Machine learning with Mathematica for Tornadoes in the U.S., 1950-2015

Zhangliang Dong
Keying Zhou
Yueran Cai

USC University of Southern California

HOMEWORK FOR USC CSCI561 FOUNDATIONS OF AI

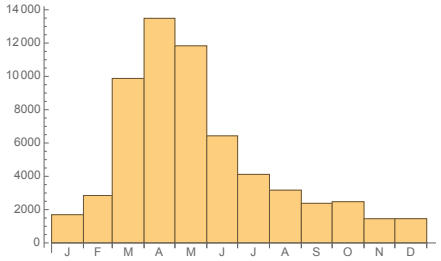
BACKGROUND: The dataset contains records of tornadoes. We are going to use 5 different methods to classify them into its magnitudes.

METHODS

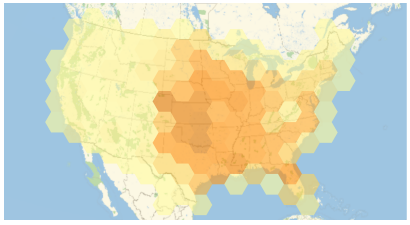
Neural Network
Random Forest
Nearest Neighbor
Naïve Bayes
Support vector machine

HOW TO BEGIN

1. Explore the dataset



Month histogram of Tornadoes



Geographical histogram of Tornadoes

2. Preprocessing the dataset. Transform the Location labels into geoPostions. Removing missing data. Extract the year, month and hours from dateObject. Replace the interval with min, max and mean.

```
Min[$Interval[{50 000, 500 000}]]
Max[$Interval[{50 000, 500 000}]]
Mean[{Min[#], Max[#]}] &[$Interval[{50 000, 500 000}]]
```

```
Select[Not[MissingQ[#EstimatedCropLoss]] &]
Select[Not[MissingQ[#EstimatedPropertyLoss]] &]
```

```
<|"EF0" → "F0", "EF1" → "F1", "EF2" → "F2", "EF3" → "F3", "EF4" → "F4", "EF5" → "F5"|>
```

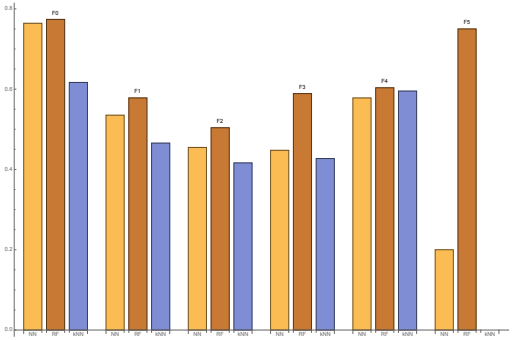
```
Tue 31 Jan 1950 12:06:10 CDT [{"Year", "Month", "Hour"}]
{1950, 1, 12}
```

```
Interpreter["Location"][Texas, United States]
GeoPosition[{31.4818, -99.3701}]
```

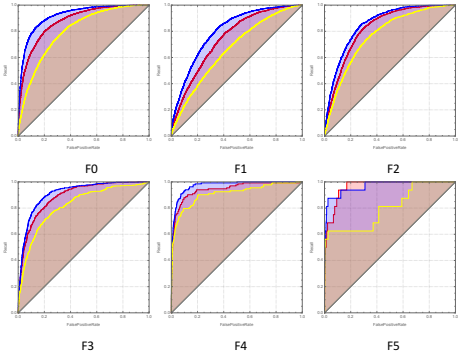
3. Training

Methods	Training Time
Neural Network	630.54 s
Random Forest	643.07 s
Naïve Bayes	755.00s
Support Vector Machine	551.15s
Nearest Neighbors	652.79 s

4. Result and comparison



The random forest model performs the best in term of precision in each class.



In these ROC curves above, the random forest shows the best.

Methods	Accuracy
Neural Network	62.54%
Random Forest	67.07%
Naïve Bayes	58.00%
SVM	49.60%
Nearest Neighbors	55.30%