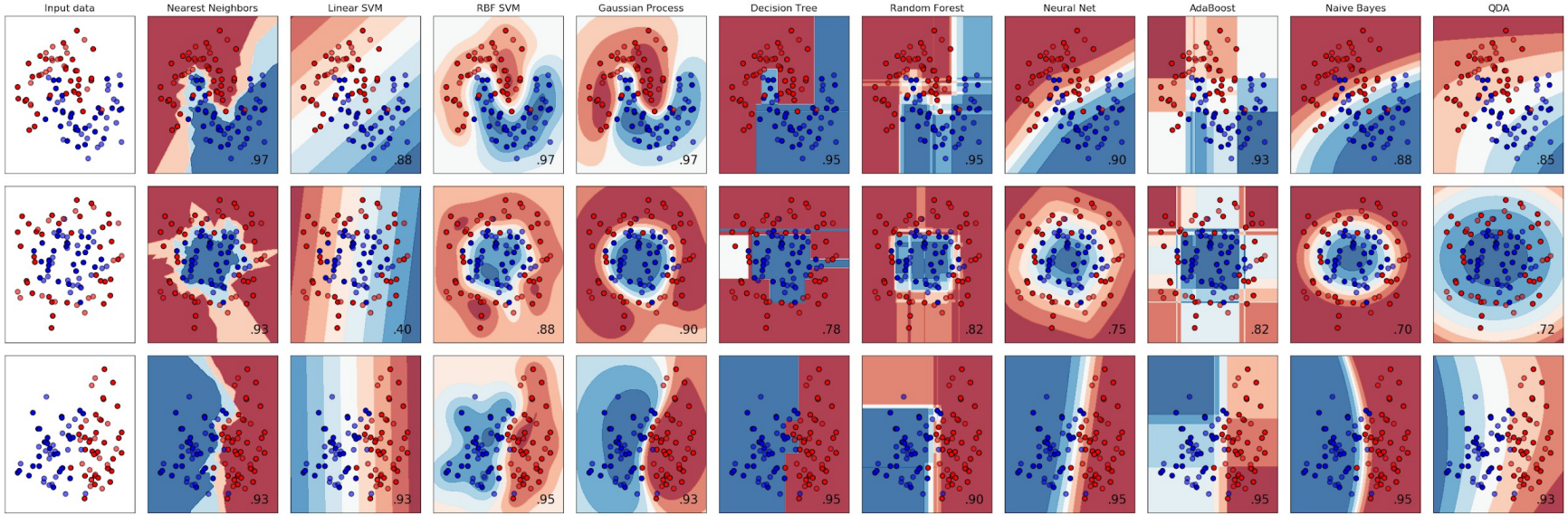
Model Selection in Machine Learning

Brendan Duke

Contents

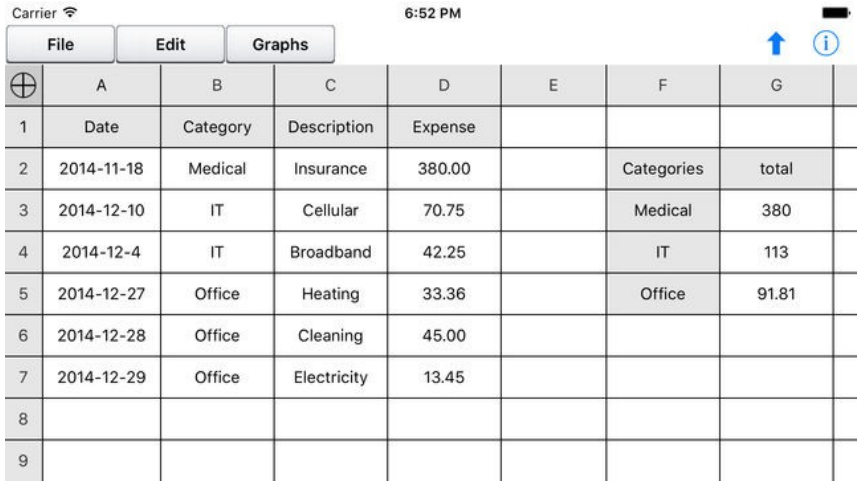
Selecting a Model

- There are many different options:
  - Gradient Boosting
  - Random Forest
  - Linear Regression
  - SVM
  - Naive Bayes
  - Gaussian Process
  - Nearest Neighbour
  - LDA/QDA
  - Neural Networks



