CS 4780 Fall '11
Final Project

Predicting Criminal Sentences

Machine learning the justice system

JUSTIN CHENG JC882 YUNCHI LUO YL477 JANE PARK JP624

Motivation

Human judges use a combination of objective knowledge of law and some subjective judgment to determine a criminal's sentence. We use machine learning techniques learned in class to create a predictor for criminal sentences using the Pennsylvania sentencing data from 1998. From these predictors, we determine the most critical factors involved in criminal sentencing.

Questions

Can we create a machine that can, with reasonable accuracy, match the decisions of the human judges and reveal the true factors that influence sentences? Are sentences determined solely and fairly according to just the crime, as the judges swore under oath, or are other variables, such as demographics, interfering with the decisions? If the judges are making purely objective decisions, is it possible to reverse-engineer parts of the law by looking at the sentences that criminals received?

Methods

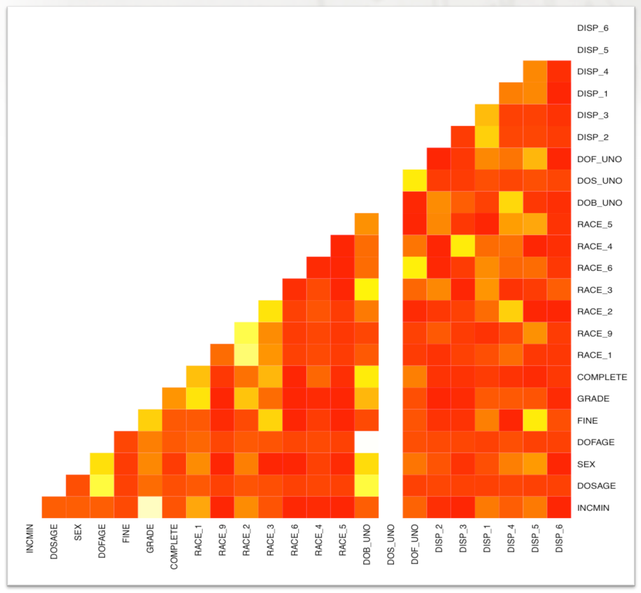
- Orange software: Bayes, Decision tree (C4.5), Linear SVM, kNN
- Label being predicted: INCMIN (minimum incarceration time)
- Variable sets used:
  - **DEMO**[graphics]: SEX, DOFAGE (age on date of offense), RACE
  - **CRIME**: FINE, GRADE (s tatutory offense grade), COMPLETE (offense completed/inchoate), DOS_UNO (date of sentence), COUNTY (sentencing county), PCSOFF (offense code), PCSSUB (offense subcode), DOF_UNO (date of offense), DISP (disposition)
  - **BASE**: DEMO + CRIME

Findings

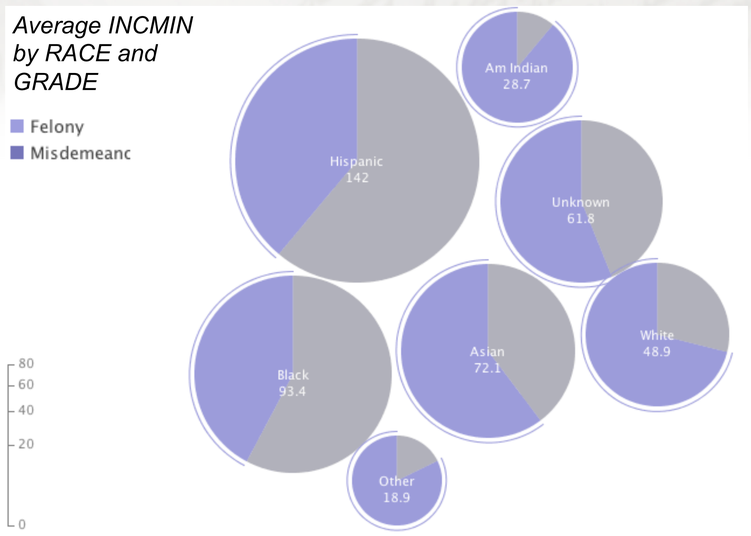
Variable Set	Method	Accuracy
DEMO	Bayes	0.629
	C4.5	0.659
	Linear SVM	0.637
	kNN	0.643
CRIME	Bayes	0.772
	C4.5	0.783
	Linear SVM	0.583
	kNN	0.769
BASE	Bayes	0.761
	C4.5	0.774
	Linear SVM	0.589
	kNN	0.773

