Weather Prediction in the US

Performance comparison between Support Vector Machines and Fully Connected Neural Networks

Jason Ngo and Emily Lin

Motivation

Motivation: Predict extreme weather conditions (drought, flooding, etc.)

Scientific questions:

- Will a support vector classifier or fully-connected neural network perform better?
- Will results vary when predicting weather for one station, nearby stations and all stations?

Data and Methods

Dataset: US Historical Climatology Network

Time: 1980 - 2010

Daily Measures:

- **Precipitation** (10th of mm)
- Min Temperature
- Max Temperature
- Snow Fall
- Snow Depth
- Others: Daily total sunshine, wind speed, etc.

Weather Stations

NOAA National Climatic Data Center

Data Processing

- Data is scattered across 119 ASCII text files
- Weather of a particular day is correlated to weather on the previous day

➔ Read input into `pandas dataframe`

➔ Combine two consecutive days into one instance

	STATION	LAT	LON	ELEV	DATE	PRCP1	TMAX1	TMIN1	SNOW1	SNOWD1	PRCP2	TMAX2	TMIN2	SNOW2	SNOWD2
0	USC00053951	37.7494	-107.0950	2727.7	731331.0	0.0	144.0	-72.0	0.0	0.0	0.0	144.0	-72.0	0.0	0.0
1	USC00024089	34.9094	-110.1544	1549.9	733420.0	0.0	144.0	-56.0	0.0	0.0	0.0	144.0	-56.0	0.0	0.0
2	USC00072730	39.1467	-75.5056	9.1	733135.0	8.0	211.0	139.0	0.0	0.0	8.0	211.0	139.0	0.0	0.0
3	USC00063207	41.3511	-72.0389	12.2	725648.0	0.0	178.0	89.0	0.0	0.0	0.0	178.0	89.0	0.0	0.0
4	USC00427729	39.6842	-111.2056	2650.8	732029.0	5.0	-6.0	-78.0	13.0	610.0	5.0	-6.0	-78.0	13.0	610.0

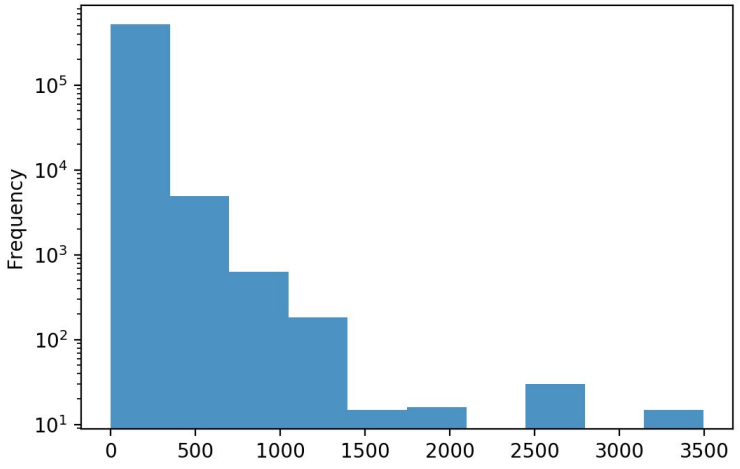
p = 14, n = 524065

Data Processing

- Regression problems are hard to implement and interpret

➔ Transform precipitation values from continuous to categorical

- Shuffle examples



Histogram for PRCP values

Experimental Setup

1. One Station
    - Predict precipitation range for each pair of days based on data from the same station only
2. Nearby Stations
    - Predict precipitation range for each pair of days for a given station based only on data from the two closest stations

One Station

Predict precipitation for one station using data from the same station

- Station: Steamboat Springs, Colorado
- Train.shape: (4404, 13)
- Test.shape: (1102, 13)
- Precipitation: 18 bins based on distribution of the values

