Let's Do Data Science

By: Allen Grimm
www.TheGrimmScientist.com
@GrimmScientist
TheGrimmScientist@gmail.com

Goals:

- Information theory crash course
- Cross Validated data as a test run.
- A few sample models
- Application of a model

A Coin Flip

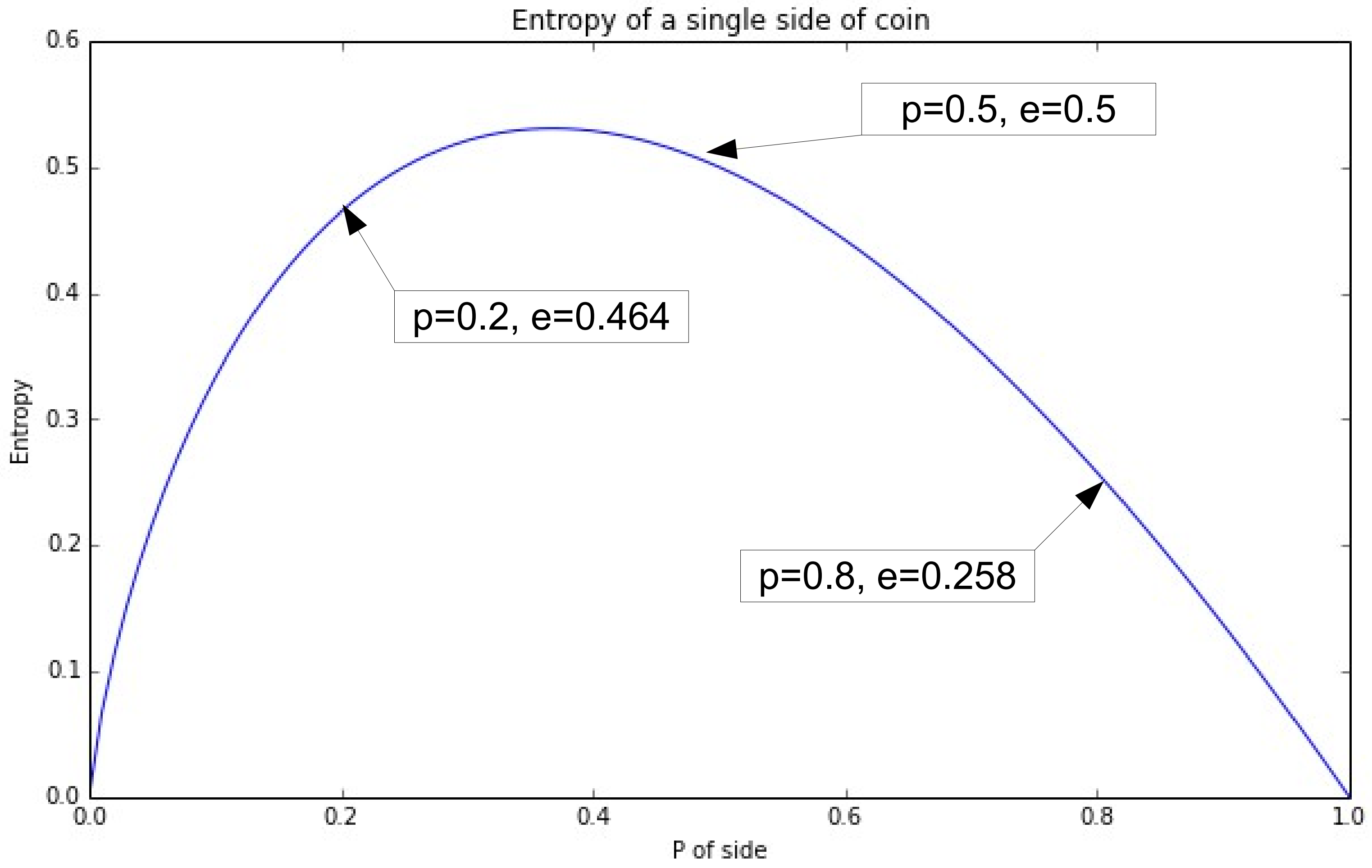
Balanced

	Heads	Tails
Probability	0.5	0.5

Biased

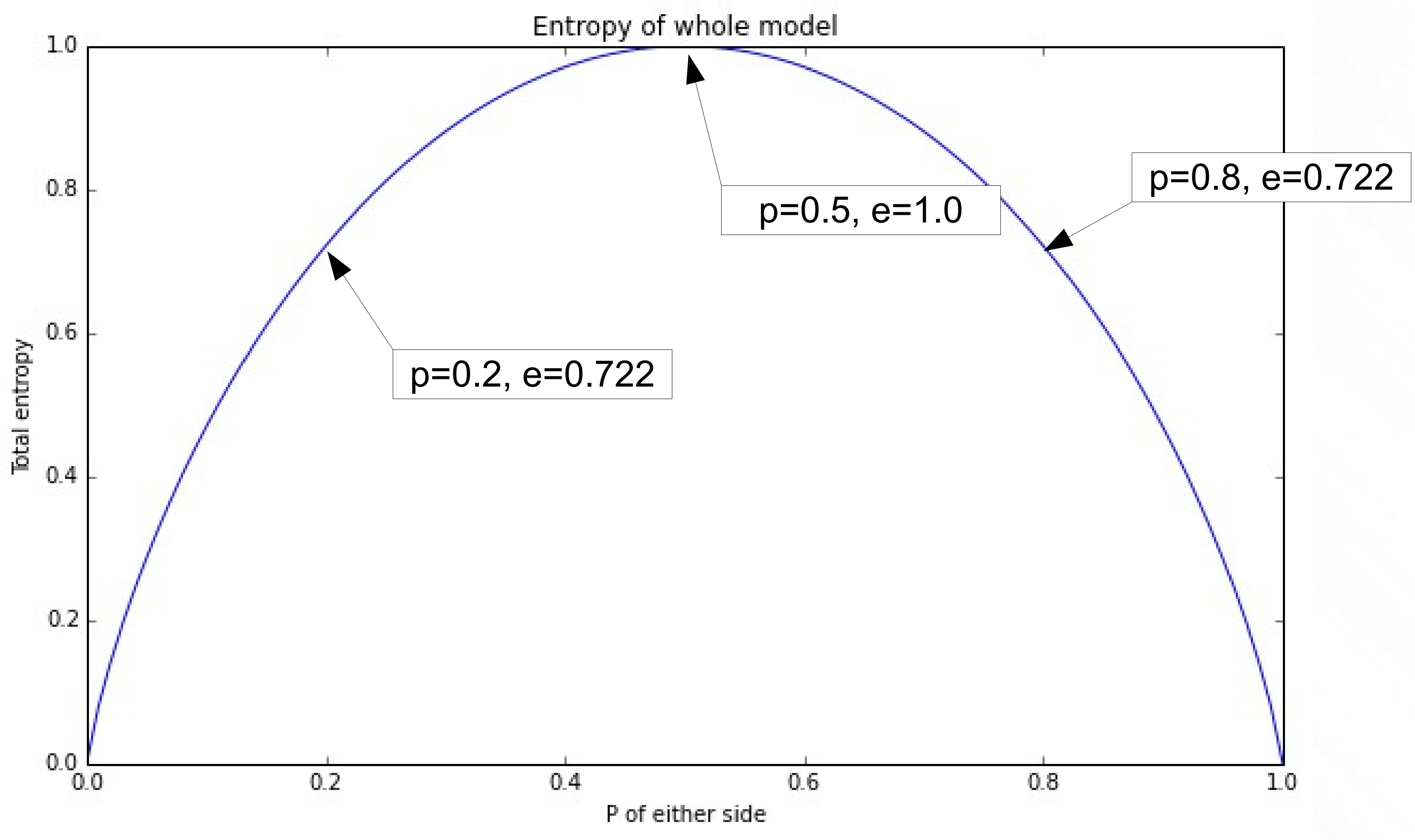
	Heads	Tails
Probability	0.2	0.8

One Side at a Time



$\text{entropy}_i = p_i * \log_2(p_i)$

