CPSC 340 Assignment 1 (due 2017-01-22 at 11:59pm)

Data Exploration, Decision Trees, Training and Testing, Naive Bayes

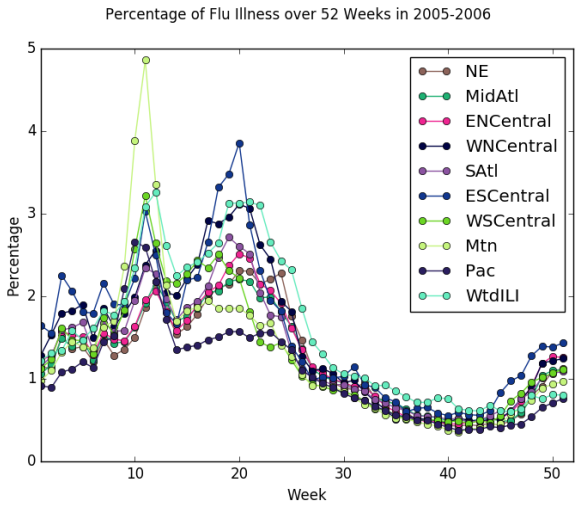
1 Data Exploration

1.1 Summary Statistics

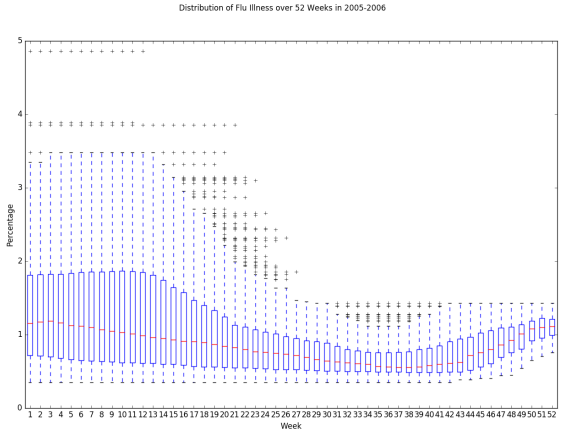
1. maximum = 4.862, minimum = 0.352, mean = 1.324625, median = 1.159, mode = 0.77
2. 10% quantile = 0.5019, 25% quantile = 0.718, 50% quantile = 1.159, 75% quantile = 1.81325, 90% quantile = 2.3154
3. highest mean = WtdILI, lowest mean = Pac, highest variance = Mtn, lowest variance = Pac
4. highest correlation between MidAtl and ENCentral regions, lowest correlation between Mtn and NE regions
5. The mode is not a reliable estimate for the most common value as this is continuous data. It is fairly unlikely that there are duplicate values exact to the n^{th} decimal place. The mode would only be sufficient for categorical data. A more meaningful measurement would be the quantile values. The IQR would be a good statistic to show where the middle 50% of values lie.