One Station - Model Summary

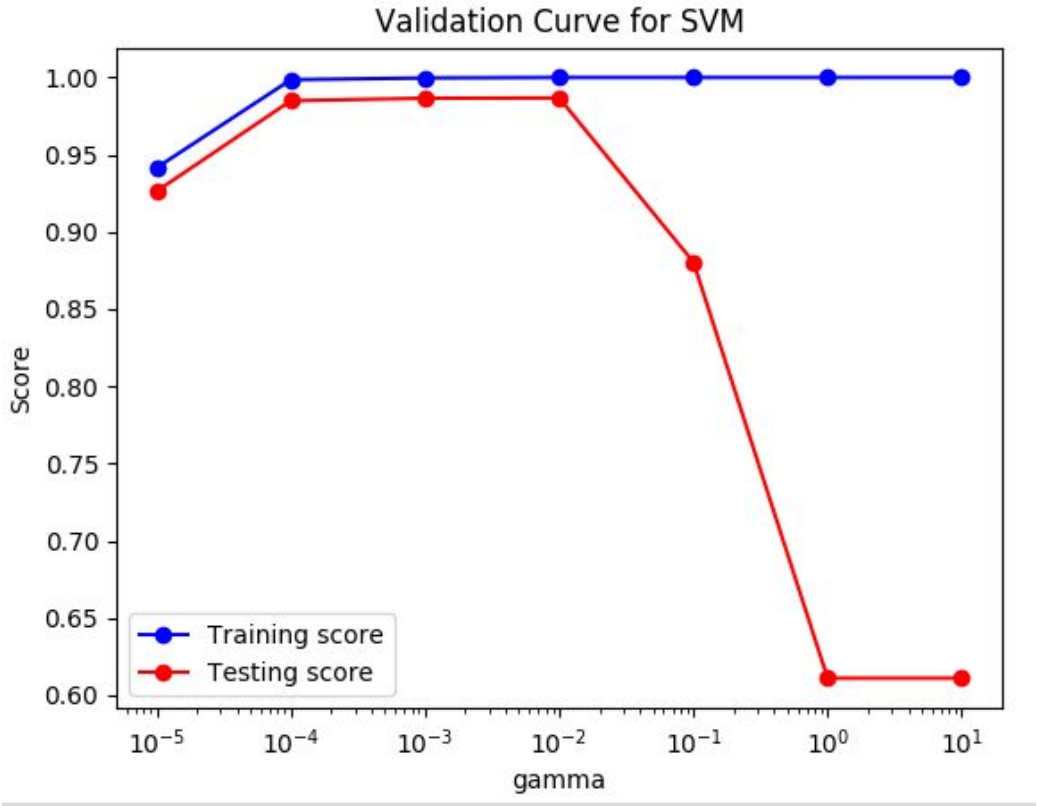
SVC hyperparameter tuning

- C = 10
- Gamma = 0.001

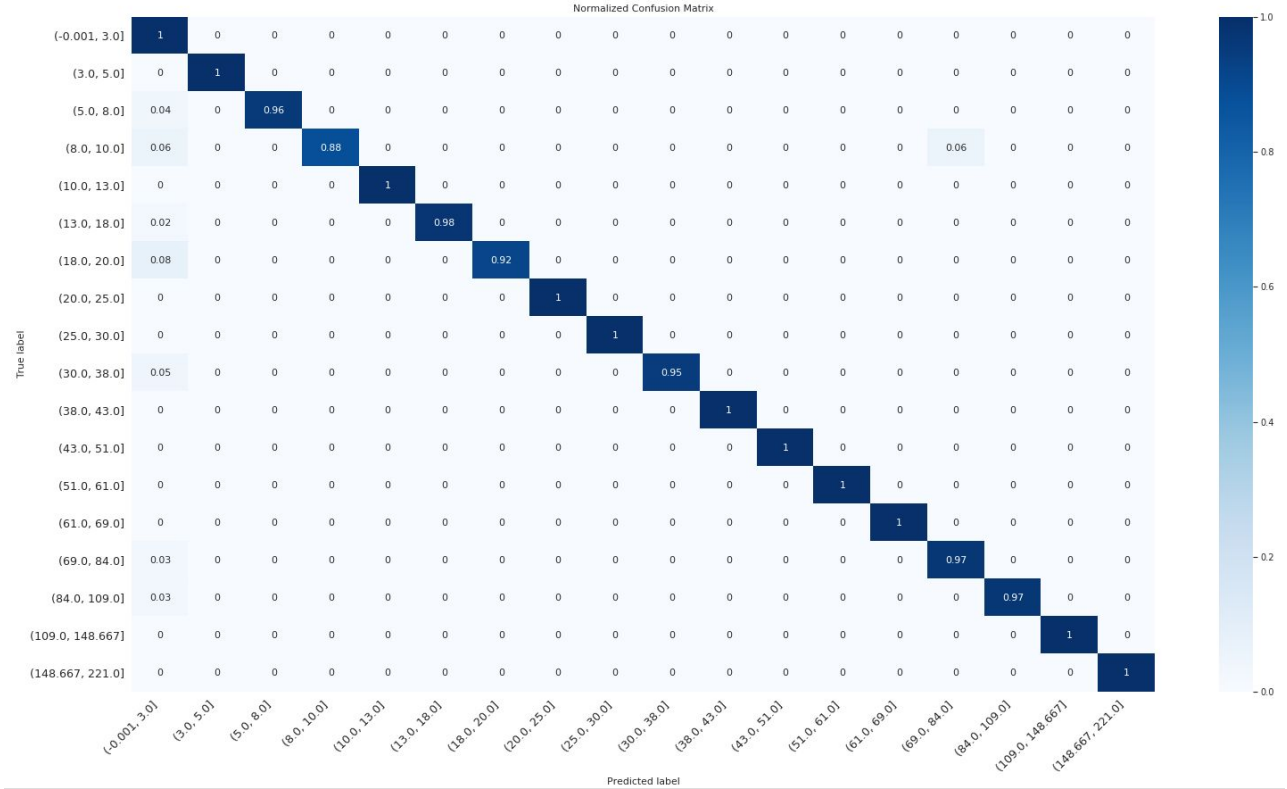
