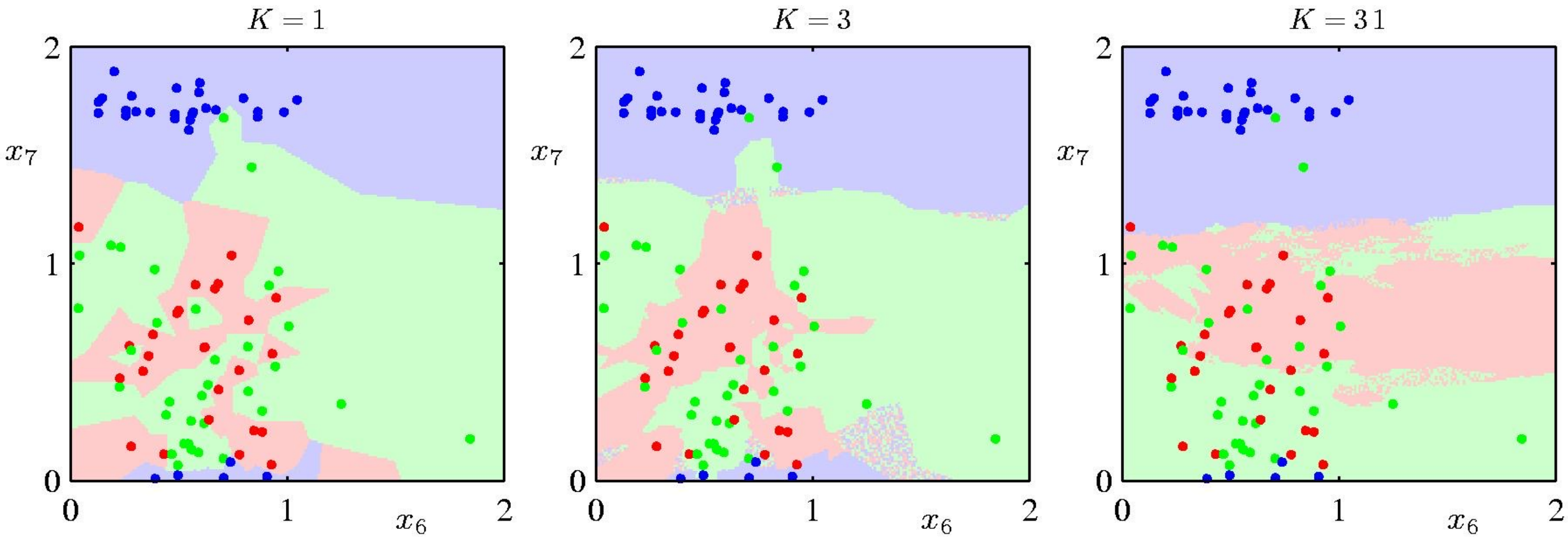


kNN



- Algorithm (to classify point x)
  - Find k nearest points to x (according to distance metric)
  - Perform majority voting to predict class of x
- Properties
  - Does not learn any model in training!
  - Instance learner (needs all data at testing time)

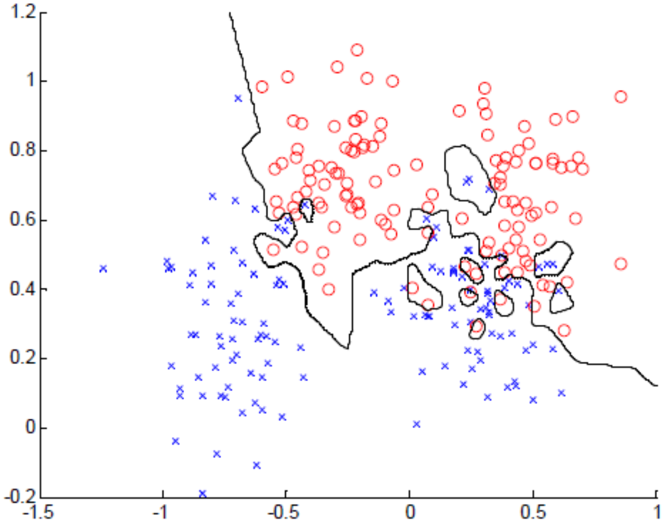
K-Nearest-Neighbours for Multi-class Classification



