TALENT COURSE 11
LEARNING FROM DATA: BAYESIAN METHODS AND MACHINE LEARNING

Lecture 10: Choosing a prior

Daniel Phillips
Ohio University
TU Darmstadt
ExtreMe Matter Institute

TALENT Course 11 is possible thanks to funding from the STFC

BAYES' THEOREM

Thomas Bayes (1701?-1761)

$$\mathrm{pr}(A\,|\,B, I) = \frac{\mathrm{pr}(B\,|\,A, I)\,\mathrm{pr}(A\,|\,I)}{\mathrm{pr}(B\,|\,I)}$$

BAYES' THEOREM

Thomas Bayes (1701?-1761)

http://www.bayesian-inference.com

$$\mathrm{pr}(A \mid B, I) = \frac{\mathrm{pr}(B \mid A, I)\,\mathrm{pr}(A \mid I)}{\mathrm{pr}(B \mid I)}$$

BAYES' THEOREM

Thomas Bayes (1701?-1761)

$$\mathrm{pr}(A \mid B, I) = \frac{\mathrm{pr}(B \mid A, I)\,\mathrm{pr}(A \mid I)}{\mathrm{pr}(B \mid I)}$$

http://www.bayesian-inference.com

$$\mathrm{pr}(\mathrm{model} \mid \mathrm{data}, I) = \frac{\mathrm{pr}(\mathrm{data} \mid \mathrm{model}, I)\,\mathrm{pr}(\mathrm{model} \mid I)}{\mathrm{pr}(\mathrm{data} \mid I)}$$

BAYES' THEOREM

Thomas Bayes (1701?-1761)

$$\mathrm{pr}(A \mid B, I) = \frac{\mathrm{pr}(B \mid A, I)\,\mathrm{pr}(A \mid I)}{\mathrm{pr}(B \mid I)}$$

http://www.bayesian-inference.com

$$\mathrm{pr}(\mathrm{model} \mid \mathrm{data}, I) = \frac{\mathrm{pr}(\mathrm{data} \mid \mathrm{model}, I)\,\mathrm{pr}(\mathrm{model} \mid I)}{\mathrm{pr}(\mathrm{data} \mid I)}$$

Posterior

BAYES' THEOREM

Thomas Bayes (1701?-1761)

$$\mathrm{pr}(A\,|\,B, I) = \frac{\mathrm{pr}(B\,|\,A, I)\mathrm{pr}(A\,|\,I)}{\mathrm{pr}(B\,|\,I)}$$

http://www.bayesian-inference.com

Likelihood

$$\mathrm{pr}(\mathrm{model}\,|\,\mathrm{data}, I) = \frac{\mathrm{pr}(\mathrm{data}\,|\,\mathrm{model}, I)\mathrm{pr}(\mathrm{model}\,|\,I)}{\mathrm{pr}(\mathrm{data}\,|\,I)}$$

Posterior

BAYES' THEOREM

Thomas Bayes (1701?-1761)

http://www.bayesian-inference.com

$$\mathrm{pr}(A \mid B, I) = \frac{\mathrm{pr}(B \mid A, I)\,\mathrm{pr}(A \mid I)}{\mathrm{pr}(B \mid I)}$$

Likelihood

Prior

$$\mathrm{pr}(\text{model} \mid \text{data}, I) = \frac{\mathrm{pr}(\text{data} \mid \text{model}, I)\,\mathrm{pr}(\text{model} \mid I)}{\mathrm{pr}(\text{data} \mid I)}$$

Posterior

BAYES' THEOREM

Thomas Bayes (1701?-1761)

http://www.bayesian-inference.com

$$\mathrm{pr}(A \mid B, I) = \frac{\mathrm{pr}(B \mid A, I)\,\mathrm{pr}(A \mid I)}{\mathrm{pr}(B \mid I)}$$

Likelihood

Prior

$$\mathrm{pr}(\mathrm{model} \mid \mathrm{data}, I) = \frac{\mathrm{pr}(\mathrm{data} \mid \mathrm{model}, I)\,\mathrm{pr}(\mathrm{model} \mid I)}{\mathrm{pr}(\mathrm{data} \mid I)}$$

Posterior

Normalization

BAYES' THEOREM

Thomas Bayes (1701?-1761)

http://www.bayesian-inference.com

$$\mathrm{pr}(A \mid B, I) = \frac{\mathrm{pr}(B \mid A, I)\,\mathrm{pr}(A \mid I)}{\mathrm{pr}(B \mid I)}$$

