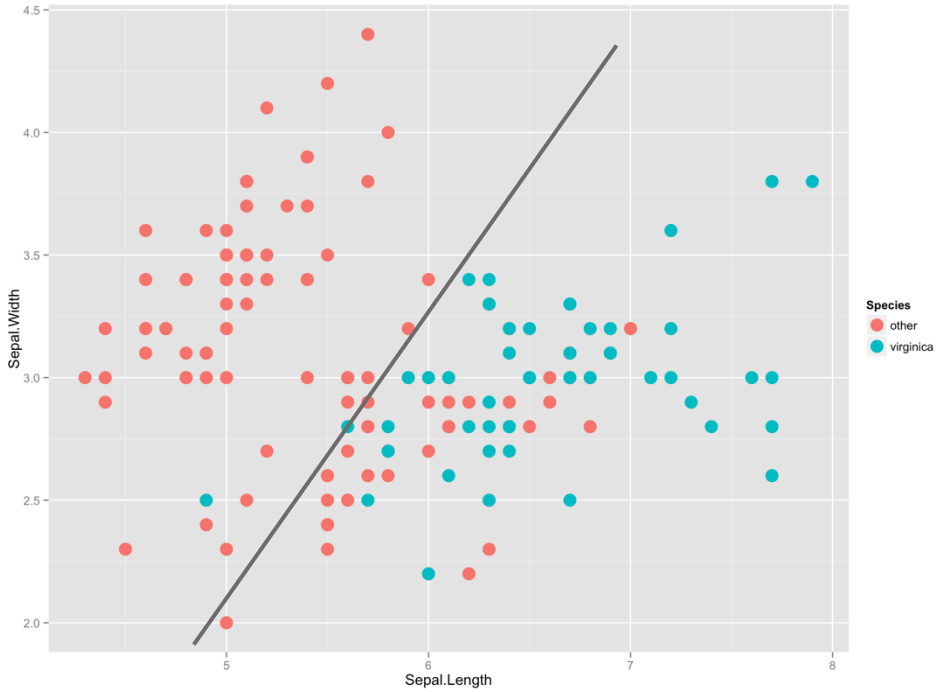
Evaluating your Classifier

Marc Light & Jay Jacobs

A Classification Problem



Use Carot along with a randomForest

```
summary(iris2$Species)
```

```
##     other virginica
##       100        50
```

```
set.seed(1)
fitControl <- trainControl(method = "repeatedcv",  number = 10,
                           repeats = 5, classProbs = TRUE, savePred=T)
rfFit <- train(Species ~ ., data = iris2,
               method = "rf",
               trControl = fitControl)
```

Evaluation

```
rfFit
## Random Forest
## 150 samples
##   2 predictor
##   2 classes: 'other', 'virginica'
##
## No pre-processing
## Resampling: Cross-Validated (10 fold, repeated 5 times)
## Summary of sample sizes: 135, 135, 135, 135, 135, 135, ...
## Resampling results
##
##   Accuracy  Kappa      Accuracy SD  Kappa SD
##   0.72      0.3729411  0.1150766    0.2564449
```

What is a reasonable baseline?

- pick majority class

What is accuracy?

- correct / test.population

What is Kappa?

- (acc - baseline) / (1 - baseline)

What if data is very skewed?

Can we do anything with the different runs that fall out of the cross validation

TP, FP, TN, FN and all that jazz

Nice Wikipedia picture

$$\text{precision} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{retrieved documents}\}|}$$

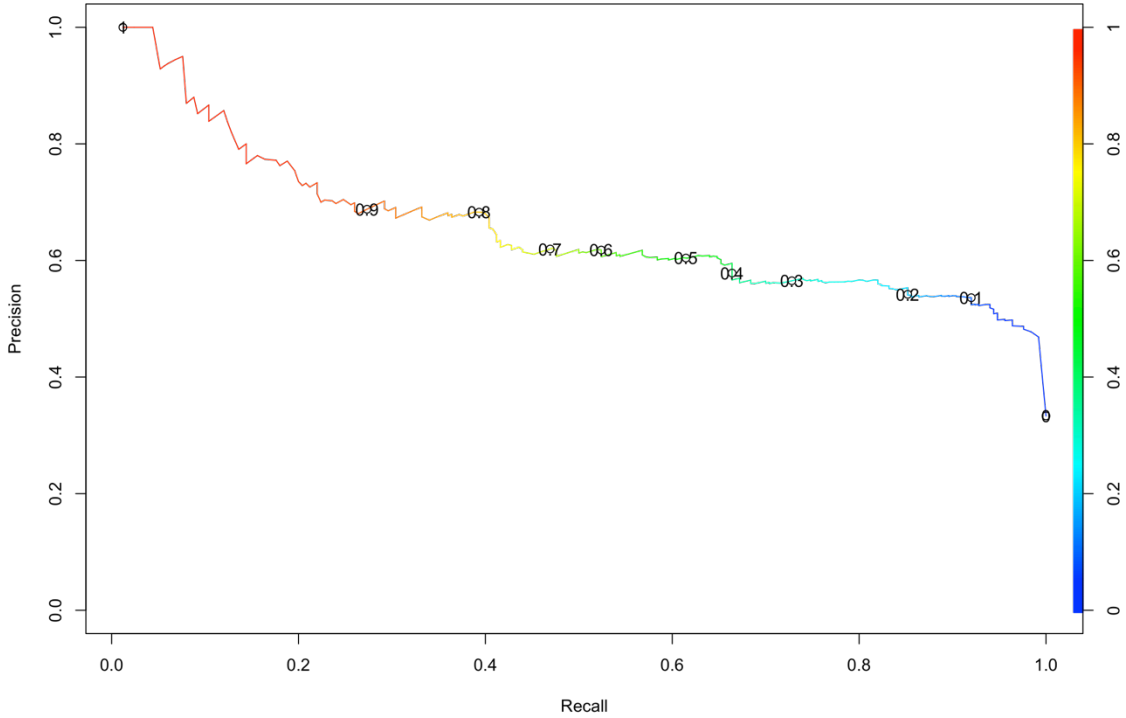
$$\text{recall} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{relevant documents}\}|}$$

