CPSC340A1

Qinglan Huang n6v9a Liu Yang v4d9

September 30, 2017

1 Summary Statistics and Data Visualization

1.1 Summary Statistics

minimum value of the dataset is 0.352
maximum value of the dataset is 4.862
mean of the dataset is 1.3246249999999997
median of the dataset is 1.1589999999999998
mode of the daataset is 0.77

mode is not a reliable estimate to measure most common value for a continuous data. Instead, finding a most common interval, like [0.7,0.8], is more reliable.

1.2 Data Visualization

1. A histogram showing the distribution of each the values in the matrix X.
Corresponds to plots D
Reason: 1. D is a histogram 2. D displays more value then C does

2. A boxplot grouping data by weeks, showing the distribution across regions for each week.
Corresponds to plots B
Reason: 1.B is the only boxplot. 2.A can only show difference between regions but cannot show distribution of each regions

3. A scatterplot between the two regions with highest correlation.
Corresponds to plots F
Reason: 1. F is a scatterplot 2. F has higher correlation compared to E

4. A single histogram showing the distribution of each column in X.
Corresponds to plots C
Reason: 1. C is a histogram 2. C displays fewer columns

5. A scatterplot between the two regions with lowest correlation.

Corresponds to plots E
Reason: 1. E is a scatterplot 2. E has lower correlation compared to F

6. A plot containing the weeks on the x-axis and the percentages for each region on the y-axis.
Corresponds to plots A
Reason: 1.A contains weeks on the x-axis 2.A shows percentages for each region

1.3 Decision Surfaces

There are 16 mis-classified training examples.
Note: there is one training example locating on the boundary,which means it can either be classified rightly or wrongly. We did not include it as a mis-clasified.

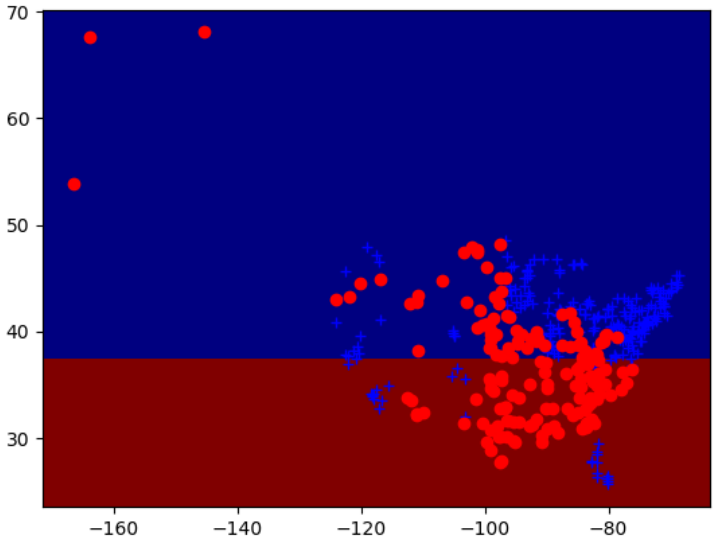
2 Decision Trees

2.1 Equality vs. Inequality Splitting Rules

Yes. Equality-based splitting rule is reasonable if applied to a categorical feature.