Likelihood

Prior

$$\mathrm{pr}(\mathrm{model} \mid \mathrm{data}, I) = \frac{\mathrm{pr}(\mathrm{data} \mid \mathrm{model}, I)\,\mathrm{pr}(\mathrm{model} \mid I)}{\mathrm{pr}(\mathrm{data} \mid I)}$$

Posterior

Normalization

Probability as degree of belief cf. frequentist view

INDIFFERENCE PRIORS: DISCRETE

INDIFFERENCE PRIORS: DISCRETE

- Consider a six-sided dice

INDIFFERENCE PRIORS: DISCRETE

- Consider a six-sided dice
- How do we assign $pr(X_i|I)$? ($i=1, 2, 3, 4, 5, 6$)

INDIFFERENCE PRIORS: DISCRETE

- Consider a six-sided dice
- How do we assign $\mathrm{pr}(X_i|I)$? (i=1, 2, 3, 4, 5, 6)
- We do know

$$\sum_i \mathrm{pr}(X_i \,|\, I) = 1$$

INDIFFERENCE PRIORS: DISCRETE

- Consider a six-sided dice
- How do we assign $\mathrm{pr}(X_i|I)$? ($i=1, 2, 3, 4, 5, 6$)
- We do know

$$\sum_i \mathrm{pr}(X_i\,|\,I) = 1$$

- Invariance under labeling$\Rightarrow\mathrm{pr}(X_i|I)=1/6$.

provided I says nothing that breaks the symmetry

INDIFFERENCE PRIORS: CONTINUOUS

INDIFFERENCE PRIORS: CONTINUOUS

- Location

$$\mathrm{pr}(x \,|\, I)\, dx = \mathrm{pr}(x + x_0 \,|\, I)\, dx$$

- Invariance under origin position⇒pr(x|I)=constant

provided I says nothing that breaks the symmetry

INDIFFERENCE PRIORS: CONTINUOUS

- Location

$$\mathrm{pr}(x\,|\,I)\,dx = \mathrm{pr}(x+x_0\,|\,I)\,dx$$

- Invariance under origin position⇒pr(x|I)=constant

provided I says nothing that breaks the symmetry

- Scale

$$\mathrm{pr}(\lambda\,|\,I)\,d\lambda = \mathrm{pr}(\beta\lambda\,|\,I)\,d(\beta\lambda)$$

- Invariance under unit choice ⇒ $\mathrm{pr}(\lambda\,|\,I) \propto 1/\lambda$

provided I says nothing that breaks the symmetry

- "Jeffreys prior": uniform prior on log(λ)

STRAIGHT LINE APPLICATION

STRAIGHT LINE APPLICATION

- First: rescale to get rid of units, i.e., express plot in terms of “natural units” for the problem

STRAIGHT LINE APPLICATION

- First: rescale to get rid of units, i.e., express plot in terms of "natural units" for the problem
- Intercept: location, scale, or something else?

STRAIGHT LINE APPLICATION

- First: rescale to get rid of units, i.e., express plot in terms of "natural units" for the problem
- Intercept: location, scale, or something else?
- Slope: location, scale, or something else?

STRAIGHT LINE APPLICATION

- First: rescale to get rid of units, i.e., express plot in terms of "natural units" for the problem
- Intercept: location, scale, or something else?
- Slope: location, scale, or something else?

$$\mathrm{pr}(m, b \,|\, I) \propto \frac{1}{(1 + m^2)^{3/2}}$$

Jaynes (1967); Jeffreys (1946)

"Symmetric prior"

STRAIGHT LINE APPLICATION

- First: rescale to get rid of units, i.e., express plot in terms of "natural units" for the problem
- Intercept: location, scale, or something else?
- Slope: location, scale, or something else?

$$\mathrm{pr}(m, b \,|\, I) \propto \frac{1}{(1+m^2)^{3/2}}$$

Jaynes (1967); Jeffreys (1946)

"Symmetric prior"

- Standard deviation: location, scale, or something else?

STRAIGHT LINE APPLICATION

- First: rescale to get rid of units, i.e., express plot in terms of "natural units" for the problem
- Intercept: location, scale, or something else?
- Slope: location, scale, or something else?

$$\mathrm{pr}(m, b \,|\, I) \propto \frac{1}{(1+m^2)^{3/2}}$$

Jaynes (1967); Jeffreys (1946)

"Symmetric prior"

- Standard deviation: location, scale, or something else?

$$\mathrm{pr}(\sigma \,|\, I) \propto \frac{1}{\sigma}$$

PRINCIPLE OF MAXIMUM ENTROPY

$$\sum_{i=1}^{6} p_i = 1$$

(something like this is always true)

$$\sum_{i=1}^{6} i p_i = 4.5$$

PRINCIPLE OF MAXIMUM ENTROPY

- Suppose we have some information on our dice, e.g.