FC model - 3 connected layers

- 2 with `relu` and one with `softmax`

One Station - SVC Validation Curve



One Station - SVC Confusion Matrix



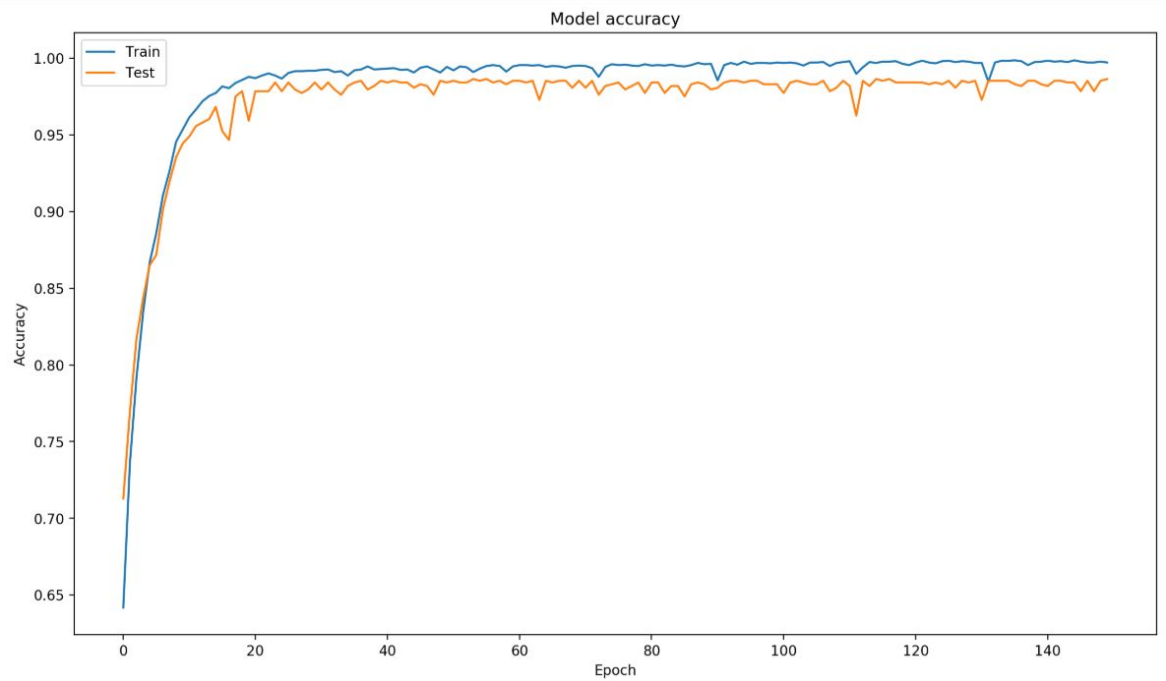
