Thinkful: Unit 3: Lesson 5: Project 6 Challenge: what model can answer this question?

Eva Rubin

December 2018

Estimated Time: 2-3 hours

You now have a fairly substantial starting toolbox of supervised learning methods that you can use to tackle a host of exciting problems. To make sure all of these ideas are organized in your mind, please go through the list of problems below. For each, identify which supervised learning method(s) would be best for addressing that particular problem. Explain your reasoning and discuss your answers with your mentor.

https://dzone.com/articles/decision-trees-vs-clustering-algorithms-vs-linear

1. Predict the running times of prospective Olympic sprinters using data from the last 20 Olympics.
    - Linear regression. Continuous variables

2. You have more features (columns) than rows in your dataset.
    - https://stats.stackexchange.com/questions/223486/modelling-with-more-variables-than-data-points
    - LASSO (l1) or Ridge Regression (l2) to impose penalties on the norm of the weights (l1 or l2) in order to reduce features by making coefficients so small they are irrelevant (Lasso) or by reducing multicollinearity (minimizing variance) by penalizing very large coefficients

3. Identify the most important characteristic predicting likelihood of being jailed before age 20.
    - Logistic Regression because we're interested in the probability of characteristics leading to a categorical outcome

4. Implement a filter to “highlight” emails that might be important to the recipient
    - Naive Bayes because it explains the relationship of an outcome to a vector of conditions rather than to a single other event - spam is an outcome as it relates to a combination of conditions (words, phrases, etc) that indicate spam

5. You have 1000+ features.
    - https://www.researchgate.net/post/Are_these_feature_selection_methods_the_state_of_the_art/3
    - PCA to reduce dimensions
    - Kselectbest to reduce features
    - Feature selection using SVM https://www.sciencedirect.com/science/article/pii/S0957417411011626?via%3Dihub

6. Predict whether someone who adds items to their cart on a website will purchase the items.
    - Decision Tree because you’re looking at binary outcome given a set of conditions

7. Your dataset dimensions are 982400 rows x 500 columns
    - https://www.analyticsvidhya.com/blog/2018/06/comprehensive-guide-for-ensemble-models/
    - https://www.researchgate.net/post/Classification_regression_with_very_large_dataset-any_thoughts
    - An ensemble method, like Random Forest or Gradient Boost Model to break down the training set hierarchically into groups, and thereby reduce the complexity.

8. Identify faces in an image.

- https://towardsdatascience.com/face-recognition-for-beginners-a7a9bd5eb5c2
- SVM [Support Vector Machines]
- PCA [Principal Component Analysis]
- LDA [Linear Discriminant Analysis]
- Kernel methods or Trace Transforms

9. Predict which of three flavors of ice cream will be most popular with boys vs girls.
   - KNN because the classification is very simple - two genders and three flavors of ice cream. There are few dimensions and they are linearly separable (otherwise SVM would be better). But they can also take a lot of time in evaluating the distance between the neighbors.
   - Random Forests can classify and handle categorical features well - also, much faster than KNN.
   - https://discuss.analyticsvidhya.com/t/which-one-to-use-randomforest-vs-svm-vs-knn/2897/3