Cornell University • Ithaca, NY

CS 4780 Final Project Report

Predicting Criminal Sentences

Justin Cheng
Yunchi Luo
Jane Jae Won Park
{jc882, yl477, jp624}@cornell.edu

December 16, 2011 [1]

[1]This version was last updated and generated on December 15, 2011.

Contents

1 Introduction

Civilizations have adopted systems of judicial guidelines to support judges and juries in decision-making. If two different individuals commit an identical or similar crime that should be given same sentences by the non-discriminatory nature of our laws, is it actually the case that the two end up receiving the same sentences? The sentence is, after all, determined by a human who uses some combination of his objective knowledge of law and subjective judgment, perhaps even some whimsical impulse. While sociologists, behavioral economists, and researchers of law and our justice system have investigated how race, age and various factors affect what kind of sentences are handed out to criminals, they have focused on the psychology and motivations behind these results, not the objective details of the observations themselves. Existing studies on criminal justice use traditional statistical techniques and limited case studies. We seek to employ machine learning techniques to provide an objective analysis regarding which factors of a criminal or the crime are most influential in determining the felon$\tilde{\text{O}}$s sentence.

2 Problem Definition and Methods

Can we recover the rules judges use to determine sentences by investigating information about the criminals and sentences they received? 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? 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?

2.1 Feature Selection

TODO: How did we select features?

COMPLETE Offense Completed/Inchoate

COUNTY Sentencing County

DAASS Drug and Alcohol Assessment

DISP Type of Disposition

DOB Date of Birth

DOF Date of Offense

DOSAGE Age at Sentencing

DOS Date of Sentence

FINE Amount of Fine Imposed

GRADE Statutory Offense Grade

INCMIN Minimum Length of Incarceration

INCMAX Maximum Length of Incarceration

INCTYPE Type of Incarceration

PCSOFF PCS Offense Code

PCSSUB PCS Offense Subcode

RACE Ethnicity of Offender

SEX Gender of Offender

2.2 Feature Processing

***b/nb* Binarization** 'Split' a multi-valued label into multiple binary variables for each label (i.e., one-hot encoding of the label)

***c/nc* Coarsification** Place continuous data into a number of evenly distributed buckets, so that there are fewer values to work with.

***_/uno* Date field separation** Process date data (MM/DD/YYYY) into month, day, year, day of week, or unix time.

_/balN Sorting label values into N buckets, and whether the sample contained equal number of examples from each bucket

2.3 Datasets

DEMO
SEX, DOFAGE, RACE (should add DAASS in the future)

CRIME
FINE, GRADE, COMPLETE, DOS_UNO, COUNTY, PCSOFF, PCSSUB, DOF_UNO, DISP

HIST
TODO

ABOUT
TODO

BASE
Features from both DEMO and CRIME

The DEMO subset only contains demographic information, while the CRIME subset contains only details about the offense committed. The BASE subset combines features from both subsets. In all subsets, the label being predicted is INCMIN.

2.4 Linear Regression

We performed linear regression on all pairs of attributes, and calculated the correlation coefficient, $\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y}$, for each of them. To validate the correlation coefficients obtained, we calculated the p-value for the null hypothesis that the correlation coefficient of two attributes is 0.

2.5 Learning Algorithms

2.5.1 Decision Tree Learning

We used an the Orange implementation of the C4.5 Decision Tree Algorithm.

Parameter Tuning Parameter tuning was carried out by trying out all possible combinations of parameters below, using the classification accuracy obtained on 10-fold cross-validation.

- Parameter m from 0.1 to 100
  - Parameter used in computing M-estimate probabilities during post-pruning
- Maximum majority from 0.5 to 1.0
  - Induction stops when a node has a majority at least this value
- Minimum examples per node from 0 to 10
  - Minimum examples in each leaf
- Minimum subset from 0 to 10
  - Minimum examples in non-null leaves
- Measures - information gain, information gain ratio, gini index, relief

2.5.2 kNN Learning

Parameter Tuning

- k from 1 to $2\sqrt{n}$
  - Since a good value of k to use is generally $\sqrt{n}$

2.6 Regression

In addition to the binary classification task detailed above, we also used regression to find out how well we could predict features like INCMIN if they were instead continuous. We used two methods - linear regression, and SVM regression. The reason for the use of these two methods is that linear regression is fast and in many cases a sufficient model in regression classification tasks. SVM regression, while significantly slower, allows the use of kernels which could possibly improve our accuracy, which we quantify in terms of R^2.