Training accuracy = 100%

Testing accuracy = 98.9%

Relative difficulty predicting ranges:

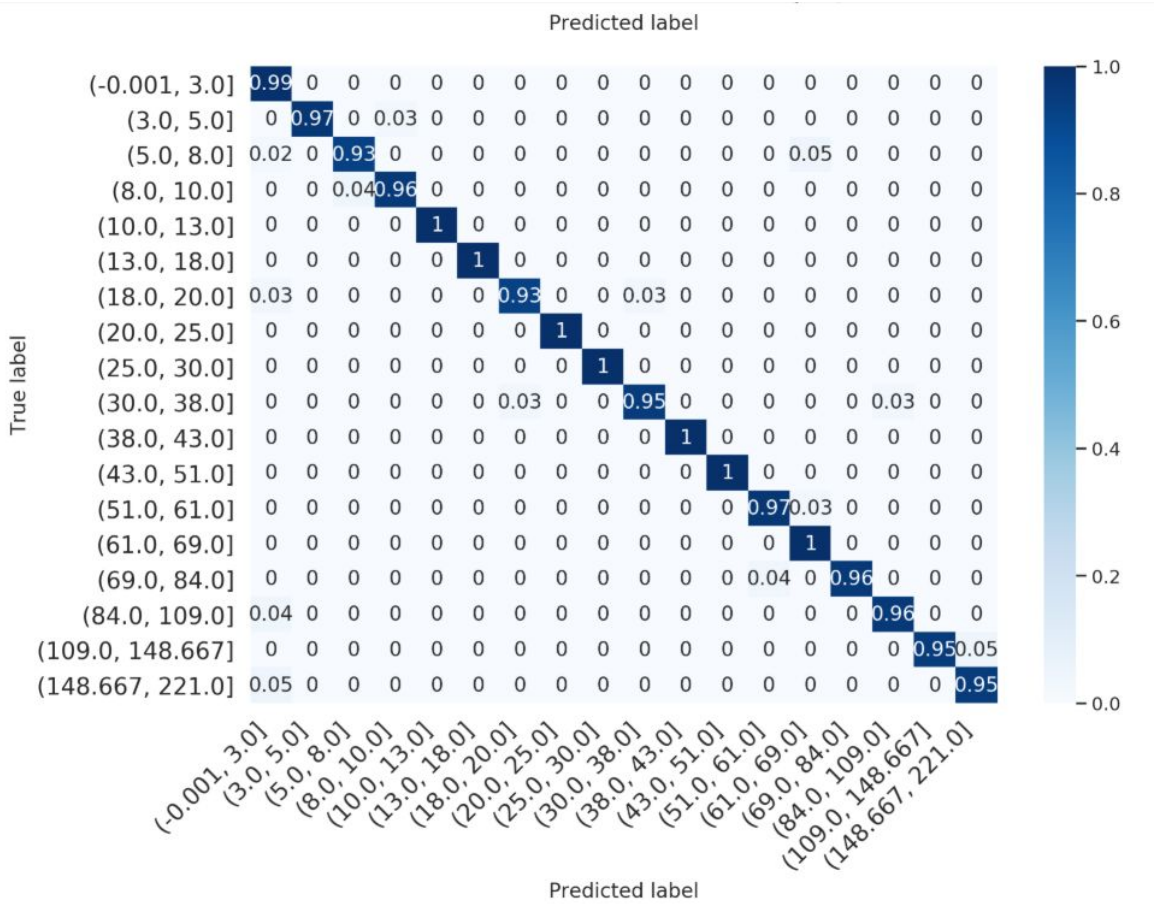
- 8-10 (0.88)
- 18-20 (0.92)
- 30-38 (0.95)

