mathematical methods - week 15

Bayesian statistics

Sara A. Solla, Northwestern University

Georgia Tech PHYS-6124
Homework HW #15 due Tuesday, December 2, 2014

Study the notes, and work out the steps indicated by "exercise"
(10 points in all).

Georgia Tech PHYS-6124

Why do we so often resort to Gaussian distributions? This is our default assumption when we see a probability density function that looks like a unimodal bump. Here we will work out the one-dimensional case of a probability distribution $p(x)$. The goal is to show that the maximal entropy distribution that has a specified mean and variance is a Gaussian. So, when we know the mean and the variance of a random variable, and we chose to represent its probability distribution by a Gaussian, we are implicitly making a maximal entropy assumption.

****

Probability distributions

Consider a one-dimensional random variable $X=\{x\}$, described by the probability distribution $p(x)$. The distribution has to satisfy a normalization condition:

$$\sum_x p(x) = 1 ,$$

the mean of the distribution is its first moment,

$$\sum_x x\, p(x) = \mu ,$$

and the variance of the distribution is its second central moment,

$$\sum_x (x-\mu)^2\, p(x) = \sigma^2 .$$

We can define higher order moments, central or not. We next define a different parameter that helps characterize the distribution: its entropy.

****

Entropy of a probability distribution

The concept of entropy of a distribution is based on Shannon's notion of information (Shannon, C. E. (1948) "A Mathematical Theory of Communication", *Bell System Technical Journal* **27** (3): 379–423). The argument is as follows: consider the event in which the random variable takes the value x. This event happens with probability $p(x)$. What is the information content of this event? Shannon argued that the information content $I(p)$ of an event that occurs with probability p had to satisfy the following properties:

1. $I(p) \geq 0$ – information is a non-negative quantity
2. $I(p=1)=0$ – the occurrence of a certain event carries no information
3. $I(p_1 p_2)= I(p_1)+ I(p_2)$ – the information carried by independent events is additive

Based on these simple requirements, the solution is unique. As shown by Shannon,

$$I(p) = \ln(1/p) = -\ln(p) .$$

The only ambiguity is the base on which the logarithm is taken; when using base 2, we talk about information measured in bits. Toss a fair coin. If you get heads, what is the information content of this event? The answer is $I= - \log_2(1/2)= \log_2(2)= 1$. You gained 1 bit of information.

We can now compute the average information content associated with the probability distribution. This is called the entropy of the distribution:

$$H(X) = \sum_x I\left(p(x)\right)p(x) = -\sum_x \left(\ln p(x)\right)p(x) = -\sum_x p(x)\ln p(x) .$$

This entropy is a measure of the information content associated with the full range of values that the random variable can take and their probabilities.

****

Example: Bernouilli process

Consider a binary random variable that can only take two values: $x=1$ with probability (p) and $x=0$ with probability ($1-p$). Let's compute the mean, variance, and entropy of the Bernouilli process:

$$\mu = \sum_x x\, p(x) = 1(p) + 0(1-p) = p$$

$$\sigma^2 = \sum_x (x-\mu)^2\, p(x) = (1-p)^2 p + (p)^2(1-p) = p(1-p)(1-p+p) = p(1-p)$$

$$H = -p\ln p - (1-p)\ln(1-p)$$

Exercise (3 points): Plot both σ^2 and H as a function of p for $0 \le p \le 1$. Comment on the relationship between variance and entropy. Comment on the value that these two quantities take at p=0 and p=1. Comment on the value of p at which a maximum is observed in these two plots.

****

Maximal entropy principle

This principle is due to Jaynes, who first clarified the connections between information theory and statistical mechanics (Jaynes, E. T. (1957) "Information Theory and Statistical Mechanics I", *Physical Review* **106** (4): 620-630; Jaynes, E. T. (1957) "Information Theory and Statistical Mechanics II", *Physical Review* **108** (2): 171-190). Jaynes argued that the entropy of statistical physics and the entropy of information theory are the same thing, and offered an information theoretical argument that justifies the Gibbs distribution.