$$\sum_{i=1}^{6} p_i = 1$$

(something like this is always true)

$$\sum_{i=1}^{6} i p_i = 4.5$$

PRINCIPLE OF MAXIMUM ENTROPY

- Suppose we have some information on our dice, e.g.

$$\sum_{i=1}^{6} p_i = 1$$

(something like this is always true)

$$\sum_{i=1}^{6} i p_i = 4.5$$

- Principle of maximum entropy seeks to maximize

$$S = -\sum_{i=1}^{6} p_i \log(p_i)$$

subject to these constraints

PRINCIPLE OF MAXIMUM ENTROPY

- Suppose we have some information on our dice, e.g.

$$\sum_{i=1}^{6} p_i = 1$$

(something like this is always true)

$$\sum_{i=1}^{6} i p_i = 4.5$$

- Principle of maximum entropy seeks to maximize

$$S = -\sum_{i=1}^{6} p_i \log(p_i)$$

subject to these constraints

- Define $Q(\{p_i\}; \lambda_0, \lambda_1) = -\sum_{i=1}^{6} p_i \log(p_i) + \lambda_0 \left(1 - \sum_i p_i\right) + \lambda_1 \left(4.5 - \sum_i i p_i\right)$

HAIR COLOR-HEIGHT EXAMPLE REVISITED

- 1/3 of all Australians are blond, and 1/4 are tall
- What proportion are blond and tall?

HAIR COLOR-HEIGHT EXAMPLE REVISITED

- 1/3 of all Australians are blond, and 1/4 are tall
- What proportion are blond and tall?

$$Q(\{p_i\};\lambda_0,\lambda_1,\lambda_2) = -\sum_{i=1}^{4} p_i \log(p_i) + \lambda_0\left(1-\sum_{i=1}^{4} p_i\right) + \lambda_1\left(\frac{1}{3}-p_1-p_2\right) + \lambda_2\left(\frac{1}{4}-p_1-p_3\right)$$

HAIR COLOR-HEIGHT EXAMPLE REVISITED

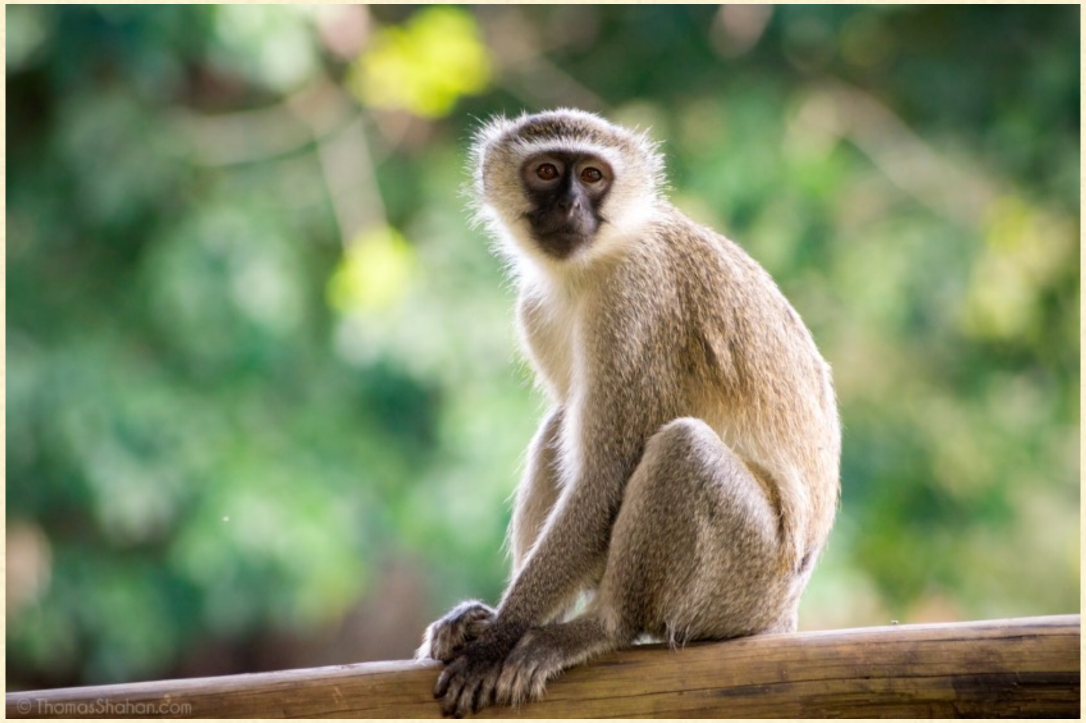
- 1/3 of all Australians are blond, and 1/4 are tall
- What proportion are blond and tall?