Vote among multiple classes

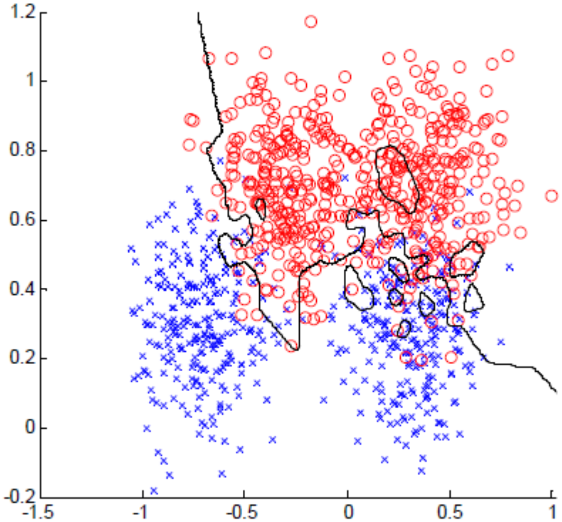
K = 1

Training data



error = 0.0

Testing data

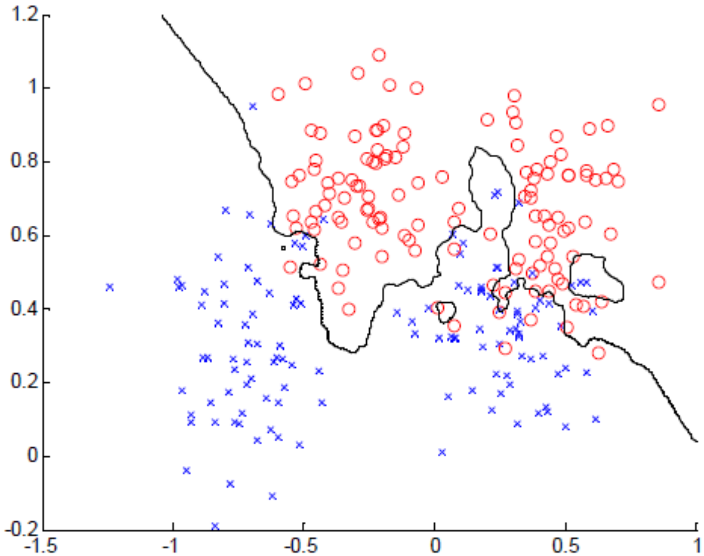


error = 0.15

How to choose k (hyper-parameter)?