1.2 Data Visualization

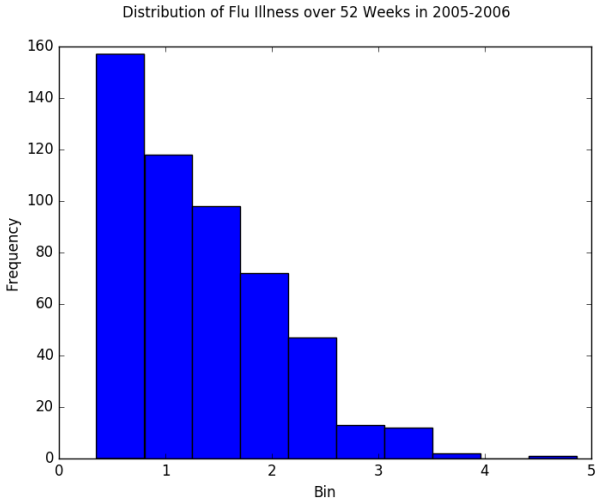
1. Weeks vs. Percentage



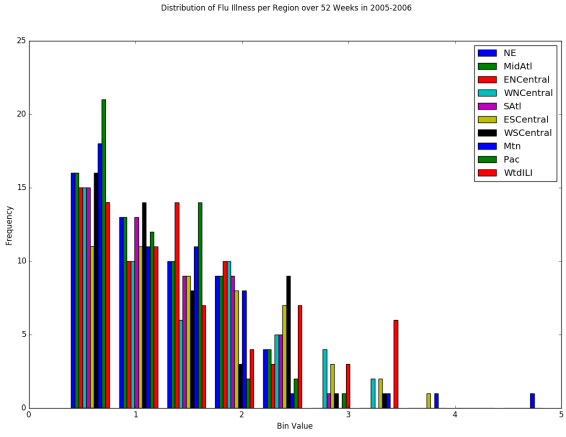
2. Boxplot by week



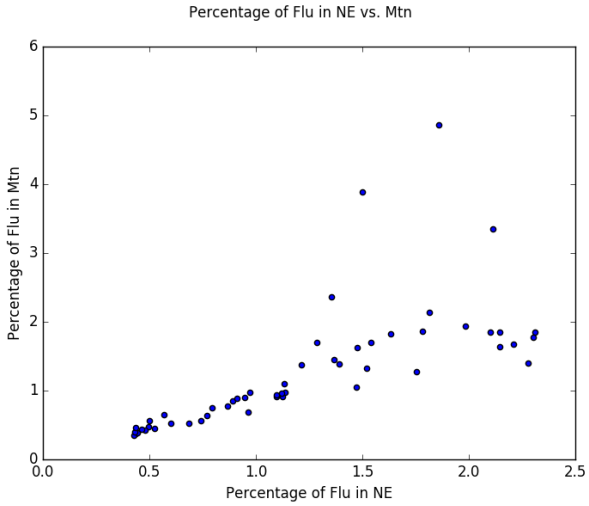
3. Histogram of all values



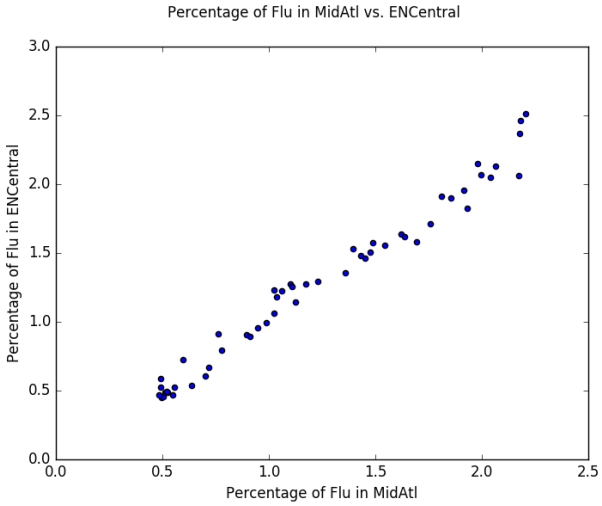
4. Histogram of each column



5. Scatterplot lowest correlation



6. Scatterplot highest correlation



2 Decision Trees

2.1 Decision Stump Implementation

`https://github.ubc.ca/cpsc340/jeanlam_nafis1_hw1/tree/master/code/decision_stump.py`
The updated error is 0.253

2.2 Constructing Decision Trees

See `https://github.ubc.ca/cpsc340/jeanlam_nafis1_hw1/tree/master/code/simple_decision.py`
The error in the scikit-learn version eventually decreases to 0 due to overfitting. Classification accuracy, however, takes the number of correct predictions divided by the total number of predictions made. However, our classification accuracy does not reach 1, or reversed, our training error does not reach 0. The way that

scikit predicts is with splitting until the entropy is equal to 0. This means that all values are definitely YES or definitely NO, resulting in a 0 training error. With our version, at a certain point, any computed errors with the threshold are not any lower than the current classification error where $y! = y_m ode$. There is no split found that improves the classification error at depth m, without increasing the overall error of the entire tree.