$$Q(\{p_i\};\lambda_0,\lambda_1,\lambda_2) = -\sum_{i=1}^{4} p_i \log(p_i) + \lambda_0\left(1-\sum_{i=1}^{4} p_i\right) + \lambda_1\left(\frac{1}{3}-p_1-p_2\right) + \lambda_2\left(\frac{1}{4}-p_1-p_3\right)$$

	BLOND	BROWN	SUMS
TALL	x	1/4-x	1/4
SHORT	1/3-x	5/12+x	3/4
SUMS	1/3	2/3	1

MONKEYS ARGUMENT

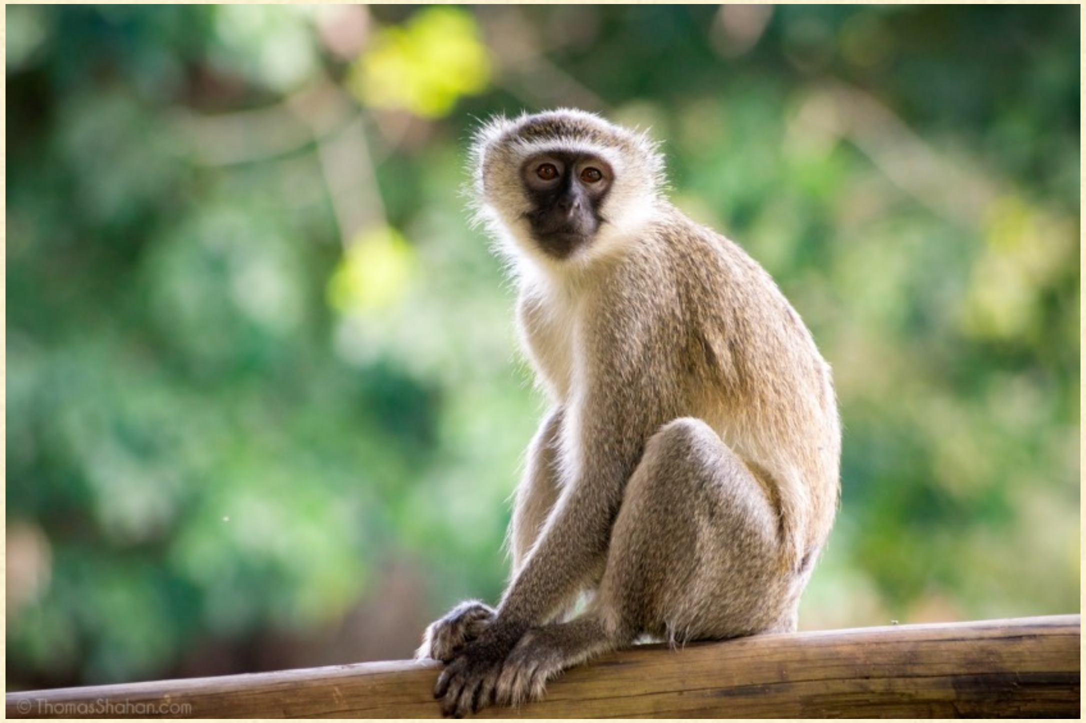
Thomas Shahan - https://www.flickr.com/photos/



- A model for assigning probabilities to M different alternatives: monkeys throwing N balls into M equally sized boxes

