ML Midterm Review II

- Agenda:
  - Midterm format
  - Example data
  - Summary
  - Q&A

Format

- **You get:**
  - Data X_train
- **We want:**
  - **Trained** classifier predict.py
  - Predict function: y_pred = predict(X)
  - Summary of your process explain.pdf (design decisions, charts/visuals, any validation you did – 2 pages MAX)

Format (cont.)

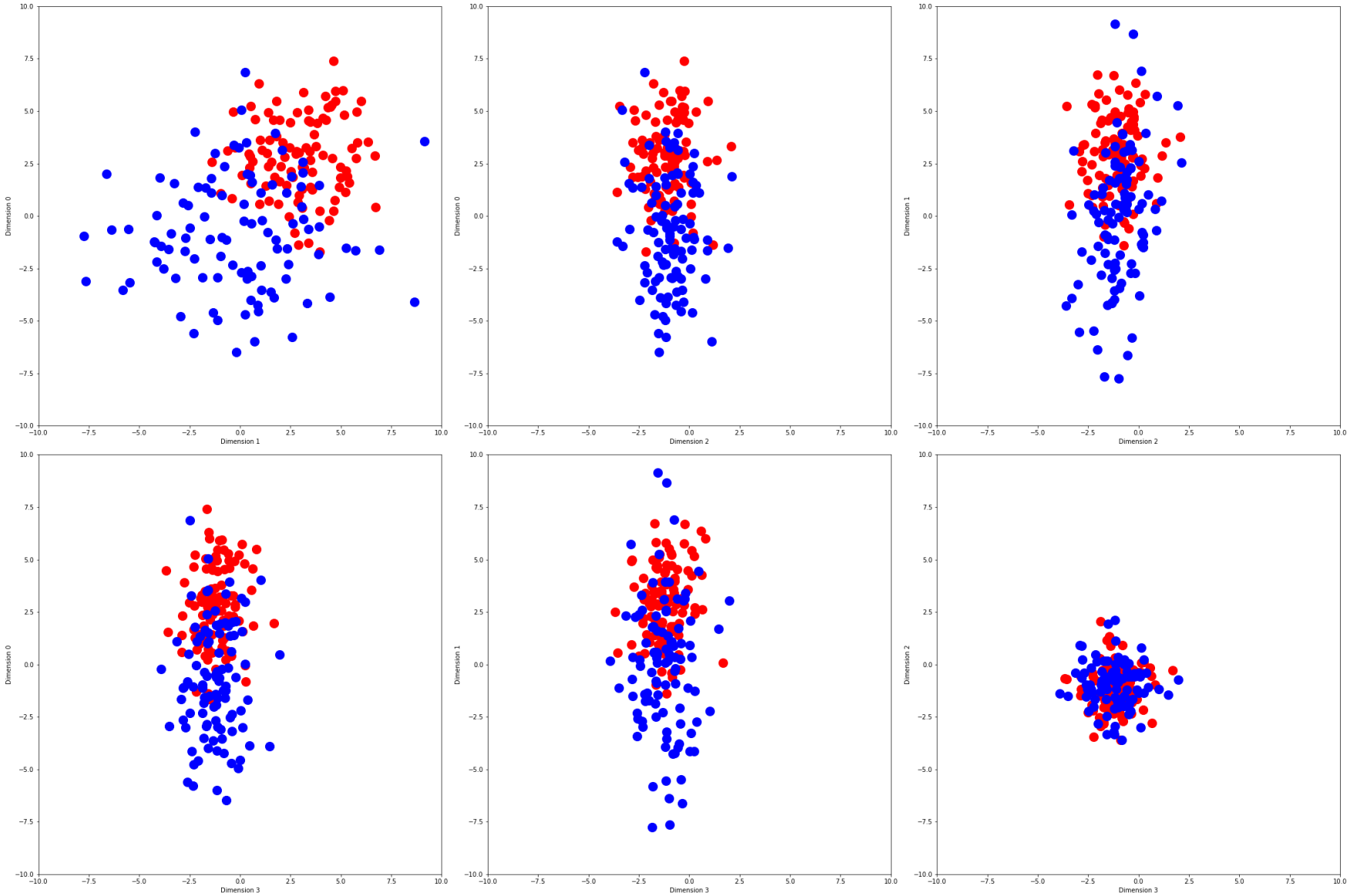
OK

- scikit learn
- piazza (private posts)
- CourseWorks @Columbia
- Dead trees

NOT OK

- Google
- stackoverflow
- Your friends

Example (200 x 4)



Example data (cont.)