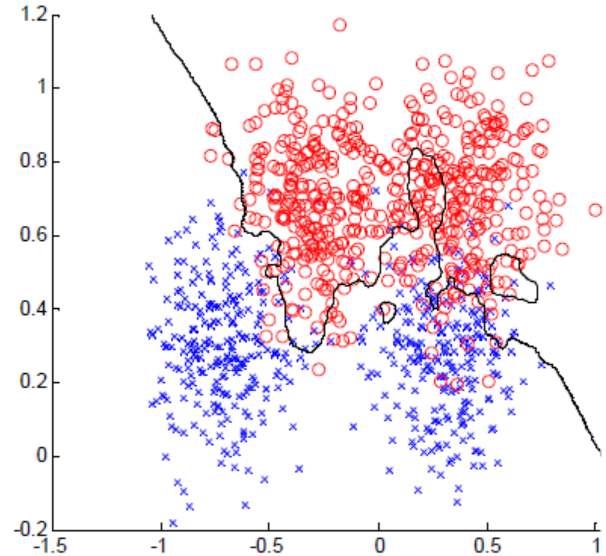
K = 3

Training data



error = 0.0760

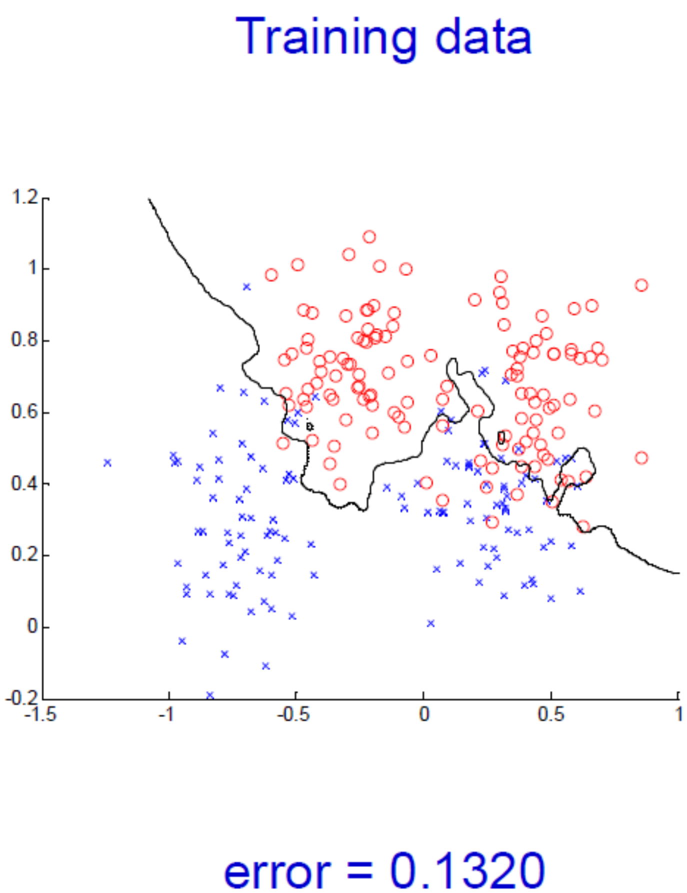
Testing data

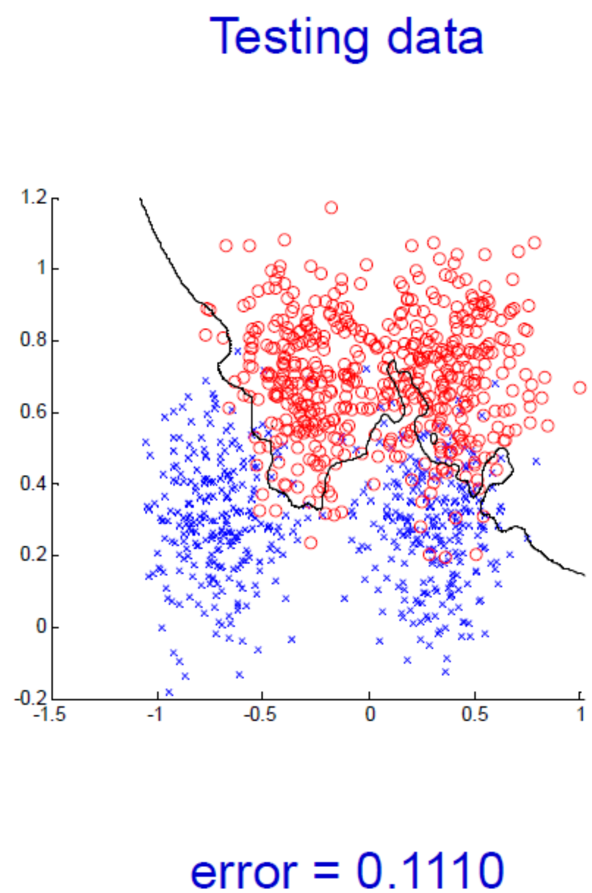


error = 0.1340

How to choose k (hyper-parameter)?