One Station - FC Validation Curve



Training/Validation Accuracy Curves

One Station - FC Confusion Matrix



Training accuracy = 99.8%

Testing accuracy = 98.1%

Relative difficulty predicting ranges:

- 5-8 (0.93)
- 18-20 (0.93)
- 30-38 (0.95)

Nearby Stations

Predict precipitation for one station using data from nearby stations

- Label: Precipitation in Steamboat Springs, CO
- 1435 instances from the two nearest stations: Cheeseman, CO and Grand Canyon, AZ
- Precipitation: 15 bins based on distribution of the values

Nearby Stations - Model Summary

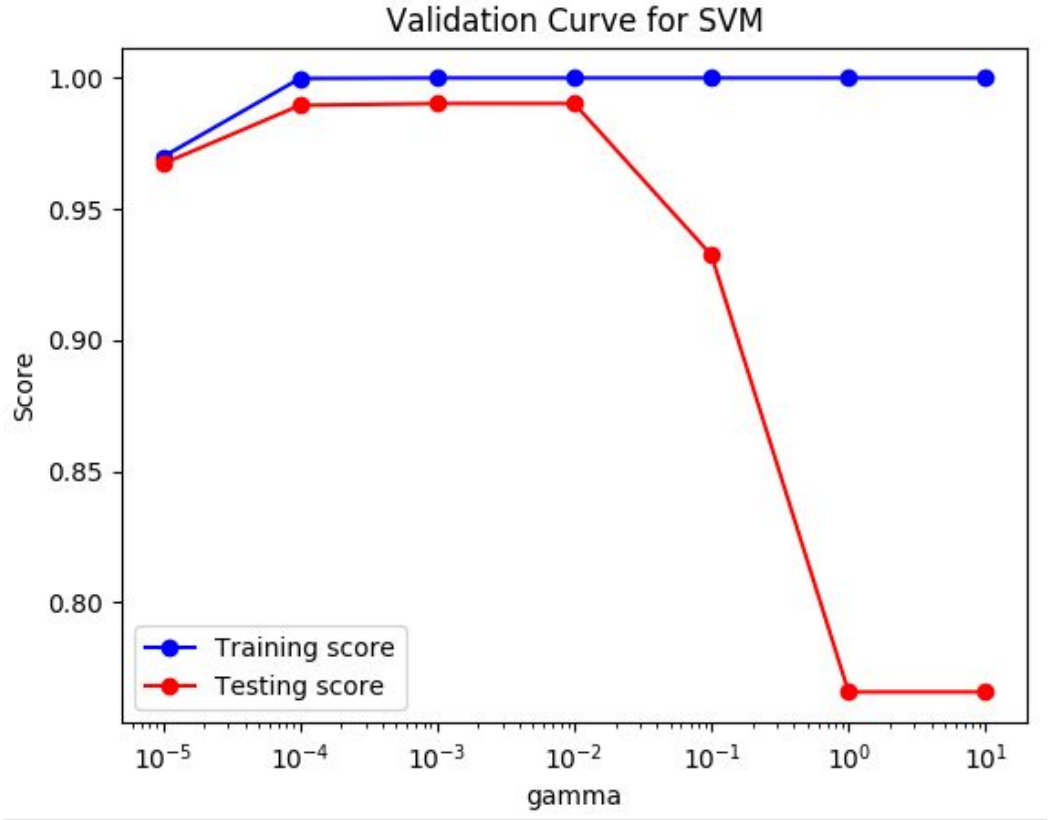
SVC hyperparameter tuning

- C = 1
- Gamma = 0.001

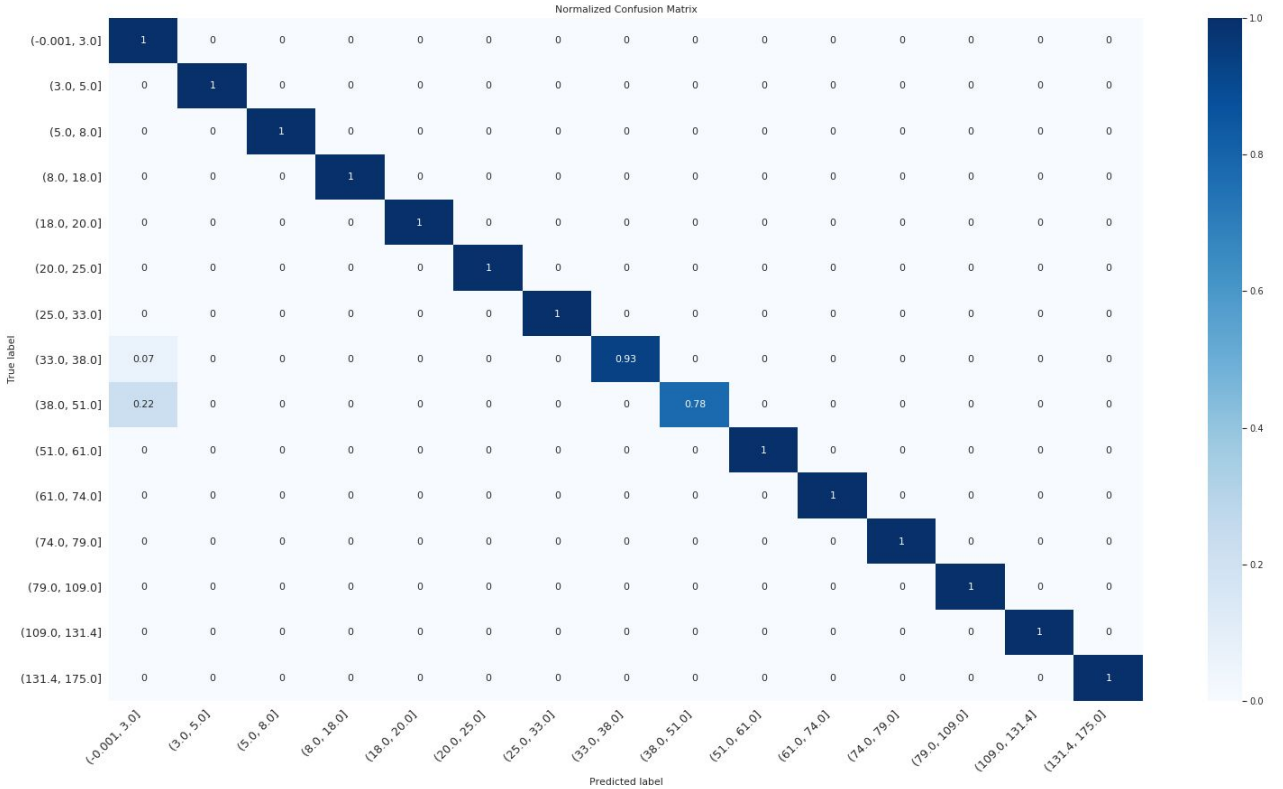
FC:

- Same three-layer fully connected neural network

Nearby Stations - SVC Validation Curve



Nearby Stations - SVC



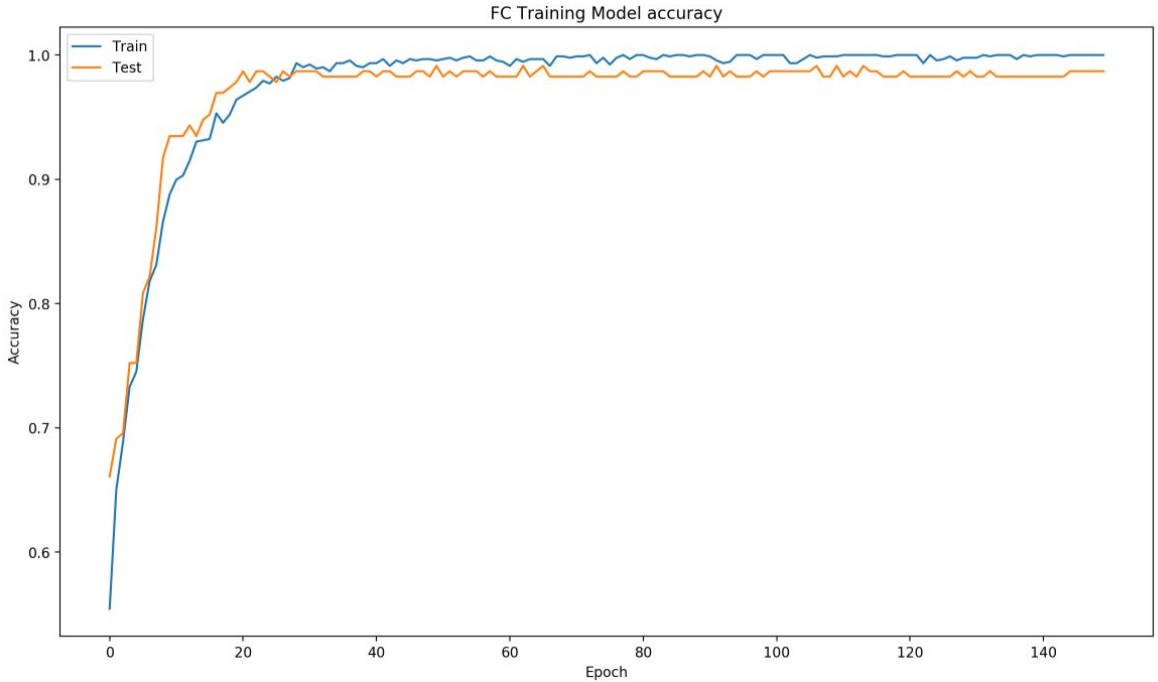
Training accuracy = 99.9%

Testing accuracy = 99.0%

Relative difficulty predicting ranges:

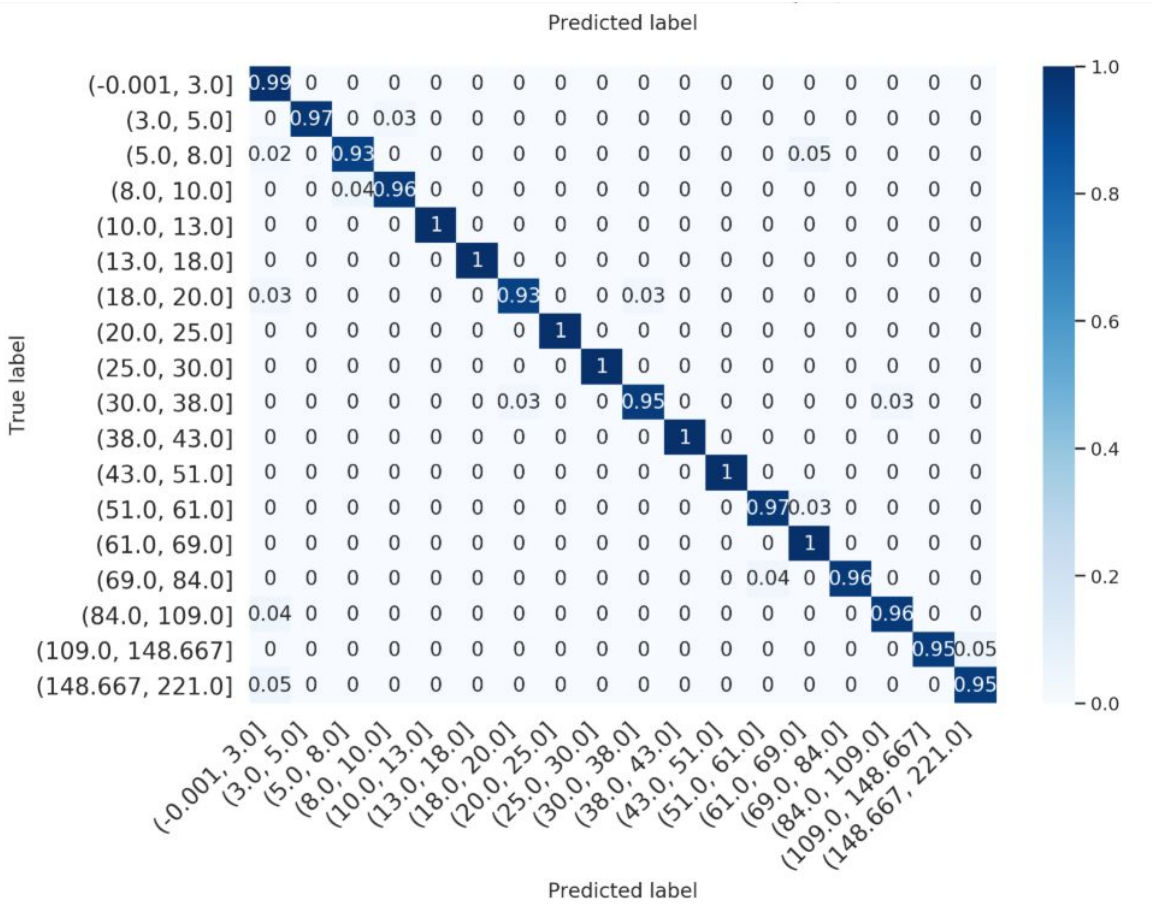
- 38-51 (0.78)
- 30-38 (0.93)

Nearby Stations - FC Validation Curve



Training/Validation Accuracy Curves

Nearby Stations - FC Confusion Matrix



Training accuracy = 99.8%

Testing accuracy = 98.6%

Relative difficulty predicting ranges:

- 5-8, 18-20 (0.93)
- 30-38 (0.95)

Conclusions

- SVC and FC performed equally well.
- Varied on which classes they predicted better or worse on.
- Results for one station vs. nearby stations were also the same.
- Trends
  - SVC predicted (-0.001, 3.0) when wrong
  - FC errors varied
  - SVC had more certainty (1.0’s in matrix)
  - FC - less certainty

Future Work

- Training and predicting using all stations
- Predicting different weather conditions
- Predicting for different stations

References

[1] Aravind, “Confusion Matrix as a Heatmap with Python,” *Data Fiction*, 12-Jun-2019. .

[2] M. J. Menne, I. Durre, R. S. Vose, B. E. Gleason, and T. G. Houston, “An overview of the global historical climatology network-daily database,” *Journal of Atmospheric and Oceanic Technology*, vol. 29, no. 7, pp. 897–910, 2012.

[3] N. Sharma, “Splitting CSV Into Train And Test Data,” *Medium*, 10-Oct-2018. [Online].

[4] “Basic classification: Classify images of clothing | TensorFlow Core,” *TensorFlow*. [Online].

[5] “Confusion matrix — scikit-learn 0.22 documentation.”

Thank you!