MONKEYS ARGUMENT

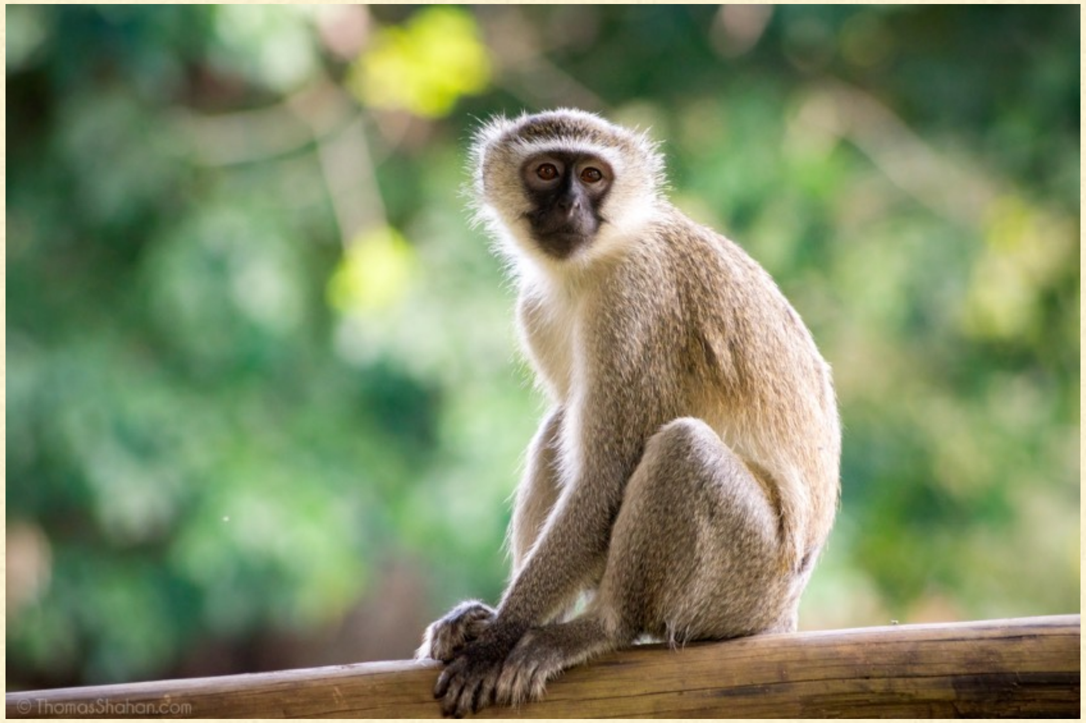
Thomas Shahan - https://www.flickr.com/photos/



- A model for assigning probabilities to M different alternatives: monkeys throwing N balls into M equally sized boxes
- Number of micro-states, W, as a function of $\{p_i\}$ is

MONKEYS ARGUMENT

Thomas Shahan - https://www.flickr.com/photos/



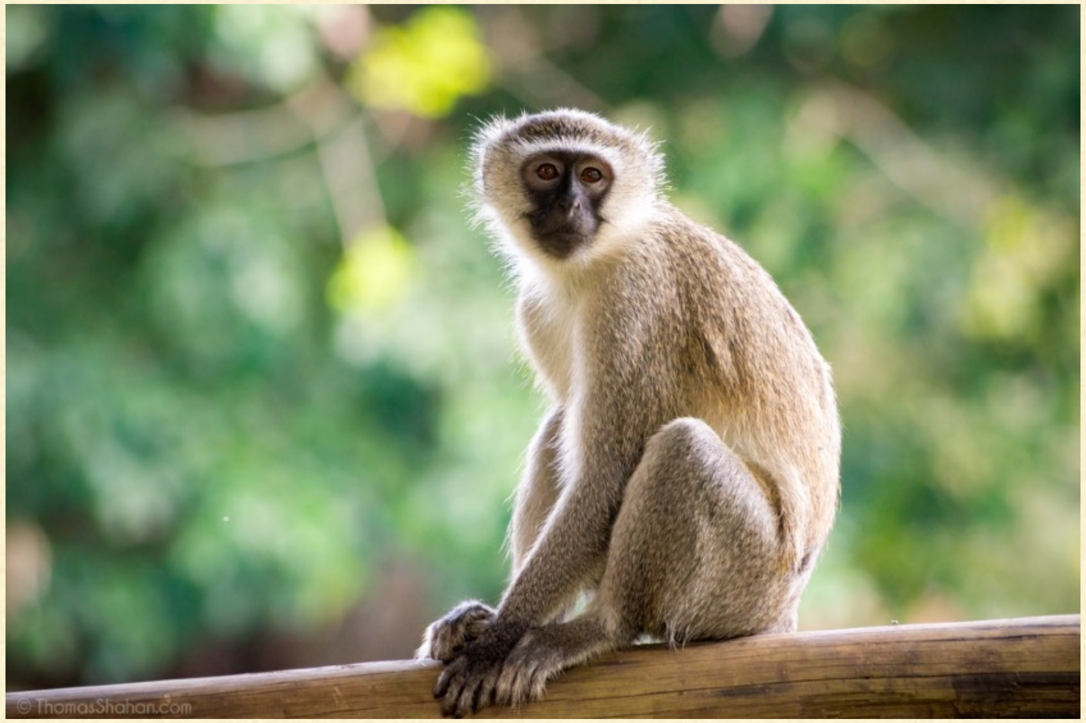
- A model for assigning probabilities to M different alternatives: monkeys throwing N balls into M equally sized boxes