We also employ the use of principle component analysis (PCA), in order to better understand which features can best predict our target variable, as well as explore the possibility of reducing the dimensionality of our data without sacrificing accuracy (i.e. principle component regression, or using the components generated by PCA in linear regression).

As in the binary task, we trained a regression classifier on 80% of the data, while validating it on the remaining 20%.

3 Experimental Evaluation

3.1 Methdology

TODO: overview of evaluation methodology, what we used to measure performance

TODO: training/test data used, why is it realistic or interesting?

3.2 Results

TODO: change all tables to charts

3.2.1 Cross-Validation Set Accuracy

The classification accuracies listed below is the accuracy obtained on 10-fold cross-validation.

DEMO

Algorithm	CA	TP	FP	FN	TN
bayes	0.629	284	153	218	345
c4.5	0.659	353	192	149	306
lin. svm	0.637	277	138	225	360
svm	0.651	?	?	?	?
kmeans	0.643	337	192	165	306

CRIME

Algorithm	CA	TP	FP	FN	TN
bayes	0.772	392	118	110	380
c4.5	0.783	404	119	98	379
lin. svm	0.583	276	191	226	307
kmeans	0.769	381	110	121	388

BASE

Algorithm	CA	TP	FP	FN	TN
bayes	0.761	390	127	112	371
c4.5	0.774	396	120	106	378
lin.svm	0.589	290	199	212	299
kmeans	0.773	400	125	102	373

3.2.2 Validation Set Accuracy

Algorithm	DEMO	CRIME	BASE
bayes	0.629	0.765	0.753
c4.5	0.610	0.762	0.754
lin.svm	0.633	0.600	0.603
svm	0.599	?	?
kmeans	0.616	0.777	0.758

3.2.3 Variable Correlation

Correlation Heat Map The correlation heat maps in the appendix show the correlation between all tested attributes.

Top Correlated Attributes Ignoring correlation between binarized attributes (ex. GRADE_1 and GRADE_2), and trivial correlations such as that between DOSAGE and DOFAGE and PCSOFF and GRADE (since unique PCSOFFs are assigned unique GRADEs in the codebook), we obtain

Attribute X	Attribute Y	$\rho_{X,Y}$	p-value
COUNTY_1	PCSOFF_182902	0.817	2.2×10^{-16}
PCSOFF_183301	ARSA	0.628	2.2×10^{-16}
COUNTY_53	PCSOFF_184905	0.577	2.2×10^{-16}
PCSOFF_757122	M1DUIC	0.576	2.2×10^{-16}
ROBA	WEA	0.547	2.2×10^{-16}
COUNTY_51	DISP_4	0.541	2.2×10^{-16}
F1A	ROBSBIC	0.534	2.2×10^{-16}
MIS	PRS	0.513	2.2×10^{-16}
PCSOFF_185506	SEXASLTC	0.500	2.2×10^{-16}
COUNTY_39	PCSOFF_183301	0.490	2.2×10^{-16}

For example, according to the high correlation between $COUNTY\_1$ and $PCSOFF\_182902$, if you're convicted in Adams county, you're likely to have unlawfully restrained someone and vice versa. If you've had prior arson adjudications, you've probably committed arson again. Somewhat unexpectedly, if you've had prior DUI convictions, you've probably altered or forged vehicle license plates. If you've had a prior robbery adjudication, naturally you've also probably had a weapon misdemeanor adjudication.