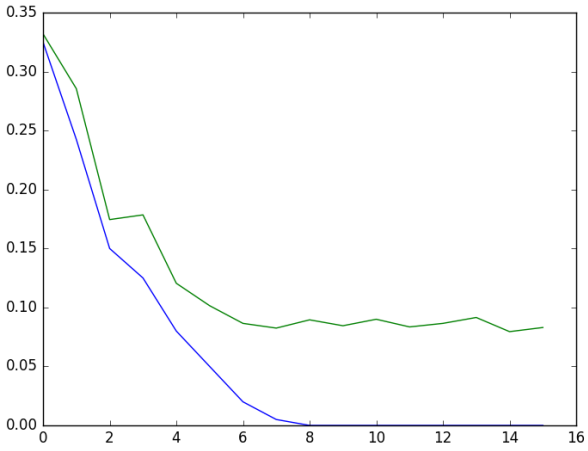
2.3 Cost of Fitting Decision Trees

2.3 The cost at each depth is O(dnlogn) as sorting costs O(nlogn) per feature.We look at the examples at each depth.To go through a depth of m we have total cost of O(mndlogn).

3 Training and Testing

3.1 Training and Testing Error Curves

Training error vs. depth of tree



Test error in green and training error in blue both goes down. The training error goes down to 0 but test error does not keep going down after depth 10.

3.2 Validation Set

We will pick a depth of 3 if validation error is minimized.Yes the answer changes to depth 6 if we switch the training and validation set.We can do n-fold cross validation to add more reliability in our result.

4 Naive Bayes

4.1 Bayes rule for drug testing

$P(D = 1|T = 1) = P(T = 1|D = 1)$*$P(D = 1)/P(T = 1|D = 1)P(D = 1) + (T = 1|D = 1)P(D = 0)$
$=(.99 * .001)/(.99)(.001) + (.01)(.999)$
$=.090$

4.2 Naive Bayes by hand

(a)

- $p(y = 1) = 6/10$
- $p(y = 0)$ = 1-(6/10) = $4/10$

(b)

- $p(x_1 = 1|y = 1)$ =$3/6$
- $p(x_2 = 1|y = 1)$ =$4/6$
- $p(x_1 = 1|y = 0)$ =$4/4$
- $p(x_2 = 1|y = 0)$ =$1/4$

(c)

If $p(y = 0|x_1 = 1, x_2 = 1) > p(y = 1|x_1 = 1, x_2 = 1)$ then the likely label is 0 otherwise its 1.

$p(y = 1|x_1 = 1, x_2 = 1)$

$= p(x_1 = 1, x_2 = 1)|p(y = 1)$We do not require the denominator as we only compare the two probabilities

$= p(x_1 = 1|y = 1)p(x_2 = 1|y = 1)p(y = 1)$ Naive Bayes assumption

$= 3/6 * 4/6 * 6/10$

$= 0.2$

$p(y = 0|x_1 = 1, x_2 = 1)$

$= p(x_1 = 1, x_2 = 1)|p(y = 0)$

$= p(x_1 = 1|y = 0)p(x_2 = 1|y = 0)p(y = 0)$

= 1 * 1/4 *4/10

= 0.1

More likely label is 1 as its probability is higher than that of 0

4.3 Naive Bayes Implementation

See `https://github.ubc.ca/cpsc340/jeanlam_nafis1_hw1/tree/master/code/naive_bayes.py`
The naive bayes error is 0.188, as compared to the decision tree validation error of 0.356

4.4 Runtime of Naive Bayes for Discrete Data

There are 3 for loops : one for d features ,one for T objects and one for k class labels. All of the loops are doing equal amount of work therefore O(dtk).