- Number of micro-states, W, as a function of $\{p_i\}$ is

$$\log(W(\{n_i\})) = \log(N!) - \sum_{i=1}^{M} \log(n_i!)$$

MONKEYS ARGUMENT

Thomas Shahan - https://www.flickr.com/photos/



- A model for assigning probabilities to M different alternatives: monkeys throwing N balls into M equally sized boxes

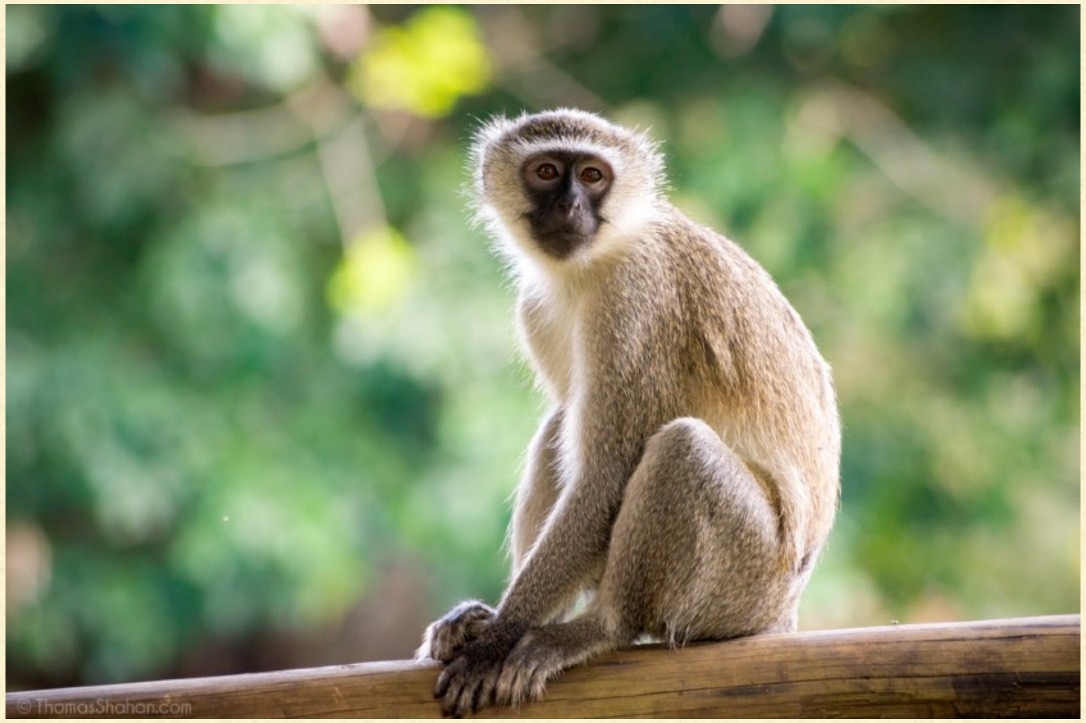
- Number of micro-states, W, as a function of $\{p_i\}$ is

$$\log(W(\{n_i\})) = \log(N!) - \sum_{i=1}^{M} \log(n_i!)$$

$$\log(W(\{n_i\})) \approx -N \sum_{i=1}^{M} p_i \log(p_i)$$

MONKEYS ARGUMENT

Thomas Shahan - https://www.flickr.com/photos/



- A model for assigning probabilities to M different alternatives: monkeys throwing N balls into M equally sized boxes