Both Sides Together



A Coin Flip

Balanced

	Heads	Tails
Probability	0.5	0.5

Total entropy: 1.0

Biased

	Heads	Tails
Probability	0.2	0.8

Total entropy: 0.722

Uniform probabilities always result in maximal entropy.
=
The more biased the probabilities in a model, the more you know about the system ahead of time.

The Data

- Cross Validated post data
- 68,386 posts
- Variables:
  - Score “S”
  - Favorite count “F”
  - Answer count “A”
  - Comment count “C”
  - Body length “B”

Our Sample Data

Two Comparable Models

Model “SF”

Score

	F = 0	F > 0
S < 0	0.0112	0.0005
S = 0	0.1916	0.0086
S > 0	0.6592	0.1289

Favorites

entropy(SF) = 1.37

Model “SA”

	A = 0	A > 0
S < 0	0.0062	0.0054
S = 0	0.1522	0.0480
S > 0	0.5285	0.2596

Answers

entropy(SF) = 1.70

Models of Different Sizes

Model “SF”

Score	S < 0	0.0112	0.0005
	S = 0	0.1916	0.0086
	S > 0	0.6592	0.1289
		F = 0	F > 0
		Favorites	

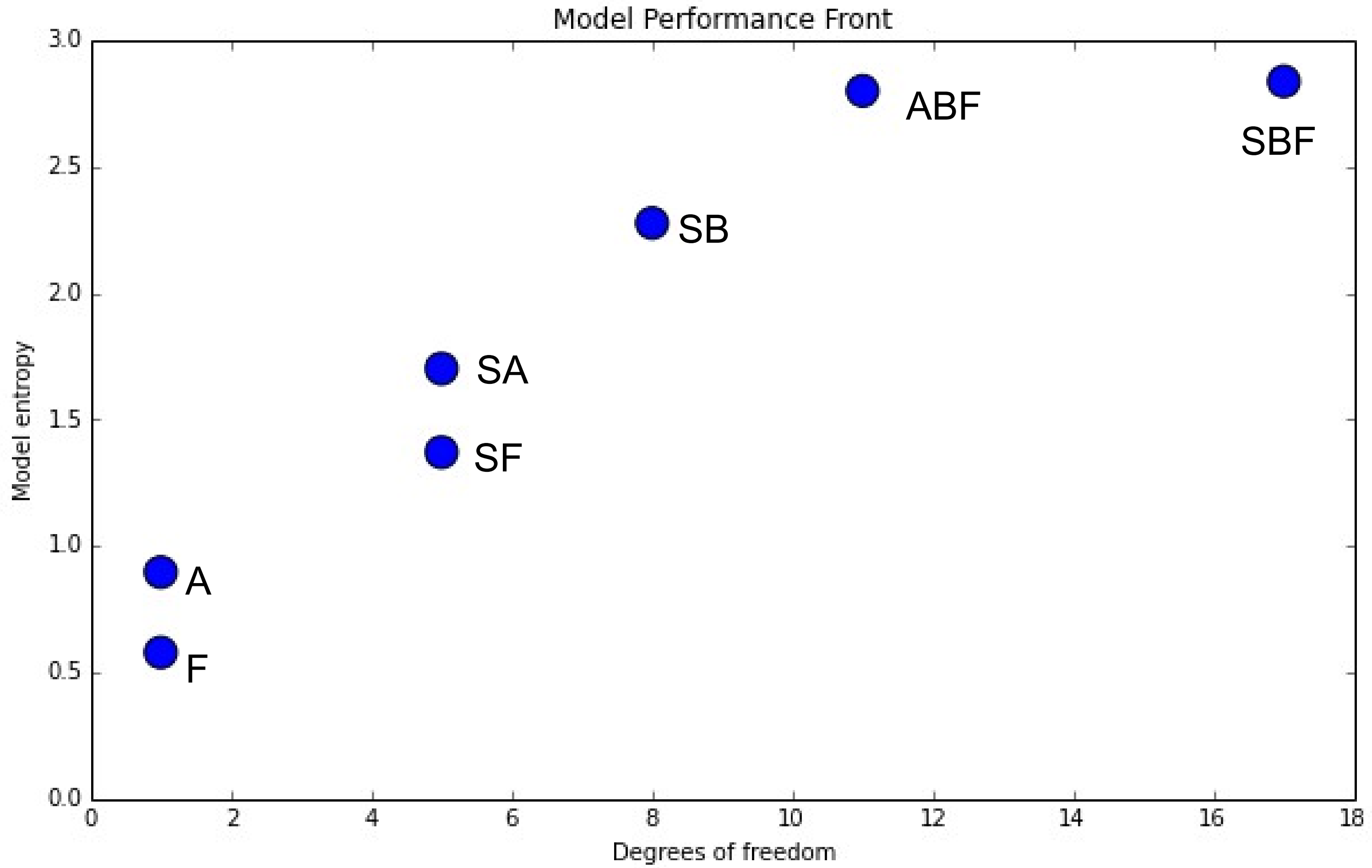
Model “SB”

