Topics of discussions:

1. Go over classification methods: what can be said about applicability in different situations.

2. Models not perfect; what can we say about robustness or other flaws in the model.

3. How do we get an intuitive feeling for the classification methods? Graphical methods?

4. Methods for variable selection? How to find the “best” set of variables? Why do we need to reduce the number of variables?

5. How many different data sets people have looked at?

6. Unisims, multisims: estimating systematic uncertainties.

7. Technical framework implementation of different classifiers.

8. Question: Is subjectivity adding to the quality of the analysis or not?

9. Can we learn about statisticians' methodology.

1. Classification methods

- Try them all (sort of)
- Don't worry so much about which one is better
- Start simple
- Could statistician prepare some comparisons
- Unlikely to get VERY precise conclusions
- some are more sensitive to tuning parameters than others, e.g.random forests are not
- Often “off the shelf” versions give largely the same results (amongst the good ones)
- Don't overlook naïve Bayes
- “Good” classifier: Decision trees with boosting, random forests, Bayesian NN

2. Robustness

- A classifier might be chosen on the basis of something other than average class errors.
- Compromise between class errors and sensitivity to parameters settings (insensitivity to parameter settings is not a sufficient condition to choose one classifier over another)
- Trade off parameter settings to sample size
- Use parameter settings as inputs to classifiers
- What to do regarding non Gaussian tails?

3. A graphical suggestion

Plot change in prediction relative to change in values of a simple variable vs those values. Look for:

- additivity or not
- interactions
- linearity or not
- intercept = 0? (i.e. useless variable)

6. Systematic uncertainties

- Unisim/multisim/designing the multisim (multisim generally prefered)
- Changing the mix of data points produced vs parameter settings

7. Technical framework implementation of different classifiers

(See Illya Narsky talk)
Suggestion to prepare consumer reports for different software

10. Mistakes not to make

See Volker's slides on normalization