- Number of micro-states, W, as a function of $\{p_i\}$ is

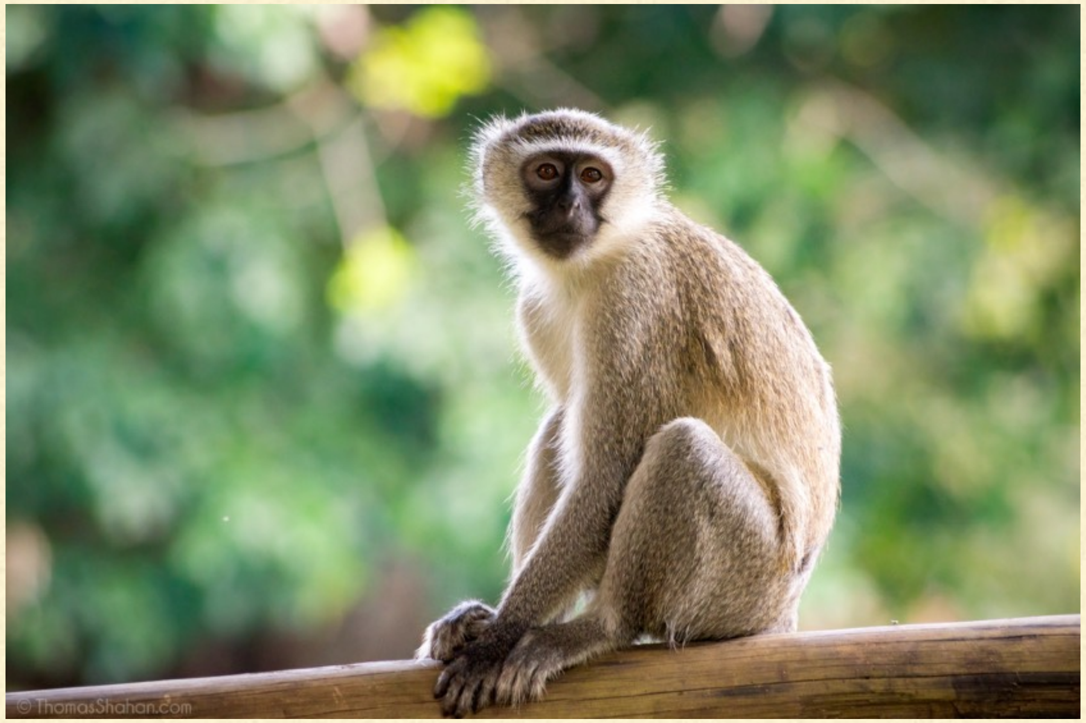
$$\log(W(\{n_i\})) = \log(N!) - \sum_{i=1}^{M} \log(n_i!)$$

$$\log(W(\{n_i\})) \approx -N \sum_{i=1}^{M} p_i \log(p_i)$$

So define S= $-\sum_{i=1}^{M} p_i \log(p_i)$

MONKEYS ARGUMENT

Thomas Shahan - https://www.flickr.com/photos/



- A model for assigning probabilities to M different alternatives: monkeys throwing N balls into M equally sized boxes

- Number of micro-states, W, as a function of $\{p_i\}$ is

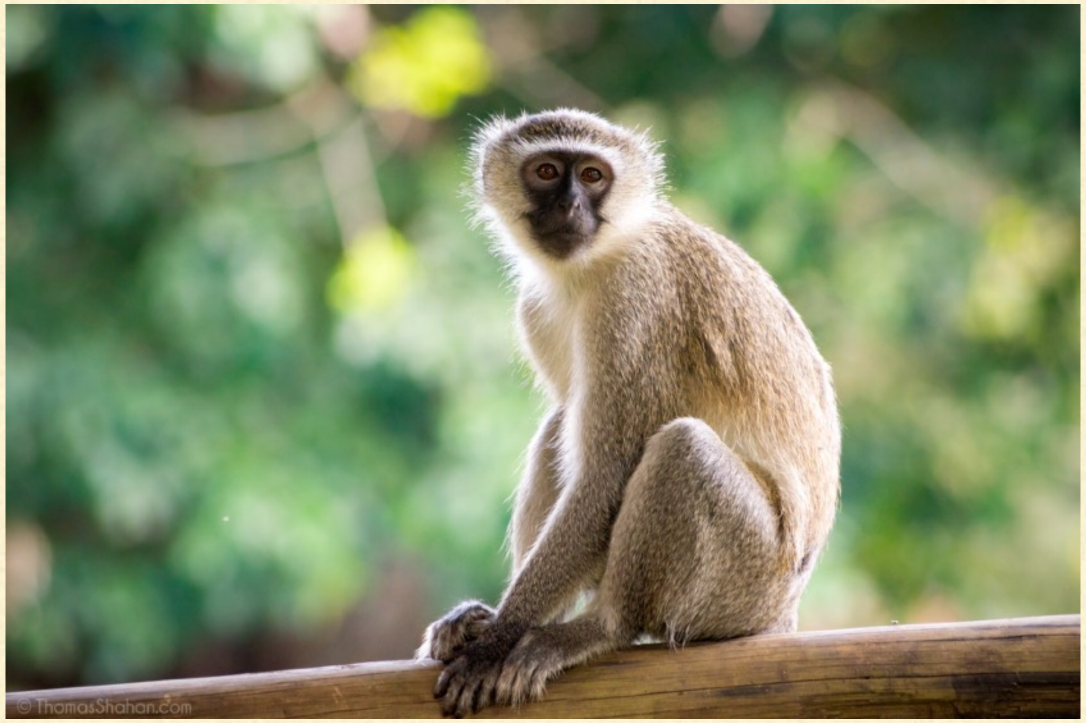
$$\log(W(\{n_i\})) = \log(N!) - \sum_{i=1}^{M} \log(n_i!)$$

$$\log(W(\{n_i\})) \approx -N \sum_{i=1}^{M} p_i \log(p_i)$$

So define $S = -\sum_{i=1}^{M} p_i \log(p_i)$ Entropy

MONKEYS ARGUMENT

Thomas Shahan - https://www.flickr.com/photos/



- A model for assigning probabilities to M different alternatives: monkeys throwing N balls into M equally sized boxes