Model Selection - Rule of Thumb

- Does the dataset contain tabular data (i.e. spreadsheet)?
  - Use gradient boosting or random forest
  - Works well for >90% of problems
  - XGBoost: the first model to try, and ultimately winning model, in Kaggle competitions
    - Feature engineering help: feature importance (https://bit.ly/2mdjOSd)



	A	B	C	D	E	F	G
1	Date	Category	Description	Expense			
2	2014-11-18	Medical	Insurance	380.00		Categories	total
3	2014-12-10	IT	Cellular	70.75		Medical	380
4	2014-12-4	IT	Broadband	42.25		IT	113
5	2014-12-27	Office	Heating	33.36		Office	91.81
6	2014-12-28	Office	Cleaning	45.00			
7	2014-12-29	Office	Electricity	13.45			
8							
9							

Gradient Boosting (Explanation)

- Decision trees: classifier with trivial splitting (e.g., $x < 2.4$)
- Build a regression tree greedily, by splitting to minimize Mean Squared Error (MSE)
  - Fast algorithm, suboptimal results
- Gradient boosting: ensemble of decision trees. (N + 1)th tree minimizes the N-tree ensemble's residual error (or in general, negative gradient of loss).
- Explanation: https://bit.ly/2Jf9bao

Gradient Boosting (Interactive Playground)

- Underfitting vs. overfitting
- High training and test error, what can you do?
  - More powerful model/algorithm
- Low training, but high test error, what can you do?
  - Regularization
  - More data/data augmentation
  - Simpler model
- Playground: https://bit.ly/2JgaGoP

Model Selection - Rule of Thumb

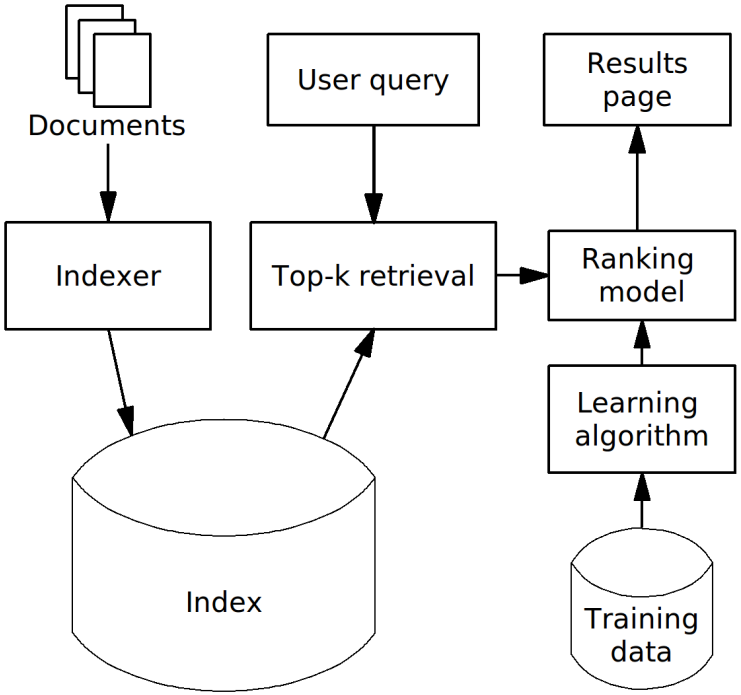
- Does your dataset contain images, audio, time series, or text data?
  - Use neural networks, if the dataset is sufficiently large
  - Works well for problems that humans are generally good at

Model Selection – Flow Charts

- Sci-kit Learn: https://bit.ly/1IxDsim
- Dlib: http://dlib.net/ml_guide.svg

Learning to Rank

- We want to use machine learning to improve a search engine
- Two parts: top-k retrieval, and ranking of k retrieved documents
- Let’s focus on the ranking model
- To the Jupyter notebook!
- https://github.com/dukebw/ml-model-selection



Realtime Vehicle Detection with Dlib

- http://blog.dlib.net/2017/08/vehicle-detection-with-dlib-195_27.html