Classification Model

1. Interpretable Models:
2. Decision Tree- how the decision boundary is determined is not interpretable
3. Logistic Regression- decision boundary is clear
4. Linear SVC -decision boundary is interpretable, this comparable to KNN, all it’s doing and see what point is closest to the line
5. SVC: low interpretability, radian basis function is in terms of this kernel that embodies the distance between all the points and maps it in this high dimensional space
6. KNN – most interpretable, a human can explain what is going on thinking of the iris data set when it groups by color,size,color
7. Random forest – looking at random features subsets, it’s not as deterministic as one decision tree
8. Naïve Bayes:
9. Multiclass:

KNN-yes

Logistic – yes (one vs rest)

Linear SVC- yes

Decision Tree –yes

Naïve Bayes-

1. Scalable with large number of features:
   1. Logistic - yes
   2. Linear SVC-yes
   3. SVC – ‘kernal trick’ harder to separate, no
   4. Decision tree and random forest – yes : ignores features with low gini
   5. Naïve bayes: YES!,ex: how many words in English language, since of the assumption that the events are independent it makes the math easy
2. Scalable with large number of data
   1. Logistic – yes
   2. Linear svc and svc – it scales at order n, it will take twice as long if you have 2n data, we don’t like it when it scales at n\*\*2,
   3. KNN : Euclidean distance
   4. DT and RF: scales pretty well, you are subsetting the data and decomposes the problem – greedy heuristic problem solver – whats the best question now only thinks one step ahead
   5. NB: yes
3. Good with diverse set of features
   1. KNN : no
   2. Logistic regression: yes, gives coefficients
   3. SVC: not really similar problem to KNN
   4. DT and RF: yes!
   5. NB: it can ask multiple questions , building conditional submodels