3.2.4 Decision Tree

Optimal Parameters

Dataset	m	Max. majority	Min. examples	Min. subset	Measure	Accuracy
DEMO	5	1.0	2	2	Gini	0.659
CRIME	100	1.0	2	2	Relief	0.783
BASE	100	0.8	1	2	Relief	0.774

Optimal Decision Tree for DEMO

```
RACE_1=0
|    SEX=1
|    |    DOFAGE<=39.500: 1 (69.05%)
|    |    DOFAGE>39.500
|    |    |    RACE_9=1: -1 (100.00%)
|    |    |    RACE_9=0
|    |    |    |    . . .
|    SEX=2
|    |    RACE_9=1: -1 (100.00%)
```

```
|    |    RACE_9=0
|    |    |    DOFAGE<=32.000: 1 (61.54%)
|    |    |    DOFAGE>32.000
|    |    |    |    . . .
RACE_1=1
|    DOFAGE<=24.500
|    |    DOFAGE<=19.500
|    |    |    DOFAGE<=17.500: 1 (80.00%)
|    |    |    DOFAGE>17.500: -1 (75.47%)
|    |    DOFAGE>19.500
|    |    |    SEX=1: 1 (57.94%)
|    |    |    SEX=2: -1 (83.33%)
|    DOFAGE>24.500
|    |    DOFAGE<=38.500: -1 (70.00%)
|    |    DOFAGE>38.500
|    |    |    DOFAGE>43.500: -1 (76.92%)
|    |    |    DOFAGE<=43.500
|    |    |    |    . . .
```

3.2.5 SVM

Dataset	C	t	t-dependent. vars	Accuracy
DEMO	5	2 (RBF)	$g = 0.05$	0.651

3.2.6 kNN

Dataset	k	Accuracy
DEMO	55	0.643
CRIME	28	0.769
BASE	42	0.773

3.2.7 Significance Tests

Binomial Sign Test Each cell in the table below corresponds to whether the learning algorithm in that row was significantly BETTER than the algorithm in that column, WORSE or similar ($\sim$).

DEMO

Algorithm	C4.5 (Tuned)	SVM	kNN (Tuned)	Bayes
C4.5 (Tuned)	-	$\sim$	$\sim$	$\sim$
SVM	-	-	$\sim$	$\sim$
kNN (Tuned)	-	-	-	$\sim$

CRIME

Algorithm	C4.5 (Tuned)	SVM	kNN (Tuned)	Bayes
C4.5 (Tuned)	-	BETTER	$\sim$	$\sim$
SVM	-	-	WORSE	WORSE
kNN (Tuned)	-	-	-	$\sim$

BASE

Algorithm	C4.5 (Tuned)	SVM	kNN (Tuned)	Bayes
C4.5 (Tuned)	-	BETTER	$\sim$	$\sim$
SVM	-	-	WORSE	WORSE
kNN (Tuned)	-	-	-	$\sim$

3.2.8 Regression Analysis

Linear Regression The results of linear regression on each dataset are summarized in Table 1. Here, we see that the Akaike information criterion (AIC) is lowest for ALL, given that every other feature set is a subset of it. Surprisingly, ABOUT, despite having only 2 features, best predicts minimum incarceration. Nevertheless, we see here that using demographic features (DEMO) results in the highest AIC score, again demonstrating that by-and-large, sentences lengths are decided independent of one's age, gender or race.

Dataset	AIC	R^2	Attr.	Coeff.	p-value
DEMO	9.00×10^{4}	1.40×10^{-2}	COUNTY48	0.249	2×10^{-16}
			COUNTY39	0.183	4.73×10^{-16}
CRIME	7.41×10^{4}	3.55×10^{-2}	PCSOFF751543	2.76	4.73×10^{-7}
			PCSOFF353009	1.34	1.19×10^{-2}
ABOUT	1.57×10^{4}	3.44×10^{-3}	SEXPRED	-3.78×10^{-1}	8.55×10^{-4}
			-	-	-
HIST	2.86×10^{4}	-7.82×10^{-3}	RAPC	7.28×10^{-1}	3.49×10^{-4}
			F1C	7.34×10^{-2}	4.43×10^{-3}
ALL	4.48×10^{3}	1.77×10^{-3}	PCSOFF185503	1.08	3.58×10^{-3}
			COMPLETE	1.97×10^{-1}	2.68×10^{-4}

Table 1: Linear regression listing components with the highest coefficients

Principle Component Regression Principle component decomposition was performed on each dataset, and the top components generated, by proportion of variance in the dataset explained (so that $\# = 1$ refers to the component that explains the most variance), are listed in Table 2, along with the attributes with the highest weights that make up each component. Table 3 shows the results of linear regression on the transformed components.