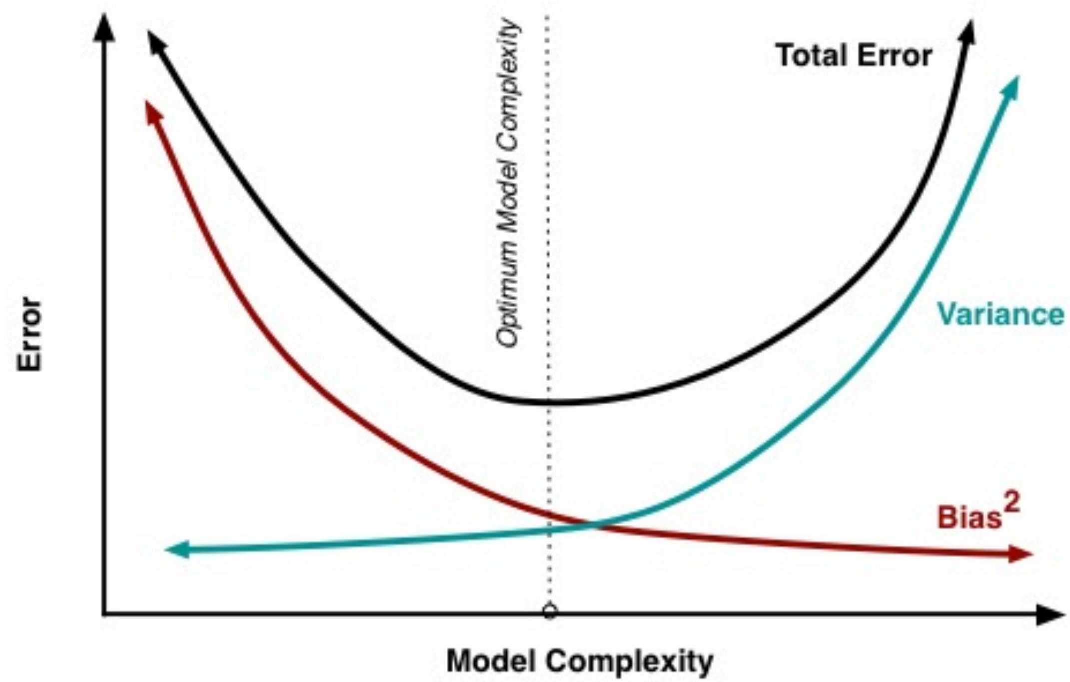
K = 7



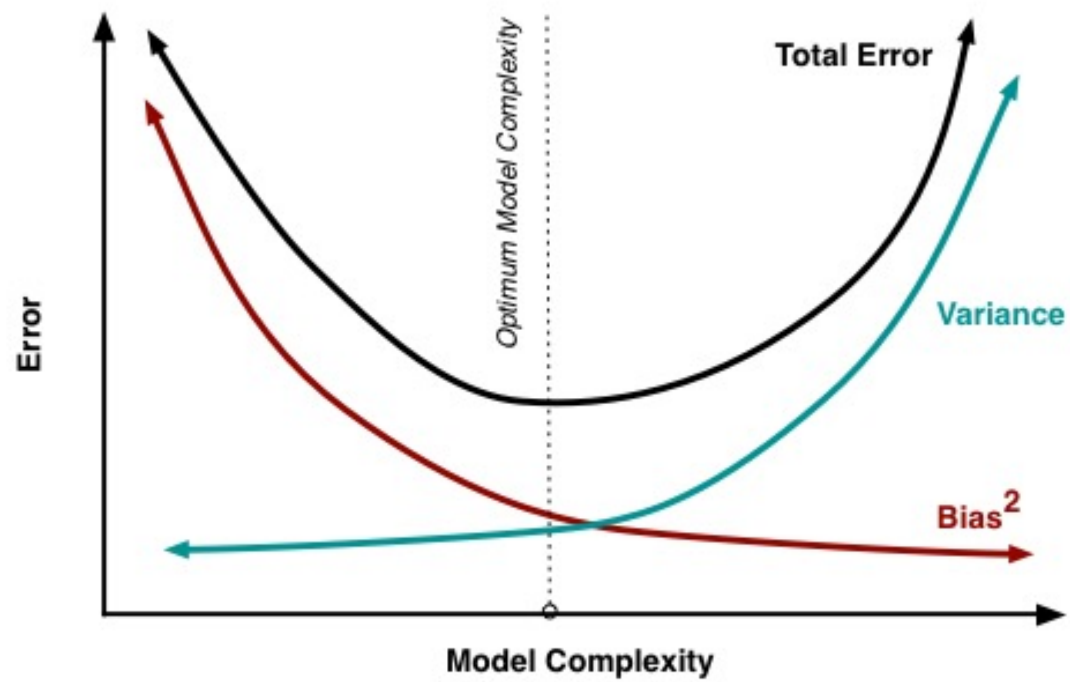


How to choose k (hyper-parameter)?