Score	S < 0	0.0045	0.0035	0.0036
	S = 0	0.0490	0.0582	0.0930
	S > 0	0.1156	0.1910	0.4816
		B < 50	50<=B<100	B <= 100
		Body Length		

entropy(SF) = 1.37
degrees of freedom = 5

entropy(SF) = 2.18
degrees of freedom = 8

Model Selection



* Assumes all models are statistically significant.
* Sample size dimension not displayed, since all models have the same sample size

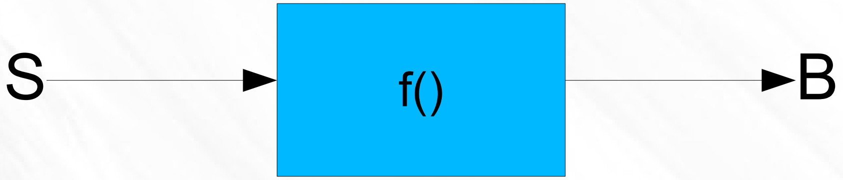
Predictions From a Model

- Given a positive, zero, or negative score, how long is your post?

Model “SB”

		B < 50	50<=B<130	B >= 130
Score	S < 0	0.0045	0.0046	0.0026
	S = 0	0.0490	0.0830	0.0682
	S > 0	0.1156	0.2916	0.3810

Body Length



S

f()

B

*Note issue of unbalanced distributions.

Predictions From a Model

- Given a zero score, how long is your post?

Model “SB”

Score		B < 50	50<=B<130	B >= 130
	S < 0	0.0045	0.0046	0.0026
	S = 0	0.0490	0.0830	0.0682
	S > 0	0.1156	0.2916	0.3810

Body Length

Body Length

B < 50	50<=B<130	B >= 130
0.0490	0.0830	0.0682

Given S = 0,
Expect 50<=B<130

Predictions From a Model

- Given a positive score, how long is your post?

Model “SB”

Score		B < 50	50<=B<130	B >= 130
	S < 0	0.0045	0.0046	0.0026
	S = 0	0.0490	0.0830	0.0682
	S > 0	0.1156	0.2916	0.3810

Body Length

Body Length

B < 50	50<=B<130	B >= 130
0.1156	0.2916	0.3810

Given S > 0,
Expect >=130

Model application

Credits, Other Resources

- www.github.com/TheGrimmScientist/DMM_Sim
  - Thanks to Ryan Price for working with me on that simulator and this example.
- http://occam.research.pdx.edu/occam/weboccam.cgi

Contact: TheGrimmScientist@gmail.com

Follow: @GrimmScientist