In most cases, the components that have the highest coefficients in the linear regression are not the ones that explain the most variance. For example, # 17 and 18 for DEMO correspond to the components which weigh $COUNTY\_48$ and $COUNTY\_54$ most.

SVM Regression SVM regression was also performed, and the results summarized in Table 4.

Dataset	#	V.E.	Top 5 Components
DEMO	1	2.70×10^{-1}	RACE1, COUNTY67, COUNTY40, COUNTY54, SEX
	2	7.16×10^{-2}	SEX, RACE2, COUNTY2, RACE1, COUNTY23
CRIME	1	1.84×10^{-1}	PCSSUB_B, OGS, PCSSUB_D, PCSOFF183502, PCSOFF183701
	2	1.57×10^{-1}	DISP2, OGS, DISP4, PCSSUB_B, DISP5
ABOUT	1	9.67×10^{-1}	DRUGDEP, SEXPRED
	2	3.27×10^{-2}	SEXPRED, DRUGDEP
HIST	1	4.13×10^{-1}	PRS, MIS, F3C, DRGC, F1C
	2	1.29×10^{-1}	MIS, AGC, M1CHILDA, WEA, M1DUIA
ALL	1	1.20×10^{-1}	RACE1, PCSOFF753731, PCSSUB_A, DISP1, DOFUNO
	2	8.55×10^{-2}	MIS, PRS, F3C, PCSSUB_A, DOFUNO

Table 2: PCA decomposition listing top components (#) by variance explained (V.E.) and makeup

Dataset	AIC	R^2	#	V.E.	Coeff.	p-value
DEMO	9.00×10^{4}	1.40×10^{-2}	17	1.19×10^{-2}	2.05×10^{-1}	$<2\times10^{-16}$
			18	1.14×10^{-2}	-2.07×10^{-1}	$<2\times10^{-16}$
CRIME	7.41×10^{4}	3.55×10^{-2}	111	1.03×10^{-4}	-1.64	2.58×10^{-15}
			113	1.02×10^{-4}	-1.89	$<2\times10^{-16}$
ABOUT	1.57×10^{4}	3.44×10^{-3}	2	3.27×10^{-2}	1.19×10^{-2}	8.55×10^{-4}
			-	-	-	-
HIST	2.86×10^{4}	-7.82×10^{-3}	37	7.21×10^{-4}	5.97×10^{-1}	7.71×10^{-3}
			30	1.14×10^{-3}	-3.54×10^{-1}	4.67×10^{-2}
ALL	4.48×10^{3}	1.77×10^{-3}	-	-	-	-
			-	-	-	-

Table 3: PCR listing significant components with the highest absolute coefficients

3.3 Discussion

TODO

4 Related Work

5 Future Work

6 Conclusion

7 References

1. Britt, Chester L., "Modeling the distribution of sentence length decisions under a guidelines system: An application of quantile regression models." Journal of Quantitative Criminology. Dec 2009, 25, (4), 341 - 370. DOI: 10.1007/s10940-009-9066-x

Dataset	R^2	Attr.	Coeff.
DEMO	???	??? ???	??? ???
CRIME	???	??? ???	??? ???
ABOUT	???	??? ???	??? ???
HIST	???	??? ???	??? ???

Table 4: SVM regression listing components with the highest coefficients (in linear case)

2. Johnson, Brian D., "Racial and ethnic disparities in sentencing departures across modes of conviction." Criminology. May 2003, 41, (2), 449 - 489. DOI: 10.1111/j.1745-9125.2003.tb00994.x

3. Ullmer, Jefferey T.; Johnson, Brian, "Sentencing in context: A multilevel analysis." Criminology. Feb 2004, 42, (1), 137 - 177. DOI: 10.1111/j.1745-9125.2004.tb00516.x

4. Ulmer, Jeffery T.; Bradley, Minda S., "Variation in trial penalties among serious violent offenses." Criminology. Aug 2006, 44, (3), 631 - 670. DOI: 10.1111/j.1745-9125.2006.00059.x