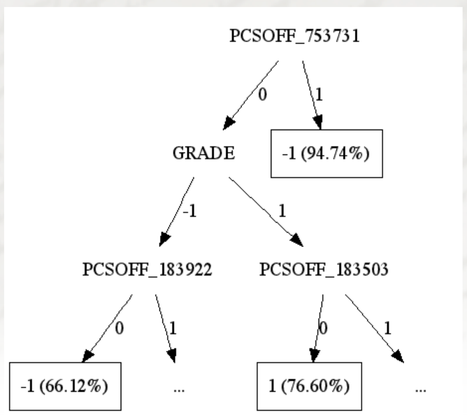
Linear SVM performed poorly across all variable sets. The other methods showed similar performance for given variable sets, with the C4.5 decision tree having the highest labeling accuracies across 10-fold cross validation. The CRIME set performed best, confirming that information about the crime itself, not the demographics of the criminal, is most important factor in determining the incarceration time. In fact, the pruned decision tree for BASE was exclusively composed of variables from CRIME.

Statistics

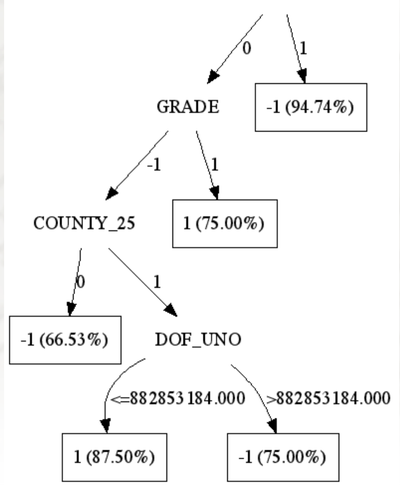


↑ *Variable Correlation Map*



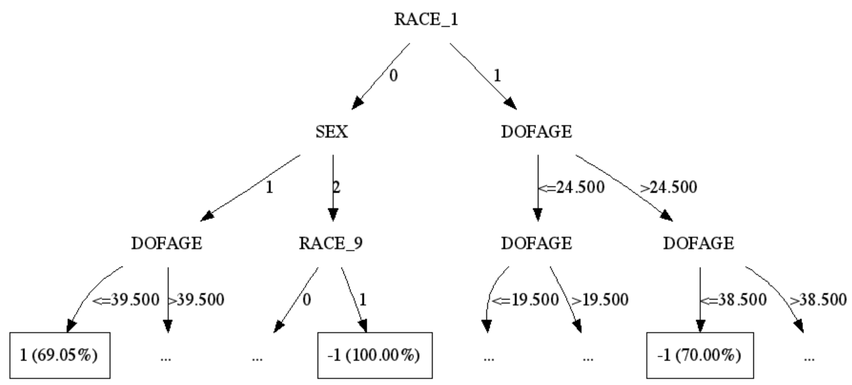


↑ *Top part of decision tree using CRIME*



↑ *Full decision tree using BASE*

↓ *Top part of decision tree using DEMO*



"I […] solemnly swear that I will administer justice without respect to persons, and do equal right to the poor and to the rich, and that I will faithfully and impartially discharge and perform all the duties incumbent upon me. […] under the Constitution and laws of the United States. So help me God."
– *Oath of United States Justices and Judges*

Conclusion & Remaining Work

"C4.5, using the CRIME variable set, was able to predict the criminal's sentence with 78.3% accuracy. "

Our preliminary results suggest that justice is indeed administered "without repsect to persons." Demographic facts about the offender were poor predictors for the sentence, while variables describing the crime were useful predictors.
It is odd that the full decision tree using BASE is not equivalent to the CRIME tree, when the BASE tree only included variables from CRIME. This is likely the result of pruning and/or tie-breaking, and an issue we need to investigate.
We will repeat our experiments using varying subsets of variables, and include more variables describing aspects of the criminal, offense, and trial that we have not incorporated yet.

References

Pennsylvania Commission on Sentencing. PENNSYLVANIA SENTENCING DATA, 1998. ICPSR version. State College, PA:Pennsylvania Commission on Sentencing, 2000. Ann Arbor, MI: Interuniversity Consortium for Political and Social Research, 2002.