Bias-Variance Tradeoff for kNN

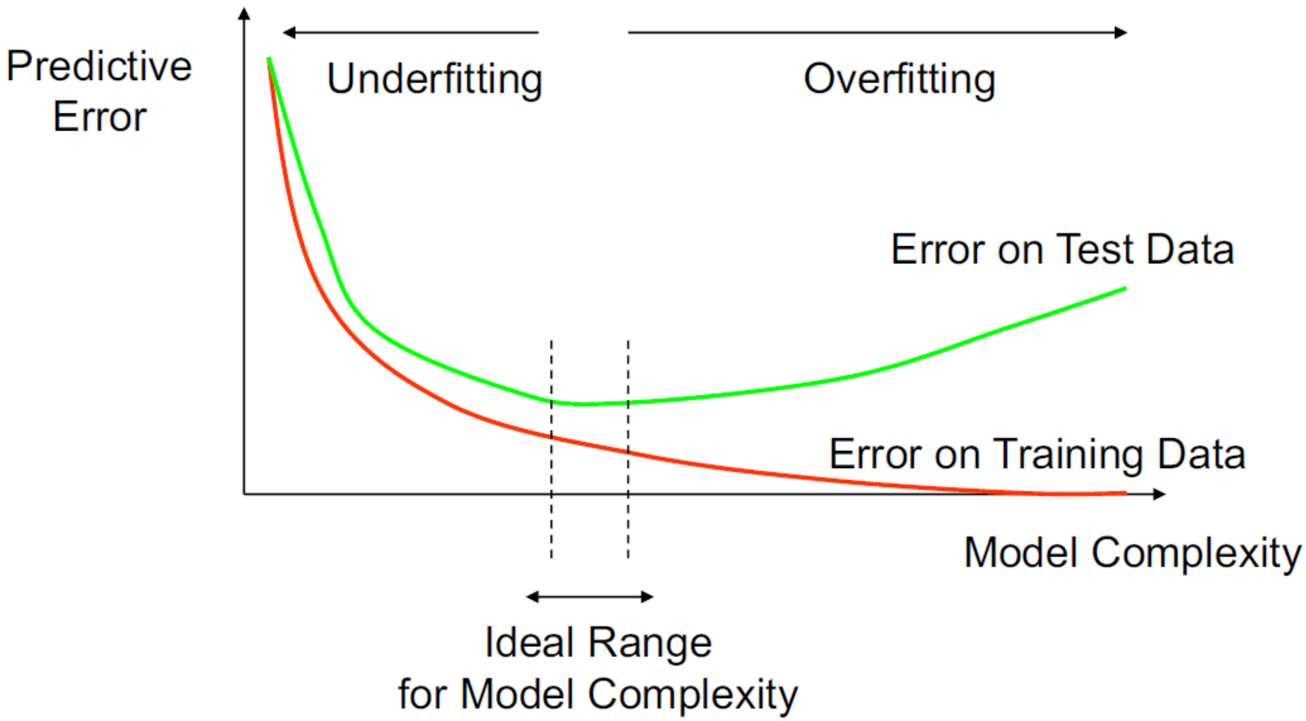


Bias-Variance Tradeoff for kNN



K decreases

How Overfitting Affects Prediction



- How to pick hyper-parameters without access to testing data?
- Goal: Reduce overfitting and variance

Important: Do not use testing data for hyper-parameter selection even if it is available

Cross Validation

As K increases:

- Classification boundary becomes smoother
- Training error can increase

Choose (learn) K by cross-validation

- Split training data into training and validation
- Hold out validation data and measure error on this

Cross Validation

As K increases:

- Classification boundary becomes smoother
- Training error can increase

Choose (learn) K by cross-validation

- Split training data into training and validation
- Hold out validation data and measure error on this

K=1

Train set 1 | Validation set 1 | Error

...