2.2 Decision Stump Implementation

Error with inequality-rule decision stump: 0.25



2.3 Constructing Decision Trees

2.4 Cost of Fitting Decision Trees

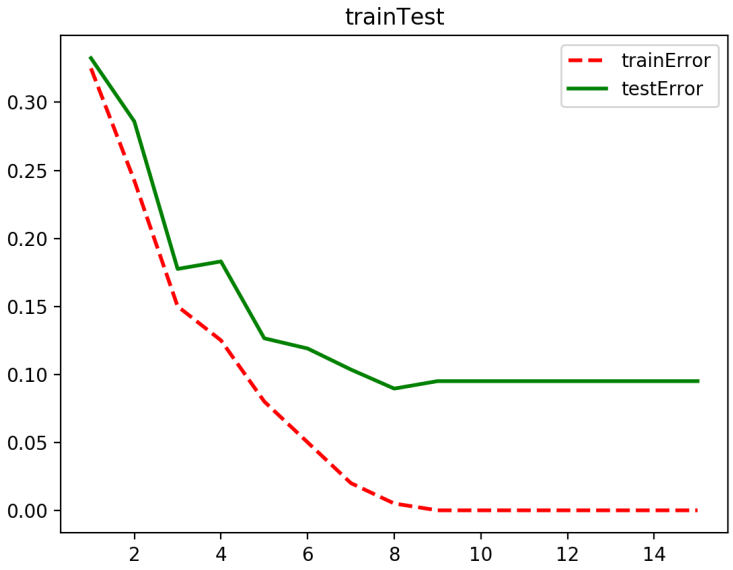
$\mathcal{O}(mdn \log n)$

3 Training and Testing

3.1 Training Error

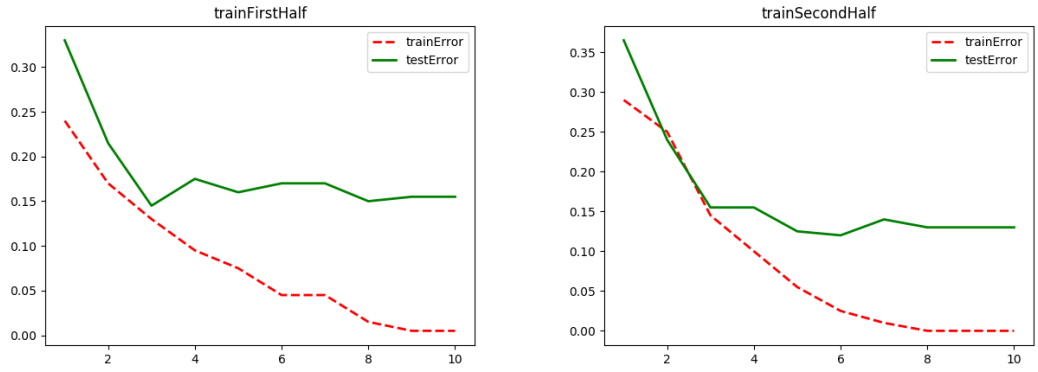
1. About training error
With smaller degree of depth, accuracy-based decision tree have less training error compared to info-gain tree. As depth increases, the training error of info-gain tree decreases to 0, however that of accuracy-based decision tree remains unchanged at 0.11
2. About training time
When depth is small, accuracy-based decision trains a little bit faster then info-gain tree. However, when I increases the depth to 30, the info-gain tree trains remarkably faster then accuracy-based decision tree.

3.2 Training and Testing Error Curves



The training error decreases monotonically along the increase of depth. The test error decreases when depth less then 9. At depth 9, test error increases a little then turns to be flat as depth increasing.

3.3 Validation Set



If we use the first half samples to train and second half samples to validate, the depth which minimizes validation error is 3.
If we use the first half samples to validate and second half samples to train, the depth which minimizes validation error is 6 .
We could use cross-validation.
I understand that testSet is different from validationSet. But I mixed them in the figures. It is just a typo mistake. Every testError in figure means validation error.

4 Naive Bayes

4.1 Naive Bayes by Hand

- $p(y = 1) = 0.6$
- $p(y = 0) = 0.4$
- $p(x_1 = 1|y = 1) = 0.5$
- $p(x_2 = 0|y = 1) = \frac{1}{3}$
- $p(x_1 = 1|y = 0) = 1$
- $p(x_2 = 0|y = 1) = \frac{3}{4}$

Compute the estimates of the 4 conditional probabilities required by naive Bayes for this example.
$p(\hat{y} = 0|\hat{x}) \propto p(\hat{x_1} = 1|\hat{y} = 0) * p(\hat{x_2} = 0|\hat{y} = 0) = 0.3$
$p(\hat{y} = 1|\hat{x}) \propto p(\hat{x_1} = 1|\hat{y} = 1) * p(\hat{x_2} = 0|\hat{y} = 1) = 0.1$

