Predicting Turbulence Using Partial Least Squares Regression and an Artificial Neural Network

Valliappa Lakshmanan[1,2]

[1]Cooperative Institute of Mesoscale Meteorological Studies
University of Oklahoma

[2]Radar Research and Development Division
National Severe Storms Laboratory

Jan. 2010

Basic Idea

1. Use partial least squares regression to reduce number of variables
2. Remove all-zero observations
3. Use ANN to estimate probability

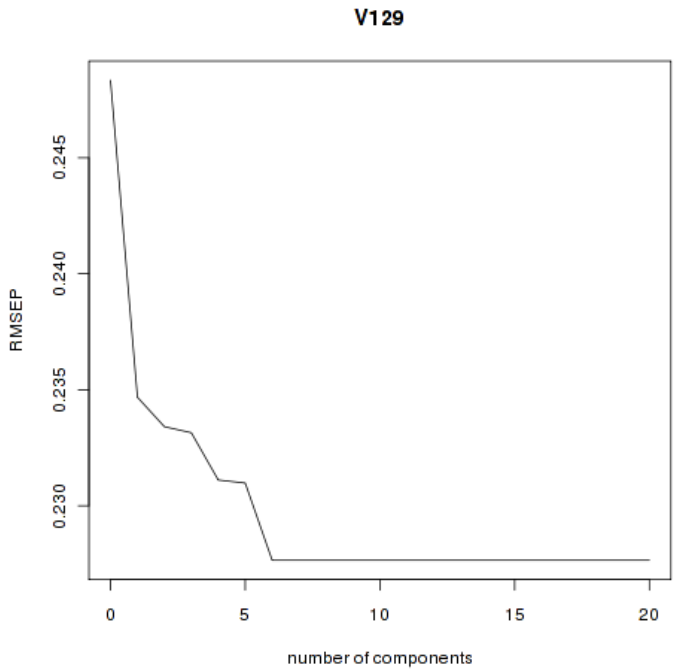
Dataset

Training dataset: 136 columns, 103,990 rows
Testing dataset: 50,127 rows
Aim: predict turbulence
Challenge: too many variables, turbulence is rare.

Partial Least Squares Regression

PLS regression is like PCA except: instead of finding the eigen vectors of the attribute matrix, we find the eigen vectors of the *covariance* of the attribute matrix and the predictand.
Better fit for the classification problem.
See [Hastie et al., 2001, Mevik and Wehrens, 2007].

RMS error by number of components



6 components are enough: each component is a linear combination of 134 attributes (many of whose weights are zero).

Significant Attributes of Significant Components

The attributes with non-zero weights of the six most significant components:

1. STONE_Linear
2. RPVORT_Linear and RSTAB_Linear
3. RICH_Linear, ROL_Linear, RPVORT_Linear, RSTAB_Linear, and SATRI_Linear
4. ROL_Linear, RSTAB_Linear and SATRI_Linear
5. PRES_SFC_Linear, ROL_Linear, RSTAB_Linear, and STONE_Linear
6. dbz_DNGood40, dbz_DNGood80 NSSL_DBZ40_Wedge0_5_mindistance, nssl_18dbz_top_DNGood0_80, PRES_SFC_Linear, ROL_Linear, RSTAB_Linear, SMHGT_Linear, STONE_Linear, and MSLMA_MSL_Linear

Significant Attributes

- dbz_DNGood40
- dbz_DNGood80
- NSSL_DBZ40_Wedge0_5_mindistance
- nssl_18dbz_top_DNGood0_80
- PRES_SFC_Linear
- RICH_Linear
- ROL_Linear
- RPVORT_Linear
- RSTAB_Linear
- SATRI_Linear
- SMHGT_Linear
- STONE_Linear
- MSLMA_MSL_Linear

Note: not in order of significance.

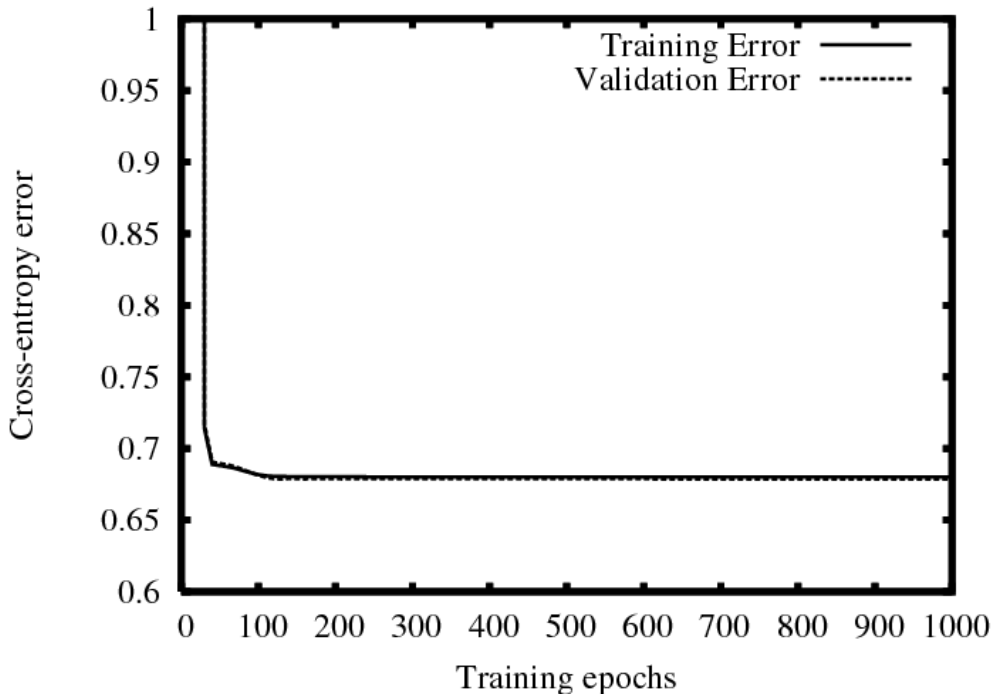
Remove zeroes

- The 103,990 input rows were randomly assigned 2:1 into a training and validation set i.e. 69327 rows were used for training and 34663 rows were used for validation.
- The transformed data had lots of rows where all the inputs were zero.
- About 6% of the all-zero observations were turbulent.
- All-zero observations were removed from the training and validation sets
- At run-time for any row whose transformed values are all zero, output will be 0.06.

