DATA 500
Machine Learning I

TuTh 3:30 - 4:45

Dr. Ousley
Office: Hammermill 413

Office Hours: **Tu Th 12:00 - 3:00 PM**
- and by appointment

Email: sousley@mercyhurst.edu
Office phone: 824-3116

A little about me…

BA, U of Maryland (Anthropology)
MA, PhD U of Tennessee (Anthropology, minor in Statistics)
- worked on anthropological databases (Paradox DOS)
- Forensic Data Bank (Paradox DOS)
- Dermatoglyphics Database (Paradox DOS)
- programmed Fordisc 1; 2 (True Basic; Visual Basic)

Forensic Anthropology
(Human skeletal variation, sex, ancestry estimation)

Biological Anthropology
(Genetic, non-metric, and metric variation in humans)

A little about me…

9 years at the Smithsonian Institution
(Repatriation Osteology Laboratory, NMNH)
- Sybase ASA databases
- Advantage DB databases
- programmed data entry application Osteoware (Delphi)
- programmed data collection from 3-D digitizer (Delphi)
- programmed Fordisc 3 (Delphi)

10 years at Mercyhurst
- created Patricia (Pediatric Radiology Interactive Atlas; MySQL)
  (PI on grant from National Institute of Justice)
- MU purchased a web server

http://math.mercyhurst.edu/~sousley/databases/radiographic_database/

A little about me…

Ongoing projects

- Upgrade Patricia interface, include more demographic data
- Migrate FDB to Mercyhurst server (MySQL)
- Set up several more biological databases on Mercyhurst server
- Program Fordisc 4 (Python, R, others)
- Program data entry and analysis of skeletal age indicators
  (co-PI on grant from National Institute of Justice)
- Program data entry and analysis of nonmetric ancestry indicators
  (consultant on grant from National Institute of Justice)
- Program data entry and visualization/analysis of archaeological data
  (Sybase ASA, MySQL)
  (database consultant to AMEC Foster Wheeler, Inc.)
- **General statistical programming, especially simulation/ Monte Carlo/bootstrapping/**

Other staff have different skills

My Philosophy

I do applied statistics and applied programming
- I don't understand everything from A to Z
- I acknowledge theory but try to not depend on theory too much
  (Lenin: "Trust, but verify") **ML is similar!**
- I test results! **ML is similar!**

There is no need to reinvent the wheel
- If code snippets are available that work, use them (after testing)
- You can improve on them later, use as a test criteria
- ***except for your homeworks! (an example of academic dishonesty)***

Sometimes you have to bang your head against the wall to learn
- After a certain point, you will get much better
- Different for everyone, but hard work gets you there

A little about you...

Where are you from?

Undergraduate major, university

Communications

I use email to communicate outside of class

Blackboard:

- You should be part of the class; if not, let me know by email

- has the most recent version of the syllabus

- has lecture notes in pdf format

- has chapter/article PDFs to be read

Overview of the Course

Using R and Rstudio
Statistical estimation
Statistical modeling
Probability
Data visualization
Data management
Error checking
Experimental design
Hypothesis testing
Correlation
Linear regression
Statistical validation
Bias and Variance
Overfitting
Feature selection, dimension reduction, and other transformations
Resampling (Cross-validation, Bootstrap, etc.)
Linear Discriminant Function Classification
Logistic regression / classification

Naïve Bayes classification (?)

What you will learn

Some traditional statistical approaches

Things to ALWAYS keep in mind in ANY analysis

The ML approach and philosophy

Some ML concepts from regression to classification

Coding in R and RStudio

Readings

The main textbook:

Practical Statistics for Data Scientists (2017) by Bruce and Bruce

We will also read chapter PDFs from other books.

Why are there so many introductory statistics books?

Statistics

Hundreds of introductory books (none is perfect)

- need not be *sadistics*

BUT

- is like learning another language
- it is a science (theory, testable hypotheses)

- is a necessary language for other sciences
- is a necessary **tool**
- it deals with TRUTH and REALITY

Statistics may not be easy, but is NOT impossible!

Numbers and Language

You are already using the "language" of numbers!

What is seven hundred fifty-seven plus five hundred thirty-eight?

What is DCCLVII plus DXXXVIII?

What is 757 + 538?

757 +

538

Applied Statistics Quiz: Numbers

1. What is 3 + 6?

2. What is 3 - 6?

3. What is 3 * 6?

4. What is 3 / 6?

5. What is 3^6 ?

6. What is 3^{-6}

7. If $x = 6$ and $y = 2$, what is x^y ?

What is Statistics?
How can statistics help us?