Trading off Precision against Recall

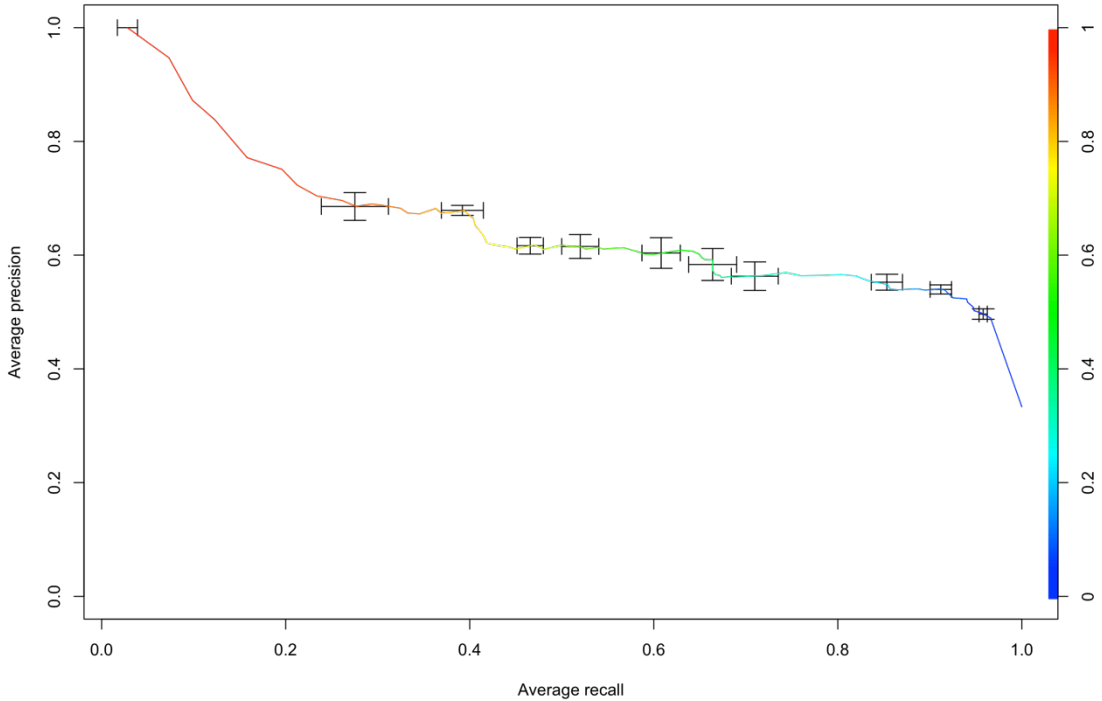
```
scores <- rfFit$pred$virginica
labels <- rfFit$pred$obs
pred <- prediction(scores, labels)
perf <- performance(pred, measure =
"prec", x.measure = "rec")
plot(perf, col=rainbow(10),
print.cutoffs.at=seq(0,1,by=0.1),
colorize=T, ylim=c(0,1))
```

ROCR R Package

Resulting ROCR generated graph



Confidence intervals from bootstrap



Reproducible Research

Jay Jacobs & Marc Light

What is Reproducible Research?

…the ultimate product of academic research is the **paper** along with the full **computational environment** used to produce the results in the paper such as the code, data, etc. that can be used to **reproduce the results** and **create new work** based on the research.

What usually breaks:

- Access to Data
- Access to Code
- Access to Thought Process
- Access to Environment

What usually breaks:

- Access to Data (raw, prep/cleaning, final)
- Access to Code (knitr/git)
- Access to Thought Process (knitr)
- Access to Environment (packrat, RevoR)

Other Guidelines

1. Whoever will reproduce your work is just as busy (and lazy) as you.
2. You will probably have to reproduce your own work (see #1).
3. Assume little-to-no knowledge or environment.

This means…

Automate everything!

- Nothing is done manually or as a one-off (excel)
- Everything is in the code and accessible
- Standard directories and file locations

This means…

Work Incrementally

- Save the data and cite the source
- Save all interim data objects (converted/cleaned)
- Run expensive blocks manually:
  - End code with a save() statement
  - wrap the block with eval=FALSE

This means…

Over Explain, Over Document

The Setup

- Start with a clear research question:
  - How well can we predict the species of iris flower given measurements of each specimen?
- Create predictive model for iris data
  - Make it interesting - show off ROCR
  - Make it reproducible

Reproducible Research

Jay Jacobs & Marc Light