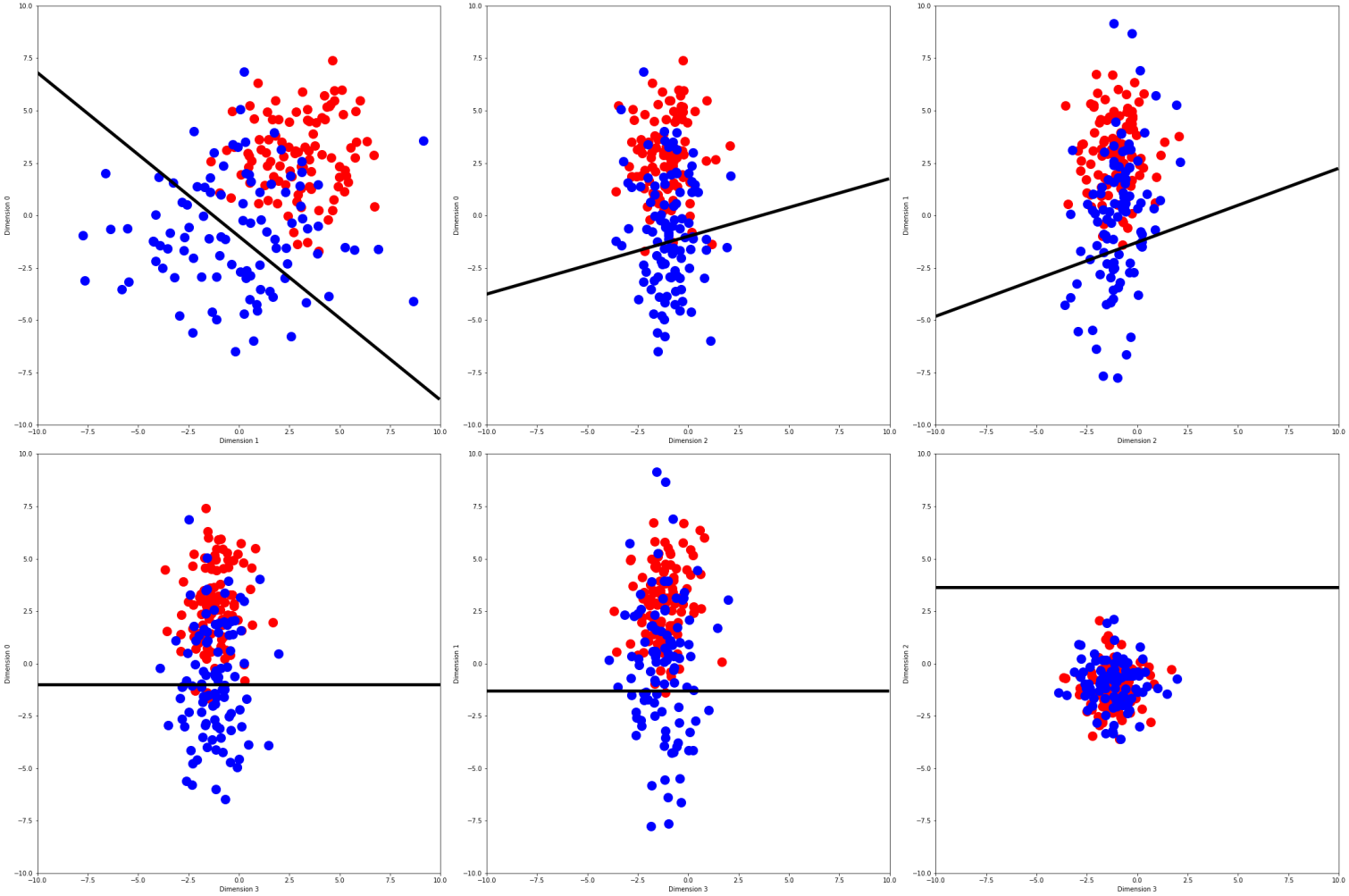
- How do we start?
  - Make some plots
  - Try to clean up data
  - Do all features matter?
- What do you notice?
  - Some dimensions “more equal”
  - Some dimensions “matter more”
  - Dim 2 vs Dim 3 uninformative

Example data (cont.)