- Number of micro-states, W, as a function of $\{p_i\}$ is

$$\log(W(\{n_i\})) = \log(N!) - \sum_{i=1}^{M} \log(n_i!)$$

$$\log(W(\{n_i\})) \approx -N \sum_{i=1}^{M} p_i \log(p_i)$$

So define $S = -\sum_{i=1}^{M} p_i \log(p_i)$ Entropy

Jaynes: derive thermo using MaxEnt to assign pdfs subject to macro constraints

WHY MAXIMIZE THE ENTROPY?

- Information theory: maximum entropy=minimum information (Shannon, 1948)
- Logical consistency (Shore & Johnson, 1960)
- Uncorrelated assignments related monotonically to S (Skilling, 1988)

WHY MAXIMIZE THE ENTROPY?

- Information theory: maximum entropy=minimum information (Shannon, 1948)
- Logical consistency (Shore & Johnson, 1960)
- Uncorrelated assignments related monotonically to S (Skilling, 1988)

VARIATIONAL FUNCTION	OPTIMAL X	IMPLIED CORRELATION
$-\sum p_i \log(p_i)$	0.0833	None
$-\sum p_i^2$	0.0417	Negative
$\sum \log(p_i)$	0.1060	Positive
$\sum \sqrt{p_i}$	0.0967	Positive

Sivia & Skilling, 2006

CONTINUOUS CASE

- Return to monkeys, but now with different probabilities for each bin. Then

$$S = -\sum_{i=1}^{M} p_i \log\left(\frac{p_i}{m_i}\right)$$

Shannon-Jaynes entropy
Kullback number
cross entropy

CONTINUOUS CASE

- Return to monkeys, but now with different probabilities for each bin. Then

$$S = -\sum_{i=1}^{M} p_i \log\left(\frac{p_i}{m_i}\right)$$

Shannon-Jaynes entropy
Kullback number
cross entropy

- Continuous case

CONTINUOUS CASE

- Return to monkeys, but now with different probabilities for each bin. Then

$$S = -\sum_{i=1}^{M} p_i \log\left(\frac{p_i}{m_i}\right)$$

Shannon-Jaynes entropy
Kullback number
cross entropy

- Continuous case

$$S[p] = -\int p(x)\,\log\left[\frac{p(x)}{m(x)}\right]$$

ONE COMMENT, TWO EXAMPLES

ONE COMMENT, TWO EXAMPLES

- $m(x)$ makes the expression invariant under a change of variables

ONE COMMENT, TWO EXAMPLES

- m(x) makes the expression invariant under a change of variables
- Example 1: maximize S subject only to normalization

ONE COMMENT, TWO EXAMPLES

- m(x) makes the expression invariant under a change of variables
- Example 1: maximize S subject only to normalization

$$Q[p;\lambda] = -\int dx\, p(x)\, \log\left[\frac{p(x)}{m(x)}\right] + \lambda\left(1 - \int dx\, p(x)\right)$$

Solution: p(x) proportional to m(x)

ONE COMMENT, TWO EXAMPLES

- m(x) makes the expression invariant under a change of variables
- Example 1: maximize S subject only to normalization

$$Q[p;\lambda] = -\int dx\, p(x)\, \log\left[\frac{p(x)}{m(x)}\right] + \lambda\left(1 - \int dx\, p(x)\right)$$

Solution: p(x) proportional to m(x)

- Example 2: maximize S[p] subject to constraint

$$\int dx\, p(x)(x-\mu)^2 = \sigma^2$$

ONE COMMENT, TWO EXAMPLES

- m(x) makes the expression invariant under a change of variables
- Example 1: maximize S subject only to normalization

$$Q[p;\lambda] = -\int dx\, p(x)\, \log\left[\frac{p(x)}{m(x)}\right] + \lambda\left(1 - \int dx\, p(x)\right)$$

Solution: p(x) proportional to m(x)

- Example 2: maximize S[p] subject to constraint

$$\int dx\, p(x)(x-\mu)^2 = \sigma^2$$

Solution: p(x) proportional to m(x)