Learning from data: Assigning probabilities

Christian Forssén

Department of Physics, Chalmers University of Technology, Sweden

Sep 29, 2019

1 Ignorance pdfs: Indifference and translation groups

- Consider a six-sided dice
- How do we assign $p(X_i|I)$, $i \in \{1, 2, 3, 4, 5, 6\}$?
- We do know $\sum_i p(X_i|I) = 1$
- Invariance under labeling $\Rightarrow p(X_i|I) = 1/6$
  - provided that the prior information I says nothing that breaks the symmetry.

1.1 Location invariance

Indifference to a shift x_0 for a location parameter x implies that

$$p(x|I)dx \approx p(x + x_0|I)d(x + x_0) = p(x + x_0|I)dx,$$

in the allowed range.

- Invariance under origin position $\Rightarrow p(x|I) = p(x + x_0|I)$, i.e., $p(x|I) = $ constant.
  - Provided that the prior information I says nothing that breaks the symmetry.
- The pdf will be zero outside the allowed range (specified by I).

1.2   Scale invariance

Indifference to a re-scaling λ of a scale parameter x implies that

$$p(x|I)dx \approx p(\lambda x|I)d(\lambda x) = \lambda p(\lambda x|I)dx,$$

in the allowed range.

- Invariance under re-scaling $\Rightarrow p(x|I) \propto 1/x$.

  - Provided that the prior information I says nothing that breaks the symmetry.

- The pdf will be zero outside the allowed range (specified by I).

- This prior is often called a *Jeffrey's prior*; it represents a complete ignorance of a scale parameter within an allowed range.

- It is equivalent to a uniform pdf for the logarithm: $p(\log(x)|I) = \text{constant}$

  - as can be verified with a change of variable $y = \log(x)$, see lecture notes on error propagation.

Example: Straight-line model. Consider the theoretical model $y_{\text{th}}(x) = \theta_1 x + \theta_0$.

- Would you consider the intercept θ_0 a location or a scale parameter, or something else?

- Would you consider the slope θ_1 a location or a scale parameter, or something else?

Consider also the statistical model for the observed data $y_i = y_{\text{th}}(x_i) + \epsilon_i$, where we assume independent, Gaussian noise $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$.

- Would you consider the standard deviation σ a location or a scale parameter, or something else?

1.3   Symmetry invariance

- In fact, by symmetry indifference we could as well have written the linear model as $x_{\text{th}}(y) = \theta_1' y + \theta_0'$

- We would then equate the probability elements for the two models $p(\theta_0, \theta_1|I)d\theta_0 d\theta_1 = q(\theta_0', \theta_1'|I)d\theta_0' d\theta_1'$.

- The transformation gives $(\theta_0', \theta_1') = (-\theta_1^{-1}\theta_0, \theta_1^{-1})$.

This change of variables implies that

$$q(\theta'_0, \theta'_1 | I) = p(\theta_0, \theta_1 | I) \left| \frac{d\theta_0 d\theta_1}{d\theta'_0 d\theta'_1} \right|,$$

where the (absolute value of the) determinant of the Jacobian is

$$\left| \frac{d\theta_0 d\theta_1}{d\theta'_0 d\theta'_1} \right| = \text{abs} \left(\begin{vmatrix} \frac{\partial \theta_0}{\partial \theta'_0} & \frac{\partial \theta_0}{\partial \theta'_1} \\ \frac{\partial \theta_1}{\partial \theta'_0} & \frac{\partial \theta_1}{\partial \theta'_1} \end{vmatrix} \right) = \frac{1}{(\theta'_1)^3}.$$

- In summary we find that $\theta_1^3 p(\theta_0, \theta_1 | I) = p(-\theta_1^{-1}\theta_0, \theta_1^{-1} | I)$.

- This functional equation is satisfied by

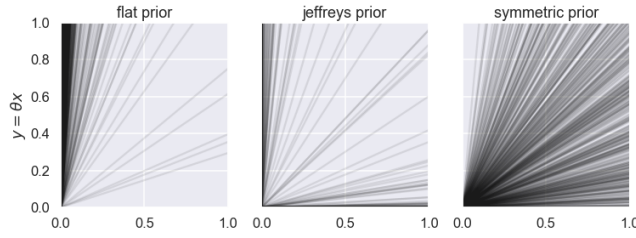$$p(\theta_0, \theta_1 | I) \propto \frac{1}{(1 + \theta_1^2)^{3/2}}.$$

Figur 1: 100 samples of straight lines with intercept 0 and slope from three different pdfs.

2 The principle of maximum entropy

Having dealt with ignorance, let us move on to a more enlightened situation.

Consider a die with the usual six faces that was rolled a very large number of times. Suppose that we were only told that the average number of dots was 2.5. What (discrete) pdf would we assign? I.e. what are the probabilities $\{p_i\}$ that the face on top had i dots after a single throw?

The available information can be summarized as follows

$$\sum_{i=1}^{6} p_i = 1, \qquad \sum_{i=1}^{6} i p_i = 2.5$$

This is obviously not a normal die, with uniform probability $p_i = 1/6$, since the average result would then be 3.5. But there are many candidate pdfs that would reproduce the given information. Which one should we prefer?

It turns out that there are several different arguments that all point in a direction that is very familiar to people with a physics background. Namely that we should prefer the probability distribution that maximizes an entropy measure, while fulfilling the given constraints.

3

2.1 The entropy of Scandinavians

Let's consider another pdf assignment problem. This is originally the *kangaroo problem* (Gull and SKilling, 1984), but adapted here to a local context. The problem is stated as follows:

Information: 70% of all Scandinavians have blonde hair, and 10% of all Scandinavians are left handed.