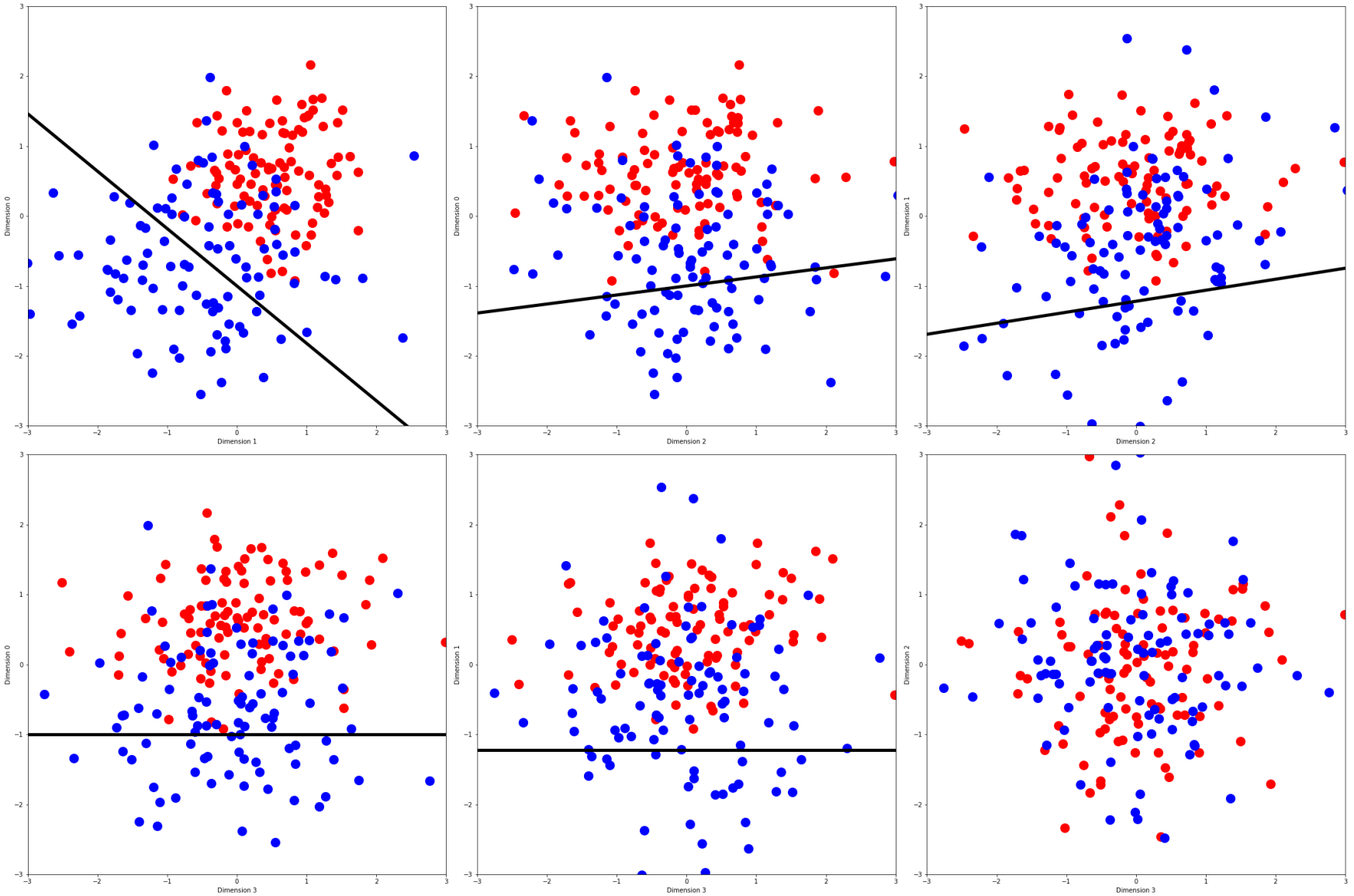
- Algorithms
  - Name some classification algorithms
  - Assumptions: must the classifier be linear? Data separable?
  - Train/test split
  - Regularization? (lasso aka L1 vs ridge aka L2)
- Ex.: Logistic regression

Logistic regression

$$f(x) = \frac{1}{1 + \exp(-\theta^T x)}$$



Logistic regression (cont.)



Summary

- DO
  - Perform cross-validation (training/testing set)
  - Think about outlier rejection, feature selection
  - Use Google Cloud
- DON’T
  - Train your classifier “on-the-fly” in predict(...)
  - Expect 99% accuracy
  - Use random algorithms we haven’t talked about
  - Use prohibited resources

Questions?