Underfitting and Overfitting Review

Lesson 5 – Section 1

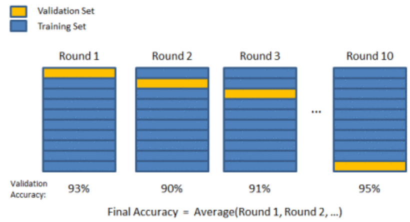
Overview

Review of underfitting and overfitting

Review of how to avoid underfitting and overfitting

Common Pitfalls in Machine Learning

- Overfitting
  - Split the data into training and validation, and only care about the performance on validation
  - Cross validation.

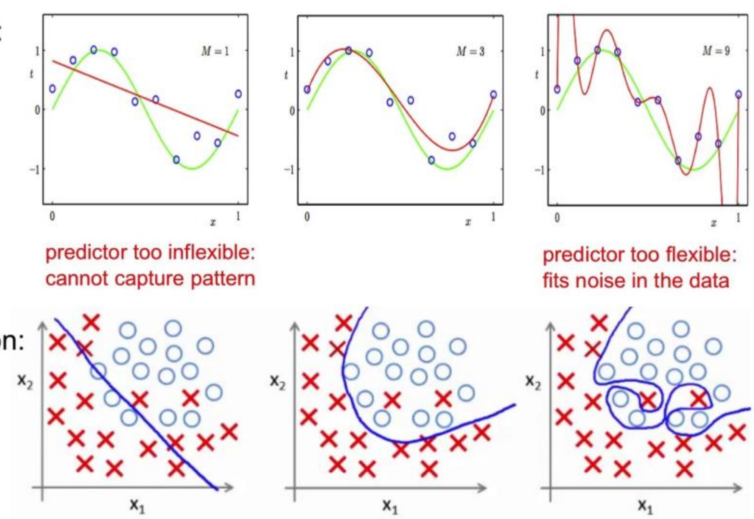


- Target leaking:
  - Predicting readmission. You have one binary variable "readmission", which is your target column. You also have columns "readmission time", "readmission location", "readmission reason".
- Model has good performance on validation, but not applicable

PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY of WASHINGTON

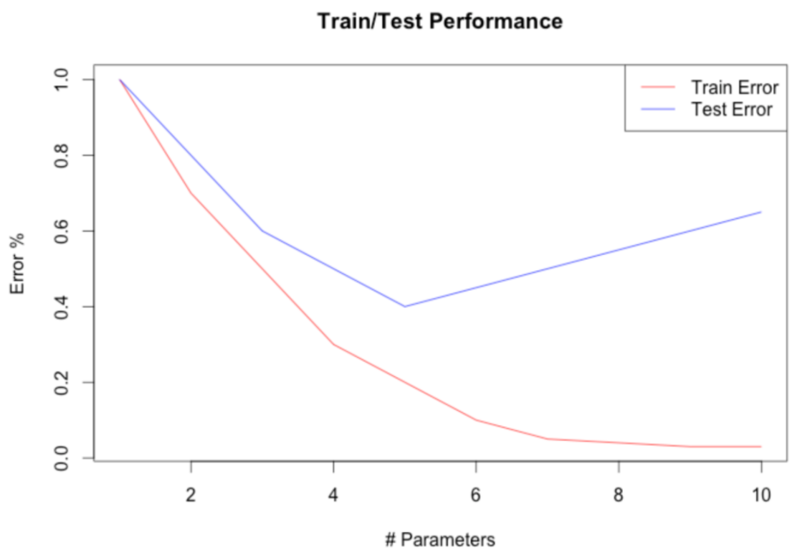
Underfitting and Overfitting Examples

Regression:



Classification:

Indicators of Underfitting and Overfitting



- Model performs poorly on both training and testing data
  - Underfitting, or
  - Not relevant data
- Model performs well on training, but poorly on testing
  - Overfitting

PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY of WASHINGTON

Reducing Underfitting

- Increase model complexity, for e.g.
  - –Increase the number of levels in a decision tree
  - –Increase the number of hidden layers in a neural network.
  - –Decrease the number of neighbors (k) in k-NN
- Increase the number of features, or create more relevant features

– In iterative training algorithms, iterate long enough so that the objective function has converged.

PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY of WASHINGTON

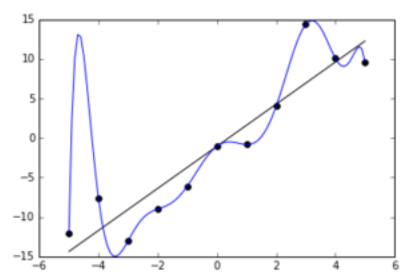
Reducing Overfitting

- Decrease model complexity, for e.g.
  - Prune a decision tree
  - Reduce the number of hidden layers in a neural network.
  - Increase the number of neighbors (k) in k-NN
- Decrease the number of features
  - More aggressive feature selection
- Regularization (control feature complexity)
  - Penalize high weights.
  - L-1 regularization (LASSO) very efficient at pushing weights of non-informative features to 0.
- Gather more training data if possible
- In iterative training algorithms, stop training earlier to prevent "memorization" of training data

Regularization: A Popular Way of Controlling Overfitting

- Loss Function of Training
  - You can almost always increase the complexity of f_θ to reduce SSE
  - Increase the risk of overfitting
- Add regularization to control overfitting
  - L1 (LASSO) or L2 (Ridge regression) regularization

$$SSE = \sum_{i=1}^{n}(y_i - f_\theta(x_i))^2$$



$$LOSS = \sum_{i=1}^{n}(y_i - f_\theta(x_i))^2 + \lambda_1 \sum_{k=1}^{m}|\theta_k| + \lambda_2 \sum_{k=1}^{m}\theta_k^2$$

$$\lambda_1, \lambda_2 \geq 0$$

$$\lambda_1 = 0, \lambda_2 > 0 : \text{Ridge regression}$$

$$\lambda_2 = 0, \lambda_1 > 0 : \text{LASSO}$$

$$\lambda_1, \lambda_2 > 0 : \text{Elastic net}$$

What to remember about classifiers

- Try simple classifiers first
- Better to have smart features and simple classifiers than simple features and smart classifiers
- Use increasingly powerful classifiers with more training data

PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY of WASHINGTON

Summary

>Reviewed what are underfitting and overfitting

>Reviewed how to avoid underfitting and overfitting
- –Adjust model complexity
- –Control the number of features/feature complexity