Question: On the basis of this information alone, what proportion of kangaroos are both blonde and left handed?

We note that for any one given Scandinavian there are four distinct possibilities:

1. Blonde and left handed (probability p_1).

2. Blonde and right handed (probability p_2).

3. Not blonde and left handed (probability p_3).

4. Not blonde and right handed (probability p_4).

The following 2x2 contingency table

	Left handed	Right handed
Blonde	p_1	p_2
Not blonde	p_3	p_4

can be written in terms of a single variable x due to the normalization condition $\sum_{i=1}^{4} p_i = 1$, and the available information $p_1 + p_2 = 0.7$ and $p_1 + p_3 = 0.1$

	Left handed	Right handed
Blonde	$0 \leq x \leq 0.1$	$0.7 - x$
Not blonde	$0.1 - x$	$0.2 + x$

But which choice of x is preferred?

2.2 The monkey argument

A model for assigning probabilities to M different alternatives that satisfy some constraint as described by I:

- Monkeys throwing N balls into M equally sized boxes.

- The normalization condition $N = \sum_{i=1}^{M} n_i$.

- The fraction of balls in each box gives a possible assignment for the corresponding probability $p_i = n_i/M$.

- The distribution of balls $\{n_i\}$ is therefore a candidate pdf $\{p_i\}$.

- The resulting pdf might not be consistent with the constraints of I, however, in which case it should be rejected as a possible candidate.

- After many such trials, some distributions will be found to come up more often than others. The one that appears most frequently (and satisfies I) would be a sensible choice for $p(\{p_i\}|I)$.

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

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

where we have used the Stirling approximation $\log(n!) \approx n\log(n) - n$ for large numbers, and a cancellation of two terms.

- There are M^N different ways to distribute the balls.

- The micro-states $\{n_i\}$ are connected to the pdf $\{p_i\}$, so the frequency of a given pdf is given by

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

Substituting $p_i = n_i/N$, and using the normalization condition finally gives

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

We note that N and M are constants so that the pdf is given by the $\{p_i\}$ that maximizes

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

You might recognise this quantity as the *entropy* from statistical mechanics. The interpretation of entropy in statistical mechanics is the measure of uncertainty, which remains about a system after its observable macroscopic properties, such as temperature, pressure and volume, have been taken into account. For a given set of macroscopic variables, the entropy measures the degree to which the probability of the system is spread out over different possible microstates. Specifically, entropy is a logarithmic measure of the number of micro-states with significant probability of being occupied $S = -k_B \sum_i p_i \log(p_i)$, where k_B is the Boltzmann constant.

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).

Consider the third argument. Let us check it empirically to the problem of hair colour and handedness of Scandinavians. We are interested in determining $p_1 \equiv p(L, B|I) \equiv x$, the probability that a Scandinavian is both left-handed and blonde. However, in this simple example we can immediately realize that the assignment $p_1 = 0.07$ is the only one that implies no correlation between left-handedness and hair color. Any joint probability smaller than 0.07 implies that left-handed people are less likely to be blonde, and any larger vale indicates that left-handed people are more likely to be blonde.

So unless you have specific information about the existence of such a correlation, you should better not build it into the assignment of the probability p_1.

Question: Can you show why $p_1 < 0.07$ and $p_1 > 0.07$ corresponds to left-handedness and blondeness being dependent variables?

Let us now empirically consider a few variational functions of $\{p_i\}$ and see if any of them gives a maximum that corresponds to the uncorrelated assignment $x = 0.07$, which implies $p_1 = 0.07$, $p_2 = 0.63$, $p_3 = 0.03$, $p_4 = 0.27$. A few variational functions and their prediction for x are shown in the following table.

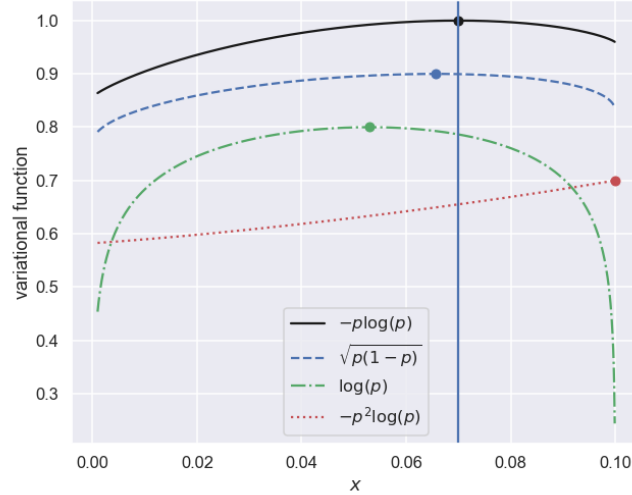| Variational function	Optimal x	Implied correlation
$-\sum_i p_i \log(p_i)$	0.070	None
$\sum_i \log(p_i)$	0.053	Negative
$-\sum_i p_i^2 \log(p_i)$	0.100	Positive
$-\sum_i \sqrt{p_i(1 - p_i)}$	0.066	Negative

Figur 2: Four different variational functions $f\left(\{p_i\}\right)$. The optimal x are shown by circles. The uncorrelated assignment $x = 0.07$ is shown by a vertical line.

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

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

which is often known as the *Shannon-Jaynes entropy*, or the *Kullback number*, or the *cross entropy* (with opposite sign).

In the continuous case

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

2.3 Derivation of common pdfs using MaxEnt

Variance and the Gaussian pdf.

Counting statistics and the Poisson distribution.