4.2 Bag of Words

1. league
2. car,engine,evidence,problem,system
3. 2

4.3 Naive Bayes Implementation

```
# We will store p(x(i,j) = 1 | y(i) = c) in p_xy(1,j,c)
# We will store p(x(i,j) = 0 | y(i) = c) in p_xy(2,j,c)
p_xy = zeros(2,d,k)

for i in 1: k
  A = y[:,1] .== i

  for j in 1: d
    p_xy[1,j,i] = sum(X[A,j]) / (n * p_y[i])

    p_xy[2,j,i] = sum(X[A,j] .== 0) / (n * p_y[i])
  end
end
```

The updated test error is 0.188.

4.4 Runtime of Naive Bayes for Discrete Data

The cost of classifying t test examples with the naive bayes modes is $\mathcal{O}(tdk)$

5 K-Nearest Neighbours

5.1 KNN Prediction

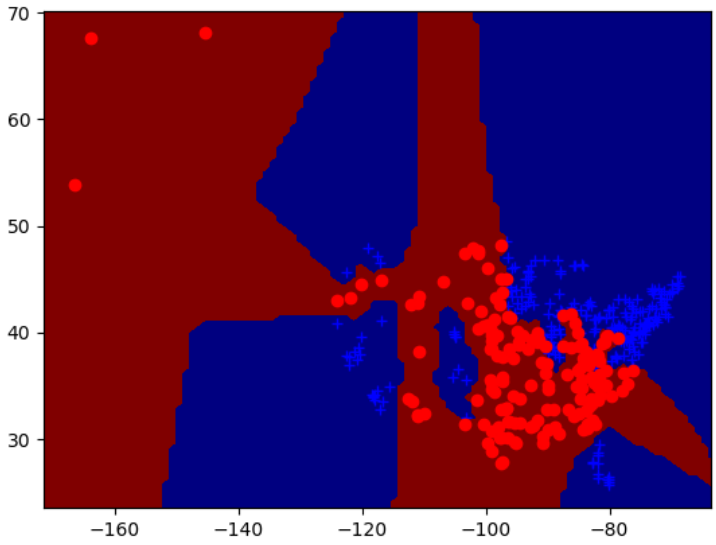
1. Hand in the predict function

```
yhat = []
for i in 1:t
        nearbyPoints = []
        for j in 1:n
                temp = 0

                for m in 1:d
                        temp += (Xhat[i,m] - X[j,m]).^2
                end
                push!(nearbyPoints, sqrt(temp))
        end
        A = sortperm(nearbyPoints)
```

```
            push!(yhat  ,mode(y[A[1:k]]))
        end

    return yhat
end
```

2. Report the training and test error obtained on the citiesSmall.mat dataset for k = 1, k = 3, and k = 10. (You can use example knn.jl to get started.)
When $k = 1$, the trainError is 0 and the testEroor is 0.065.
When $k = 3$, the trainError is 0.028 and the testEroor is 0.066.
When $k = 10$, the trainError is 0.072 and the testEroor is 0.097.

3. Hand in the plot generatied by classifier2Dplot on the citiesSmall.mat dataset for k = 1 on the training data. (Note that this version of the function also plots the test data.)



4. Why is the training error 0 for k = 1?
Because every example is 1-nearest neighborhood of itself. KNN would predict $\hat{y}$ based on its y value only. Therefor $\hat{y} = y$ for all rows which means training error is 0. (If we have two exactly same rows but with different y values, we may have a different training error).