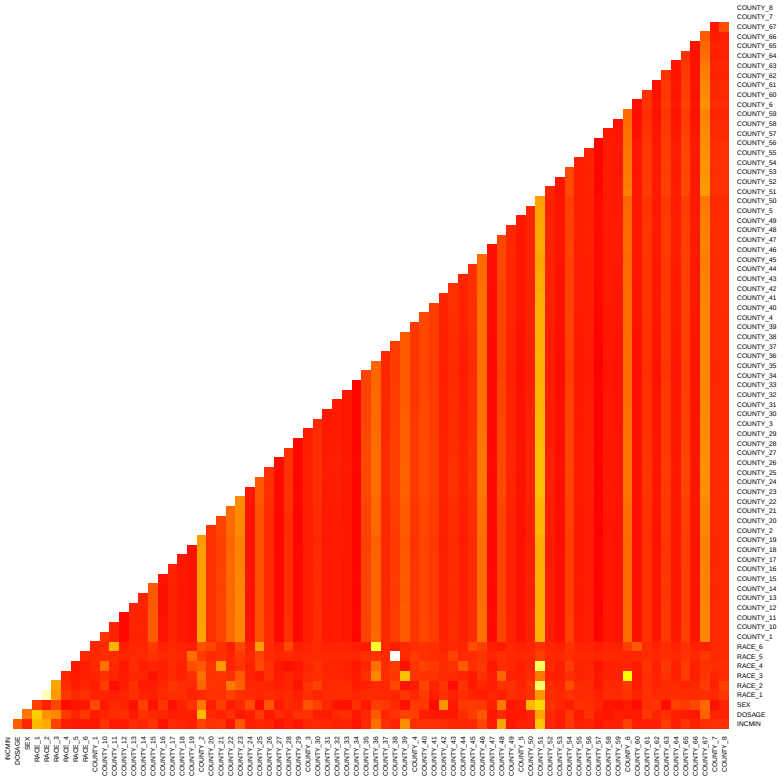
5. Uri J. Schild and Ruth Kannai. 2003. Intelligent computer evaluation of offender's previous record. In Proceedings of the 9th international conference on Artificial intelligence and law (ICAIL '03). ACM, New York, NY, USA, 185-194. DOI=10.1145/1047788.1047831 http://doi.acm.org/10.1145/1047788.1047831

8 Appendix

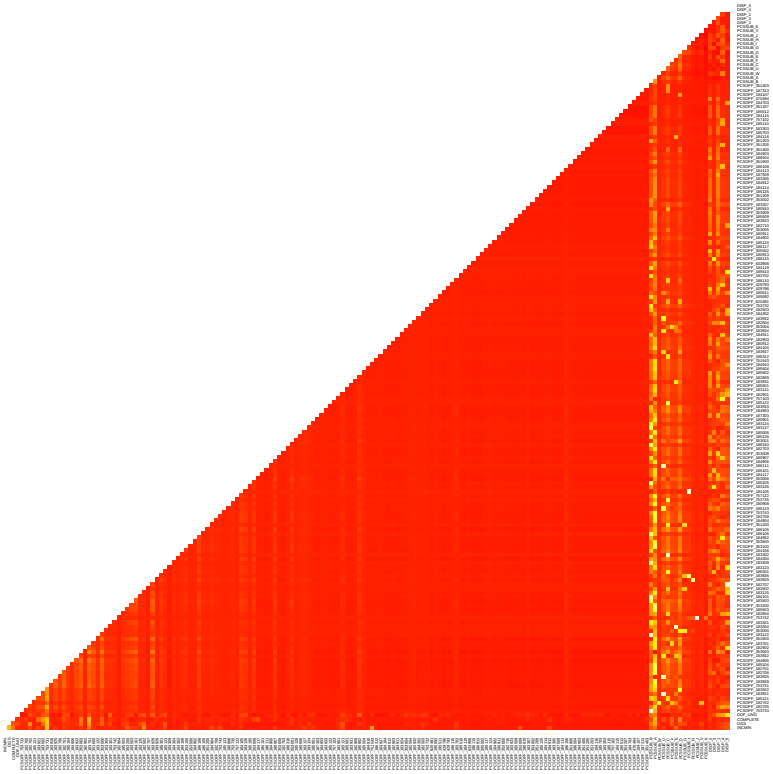
8.1 Correlation Heat Maps

Red indicates low correlation, light yellow indicates high correlation.