Train set 10 | Validation set 10 | Error

Avg Validation Error

K=7

Train set 1 | Validation set 1 | Error

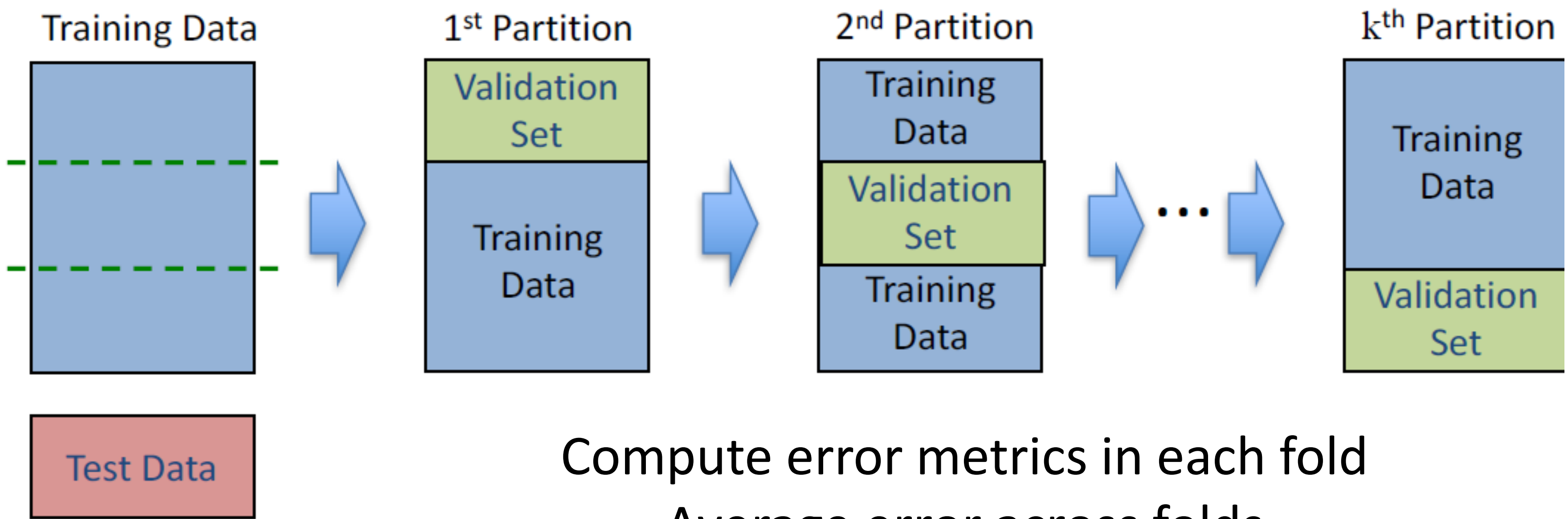
...

Train set 10 | Validation set 10 | Error

Avg Validation Error

Minimum error!

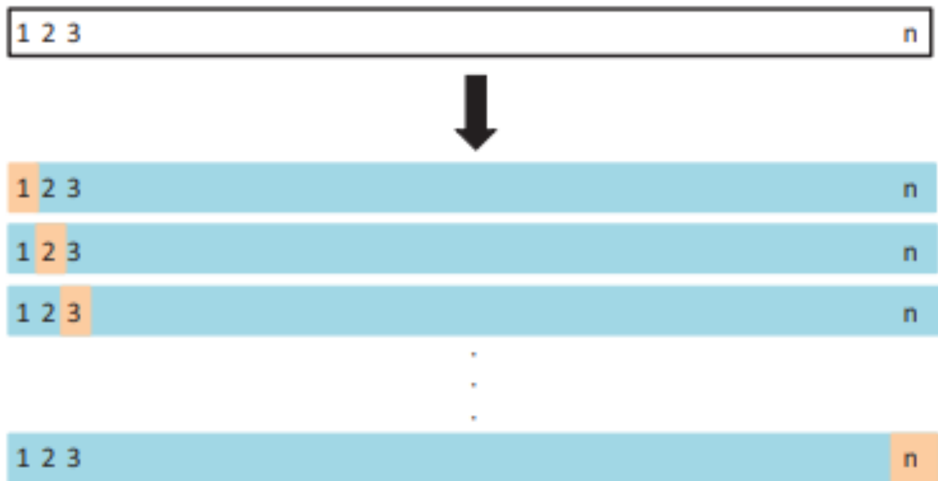
Cross Validation



Compute error metrics in each fold
Average error across folds

1. k-fold CV
   – Split training data into k partitions (folds) of equal size
   – Pick the optimal value of hyper-parameter according to error metric averaged over all folds

Cross Validation



2. Leave-one-out CV (LOOCV)
   – k=n (validation set only one point)

- Pros: Less bias
- Cons: More expensive to implement, higher variance
- Used for small training sets

Recommendation: perform k-fold CV with k=5 or k=10

Cross-Validation Takeaways

- General method to estimate performance of ML model at testing and select hyper-parameters
  - Improves model generalization
  - Avoids overfitting to training data
- Techniques for CV: k-fold CV and LOOCV
- Compare to regularization

Cross-Validation Takeaways

- General method to estimate performance of ML model at testing and select hyper-parameters
  – Improves model generalization
  – Avoids overfitting to training data
- Techniques for CV: k-fold CV and LOOCV
- Compare to regularization
  – Regularization works when training with GD
  – Cross-validation can be used for hyper-parameter selection
  – The two methods can be combined