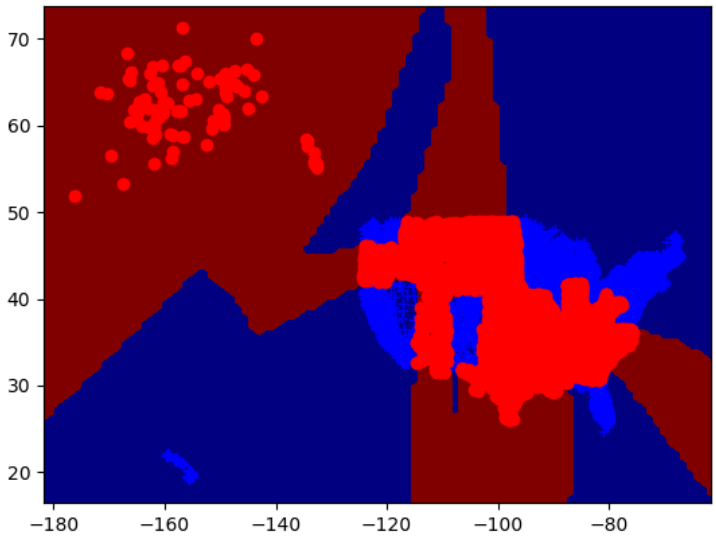
5. If you didn't have an explicit test set, how would you choose k?
Use cross validation, to choose k with lowest validation error.

5.2 Condensed Nearest Neighbours

1. the running time of CKNN is 19.94. The running time of KNN is much more longer

2. training error 0.008, test error 0.018, variables included 457

3. Hand in the plot generatied by classifier2Dplot on the citiesSmall.mat dataset for k = 1 on the training data.



4. Because we no longer include every training objects in our Xcond(new train set). The nearest one neighbour is no longer itself if it is not included in Xcond. Therefor, the yvalue may be different.

5. If you have s examples in the subset, the cost of running the predict function on t test examples is $\mathcal{O}(std)$.

6. The test error is high because both training set and test set are ordered by state, conflicting IID assumption.
   The train error is also high. For example, if first 200 objects are all labeled y=1 and remaining 100 objects are labeled y=2, then only the first object would be added into Xcond. However, when we concentrate on the remaining 100 objects, more than one would be added into Xcond. And when we calculate train error, the first 200 objects can be hardly classified correctly(since we only have 1 y=1 object in Xcond).

6 Very-Short Answer Questions

1. What is one reason we would want to look at scatterplots of the data before doing supervised learning?
   We can have an idea of how data distributed so that we can choose appropriate algorithm.

2. What is a reason that the examples in a training and test set might not be IID?

3. What is the difference between a validation set and a test set?
   Validation set is a part of train set. It can be used to choose from different model and be used to approximate the text error.
   Test set can not be used to train and is to check the behaviour of a given model.

4. Why is naive Bayes called "naive"?
   Because we assume every feature is independent conditioned on y = i.

5. What is a situation where the naive Bayes assumption could lead to poor performance?
   When a dataset has too many attributes(features)

6. What is the main advantage of non-parametric models?
   Non-parametric models will become more accurate if more data is given.

7. A standard pre-processing step is "standardization" of the features: for each column of X we subtract its mean and divide by its variance. Would this pre-processing change the accuracy of a decision tree classifier? Would it change the accuracy of a KNN classifier?
   Standardization does not change accuracy of a decision tree in any circumstance. However it will improve KNN's accuracy if its dataset has numerical attribute.

8. Does increasing k in KNN affect the training or prediction asymptotic runtimes?
   Nope. The prediction runtime is $\mathcal{O}(tnd)$ which is irrelevant to k.Also KNN is a lazy model and has no training phase. Therefore K does not change the training time of kNN either.

9. How does increase the parameter k in k-nearest neighbours affect the two parts (training error and approximation error) of the fundamental trade-off (hint: think of the extreme values).
   As k increases, the train error and test error decrease and then increase together(except when k =1, train eror is 0).

10. For any parametric model, how does increasing number of training examples n affect the two parts of the fundamental trade-off.
    Training error will decrease monotonically to 0 as the data size increases.

However, the test error will initially decrease and increase as the data size increases.