Neural Network Architecture

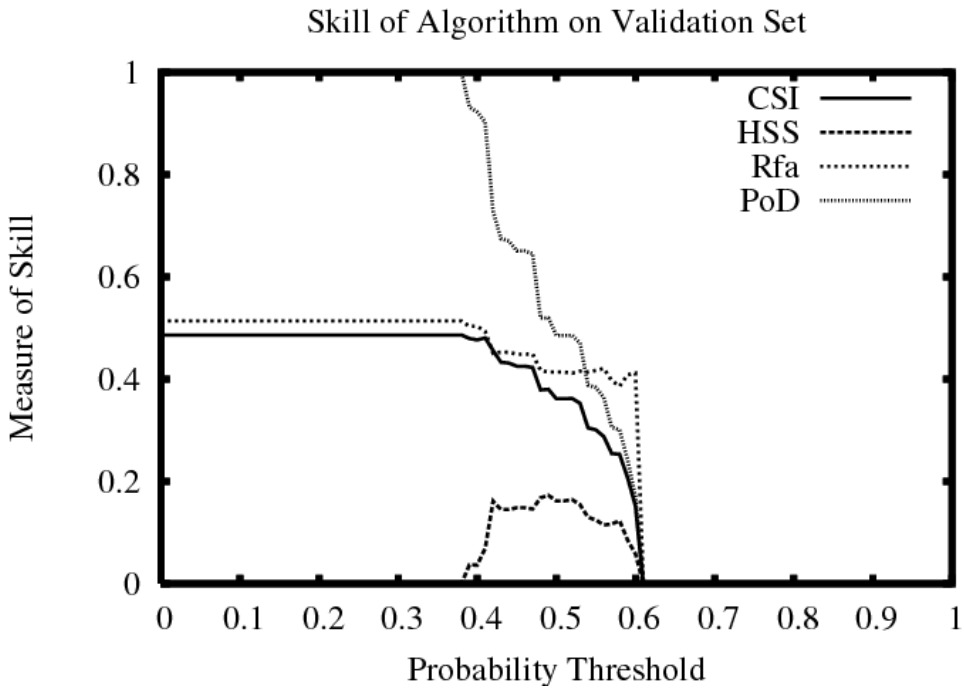
- 6 inputs (the 6 transformed components)
- 1 output (0/1 for turbulence)
- 3 hidden nodes in one layer
- Input and output nodes had a sigmoid transfer function
- Hidden nodes had a tanh transfer function
- Cross-entropy minimized so that predicted value is a probability

Validation error



No overfitting ...

NN Skill on validation set

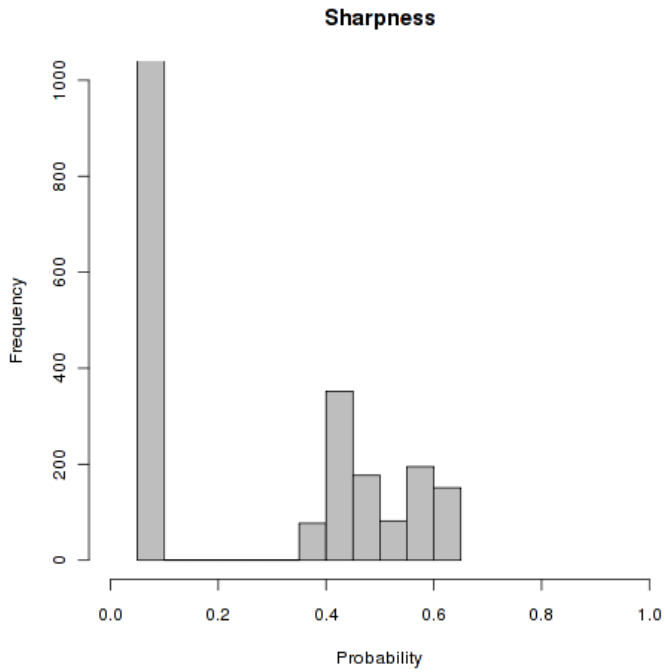


The neural network outputs are clustered between 0.4 and 0.6 indicating that the neural network was not able to separate out the two classes very well.

Analysis

- The neural network was not able to separate out the no-turbulence observations from the turbulence observations very well.
- Only 1-2% of the provided observations were useful data: all-zero observations consisted of 98-99% of the data.
- A better prediction algorithm may be possible, but only if there are more usable observations of no-turbulence *in the presence of weather*.

Sharpness on test dataset



In the absence of ground truth, the only measure we can compute is sharpness.

Acknowledgements

Funding for this research was provided under NOAA-OU Cooperative Agreement NA17RJ1227.
The author also thanks Gillian Peguero and John Williams for organizing the contest this year. This dataset provided me with the motivation to learn to apply PLS regression.

References

Hastie, T., Tibshirani, R., and Friedman, J. (2001).

The Elements of Statistical Learning.
Springer.

Mevik, B. and Wehrens, R. (2007).

The pls package: Principal component and partial least squares regression in R.
J. Statistical Software, 18(2):1–24.