The basic idea of `maximal entropy' is that the probability distribution that best represents the current state of knowledge about a random variable is the maximal entropy distribution that is consistent with all available information about the random variable. This is a `maximum ignorance' principle: the selected distribution is the one that makes the least claim to being informed beyond the available information, that is to say the one that admits the most ignorance beyond the available information. Usually, the available information about the random variable is in the form of known moments, but it could also be about symmetries that the probability distribution should obey.

****

Example: Gaussian distribution

Consider a continuous random variable whose mean is constrained to be equal to a given value μ and whose variance is constrained to a given value σ^2. What is the maximal entropy distribution that satisfies these constraints?

Let's pose the problem. We need to find a probability distribution $p(x)$ that maximizes

$$H = -\int dx\, p(x) \ln p(x),$$

while satisfying three constraints:

$$\int dx\ p(x) = 1$$
$$\int dx\, x\, p(x) = \mu \qquad .$$
$$\int dx\, (x-\mu)^2 p(x) = \sigma^2$$

This constrained optimization problem is solved through the use of Lagrange multipliers. The function to be maximized is redefined to be:

$$\tilde{H} = -\int dx\, p(x) \ln p(x) + \lambda_1 \left(\int dx\, p(x) - 1 \right) + \lambda_2 \left(\int dx\, x\, p(x) - \mu \right) + \lambda_3 \left(\int dx\, (x-\mu)^2\, p(x) - \sigma^2 \right)$$

Note the introduction of three Lagrange multipliers, $\{\lambda_1, \lambda_2, \lambda_3\}$, each one associated with one of the constraints. Note also that $\tilde{H}$ is a not a function but a functional: it is not a function of a variable, but a function of the function $p(x)$, the probability distribution that we are trying to find.

Exercise (7 points):

To maximize $\tilde{H}$, take its functional derivative with respect to $p(x)$ at a specific point x, and set it to zero. We can consider the value of p at x to be independent of its value at another point x' because all constraints that would violate this independence have been explicitly incorporated in $\tilde{H}$. Show that this procedure leads to:

$$\ln p(x) = -1 + \lambda_1 + \lambda_2\, x + \lambda_3 (x-\mu)^2 ,$$

thus $p(x)$ is of the form

$$p(x) = \exp\left\{ -1 + \lambda_1 + \lambda_2\, x + \lambda_3 (x-\mu)^2 \right\}.$$

Next, determine the Lagrange multipliers by imposing the constraints. Clearly $\lambda_3 < 0$ is needed for $p(x)$ to be bounded. In order to carry out the calculations needed to impose the constraints, remember that for $\lambda_3 < 0$,

$$\int dy \exp\left\{ \lambda_2\, y + \lambda_3\, y^2 \right\} = \sqrt{\frac{\pi}{|\lambda_3|}} \exp\left\{ -\frac{(\lambda_2)^2}{4|\lambda_3|} \right\}$$

Using this identity, show that the normalization constraint results in

$$p(x) = \frac{1}{Z} \exp\left\{ \lambda_2 (x-\mu) + \lambda_3 (x-\mu)^2 \right\},$$

with

$$Z = \sqrt{\frac{\pi}{|\lambda_3|}} \exp\left\{ -\frac{(\lambda_2)^2}{4|\lambda_3|} \right\}.$$

Next, show that the constraint on the mean leads to λ_2=0. This part will require a bit of careful algebra. Finally, show that the constraint on the variance leads to

$$|\lambda_3| = 1 \Big/ \left(2\sigma^2\right),$$

which, together with λ_2=0, results in

$$Z = \sqrt{2\pi\sigma^2}\ ,$$

and

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{\left(x-\mu\right)^2}{2\sigma^2}\right\}.$$

a properly normalized Gaussian with mean μ and variance σ^2 !!!