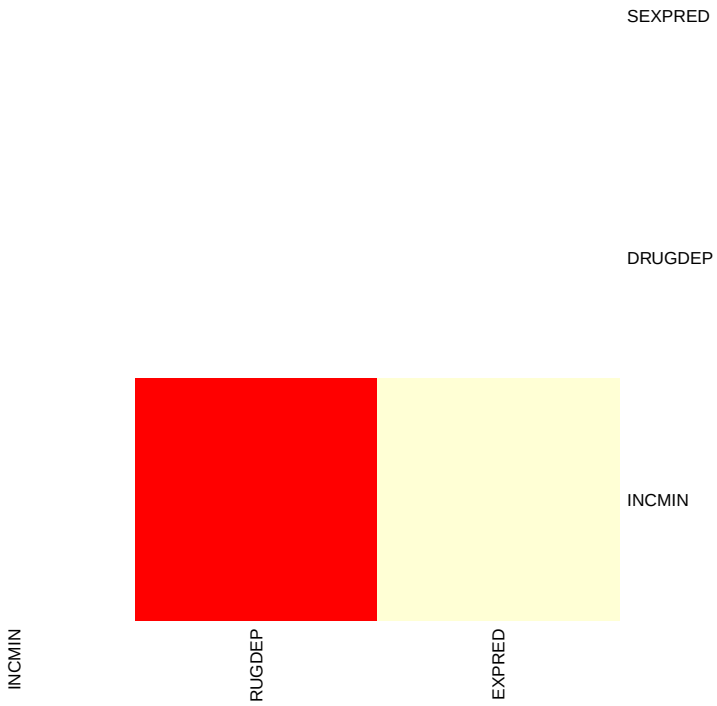
8.1.1 DEMO Correlation Heat Map



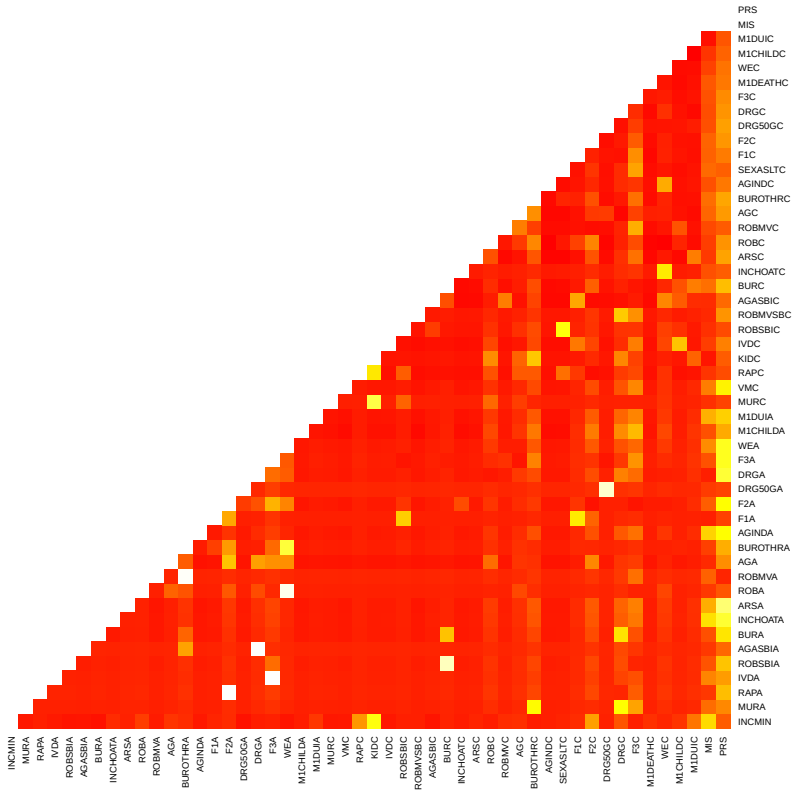
8.1.2 CRIME Correlation Heat Map



8.1.3 ABOUT Correlation Heat Map



8.1.4 HIST Correlation Heat Map



8.1.5 ALL Correlation Heat Map