Problem Set 7: Prediction Contest

Shane T. Mueller shanem@mtu.edu

January 9, 2017

Modeling of Data

The file masking.csv describes an experiment in which people saw different letter stimuli (A..Z) briefly, and had to make a forced choice identification, each time between the actual letter and an alternative. The data summarizes the median response time for each stimulus letter for each person. Two different masks were given, a @ in one condition (mask==0) and a # in a second condition (mask==1). A mask is a image that appears immediately before and after the stimulus to disrupt processing. Masks will have different impacts on different letters, and so we may be able to detect which mask a person had by their response patterns. Can you identify which mask was given based only on the median response times?

```
dat.all <- read.csv("masking.csv")
head(dat.all)
```

```
##     A   B   C   D   E   F   G   H   I   J   K   L   M   N   O   P   Q   R
## 1 476 597 494 473 577 542 577 476 417 561 555 458 531 437 455 497 634 541
## 2 744 804 629 871 647 749 588 461 548 666 469 585 543 487 690 725 682 749
## 3 456 472 395 373 398 356 392 417 213 452 336 401 400 421 297 292 341 332
## 4 472 437 177 401 448 444 469 370 493 431  18 350 457 448 433  10 491 402
## 5 453 449 557 511 443 474 534 453 469 450 443 449 429 452 482 443 531 434
## 6 615 523 632 658 511 631 710 498 455 598 571 478 513 522 673 498 797 631
##     S   T   U   V   W   X   Y   Z mask
## 1 534 436 538 472 561 434 542 597    0
## 2 730 527 521 470 486 581 464 608    1
## 3 356 356 457 455 393 314 313 193    0
## 4 147 403 430 231 474 474 409 391    1
## 5 492 439 511 450 448 449 471 451    1
## 6 473 470 711 497 609 590 473 488    1
```

In this problem set, you must build at least six classification models, from the models we have learned about in class. These include:

- Logistic regression
- LDA
- QDA
- Naive Bayes
- Decision/partition Tree
- Random Forest
- K-nearest Neighbor
- SVM
- Neural Network

Build six models of the data. In each case, you must described the model, use some type of cross-validation or complexity scheme (such as AIC) to select parameters, and identify a justifiably good model within the category. When necessary, be sure fit the model enough times so that you know you have produced a good solution (i.e., when the optimization may end up in a local optimum). Write a brief description (2-4 sentences) about each of the six models, describing whether/why/how the model is doing, and provide evidence to back up your conclusions.

Prediction contest

Make predictions with each of your selected models about the twelve NEW cases in the testcases file. Show a table summarizing the predictions across models. Then, using whatever criteria you want, select your best guess about each of the twelve cases (this could be a consensus of models, a single prediction of your best model, or even just flipping a coin). Your predictions will be compared to other students in class, and we will declare a "winner".

Discuss Comparing Models

Write a discussion comparing the six different models you have fit. Describe the advantages and disadvantages of each model for this problem. Discuss why you chose the particular prediction that you did, in the context of the different model's strengths and weaknesses.