Statistics is a scientific tool

Statistics is the science of prediction, approximation, and (un)certainty

Approximation
Certainty
Prediction
Probability

Statistics provides the best **model** of "**R**eality"

“reality” must be based on SAMPLES

ID	ContNum	Sex	Pop	Ethnicity	BirthYear	Age	GOL	NOL	BNL	BBH	XCB	XFB	WFB	ZYB	AUB	WCB	ASB	BPL	NPH	NLH	JUB	NLB
F1019	T-1605	F	B		1921	41	183	182	100	131	140	120	105	129	119	71	104	102	68	50	113	24
F102	OMI-524-74	M	AMIN	NAVAJO	1930	44	172	166	103	145	155	129	100	141	137	75	117	92	68	53	120	25
F1020	T-1606	M	B		1929	34	179	176	102	132	130	113	96	122	112	67	106	103	63	47	108	22
F1022	T-1615	M	B		1927	27	192	191	108	135	139	121	99	133	122	71	108	110	71	53	114	25
F103	OMI-7049-76	M	AMIN	AMIN	1926	50	170	165	94	141	155	128	88	138	134	78	115	90	69	51	124	30
F1033	W-1949	F	B		1911	19	178	177	95	126	131	114	97	123	111	69	99	94	67	48	110	22
F1035	W-2065	F	B		1912	19	175	173	85	119	137	118	95	125	116	71	108	88	64	47	110	23
F104	OMI-7044-77	M	AMIN	NAVAJO	1955	22	156	150	90	132	141	109	87	125	123	68	109	89	60	46	105	21
F1042	W-2939	F	W		1911	25	168	164	93	124	139	112	91	120	114	70	106	88	57	45	104	22
F105	UNM-4	M	W		1924	52	187	184	105	145	153	134	107	136	132	80	116	90	68	52	115	26
F1067	T-11R	F	B		1918	24	171	167	94	129	129	112	91	117	108	69	105	95	64	44	104	20
F1069	T-13R	M	B		1910	28	172	170	99	131	127	105	90	126	118	78	99	98	72	54	115	29
F1073	T-45R	M	H	MEXICAN	1910	32	191	189	105	140	138	122	101	136	127	70	116	97	73	54	115	26
F1080	T-171R	F	B		1910	29	170	170	96	123	132	112	93	121	115	69	102	92	64	46	104	24
F1082	T-187R	M	W		1917	29	182	178	95	142	128	110	89	123	118	73	111	85	68	49	107	22
F1083	SI-910;90-7261	F	W		1971	19	182	178	101	137	134	113	92	115	116	71	105	94	69	50	98	22
F1084	UT90-4F;FA89-172	F	W		1935	52	183	181	98	120	136	112	97	124	121	67	114	104	71	50	110	23
F1085	UT90-28F	M	W		1928	61	196	192	110	155	144	125	97	136	129	83	122	97	78	57	118	27
F1086	UT90-29F	M	W		1964	26				129	132				116							
F1087	UT89-23F	M	W		1936	53	178	175	102	144	148	131	97	130	121	71	116	94	66	50	116	25
F1088	UT90-34F	F	W		1965	24	175	174	94	138	135	117	92	115	110	78	108	82	69	55	106	22
F1089	UT91-10F	M	W		1963	27					144		97									
F109	LA86-6	M	B		1913	72	192	187	113	144	147	130			131							32
F1090	UT91-17F	F	W		1964	18	192	190	104	141	136	120	100	117	113		110	98	69	50	106	25
F1091	UT91-30F	M	W	GREEK	1935	54	181	178	106	140	143	123	94	128	122	73	109	98	68	52	108	23
F1092	UT91-36F	M	B		1961	30	193	188	107	137	129	108	89	129	118	70	117	108	66	49	116	31
F1093	UT89-24F	F	W		1941	48				135	138				120		113	90				24
F1094	UT90-40F	M	AMIN		1953	37				143	140				127		118					
F1095	UT91-42F	M	B		1939	40	195	191	102	136	133	114	98	130	121	80	115	95	73	53	122	24
F1096	LML26584-91	F	W		1945	46	181		101	136	129		94	115	108			87	70	53		21

What does Statistics do for us?

Statistics provides an abbreviation, an approximation, a model, of something in the real world, usually a population

A population may have millions of individual values, but thanks to sampling, we can boil down the most important information into a few numbers.

Important Questions need Statistics

Does smoking cause cancer?
Does smoking marijuana cause cognitive deficits?
Do cell phones cause cancer?
Does fracking harm the environment and human health?
How tall will my children be?

Does the Zika virus cause microcephaly?

Does prayer help healing?
Does spending money on social programs help the poor?

How should I invest money so I don't go broke?
Which hospitals are the safest?

The answers are often in terms of probabilities

Grading:

Homeworks

Take-home exams

Participation

Project?

90+: A; 80-90: B; etc.

WARNING: Academic dishonesty may result in an F

How to do well in this course

Set aside some time to:

Read the assigned chapter(s)

(study)

When you "read", experiment using the computer!

(experiment)

Do the homework

(apply concepts and methods)

Participate!

(ask good questions and provide thoughtful answers)

You will do well and pick up skills

For Thursday

Read Verzani chapters on Blackboard and experiment!